

DOI: <https://doi.org/10.17721/1812-5409.2026/1>

Наведено результати теоретичних, прикладних і науково-експериментальних досліджень з актуальних питань фізики та математики.

Для науковців, викладачів, студентів та аспірантів.

ГОЛОВНИЙ РЕДАКТОР РЕДАКЦІЙНА КОЛЕГІЯ	Розора Ірина, д-р фіз.-мат. наук, проф. (Україна) Анікушин Андрій, канд. фіз.-мат. наук, доц. (Україна); Анісімов Анатолій, д-р фіз.-мат. наук, проф., акад. НАН України, (Україна); Анісімов Ігор, д-р фіз.-мат. наук, проф. (Україна); Бавула Володимир, проф. (Велика Британія); Бартлова Мілада, д-р філософії (Чеська Республіка); Бегін Луїза, проф. (Італія); Безущак Оксана, д-р фіз.-мат. наук, проф. (Україна); Гаврюшенко Дмитро, д-р фіз.-мат. наук, проф. (Україна); Горячко Андрій, д-р фіз.-мат. наук, доц. (Україна); Джуліано Ріта, проф. (Італія); Дмитрук Ігор, д-р фіз.-мат. наук, проф. (Україна); Жук Ярослав, д-р фіз.-мат. наук, проф. (Україна); Заславський Володимир, д-р техн. наук, проф. (Україна); Затула Дмитро, канд. фіз.-мат. наук, доц. (заст. голов. ред.) (Україна); Капустян Олексій, д-р фіз.-мат. наук, проф. (Україна); Капустян Олена, д-р фіз.-мат. наук, ст. наук. співроб. (Україна); Кашпур Олена, д-р фіз.-мат. наук, доц. (Україна); Коган Ніколас, проф. (США); Кондратенко Сергій, д-р фіз.-мат. наук, проф. (Україна); Кудін Володимир, д-р техн. наук, ст. наук. співроб. (Україна); Леоненко Микола, проф. (Велика Британія); Львов Віктор, д-р фіз.-мат. наук, проф. (Україна); Майборода Ростислав, д-р фіз.-мат. наук, проф. (Україна); Макарець Микола, д-р фіз.-мат. наук, проф. (Україна); Медвідс Артурс, д-р фіз.-мат. наук, проф. (Латвія); Мішура Юлія, д-р фіз.-мат. наук, проф. (Україна); Оленко Андрій, проф. (Австралія); Орсігер Енцо, проф. (Італія); Пашко Анатолій, д-р фіз.-мат. наук, проф. (Україна); Петравчук Анатолій, д-р фіз.-мат. наук, проф. (Україна); Погані Тібор, проф. (Хорватія); Погорілий Сергій, д-р техн. наук, проф. (Україна); Ронто Мікрос, д-р наук, проф. (Угорщина); Савенков Сергій, д-р фіз.-мат. наук, доц. (Україна); Сільвестров Дмитро, д-р фіз.-мат. наук, проф. (Швеція); Скришевський Валерій, д-р фіз.-мат. наук, проф. (Україна); Сливка-Тилищак Ганна, д-р фіз.-мат. наук, проф. (Україна); Сотнін Томмі, проф. (Фінляндія); Тору Аокі, д-р філософії, проф. (Японія); Трофімчук Сергій, проф. (Чилі); Футорний В'ячеслав, проф. (Бразилія); Худак Стефан, д-р наук, проф. (Словацька Республіка); Яневич Тетяна, канд. фіз.-мат. наук, ст. наук. співроб. (відп. ред.) (Україна)
Адреса редколегії	Факультет комп'ютерних наук та кібернетики просп. Академіка Глушкова, 4д, м. Київ, 03022 ☎ (38044) 259 01 49 e-mail: physmath.bulletin@knu.ua web: https://bphm.knu.ua
Затверджено	вченою радою факультету комп'ютерних наук та кібернетики 21.04.26 (протокол № 13)
Зареєстровано	Національною радою України з питань телебачення і радіомовлення Рішення № 1089 від 28.03.24 Ідентифікатор друкованого медіа: R30-03805
Атестовано	Міністерством освіти і науки України (категорія А) Наказ № 641 від 28.04.25
Індексування	Scopus, DOAJ, Zentralblatt MATH (zbMATH), JournalTOCs, Crossref, ROAD, Google Scholar, Наукова періодика України
Засновник та видавець	Київський національний університет імені Тараса Шевченка Видавничо-поліграфічний центр "Київський університет" Свідоцтво про внесення до Державного реєстру ДК № 1103 від 31.10.02
Адреса видавця	ВПЦ "Київський університет" б-р Тараса Шевченка, 14, м. Київ, 01601 ☎ (38044) 239 32 22, 239 31 58, 239 31 28 e-mail: vpcc@knu.ua

BULLETIN

TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

PHYSICS AND MATHEMATICS

1(82)/2026

Established in 2005

DOI: <https://doi.org/10.17721/1812-5409.2026/1>

The Bulletin highlights the results of theoretical, scientific, applied and scientific-experimental researches on the current issues of Physics and Mathematics.

For scientists, researchers, lecturers, PhD students and students.

EDITOR-IN-CHIEF	Rozora Iryna, DSc (Phys. & Math.), Prof. (Ukraine)
EDITORIAL BOARD	Anikushyn Andrii, PhD (Phys. & Math.), Assoc. Prof. (Ukraine); Anisimov Anatoliy, DSc (Phys. & Math.), Prof., Corresponding Member of the National Academy of Sciences of Ukraine, (Ukraine); Anisimov Igor, DSc (Phys. & Math.), Prof. (Ukraine); Bartlova Milada, PhD (Czech Republic); Bavula Vladimir, Prof. (Great Britain); Beghin Luisa, Prof. (Italy); Bezuschak Oksana, DSc (Phys. & Math.), Prof., (Ukraine); Cogan Nicholas, Prof. (USA); Dmitruk Igor, DSc (Phys. & Math.), Prof. (Ukraine); Futorny Vyacheslav, Prof., (Brazil); Gavryushenko Dmytro, DSc (Phys. & Math.), Prof. (Ukraine); Giuliano Rita, Prof. (Italy); Goriachko Andrii, DSc (Phys. & Math.), Assoc. Prof. (Ukraine); Hudak Stefan, DSc, Prof. (Slovak Republic); Ianevych Tetiana, PhD (Phys. & Math.), Senior Researcher (Managing Editor) (Ukraine); Kashpur Olena, DSc (Phys. & Math.), Assoc. Prof. (Ukraine); Kapustian Olena, DSc (Phys. & Math.), Senior Researcher (Ukraine); Kapustyan Oleksiy, DSc (Phys. & Math.), Prof. (Ukraine); Kondratenko Serhiy, DSc (Phys. & Math.), Prof. (Ukraine); Kudin Volodymyr, DSc (Engin.), Senior Researcher (Ukraine); Leonenko Nikolay, Prof. (Great Britain); Lvov Victor, DSc (Phys. & Math.), Prof. (Ukraine); Maiboroda Rostyslav, DSc (Phys. & Math.), Prof. (Ukraine); Makarets' Nikolay, DSc (Phys. & Math.), Prof. (Ukraine); Medvids Arturs, DSc (Phys. & Math.), Prof. (Latvia); Mishura Yuliya, DSc (Phys. & Math.), Prof. (Ukraine); Olenko Andriy, Prof. (Australia); Orsingher Enzo, Prof. (Italy); Pashko Anatolij, DSc. (Phys. & Math), Prof. (Ukraine); Petravchuk Anatolij, DSc (Phys. & Math.), Prof. (Ukraine); Pogány Tibor, Prof. (Croatia); Pogorily Serhiy, DSc (Engin.), Prof. (Ukraine); Rontó Miklós, DSc, Prof. (Hungary); Savenkov Serhij, DSc (Phys. & Math.), Associate Prof. (Ukraine); Silvestrov Dmitrii, DSc (Phys. & Math.), Prof. (Sweden); Skryshevsky Valery, DSc (Phys. & Math.), Prof. (Ukraine); Slyvka-Tylyshchak Anna, DSc (Phys. & Math.), Prof. (Ukraine); Sottinen Tommi, Prof. (Finland); Toru Aoki, PhD., Prof. (Japan); Trofimchuk Sergey, Prof. (Chile); Zaslavs'kyj Volodymyr, DSc, Prof. (Ukraine); Zatula Dmytro, PhD (Phys. & Math.), Assoc. Prof. (Deputy Editor-in-Chief) (Ukraine); Zhuk Yaroslav, DSc (Phys. & Math.), Prof. (Ukraine)
Address	Faculty of Computer Science and Cybernetics 4d, Hlushkov ave., Kyiv, 03022 ☎ (38044) 259 01 49 e-mail: physmath.bulletin@knu.ua web: https://bphm.knu.ua
Approved by	the Academic Council of the Faculty of Computer Science and Cybernetics 21.04.26 (protocol № 13)
Registered by	the National Council of Television and Radio Broadcasting of Ukraine Decision № 1089 of 28.03.24 Identifier of printed media: R30-03805
Certified by	the Ministry of Education and Science of Ukraine (category A) Order № 641 dated 28.04.25
Indexing	Scopus, DOAJ, Zentralblatt MATH (zbMATH), JournalTOCs, Crossref, ROAD, Google Scholar, and Scientific periodicals of Ukraine
Founder and publisher	Taras Shevchenko National University of Kyiv Publishing and Polygraphic Center "Kyiv University" Certificate of entry into the State Register ДК № 1103 dated 31.10.02
Address	PPC "Kyiv University" 14, Taras Shevchenko blvd., Kyiv, 01601 ☎ (38044) 239 32 22, 239 31 58, 239 31 28 e-mail: vpc@knu.ua

ЗМІСТ

КАШПУР Олена, ЗГУРОВСЬКИЙ Михайло, КАПУСТЯН Олена, ПАНКРАТОВА Наталія, ЗІНЬКО Петро, ЛОСЄВА Марія

Славетний ювілей видатної постаті: до 80-річчя від дня народження Олександра Григоровича Наконечного.....5

АЛГЕБРА, ГЕОМЕТРІЯ ТА ТЕОРІЯ ЙМОВІРНОСТЕЙ

АНУРАГ, В'ЯС Апарна, КІМ Гібак

Розширена теорема про єдину нерухому точку в повному квазіметричному просторі.....13

БОРИСЕНКО Олександр, БОРИСЕНКО Ольга

Динаміка неавтономної стохастичної моделі мутуалізму зі стрибками.....17

ВАСИЛИК Ольга, ОКУНЄВ Єгор

Моделювання епідемії ВІЛ/СНІД в Україні як марковський процес із неперервним часом27

ВОЛКОВ Олександр, ВОЛКОВ Юрій, ВОЙНАЛОВИЧ Наталія

Клас експоненціальних статистичних структур типу В34

КАРНАУХ Євген

Моделювання меж для моделі потоку рідини другого порядку43

КУДІН Володимир, ОНОЦЬКИЙ В'ячеслав

Аналіз лінійних систем із комплексними елементами методом базисних матриць55

МОКАНУ Яна

Ергодичність деяких процесів типу Леві в $\mathbb{R}^d$ 63

МОРОЗ Микола

Збіжність числових послідовностей у термінах додатних і знакозмінних розкладів Перрона72

ПЛАТОНОВ Денис

Аналіз слабкої збіжності динамічної зрізки для процесів типу Леві78

УАРКА Ксав'єр М.

Ймовірнісні розподіли типу Н87

ДИФЕРЕНЦІАЛЬНІ РІВНЯННЯ, МАТЕМАТИЧНА ФІЗИКА ТА МЕХАНІКА

АЛЕКСАНДРОВИЧ Ірина, АДЖУБЕЙ Лариса, РИЖЕНКО Андрій, ЛЯШКО Віктор

Задача Робена для гіперболічного рівняння четвертого порядку.....101

АЛЕКСЄЄНКО Сергій, НАКАШИДЗЕ Лілія, ЗОЛОТАРЕНКО Сергій

Чисельне моделювання аероакустичних процесів під час обтікання сегментів лопатей вітряних турбін106

ГРИГОРЕНКО Олександр, ТРАЧ Володимир, ПОДВОРНИЙ Андрій, ЖУКОВА Наталія

Напружено-деформований стан анізотропних товстостінних циліндричних оболонок під змінним навантаженням115

ДЕНИСЮК Володимир, ОЛЕШКО Тетяна, РИБАЧУК Людмила

Неповні тригонометричні сплайни в задачах побудови наближених розв'язків лінійних диференціальних рівнянь другого порядку.....119

ЖУК Олександр, ЖУК Ярослав, КАШТАЛЯН Марія

Дія сил акустичного випромінювання в ідеальній рідині на в'язку краплю126

КАПУСТЯН Олексій, СУКРЕТНА Анна, КАПУСТЯН Дмитро, ОСПАНОВ Мирзгалі

Про стійкість майже періодичного вільного від інфекції розв'язку в імпульсній SEIR-моделі.....130

ЛОВЕЙКІН Андрій, КРЕНЕВИЧ Андрій

Особливість потенціального поля у вершині симетричного наконечника із чотирма однаковими, перпендикулярними клиноподібними крилами.....136

МІШУРІС Генадій, ПРОЦЕРОВ Юрій, ВАЙСФЕЛЬД Наталія, ЖУРАВЛЬОВА Зінаїда

Задача теплопровідності для складеного циліндру з неідеальними умовами контакту між його шарами143

НАЗАРЧУК Вікторія

Керованість псевдопараболічними інтегро-диференціальними системами з пам'яттю147

УЛІТКО Ігор, БОРИСЕЙКО Олександр, ЛЕБЕДЄВА Ірина, КУРИЛКО Олександр Резонансні властивості п'єзоелектричного гіроскопічного сенсора на об'ємних хвилях зсуву	161
--	-----

ШИКУЛА Олена, ЖУКОВА Наталія, ВИШНІВСЬКИЙ Віктор, БІЛОУСОВА Світлана Модель деформування армованих у площині волокнистих матеріалів із фізично нелінійною матрицею	165
--	-----

КОМП'ЮТЕРНІ НАУКИ ТА ІНФОРМАТИКА

АФАНАСЬЄВ Олег, КІЗІЛОВА Наталія, ПОСЛАВСЬКИЙ Сергій Випадкові мережі у моделюванні тепломасопереносу в пористих матеріалах	170
---	-----

БАШУК Олексій Порівняльний аналіз постквантових протоколів цифрового підпису, базованих на геш-функціях	176
---	-----

ГАЛУШКО Олександр Інтегрована модель факторного відбору та категоріальної регресії для прогнозування рівня ризику імплантації діагностично-терапевтичних пристроїв при патології шлунково-стравохідного з'єднання	187
---	-----

ГЛИБОВЕЦЬ Микола, ВАНІН Данило Використання одно- та багатомовних моделей на базі Bert для розв'язання задач автоматичної обробки текстів українською мовою	195
---	-----

ДОЦЕНКО Сергій, ДІМІТРОВ Георгій, ВСЧЕРКОВСЬКА Анастасія, МАКУШЕНКО Ігор Алгоритмічні підходи до моделювання та аналізу фінальних позицій у грі в нарди	206
---	-----

КУБИЦЬКИЙ Володимир, БОЖОК Артем Конфігурації проміжних шарів для задач подібності зображень: оцінка багаторівневого векторного Представлення, отриманого з архітектур CNN і трансформерів	213
--	-----

КРИВОЛАП Андрій, ШИШАЦЬКА Олена, РУСІНА Наталія Рамкове правило в монотонному розширенні логіки Флойда – Хоара	218
--	-----

ПАШКО Анатолій, ПАПРОЦЬКИЙ Ігор, ГЛИБОВЕЦЬ Андрій, ТЕРЕНЧУК Світлана Удосконалення предиктивної моделі патерну circuit breaker для підвищення відмовостійкості мікросервісів	225
--	-----

ПИХТАР Микола, ЛИТВИН Мирослава Статистичний аналіз факторів, що впливають на захворюваність цукровим діабетом	237
--	-----

САВЧЕНКО Тетяна, ЛУЦЬКА Наталія, ВЛАСЕНКО Лідія, ПАРХОМЕНКО Іван Цифровий двійник у структурі сучасного освітнього середовища	242
---	-----

СМОЛІН Ярослав, ЛІСКІН Вячеслав Метааналіз Парето-оптимальних моделей виявлення об'єктів для безпілотних літальних апаратів на наборі даних VisDrone2019-DET	255
--	-----

ЧІГІДІНА Марина, ЛАЗАРЕНКО Ірина, НОМІРОВСЬКИЙ Дмитро, РУБЛЬОВ Богдан, СЕМЕНОВ Володимир, СВІТОВЕНКО Катерина Основні властивості метрики дотичних	265
--	-----

ЩЕСТЮК Наталія, СЛУЧИНСЬКИЙ Дмитро, СЕРДЮК Федір Субдифузійні моделі з локальною волатильністю для неліквідних ринків	273
---	-----

СУЧАСНА ФІЗИКА

БЕХ Ігор, БОГДАНОВ Роман Вплив опору резистора позитивного зворотного зв'язку на амплітудно-частотну характеристику гіратора на операційному підсилювачі	280
--	-----

ЗАБАШТА Юрій, ЛАЗАРЕНКО Ярослав, АЛЕКСЄЄВ Олександр, ВЕРГУН Лена, ГНАТЮК Катерина, АНДРУСЕНКО Дмитро, ЯБЛОЧКОВА Катерина, ДІНЖОС Роман, КОСЄВА Нелі, ЛАЗАРЕНКО Максим Вплив каркаса на дифузію сторонніх частинок в полімерних гідрогелях	285
---	-----

CONTENTS

KASHPUR Olena, ZGUROVSKY Michael, KAPUSTIAN Olena, PANKRATOVA Nataliya, ZINKO Petro, LOSIEVA Mariia	
Glorious anniversary of an outstanding figure: to the 80th anniversary of the birth of Oleksandr Grigorovych Nakonechnyi	5

ALGEBRA, GEOMETRY AND PROBABILITY THEORY

ANURAG, VYAS Aparna, KIM Gibak	
Extended unique fixed point theorem in complete quasi-metric space	13
BORYSENKO Oleksandr, BORYSENKO Olga	
Dynamics of a non-autonomous stochastic mutualism model with jumps	17
VASYLYK Olga, OKUNIEV Yehor	
Modelling HIV/AIDS epidemic in Ukraine as a continuous-time Markov process.....	27
VOLKOV Oleksandr, VOLKOV Yurii, VOINALOVYCH Nataliia	
On the class of exponential statistical structures of type B	34
KARNAUKH Ievgen	
Bounds simulation for fluid flow model of the second order	43
KUDIN Volodymyr, ONOTSKYI Viacheslav	
Analysis of linear systems with complex elements by the basis matrix method.....	55
MOKANU Yana	
On ergodic property of some Lévy-type processes in $\mathbb{R}^d$	63
MOROZ Mykola	
The convergence of sequences in terms of positive and alternating Perron expansions.....	72
PLATONOV Denis	
Weak convergence analysis of dynamic cutting for Lévy-type processes	78
HUARCA Javier M.	
The H probability distributions.....	87

DIFFERENTIAL EQUATIONS, MATHEMATICAL PHYSICS AND MECHANICS

ALEXANDROVICH Iryna, ADZHUBEY Larisa, RYZHENKO Andrii, LYASHKO Viktor	
Robin's problem for a fourth-order hyperbolic equation	101
ALEKSIEIENKO Serhii, NAKASHYDZE Liliya, ZOLOTARENKO Serhii	
Numerical modeling of aeroacoustic processes during the flow around wind turbine blade segments.....	106
GRIGORENKO Alexander, TRACH Volodymyr, PODVORNYI Andrii, ZHUKOVA Natalia	
Stress-strain state of anisotropic thick-walled cylindrical shells under variable loading.....	115
DENYSIUK Volodymyr, OLESHKO Tetiana, RYBACHUK Liudmyla	
Incomplete trigonometric splines in the problems of constructing approximate solutions of the second order linear differential equations	119
ZHUK Oleksandr, ZHUK Yaroslav, KASHTALYAN Maria	
The effect of acoustic radiation in an ideal fluid on the viscosity of a drop.....	126
KAPUSTYAN Oleksiy, SUKRETN Anna, KAPUSTIAN Dmytro, OSPANOV Myrzagali	
Stability of infection-free almost periodic solution in impulsive SEIR-model	130
LOVEIKIN Andrii, KRENEVYCH Andrii	
Singularity of the potential field at the vertex of a symmetric tip with four identical perpendicular wings of wedge-shaped form	136
MISHURIS Gennady, PROTSEROV Yurii, VAYSFELD Natalya, ZHURAVLOVA Zinaida	
Heat conduction problem for a composite cylinder with imperfect contact conditions between its layers	143
NAZARCHUK Viktoriia	
Controllability of pseudoparabolic integro-differential systems with memory	147

CONTENTS

ULITKO Igor, BORYSEIKO Oleksandr, LEBEDYEVA Iryna, KURLKO Oleksandr Resonance properties of a piezoelectric gyroscopic sensor based on bulk acoustic shear waves	161
SHIKULA Elena, ZHUKOVA Nataliia, VYSHNIVSKYI Victor, BILOUSOVA Svitlana Model of deformation of in-plane reinforced fibrous materials with physically nonlinear matrix	165

COMPUTER SCIENCES AND INFORMATICS

AFANASIEV Oleg, KIZILOVA Natalya, POSLAVSKYI Sergii Random networks in modeling heat and mass transfer in porous materials	170
BASHUK Oleksii Comparative analysis of post-quantum digital signature protocols based on hash functions	176
HALUSHKO Oleksandr Integrated model of factor selection and categorical regression for predicting the implantation risk of diagnostic-therapeutic devices in gastroesophageal junction pathology.....	187
GLYBOVETS Mykola, VANIN Danylo The application of monolingual and multilingual BERT-based models for text automation tasks in Ukrainian.....	195
DOTSENKO Serhii, DIMITROV Georgi, VECHERKOVSKAYA Anastasiya, MAKUSHENKO Igor Algorithmic approaches to modeling and analyzing endgame positions in backgammon.....	206
KUBYTSKYI Volodymyr, BOZHOK Artem Intermediate layer configurations for image similarity: evaluating multi-level feature extraction across CNN and transformer architectures	213
KRYVOLAP Andrii, SHYSHATSKA Olena, RUSINA Nataliia Frame rule in monotone Floyd – Hoare logic extension.....	218
PASHKO Anatolii, PAPROTSKYI Ihor, HLYBOVETS Andrii, TERENCECHUK Svitlana Improvements to the circuit breaker prediction model for increasing the fault tolerance of microservices.....	225
PYCHTAR Mykola, LYTVYN Myroslava Statistical analysis of factors affecting the incidence of diabetes mellitus.....	237
SAVCHENKO Tetiana, LUTSKA Nataliia, VLASENKO Lidia, PARKHOMENKO Ivan Digital twin in the structure of the modern educational environment.....	242
SMOLIN Yaroslav, LISKIN Viacheslav Meta-analysis of Pareto-optimal object detection models for unmanned aerial vehicles on the VisDrone2019-DET dataset	255
CHIGIDINA Maryna, LAZARENKO Iryna, NOMIROVSKII Dmytro, RUBLYOV Bogdan, SEMENOV Vladimir, SVITOVENKO Kateryna Basic properties of the metric of tangents.....	265
SHCHESTYUK Nataliya, SLUCHYNSKYI Dmytro, SERDUK Fedir Subdiffusive models with local volatility for illiquid markets	273

MODERN PHYSICS

BEKH Ihor, BOHDANOV Roman Influence of the positive feedback resistor resistance on the amplitude-frequency response of the gyrator based on the operational amplifier	280
ZABASHTA Yuri, LAZARENKO Yaroslav, ALEKSEEV Oleksandr, VERGUN Lena, HNATIUK Kateryna, ANDRUSENKO Dmytro, YABLOCHKOVA Kateryna, DINZHOS Roman, KOSEVA Neli, LAZARENKO Maksym Effect of the polymer network on the diffusion of foreign particles in hydrogels.....	285

УДК 519.21

DOI: <https://doi.org/10.17721/1812-5409.2026/1.1>

Олена КАШПУР¹, д-р фіз.-мат. наук, доц.
ORCID ID: 0000-0001-7091-8851
e-mail: olena.kashpur@knu.ua

Михайло ЗГУРОВСЬКИЙ², д-р техн. наук, проф., акад. НАН України
ORCID ID: 0000-0001-5896-7466
e-mail: mzz@kpi.ua

Олена КАПУСТЯН¹, д-р фіз.-мат. наук, ст. наук. співроб.
ORCID ID: 0000-0002-2629-0750
e-mail: olenakapustian@knu.ua

Наталія ПАНКРАТОВА², д-р техн. наук, проф., чл.-кор. НАН України
ORCID ID: 0000-0002-6372-5813
e-mail: natalidmp@gmail.com

Петро ЗІНЬКО¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-5111-4417
e-mail: petro.zinko@knu.ua

Марія ЛОСЄВА¹, провід. інж.
ORCID ID: 0000-0002-2282-206X
e-mail: mariia.losieva@knu.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Національний технічний університет України

"Київський Політехнічний інститут імені Ігоря Сікорського", Київ, Україна

СЛАВЕТНИЙ ЮВІЛЕЙ ВИДАТНОЇ ПОСТАТІ:

ДО 80-РІЧЧЯ ВІД ДНЯ НАРОДЖЕННЯ ОЛЕКСАНДРА ГРИГОРОВИЧА НАКОНЕЧНОГО

Присвячено 80-річчю від дня народження видатного українського вченого-математика та педагога професора Київського національного університету імені Тараса Шевченка Олександра Григоровича Наконечного. Окреслено його життєвий і професійний шлях. Основну увагу приділено огляду фундаментальних наукових результатів О. Г. Наконечного. Наведено перелік вибраних наукових праць ювіляра.

Ключові слова: Олександр Наконечний, випадкові процеси, системний аналіз, теорія прийняття рішень, умови невизначеності.

Класифікація відповідно до AMS 2020: 60G20, 39A50, 01A70.

У національній науковій традиції є постаті, чия праця не лише визначає розвиток окремих галузей, а й формує інтелектуальні горизонти країни. До таких людей належить Олександр Григорович Наконечний – український учений, педагог, доктор фізико-математичних наук, професор, академік і третій президент Академії наук вищої школи України.

Вражає життєвий і професійний шлях ученого. Це переконливий приклад того, як наполегливість, інтелект і відданість справі здатні створити фундамент для нових наукових шкіл і цілих напрямів досліджень. Народившись 9 січня 1946 р. в с. Красне Обухівського р-ну Київської обл., він змалку пізнав цінну працю та відповідальності. Закінчивши у 1962 р. міське професійно-технічне училище № 4 при Київському заводі верстатів-автоматів імені Горького, а згодом – вечірню школу робітничої молоді в Бучі, молодий Олександр зумів поєднати фахову підготовку з високою освітньою метою. Визначальним у його житті стало навчання на механіко-математичному факультеті Київського університету імені Тараса Шевченка, яке він завершив у 1969 р., спеціалізуючись у галузі теорії ймовірностей.

Уже в перші роки після отримання диплому Олександр Григорович виявив себе як талановитий дослідник і фахівець у сфері математичного програмування та кібернетики. Робота інженером-програмістом у лабораторії математичного програмування, а згодом викладацька діяльність на факультеті кібернетики – від асистента до професора кафедри моделювання складних систем – стали тим ґрунтом, на якому виріли його майбутні наукові досягнення.

Захист кандидатської дисертації на тему "Моментні функції старших порядків випадкових процесів" у 1973 р. під керівництвом академіка А. В. Скорохода та докторської "Мінімаксне оцінювання функціоналів від розв'язків рівнянь з частинними похідними" – у 1981 р. за спеціальністю "математична кібернетика" – утвердили його як одного з провідних учених у галузі випадкових процесів, системного аналізу та оптимального оцінювання складних математичних моделей. Професійний авторитет Олександра Григоровича закономірно привів його до керівництва кафедрою системного



аналізу та теорії прийняття рішень, яка з 1996 р. стала одним із ключових науково-освітніх центрів факультету кібернетики КНУ імені Тараса Шевченка.

О. Г. Наконечний є видатним ученим у галузі теорії прийняття рішень в умовах невизначеності та системного аналізу процесів різної природи (Наконечний, & Качинський, 1996; Marzeniuk, & Nakonechny, 2003; Наконечний, & Верес, 2007; Наконечний, & Жук, 2009; Марценюк, & Наконечний, 2009; Mallet, Nakonechny, & Zhuk, 2013; Наконечний, Марценюк, & Андрущак, 2014; Капустян, Капустян, & Наконечний, 2021; Bekesiene, et al., 2023a; Bekesiene, et al., 2023b; Bekesiene et al., 2024). Фактично, він є одним із творців та ідейним натхненником упровадження спеціальності "Системний аналіз" в освітній та науковий ландшафт України. Автор понад 421 наукових праць (98 його наукових праць входять до наукометричної бази Scopus), з них 13 монографій, 8 навчальних посібників. У галузі прикладної математики та системного аналізу створив й очолює власну наукову школу "Математичне моделювання та теорія оптимальних рішень". Серед його учнів 20 кандидатів і 6 докторів наук, які працюють у провідних університетах і наукових інститутах світу, розвиваючи ідеї свого вчителя та примножуючи інтелектуальний потенціал української науки.

О. Г. Наконечним було одержано фундаментальні результати в галузі оцінювання розв'язків і параметрів рівнянь із частинними похідними з детермінованими та стохастичними збуреннями (Nakonechny, & Podlipenko, 2021; Mallet, Nakonechny, & Zhuk, 2013; Zhuk, Polyakov, & Nakonechny, 2017; Nakonechny, Kapustian, & Chikrii, 2019; Nakonechny, & Podlipenko, 2022). Для побудови конструктивної теорії оцінювання О. Г. Наконечний застосував мінімаксий підхід (Nakonechny, & Zinko, 2023; Nakonechny et al., 2023a; Nakonechny et al., 2023b; Kapustian, & Nakonechny, 2020; Nakonechny, Kapustian, & Podlipenko, 2019; Nakonechny, Zinko, & Shevchuk, 2018; Zhuk, & Nakonechny, 2018; Наконечний, 2000). Виходячи з варіаційної теорії рівнянь еліптичного, параболического та гіперболического типів у гільбертових просторах, ним були розроблені оригінальні методи оптимізації в умовах невизначеності (Nakonechny, Shestopalov, & Podlipenko, 2014; Zhuk, Polyakov, & Nakonechny, 2017; Nakonechny, Kapustian, & Chikrii, 2019; Nakonechny, & Podlipenko, 2022; Nakonechny, Podlipenko, & Shestopalov, 2021a; Nakonechny, Podlipenko, Shestopalov, 2021b; Nakonechny, Kapustian, Chikrii, 2019), що суттєво узагальнювали результати Калмана-Б'юсі, Н. Н. Красовського, А. Б. Куржанського, Ж.-Л. Ліонса. Створення такої теорії відкрило можливість її застосування під час розроблення обчислювальних систем автоматизованої обробки результатів експериментів із визначення параметрів середовищ, інтенсивності джерел тощо. Подібні задачі виникають в акустиці, гідродинаміці, оптиці, радіолокації, астрономії, метеорології, сейсмології, теплофізиці (Наконечний, & Качинський, 1996; Marzeniuk, & Nakonechny, 2003; Марценюк, & Наконечний, 2009; Наконечний, Марценюк, & Андрущак, 2014; Mallet, Nakonechny, & Zhuk, 2013; Nakonechny, Zinko, & Zinko, 2025a; Nakonechny, Zinko, & Zinko, 2025b).

Надалі наукова діяльність О. Г. Наконечного була пов'язана з моделюванням й оптимізацією складних систем. Добре відомі фахівцям його праці з моделювання й оптимізації звичайних диференціальних рівнянь і рівнянь із частинними похідними. Ці дослідження були розпочаті в 1974 р. і суттєво розвинули теорію оцінювання та керування в умовах невизначеності (Nakonechny, & Podlipenko, 2021; Zhuk, Polyakov, & Nakonechny, 2017; Nakonechny, Podlipenko, & Shestopalov, 2021b; Nakonechny, Kapustian, & Chikrii, 2019). Узагальнивши принцип дуальності Р. Калмана, він розробив методи одержання рекурентних оцінок станів і параметрів стохастичних рівнянь в умовах невизначеності. Професор Наконечний та очолювані ним науковці здобули значні результати у таких напрямках:

- стохастичні рівняння зі стрибками,
- стохастичні рівняння зі змінною структурою,
- стохастичні рівняння з інформаційними критеріями оцінювання,
- диференціальні рівняння із запізненням,
- планування оптимальних стратегій спостереження для рівнянь із частинними похідними,
- задачі оцінювання та керування в гідроакустиці,
- задачі прогнозування розв'язків рівнянь переносу та дифузії.

Особливу увагу Олександр Григорович приділяє впровадженню одержаних наукових результатів у виробництво. З 1976 р. він був керівником робіт зі створення автоматизованої системи обробки результатів льотного експерименту. Із цією метою вперше був розроблений пакет програм "Модель", куди ввійшли нові алгоритми рекурентного оцінювання. Цей пакет знайшов застосування на заводі ім. Антонова (м. Київ).

З 1980 р. за керівництва О. Г. Наконечного виконувалися роботи зі створення системи моделювання й обробки даних гідроакустичного експерименту (СМОГАС), яка знайшла застосування в оборонній промисловості.

У 1997–1999 рр. разом із вченими Єльського університету (США) велася робота, очолював яку О. Г. Наконечний, зі створення системи моделювання й обробки даних ультразвукових експериментів з метою дистанційного керування інтелектуальними роботами нового покоління типу "Кажан" і "Дельфін". У 2011–2012 р. Олександр Григорович керував українсько-французьким проектом "Апостеріорне мінімакське оцінювання руху рідин" разом із науковцями з Національного дослідницького інституту комп'ютерних наук і керування (INRIA) у межах Програми спільних дій між Україною та Францією в галузі науково-технологічного співробітництва "Дніпро". У 2022–2023 рр. колектив під орудою Олександра Григоровича виконував спільний українсько-литовський проект "Моделювання ролі людського потенціалу для забезпечення оборони країни під час новітніх загроз" разом з науковцями Литовської військової академії імені генерала Йонаса Жемайтиса.

Науковий ентузіазм О. Г. Наконечного знайшов своє втілення у міжнародній науковій конференції "Проблеми прийняття рішень в умовах невизначеності" (PDMU), яку він разом з однодумцями започаткував у 2001 р. Конференція має значний резонанс у науковому світі, об'єднує учених з багатьох країн і щорічно проводиться як в Україні, так і за її межами (Чехія, Польща, Грузія, Азербайджан, Литва, Єгипет) (фото 1, 2, 3).

О. Г. Наконечний послідовно обстоює принципи академічної доброчесності та зміцнення партнерських зв'язків з провідними українськими освітньо-науковими установами у галузі системного аналізу – Навчально-науковим комплексом "Інститут прикладного системного аналізу" Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського" МОН України і НАН України, Львівським національним університетом ім. Івана Франка,

Національним університетом "Львівська Політехніка", активно долучаючись до формування наукових програм, роботи спеціалізованих рад й експертних комісій.



Фото 1. Відкриття міжнародної конференції PDMU у м. Бердянськ (2005)



Фото 2. Учасники конференції PDMU у м. Бердянськ (2005)



Фото 3. Учасники конференції PDMU у м. Бельсько-Бяла, Польща (2025)

Особливу увагу професор Наконечний приділяє вихованню наукових кадрів вищої кваліфікації та методичній роботі. З 1996 по 2021 р. – голова спеціалізованої вченої ради Д 26.001.35 із захисту докторських дисертацій, а із 2024 р. – заступник голови спеціалізованої вченої ради Д 26.001.09. У 1991–2010 рр. працював в експертній раді ВАК України. Із 2007 по 2010 р. був заступником голови Науково-методичної комісії з інформатики та прикладної математики, а також заступником голови Науково-методичної комісії із системного аналізу при МОН України. У 2010–2015 рр. – заступник голови Науково-методичної комісії із системного аналізу, голова підкомісії із системного аналізу класичних університетів при МОН України. Є одним з авторів розробки галузевих стандартів вищої освіти напрямку 040303 "Системний аналіз". У 2016–2019 рр. був членом колегії Міністерства освіти і науки України та членом експертної ради Державного фонду фундаментальних досліджень.

Олександр Григорович входить до редколегій українських наукових журналів: "Журнал обчислювальної та прикладної математики", "Mathematical modeling and computing" (входить до Scopus із 2017 р.) та закордонного журналу "Transactions of National Academy of Sciences of Azerbaijan. Series of Physical-Technical and Mathematical Sciences, Issue Mathematics".

Натхнення науково-педагогічна праця та послідовна громадянська позиція професора О. Г. Наконечного були відзначені низкою нагород: Почесною грамотою Міністерства освіти і науки України (2006), відзнакою Вченої ради Київського національного університету імені Тараса Шевченка (2008), Грамотою Вищої атестаційної комісії України (2010), нагородою Святого Володимира АН ВШ України (2016). Олександр Григорович – Заслужений працівник освіти України (2009), лауреат Державної премії України в галузі науки і техніки (2011) (фото 4).

У 2015 р. вчений отримав Премію імені Тараса Шевченка КНУ імені Тараса Шевченка. У 2015 р. О. Г. Наконечному присвоєно почесне звання "Заслужений професор Київського національного університету імені Тараса Шевченка". У 2018 р. він обраний Почесним професором Ланкаранського державного університету (Азербайджан). О. Г. Наконечний – Почесний доктор Українсько-американського університету Конкордія. Із 2016 р. він є Почесним членом Міжнародної академії наук з інформатизації освіти (Грузія).



Фото 4. На врученні Державної Премії України в галузі науки і техніки (2011)

Професор О. Г. Наконечний успішно поєднує науково-педагогічну роботу з активною громадською діяльністю. Із 2003 р. – академік-секретар відділення інформатики та системного аналізу, а із 2014 р. – президент громадської організації "Академія наук вищої школи України", що об'єднує вчених закладів вищої освіти України для сприяння в проведенні фундаментальних і прикладних досліджень у галузі природничих, технічних та соціо-гуманітарних наук, обміну науковим та педагогічним досвідом, захисту їх фахових інтересів.

Відзначаючи 80-річний ювілей Олександра Григоровича Наконечного, ми вшановуємо не лише видатного вченого, а й людину, яка зуміла поєднати глибокий інтелект із рідкісною людяністю, вимогливість – із доброзичливістю, академічність – із відкритістю до світу та нових ідей. Його внесок у розвиток української кібернетики, прикладної математики й підготовку наукових кадрів залишиться важливою сторінкою історії вітчизняної науки. Колеги й учні, віддаючи шану здобуткам професора О. Г. Наконечного, щиро вітають його з ювілеєм, зичать нових творчих здобутків, довгих, плідних і щасливих років життя.

Внесок авторів: Олена Кашпур, Михайло Згуровський – концептуалізація; методика; Олена Капустян – збір і перевірка емпіричних даних; Наталія Панкратова – емпіричне дослідження; Петро Зінко, Марія Лосева – аналіз джерел, підготовка огляду літератури або теоретичних основ дослідження.

Джерела фінансування. Фінансування дослідження частково забезпечено Київським національним університетом імені Тараса Шевченка.

Вибрані роботи О. Г. Наконечного

- Капустян, О. А., Капустян О. В., & Наконечний, О. Г. (2021). *Оптимальне керування та гарантоване оцінювання у розподілених системах з малим параметром*. ВПЦ "Київський університет".
- Марценюк, В. П., & Наконечний, О. Г. (2009). *Моделі та методи популяційної динаміки в програмному середовищі підтримки системних медичних досліджень*. ТДМУ, "Укрмедкнига".
- Наконечний, О. Г. (2000). Мінімаксні оцінки функціоналів від правих частин варіаційних рівнянь з параметром. *Вісник Київ. ун-ту. Серія: Кібернетика*, 1, 27–34.
- Наконечний, О. Г., & Верес, М. М. (2007). *Мінімаксні методи оцінювання в лінійних задачах із параметром*. "Логос".
- Наконечний, О. Г., & Жук, С. М. (2009). *Оцінювання розв'язків алгебраїчно-диференціальних рівнянь в умовах невизначеності*. Національний університет водного господарства та природокористування.
- Наконечний, О. Г., & Качинський А. В. (1996). *Стійкість екосистем та проблема нормування в екологічній безпеці України*. Національний інститут стратегічних досліджень.
- Наконечний, О. Г., Марценюк, В. П., & Андрущак І. Є. (2014). *Інформаційні технології прийняття рішень, оптимізації та керування в системних медичних дослідженнях*. Луцький національний технічний університет.
- Bekesiene, S., Nakonechnyi, O., Kapustyan, O., Smaliukiene, R., & Vaičaitienė, R. (2023a). Determining the main resilience competencies by applying fuzzy logic in military organization. *Mathematics*, 11(10), 2270.
- Bekesiene S., Nakonechnyi O., Kapustian O., Shevchuk I., & Loseva M. (2023b). Forecast and optimization of dynamics of number of persons with stress syndrome under uncertainties. *Вісник Київського національного ун-ту. Серія: Фізико-математичні науки*, 2, 195–199.
- Bekesiene, S., Smaliukienė, R., Vaičaitienė, R., Bagdžiūnienė, D., Kanapekaitė, R., Kapustian, O., & Nakonechnyi, O. (2024). Prioritizing competencies for soldier's mental resilience: an application of integrative fuzzy-trapezoidal decision-making trial and evaluation laboratory in updating training program. *Frontiers in Psychology*, 14, 1239481.
- Kapustian, O. A., Nakonechnyi, O. G. (2020). Approximate minimax estimation of functionals of solutions to the wave equation under nonlinear observations. *Cybernetics and Systems Analysis*, 56(5), 793–801.
- Mallet, V., Nakonechny, O., & Zhuk, S. (2013). Minimax filtering for sequential aggregation: Application to ensemble forecast of ozone analyses. *Journal of Geophysical Research Atmospheres*, 118(19), 11–303.
- Marzeniuk, V. P., & Nakonechny, O. G. (2003). *System analysis methods of medical and biological processes*. Ternopil, Ukrmedknyga.
- Nakonechnyi, O. G., Kapustian, O. A., & Chikrii, A. O. (2019). Approximate guaranteed mean square estimates of functionals on solutions of parabolic problems with fast oscillating coefficients under nonlinear observations. *Cybernetics and Systems Analysis*, 55(5), 785–795.
- Nakonechnyi, O., Kapustian, O., Podlipenko, Y. (2019). Minimax estimation of solutions of the first order linear hyperbolic systems with uncertain data. *Statistics, Optimization and Information Computing*, 7(4), 695–708.
- Nakonechnyi, O. G., Kudin, G. I., Zinko, P. M., & Zinko, T. P. (2023a). Guaranteed root mean square estimates of linear matrix equations solutions under conditions of uncertainty. *Mathematical Modeling and Computing*, 10(2), 474–486.
- Nakonechnyi, O. G., Kudin, G. I., Zinko, P. M., & Zinko, T. P. (2023b). Guaranteed root-mean-square estimates of the forecast of matrix observations under conditions of statistical uncertainty. *System research and Information technologies*, 2, 86–103.
- Nakonechnyi, O., Martsenyuk, V., Karpinski, M., & Klos-Witkowska, A. (2021). On qualitative analysis of lattice dynamical system of two-and three-dimensional biopixels array: bifurcations and transition to "chaos". *Springer Proceedings in Mathematics and Statistics*, 364, 23–43.
- Nakonechnyi, O., Mashchenko, S., Chikrii, V. (2018). Motion control under conflict condition. *Journal of Automation and Information Sciences*, 50(1), 54–75.
- Nakonechnyi, O., & Podlipenko, Y. (2021). *Guaranteed estimation problems in the theory of linear ordinary differential equations with uncertain data*. River Publishers Series in Mathematical and Engineering Sciences, River Publishers.
- Nakonechnyi, O., & Podlipenko, Y. (2022). Guaranteed a posteriori estimation of unknown right-hand sides of linear periodic systems of ODEs. *Applicable Analysis*, 101(17), 6212–6221.
- Nakonechnyi, O., Podlipenko, Y., Shestopalov, Y. (2021a). Guaranteed estimation of solutions to the cauchy problem when the restrictions on unknown initial data are not posed. *Mathematics*, 9(24), 3218.
- Nakonechnyi, O., Podlipenko, Y., & Shestopalov, Y. (2021b). Guaranteed a posteriori estimation of uncertain data in exterior Neumann problems for Helmholtz equation from inexact indirect observations of their solutions. *Inverse Problems in Science and Engineering*, 29(4), 525–535.
- Nakonechnyi, O., Shestopalov, Y., & Podlipenko, Y. (2014). Estimation of parameters of boundary value problems for linear ordinary differential equations with uncertain data. *Advances in Pure Mathematics*, 4(4), 118–146.
- Nakonechnyi, O. G., & Zinko, P. M. (2023). Estimates of matrix solutions of operator equations with random parameters under uncertainties. *Matematychni Studii*, 60(2), 208–222.
- Nakonechnyi, A., Zinko, P., & Shevchuk, Y. (2018). Guaranteed estimation of nonstationary parameters of difference equations under uncertainty. *Journal of Automation and Information Sciences*, 50(11), 1–18.
- Nakonechnyi, O. G., Zinko P. M., & Zinko, T. P. (2025a). Estimates of solutions for ambiguously solvable linear matrix equations. *Mathematical Modeling and Computing*, 12(1), 10–21.
- Nakonechnyi, O. G., Zinko, P. M., & Zinko, T. P. (2025b). Linear mean square estimates of the spectral functions of mean values of nonstationary random processes under uncertainty. *Cybernetics and Systems Analysis*, 61, 782–792.
- Zhuk, S., & Nakonechnyi, O. (2018). Minimax state estimates for abstract Neumann problems. *Minimax Theory and its Applications*, 3(1), 1–21.
- Zhuk, S., Polyakov, A., & Nakonechnyi, O. (2017). note on minimax sliding mode control design for linear systems. *IEEE Transactions on Automatic Control*, 62(7), 3395–3400.

Отримано редакцією журналу / Received: 25.11.25

Прорецензовано / Revised: 29.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Olena KASHPUR¹, DSc (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-7091-8851
e-mail: olena.kashpur@knu.ua

Michael ZGUROVSKY², DSc (Tech.), Prof., Active Member of the National Academy of Sciences of Ukraine
ORCID ID: 0000-0001-5896-7466
e-mail: mzz@kpi.ua

Olena KAPUSTIAN¹, DSc (Phys. & Math.), Senior Researcher
ORCID ID: 0000-0002-2629-0750
e-mail: olenakapustian@knu.ua

Nataliya PANKRATOVA², DSc (Tech.), Prof., Corresponding Member of the National Academy of Sciences of Ukraine
ORCID ID: 0000-0002-6372-5813
e-mail: natalidmp@gmail.com

Petro ZINKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-5111-4417
e-mail: petro.zinko@knu.ua

Mariia LOSIEVA¹, Leading Engineer
ORCID ID: 0000-0002-2282-206X
e-mail: mariia.losieva@knu.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

GLORIOUS ANNIVERSARY OF AN OUTSTANDING FIGURE: TO THE 80TH ANNIVERSARY OF THE BIRTH OF OLEKSANDR GRIGOROVYCH NAKONECHNYI

The article is dedicated to the 80th anniversary of the birth of Oleksandr Hryhorovych Nakonechny, the outstanding Ukrainian scientist, mathematician and teacher, professor at the Taras Shevchenko National University of Kyiv. His life and professional path are outlined. The article highlights his academic career, scientific and pedagogical activities, as well as his significant contribution to the development of mathematics education and science in Ukraine, in particular, in the systems analysis, decision-making theory, stochastic processes with uncertainty. The main attention is paid to the review of the fundamental scientific results of O. G. Nakonechny. A list of selected scientific works of the jubilee is provided.

Keywords: *Oleksandr Nakonechnyi, stochastic processes, systems analysis, decision-making theory, uncertainty conditions.*

Автори (О. Кашпур та О. Капустян) є членами редколегії видання, тому вони не брали участі у рецензуванні та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors (O. Kashpur and O. Kapustian) are members of the journal editorial board; therefore, they did not take part in the peer-review process or in the decision to publish this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

АЛГЕБРА, ГЕОМЕТРІЯ ТА ТЕОРІЯ ЙМОВІРНОСТЕЙ

UDC 515.47

DOI: <https://doi.org/10.17721/1812-5409.2026/1.2>

ANURAG, Master's Student
ORCID ID: 0009-0003-8391-7622
e-mail: anurag_kumar23@mru.ac.in
Manav Rachna University, Faridabad, India

Aparna VYAS, PhD, Prof.
ORCID ID: 0000-0002-5301-8234
e-mail: aparna@mru.edu.in
Manav Rachna University, Faridabad, India

Gibak KIM, PhD, Prof.
ORCID ID: 0000-0001-5114-4117
e-mail: imkgb27@ssu.ac.kr
Soongsil University, Seoul, Republic of Korea

EXTENDED UNIQUE FIXED POINT THEOREM IN COMPLETE QUASI-METRIC SPACE

The primary goal of this research is to present novel fixed point theorems within the framework of complete quasi-metric spaces. Fixed point theory is a critical area connecting pure and applied mathematics. While classical results like the Banach contraction principle are well-established in metric spaces, this paper extends these concepts to the more generalized quasi-metric space, which lacks the symmetric condition of a standard metric space. We prove a unique fixed point theorem for continuous mappings in a complete quasi-metric space under a newly defined, comprehensive contractive condition. This result significantly generalizes and extends several existing fixed-point theorems in the literature. An example is provided to illustrate the applicability of our findings. Our work contributes to the development of fixed point theory in asymmetric spaces, which have applications in diverse fields such as shape-memory alloys and Hamilton - Jacobi equations.

Key words: fixed point, contraction mapping, complete metric space, complete quasi metric space.

AMS 2020 classification: 47H10, 54H25.

Introduction

Fixed point theory is a fascinating branch of mathematics with deep connections across pure and applied disciplines. Its origins trace back to the early 20th century, notably with Banach's pioneering work in 1922 on the Banach contraction principle (Banach, 1922). This principle has wide-ranging applications in areas like differential and integral equations, metric spaces, and numerical analysis. Inspired by this, many researchers have proven various fixed point theorems in different mathematical spaces.

Following Banach, Kannan in 1968 (Kannan, 1968) and Chatterjea in 1972 (Chatterjea, 1972) introduced similar results on metric spaces, which were later generalized for complete metric spaces by Jojar et al. (Jojar et al., 2022). A pivotal development in the generalization of metric spaces was the introduction of quasi-metric spaces, also known as asymmetric metric spaces (Wilson, 1931). These spaces are defined without the symmetric condition of a standard metric, making them suitable for modeling phenomena where directional distance matters.

The relevance of quasi-metric spaces extends far beyond theoretical mathematics, finding numerous recent and significant applications in both pure and applied domains. These applications include, but are not limited to, the modeling of complex behaviors in shape-memory alloys (Mielke, & Roubíček, 2003), addressing questions of existence and uniqueness in Hamilton-Jacobi equations (Mennucci, 2013), facilitating automated taxonomy construction (Woon, & Madnick, 2009), providing robust models for material failure (Rieger, & Zimmer, 2009), and analyzing rate-independent models for plasticity (Mainik, & Mielke, 2005). The utility of fixed point theorems in these contexts often lies in proving the existence and uniqueness of solutions to equations that arise from such models.

This research article aims to further advance fixed point theory by establishing new fixed point theorems for continuous mappings in complete quasi-metric spaces. Our results introduce a novel contractive condition that extends and generalizes several well-known existing results in the literature. We provide a suitable example to demonstrate the applicability of our findings. This work contributes to the theoretical understanding of fixed points in asymmetric spaces and their potential use in various scientific and engineering problems.

1. Preliminaries

In this section, we introduce some basic definitions and an example that will be used further.

Definition 1. (Wilson, 1931): Let X be a nonempty set and let $d : X \times X \rightarrow [0, \infty)$ be a function satisfying the following conditions:

1. $d(x, y) = d(y, x) = 0 \implies x = y$;
2. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in X$.

Then, d is called quasi-metric (or q-metric) on X and the ordered pair (X, d) is called a quasi-metric space.

Example. (Ayoob, Chuan, & Mlaiki, 2023) An example of a quasi-metric space is a circle, where the distance d between two points A & B on the circle is defined as the length of the shortest arc measured in a clockwise direction.

Definition 2. (Jleli, & Samet, 2018) A sequence $\{x_n\}$ in a quasi-metric space (X, d) is said to be q -convergent to z if $\lim_{n \rightarrow \infty} d(x_n, z) = 0 = \lim_{n \rightarrow \infty} d(z, x_n)$. Here, z is the q -limit point of a sequence $\{x_n\}$.

Definition 3. (Jleli, & Samet, 2018) A sequence $\{x_n\}$ in a quasi-metric space (X, d) is said to be Cauchy sequence if for a given $\epsilon > 0$, there exists a $n_0 \in \mathbb{N}$ such that for all $m, n \geq n_0$,

$$d(x_n, x_m) < \epsilon \text{ or } d(x_m, x_n) < \epsilon.$$

Definition 4. (Wilson, 1931) A quasi-metric space (X, d) is considered complete if every Cauchy sequence in X converges to a point in X .

Definition 5. (Banach, 1922) Let (X, d) be a metric space and let $f : X \rightarrow X$ be a function. Then a point $x \in X$ is a fixed point of f if $f(x) = x$.

2. Main Results

In this section, we establish some fixed point theorems for continuous mappings in a complete quasi-metric space. Our results extend and generalize some well-known existing results in the literature in (Banach, 1922; Chatterjea, 1972; Jojar et al., 2022; Kannan, 1968; Mittal et al., 2023).

Theorem 1. Let (X, d) be a complete quasi-metric space and $T : X \rightarrow X$ be a continuous self-mapping satisfying the condition

$$\begin{aligned} d(Tx, Ty) \leq & a_1 d(x, y) + a_2 [d(x, Tx) + d(y, Ty)] + a_3 [d(x, Ty) + d(y, Tx)] + \\ & + a_4 \left[\frac{d(x, y)d(x, Ty)}{d(x, y) + d(y, Ty)} \right] + a_5 \left[\frac{d(y, Ty)(1 + d(x, Tx))}{1 + d(x, y)} \right] + a_6 \left[\frac{d(x, Ty)d(y, Ty)}{d(x, y) + d(y, Ty)} \right] \end{aligned} \quad (1)$$

for all $x, y \in X$, $a_1, a_2, a_3, a_4, a_5, a_6 \geq 0$ and $a_1 + 2a_2 + 2a_3 + a_4 + a_5 + a_6 < 1$. Thus, T has a unique fixed point in X .

Proof. Let $x_0 \in X$ be an arbitrary element in X and define an iterative sequence as

$$x_1 = Tx_0, x_2 = Tx_1 = T^2x_0, \dots, x_n = Tx_{n-1} = T^n x_0$$

for all $n \in \mathbb{N}$. Then

$$\begin{aligned} d(x_1, x_2) &= d(Tx_0, Tx_1) \leq \\ &\leq a_1 d(x_0, x_1) + a_2 [d(x_0, Tx_0) + d(x_1, Tx_1)] + a_3 [d(x_0, Tx_1) + d(x_1, Tx_0)] + \\ &+ a_4 \left[\frac{d(x_0, x_1)d(x_0, Tx_1)}{d(x_0, x_1) + d(x_1, Tx_1)} \right] + a_5 \left[\frac{d(x_1, Tx_1)(1 + d(x_0, Tx_0))}{1 + d(x_0, x_1)} \right] + a_6 \left[\frac{d(x_0, Tx_1)d(x_1, Tx_1)}{d(x_0, x_1) + d(x_1, Tx_1)} \right] = \\ &= a_1 d(x_0, x_1) + a_2 [d(x_0, x_1) + d(x_1, x_2)] + a_3 [d(x_0, x_2) + d(x_1, x_1)] + \\ &+ a_4 \left[\frac{d(x_0, x_1)d(x_0, x_2)}{d(x_0, x_1) + d(x_1, x_2)} \right] + a_5 \left[\frac{d(x_1, x_2)(1 + d(x_0, x_1))}{1 + d(x_0, x_1)} \right] + a_6 \left[\frac{d(x_0, x_2)d(x_1, x_2)}{d(x_0, x_1) + d(x_1, x_2)} \right] \leq \\ &\leq a_1 d(x_0, x_1) + a_2 d(x_0, x_1) + a_2 d(x_1, x_2) + a_3 [d(x_0, x_1) + d(x_1, x_2)] + a_4 d(x_0, x_1) + a_5 d(x_1, x_2) + a_6 d(x_1, x_2) = \\ &= [a_1 + a_2 + a_3 + a_4]d(x_0, x_1) + [a_2 + a_3 + a_5 + a_6]d(x_1, x_2). \end{aligned}$$

Therefore, $[1 - (a_2 + a_3 + a_5 + a_6)]d(x_1, x_2) \leq [a_1 + a_2 + a_3 + a_4]d(x_0, x_1)$. This implies

$$d(x_1, x_2) \leq \left[\frac{a_1 + a_2 + a_3 + a_4}{1 - (a_2 + a_3 + a_5 + a_6)} \right] d(x_0, x_1).$$

Let

$$k = \left[\frac{a_1 + a_2 + a_3 + a_4}{1 - (a_2 + a_3 + a_5 + a_6)} \right].$$

Then, $d(x_1, x_2) \leq kd(x_0, x_1)$. Similarly,

$$d(x_2, x_3) \leq k^2 d(x_0, x_1), \quad d(x_3, x_4) \leq k^3 d(x_0, x_1).$$

Proceeding like this we get for any positive integer n , $d(x_n, x_{n+1}) \leq k^n d(x_0, x_1)$.

Now, if p is any positive integer,

$$\begin{aligned} d(x_n, x_{n+p}) &\leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \dots + d(x_{n+p-1}, x_{n+p}) \leq \\ &\leq k^n d(x_0, x_1) + k^{n+1} d(x_0, x_1) + \dots + k^{n+p-1} d(x_0, x_1) = \\ &= k^n (1 + k + \dots + k^{p-1}) d(x_0, x_1) \leq \\ &\leq \left[\frac{k^n}{1 - k} \right] d(x_0, x_1) \leq \left[\frac{k^n}{1 - k} \right] d(x_0, Tx_0). \end{aligned}$$

Since $a_1 + 2a_2 + 2a_3 + a_4 + a_5 + a_6 < 1$, then $0 \leq k < 1$. So, $d(x_n, x_{n+p}) \rightarrow 0$ as $n \rightarrow \infty$. Therefore, $\{x_n\}$ is a Cauchy sequence. As X is a complete quasi-metric space, we have

$$\lim_{n \rightarrow \infty} x_n = p \in X.$$

Since T is continuous, $\lim_{n \rightarrow \infty} Tx_n = Tp \Rightarrow \lim_{n \rightarrow \infty} Tx_{n+1} = Tp$. Thus,

$$Tp = p.$$

Hence, p is a fixed point of T .

Uniqueness: If $p \in X$ is a fixed point of T . Then by Equation (1)

$$\begin{aligned} d(p, p) &= d(Tp, Tp) \leq \\ &\leq a_1 d(p, p) + a_2 [d(p, Tp) + d(p, Tp)] + a_3 [d(p, Tp) + d(p, Tp)] + \\ &+ a_4 \left[\frac{d(p, p)d(p, Tp)}{d(p, p) + d(p, Tp)} \right] + a_5 \left[\frac{d(p, Tp)(1 + d(p, Tp))}{1 + d(p, p)} \right] + a_6 \left[\frac{d(p, Tp)d(p, Tp)}{d(p, p) + d(p, Tp)} \right] \leq \\ &\leq \left[a_1 + 2a_2 + 2a_3 + \frac{a_4}{2} + a_5 + \frac{a_6}{2} \right] d(p, p). \end{aligned}$$

Since $a_1 + 2a_2 + 2a_3 + \frac{a_4}{2} + a_5 + \frac{a_6}{2} < 1$ and $d(p, p) \geq 0$, we have $d(p, p) = 0$. Thus, $d(p, p) = 0$ if p is a fixed point of T . Suppose that p and q are two fixed points of T ($p \neq q$); that is, $p = Tp$ and $q = Tq$. We have

$$\begin{aligned} d(p, q) &= d(Tp, Tq) \leq \\ &\leq a_1 d(p, q) + a_2 [d(p, Tp) + d(q, Tq)] + a_3 [d(p, Tq) + d(q, Tp)] + \\ &+ a_4 \left[\frac{d(p, q)d(p, Tq)}{d(p, q) + d(q, Tq)} \right] + a_5 \left[\frac{d(q, Tq)(1 + d(p, Tp))}{1 + d(p, q)} \right] + a_6 \left[\frac{d(p, Tq)d(q, Tq)}{d(p, q) + d(q, Tq)} \right]. \end{aligned}$$

$$d(p, q) \leq [a_1 + a_3 + a_4]d(p, q) + a_3 d(q, p). \quad (2)$$

Similarly,

$$d(q, p) \leq [a_1 + a_3 + a_4]d(q, p) + a_3 d(p, q). \quad (3)$$

Subtracting equation (3) from (2), we get

$$|d(p, q) - d(q, p)| \leq |a_1 + a_4| |d(p, q) - d(q, p)|.$$

Since, $|a_1 + a_4| < 1$, we have $|d(p, q) - d(q, p)| = 0$. This implies

$$d(p, q) = d(q, p). \quad (4)$$

Putting equation (4) in (2), we get

$$d(p, q) \leq [a_1 + a_3 + a_4]d(p, q).$$

Since, $0 < a_1 + a_3 + a_4 < 1$, we have $d(p, q) = 0$. Similarly, $d(q, p) = 0$. Hence, $p = q$. $\square$

Example. Let $X = \mathbb{R}$ with q-metric defined by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ x^2 & \text{if } x \neq y \end{cases} \quad \forall x, y \in X$$

and let $T : X \rightarrow X$ such that $Tx = \frac{x}{2}$ be a continuous self-mapping. Suppose $a_1 = \frac{1}{3}$, $a_2 = \frac{1}{6}$, $a_3 = \frac{1}{12}$, $a_4 = \frac{1}{15}$, $a_5 = \frac{1}{20}$, $a_6 = \frac{1}{25}$, then T satisfies all the conditions of Theorem 1, and $x = 0$ is the unique fixed point of T in X .

Remark 1. In Theorem 1,

1. If $a_2 = a_3 = a_4 = a_5 = a_6 = 0$, then it extends the result of (Banach, 1922).
2. If $a_1 = a_2 = a_4 = a_5 = a_6 = 0$, then it extends the result of (Chatterjea, 1972).
3. If $a_1 = a_4 = a_5 = a_6 = 0$, then it extends the result of (Jojar et al., 2022).
4. If $a_1 = a_3 = a_4 = a_5 = a_6 = 0$, then it extends the result of (Kannan, 1968).
5. If $a_5 = 0$, then it extends the result of (Mittal et al., 2023).

Discussions and conclusion

In this paper, we proved a unique fixed point theorem for continuous self-mappings on complete quasi-metric spaces. The result is based on a generalized contraction condition involving several distance terms and extends classical fixed point theorems such as those by Banach, Kannan, and Chatterjea. Unlike many earlier results, our theorem does not require

additional structures like convexity or dislocation. An example was also given to support the main result. We hope this work will be useful for further studies in fixed point theory on asymmetric spaces.

Authors' contribution: Anurag – conceptualization, methodology, validation, formal analysis, original draft preparation; Aparna, Vyas – conceptualization, formal analysis, original draft preparation; Gibak, Kim – review and editing.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Ayoob, I., Chuan, N. Z., & Mlaiki, N. (2023). Quasi m-metric spaces. *AIMS Mathematics*, 8(5), 10228–10248. <https://doi.org/10.3934/math.2023518>
- Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3, 133–181. <https://doi.org/10.4064/fm-3-1-133-181>
- Chatterjea, S. K. (1972). Fixed point theorem. *Comptes Rendus de l'Académie Bulgare des Sciences*, 25, 727–730.
- Jleli, M., & Samet, B. (2018). Fixed point theorem. *Comptes Rendus de l'Académie Bulgare des Sciences*, 9, 118–133. <https://doi.org/10.22436/jnsa.009.06.31>
- Jogar, G., Dolhare, U., Basude, S., Darkunde, N., & Swami, P. (2022). Some results on unique fixed point theorems in complete metric space. *International Journal of Mathematics Trends and Technology*, 68(6), 111–116. <https://doi.org/10.14445/22315373/IJMTT-V68I6P513>
- Kannan, R. (1968). Some results on fixed points. *Bulletin of the Calcutta Mathematical Society*, 60, 71–76.
- Mainik, A., & Mielke, A. (2005). Existence results for energetic models for rate-independent systems. *Calculus of Variations and Partial Differential Equations*, 22, 73–99. <https://doi.org/10.1007/s00526-004-0267-8>
- Mennucci, A. C. G. (2013). On asymmetric distances. *Analysis and Geometry in Metric Spaces*, 1, 200–231. <https://doi.org/10.2478/agms-2013-0004>
- Mielke, A., & Roubíček, T. (2003). A rate-independent model for inelastic behavior of shape memory alloys. *Multiscale Modeling & Simulation*, 1, 571–597. <https://doi.org/10.1137/S1540345903422860>
- Mittal, N., Kumar, P., Ansari, Z. K., & Vyas, A. (2023). Some fixed point theorems satisfying generalized contraction conditions in dislocated quasi metric space. *International Journal of Nonlinear Analysis and Applications*, 16(1), 1–7. <http://dx.doi.org/10.22075/ijnaa.2023.28150.3816>
- Rieger, M. O., & Zimmer, J. (2009). Young measure flow as a model for damage. *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, 60, 1–32. <https://doi.org/10.1007/s00033-008-7016-3>
- Wilson, W. A. (1931). On quasi metric spaces. *American Journal of Mathematics*, 53, 675–684. <https://doi.org/10.2307/2371174>
- Woon, W. L., & Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. *Analysis and Geometry in Metric Spaces*, 21, 91–111. <https://doi.org/10.2139/ssrn.1256562>

Отримано редакцією журналу / Received: 05.08.25

Прорецензовано / Revised: 26.11.25

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

АНУРАГ, магістр
ORCID ID: 0009-0003-8391-7622
e-mail: anurag_kumar23@mru.ac.in
Університет Манава Рачна, Фаридабад, Індія

Апарна В'ЯС, д-р філософії, проф.
ORCID ID: 0000-0002-5301-8234
e-mail: aparna@mru.edu.in
Університет Манава Рачна, Фаридабад, Індія

Гібак КІМ, д-р філософії, проф.
ORCID ID: 0000-0001-5114-4117
e-mail: imkgb27@ssu.ac.kr
Університет Сунсіль, Сеул, Республіка Корея

РОЗШИРЕНА ТЕОРЕМА ПРО ЄДИНУ НЕРУХОМУ ТОЧКУ В ПОВНОМУ КВАЗІМЕТРИЧНОМУ ПРОСТОРІ

Основною метою дослідження є представлення нових теорем про нерухому точку в повних квазіметричних просторах. Теорія нерухомих точок є важливою галуззю, що поєднує теоретичну та прикладну математику. Хоча класичні результати, такі як принцип стискальних відображень Банаха, добре відомі для метричних просторів, у цій роботі зазначені концепції поширюються на більш загальний квазіметричний простір, що не задовольняє симетричну умову стандартного метричного простору. Ми доводимо теорему про єдину нерухому точку для неперервних відображень у повному квазіметричному просторі за нововизначеною, узагальненою умовою стискання. Цей результат суттєво узагальнює та розширює низку відомих теорем про нерухомі точки, згаданих у літературі. Для ілюстрації застосовності наших результатів наведено приклад. Наша робота робить внесок у розвиток теорії нерухомих точок в асиметричних просторах, які знаходять застосування у різних галузях, зокрема в дослідженні сплавів з ефектом пам'яті форми та рівняннях Гамільтона - Якобі.

Ключові слова: нерухома точка, стискальне відображення, повний метричний простір, повний квазіметричний простір.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.3>

Oleksandr BORYSENKO, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-6670-8605
e-mail: oleksandr_borysenko@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Olga BORYSENKO, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-2099-2911
e-mail: oborisenko1373@gmail.com
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

DYNAMICS OF A NON-AUTONOMOUS STOCHASTIC MUTUALISM MODEL WITH JUMPS

In nature, we can find many examples where the interaction of two or more species is to the advantage of all. Mutualism occurs when one species provides some benefit in exchange for some benefit, for example, ants and aphids, in which the ants obtain honeydew food resources excreted by aphids while the aphids obtain increased survival by the non-trophic service of ant defense against natural enemies of the aphids. Mathematical deterministic mutualism models are widely used to study the dynamics of such population systems. In real life, there is a lot of randomness and stochasticity, such as environmental noise, so using stochastic models is more suitable. Suppose we want to take into account abrupt environmental perturbations, such as epidemics, fires, earthquakes, etc., in the considered models. In that case, we must introduce Poisson noises into population models to describe such discontinuous systems. So, we take into account not only "small" jumps, corresponding to the centered Poisson measure, but also the "large" jumps, corresponding to the non-centered Poisson measure. The existence and uniqueness of the global positive solution are proved for the system of stochastic differential equations describing a non-autonomous mutualism model disturbed by white noise, centered, and non-centered Poisson noises. We obtain sufficient conditions of stochastic ultimate boundedness, stochastic permanence, non-persistence in the mean, weak and strong persistence in the mean, and extinction of the solution to the considered system.

Key words: *stochastic mutualism model, global solution, stochastic ultimate boundedness, stochastic permanence, extinction, non-persistence in the mean, weak and strong persistence in the mean.*

AMS 2020 classification: 92D25, 60H10, 60H30.

Introduction

In nature we can find many examples where the interaction of two or more species is to the advantage of all. Mutualism occurs when one species provides some benefit in exchange for some benefit, for example, ants and aphids, in which the ants obtain honeydew food resources excreted by aphids while the aphids obtain increased survival by the non-trophic service of ant defense against natural enemies of the aphids. Mathematical deterministic mutualism models are widely used to study the dynamics of such population systems.

In (May, 1976), the following deterministic two-species mutualism model

$$\begin{aligned} dx(t) &= x(t) \left(r_1 - \frac{b_1 x(t)}{K_1 + y(t)} - \varepsilon_1 x(t) \right) dt, \\ dy(t) &= y(t) \left(r_2 - \frac{b_2 y(t)}{K_2 + x(t)} - \varepsilon_2 y(t) \right) dt \end{aligned} \quad (1)$$

is proposed. Here $x(t), y(t)$ denote the population densities of each species at time t , $r_i, K_i, b_i, \varepsilon_i$ ($i = 1, 2$) are positive constants, r_1, r_2 denote the intrinsic growth rate of species $x(t), y(t)$ respectively, K_1 is the capability of species $x(t)$ being short of $y(t)$, similarly K_2 is the capability of species $y(t)$ being short of $x(t)$.

In the paper (Li et al., 2015) the stochastic version of the model (1) is considered in the following form

$$\begin{aligned} dx(t) &= x(t) \left(r_1 - \frac{b_1 x(t)}{K_1 + y(t)} - \varepsilon_1 x(t) \right) dt + \alpha_1 x(t) dw_1(t), \\ dy(t) &= y(t) \left(r_2 - \frac{b_2 y(t)}{K_2 + x(t)} - \varepsilon_2 y(t) \right) dt + \alpha_2 y(t) dw_2(t), \end{aligned} \quad (2)$$

where $w_1(t), w_2(t)$ are mutually independent Wiener processes. The authors proved the existence and uniqueness of a global positive solution to the system (2). It is proven that the solution to the system (2) is stochastically ultimate bounded, stochastically permanent, uniformly Hölder-continuous. The sufficient conditions for persistence in the mean and extinction of the solution to the considered system are obtained.

Population systems may suffer abrupt environmental perturbations, such as epidemics, fires, earthquakes, etc. So, it is natural to introduce Poisson noises into the population model for describing such discontinuous systems. In the paper (Li, Gao, & Wang, 2016), the authors consider the stochastic mutualism model with white and centered Poisson noises

$$\begin{aligned} dx(t) &= x(t^-) \left[\left(r_1(t) - \frac{b_1(t)x(t)}{K_1(t)+y(t)} - \varepsilon_1(t)x(t) \right) dt + \alpha_1(t)dw_1(t) + \int_{\mathbb{Y}} \gamma_1(t, z) \tilde{\nu}(dt, dz) \right], \\ dy(t) &= y(t^-) \left[\left(r_2(t) - \frac{b_2(t)y(t)}{K_2(t)+x(t)} - \varepsilon_2(t)y(t) \right) dt + \alpha_2(t)dw_2(t) + \int_{\mathbb{Y}} \gamma_2(t, z) \tilde{\nu}(dt, dz) \right]. \end{aligned} \quad (3)$$

Here $x(t^-)$, $y(t^-)$ are the left limit of $x(t)$ and $y(t)$ respectively, $r_i(t)$, $b_i(t)$, $K_i(t)$, $\varepsilon_i(t)$, $\alpha_i(t)$, $i = 1, 2$ are all positive, continuous and bounded functions, $\mathbb{Y}$ is a measurable subset of $(0, +\infty)$, $w_i(t)$, $i = 1, 2$ are independent standard one-dimensional Wiener processes, $\tilde{\nu}(t, A) = \nu(t, A) - t\Pi(A)$ is the centered Poisson measure independent on $w_i(t)$, $i = 1, 2$, $E[\nu(t, A)] = t\Pi(A)$, $\Pi(\mathbb{Y}) < \infty$, $\gamma_i(t, z)$, $i = 1, 2$ are random, measurable, bounded, continuous on t . The global existence and uniqueness of the positive solution to the problem (3) are proved. The sufficient conditions of stochastic ultimate boundedness, stochastic permanence, persistence in the mean, and extinction of the solution to the system (3) are obtained.

In this paper, we consider the stochastic mutualism model with jumps generated by centered and non-centered Poisson measures. So, we take into account not only "small" jumps, corresponding to the centered Poisson measure, but also the "large" jumps, corresponding to the non-centered Poisson measure. This model is driven by the system of stochastic differential equations

$$\begin{aligned} dx_i(t) = & x_i(t) \left[r_i(t) - \frac{b_i(t)x_i(t)}{K_i(t) + x_{3-i}(t)} - c_i(t)x_i(t) \right] dt + \sigma_i(t)x_i(t)dw_i(t) + \\ & + \int_{\mathbb{R}} \gamma_i(t, z)x_i(t^-)\tilde{\nu}_1(dt, dz) + \int_{\mathbb{R}} \delta_i(t, z)x_i(t^-)\nu_2(dt, dz), \quad x_i(0) = x_{i0} > 0, \quad i = 1, 2, \end{aligned} \quad (4)$$

where $w_i(t)$, $i = 1, 2$ are independent standard one-dimensional Wiener processes, $\tilde{\nu}_1(t, A) = \nu_1(t, A) - t\Pi_1(A)$, $\nu_i(t, A)$, $i = 1, 2$ are independent Poisson measures, which are independent on $w_i(t)$, $i = 1, 2$, $E[\nu_i(t, A)] = t\Pi_i(A)$, $i = 1, 2$, $\Pi_i(A)$, $i = 1, 2$ are finite measures on the Borel sets A in $\mathbb{R}$.

We prove that system (4) has a unique, positive, global solution for any positive initial value, and this solution is stochastically ultimately bounded. The sufficient conditions for stochastic permanence, extinction, non-persistence in the mean, weak and strong persistence in the mean of the solution to the system (4) are derived.

It is worth noting that the impact of centered and non-centered Poisson noises to the stochastic predator-prey model that depends on the population density of the predator is studied in the paper (Borysenko, & Borysenko, 2023).

In the following we will use the notations $X(t) = (x_1(t), x_2(t))$, $X_0 = (x_{10}, x_{20})$, $|X(t)| = \sqrt{x_1^2(t) + x_2^2(t)}$, $\mathbb{R}_+^2 = \{X \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$,

$$\beta_i(t) = \frac{\sigma_i^2(t)}{2} + \int_{\mathbb{R}} [\gamma_i(t, z) - \ln(1 + \gamma_i(t, z))] \Pi_1(dz) - \int_{\mathbb{R}} \ln(1 + \delta_i(t, z)) \Pi_2(dz),$$

$i = 1, 2$. For the bounded, continuous functions $f_i(t)$, $t \in [0, +\infty)$, $i = 1, 2$, let us denote

$$f_i \sup = \sup_{t \geq 0} f_i(t), \quad f_i \inf = \inf_{t \geq 0} f_i(t), \quad i = 1, 2, \quad f_{\max} = \max\{f_{1 \sup}, f_{2 \sup}\}, \quad f_{\min} = \min\{f_{1 \inf}, f_{2 \inf}\}.$$

The rest of this paper is organized as follows. In Section 2, we prove the existence of a unique global positive solution to the system (4) and derive some auxiliary results. In Section 3, we prove the stochastic ultimate boundedness of the solution to the system (4), and obtain sufficient conditions under which the solution is stochastically permanent. The sufficient conditions for extinction, non-persistence in the mean, strong and weak persistence in the mean of the solution are derived.

1. Existence of a global solution and some auxiliary lemmas

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $w_i(t)$, $i = 1, 2$, $t \geq 0$ are independent standard one-dimensional Wiener processes on $(\Omega, \mathcal{F}, P)$, and $\nu_i(t, A)$, $i = 1, 2$ are independent Poisson measures defined on $(\Omega, \mathcal{F}, P)$ independent on $w_i(t)$, $i = 1, 2$. Here $E[\nu_i(t, A)] = t\Pi_i(A)$, $i = 1, 2$, $\tilde{\nu}_i(t, A) = \nu_i(t, A) - t\Pi_i(A)$, $i = 1, 2$, $\Pi_i(\cdot)$, $i = 1, 2$ are finite measures on the Borel sets in $\mathbb{R}$. On the probability space $(\Omega, \mathcal{F}, P)$ we consider an increasing, right continuous family of complete sub- σ -algebras $\{\mathcal{F}_t\}_{t \geq 0}$, where $\mathcal{F}_t = \sigma\{w_i(s), \nu_i(s, A), s \leq t, i = 1, 2\}$.

We need the following assumption.

Assumption 1. It is assumed, that $r_i(t) > 0$, $b_i(t) > 0$, $K_i(t) > 0$, $c_i(t) > 0$, $\sigma_i(t)$, $i = 1, 2$ are bounded, continuous on t functions, $c_{\min} > 0$, $K_{\min} > 0$, $\gamma_i(t, z)$, $\delta_i(t, z)$, $i = 1, 2$ are continuous on t functions and $\ln(1 + \gamma_i(t, z))$, $\ln(1 + \delta_i(t, z))$ are well-defined and bounded, $\Pi_i(\mathbb{R}) < \infty$, $i = 1, 2$.

In what follows, we will assume that Assumption 1 holds.

Theorem 1. *There exists a unique global solution $X(t)$ to the system (4) for any initial value $X(0) = X_0 \in \mathbb{R}_+^2$, and $P\{X(t) \in \mathbb{R}_+^2\} = 1$.*

Proof. Let us consider the system of stochastic differential equations

$$\begin{aligned} dv_i(t) = & \left(r_i(t) - \frac{b_i(t) \exp\{v_i(t)\}}{K_i(t) + \exp\{v_{3-i}(t)\}} - c_i(t) \exp\{v_i(t)\} - \beta_i(t) \right) dt + \sigma_i(t) dw_i(t) + \\ & + \int_{\mathbb{R}} \ln(1 + \gamma_i(t, z)) \tilde{\nu}_1(dt, dz) + \int_{\mathbb{R}} \ln(1 + \delta_i(t, z)) \tilde{\nu}_2(dt, dz), \quad v_i(0) = \ln x_{i0}, \quad i = 1, 2. \end{aligned} \quad (5)$$

The coefficients of the system (5) are locally Lipschitz continuous. Therefore, for any initial value $(v_1(0), v_2(0))$ there exists a unique local solution $V(t) = (v_1(t), v_2(t))$ on $[0, \tau_e)$, where $\sup_{t < \tau_e} |V(t)| = +\infty$ (cf. Theorem 6, p. 246, (Gikhman, & Skorokhod, 1982)). So, from the Itô's formula, we derive that the process $X(t) = (\exp\{v_1(t)\}, \exp\{v_2(t)\})$ is a unique,

positive local solution to the equation (4). To show this solution is global, we need to show that $\tau_e = +\infty$ a.s. Let $n_0 \in \mathbb{N}$ be sufficiently large for $x_{i0} \in [1/n_0, n_0], i = 1, 2$.

For any $n \geq n_0$ we define the stopping time $\tau_n = \inf \left\{ t \in [0, \tau_e) : X(t) \notin \left(\frac{1}{n}, n \right) \times \left(\frac{1}{n}, n \right) \right\}$. It is easy to see that τ_n is increasing as $n \rightarrow +\infty$. Denote $\tau_\infty = \lim_{n \rightarrow \infty} \tau_n$, from which $\tau_\infty \leq \tau_e$ a.s. If we prove that $\tau_\infty = \infty$ a.s., then $\tau_e = \infty$ a.s. and $X(t) \in \mathbb{R}_+^2$ a.s. for all $t \in [0, +\infty)$. So we need to show that $\tau_\infty = \infty$ a.s. If it is not true, there are constants $T > 0$ and $\varepsilon \in (0, 1)$, such that $P\{\tau_\infty < T\} > \varepsilon$. Hence, there is $n_1 \geq n_0$ such that

$$P\{\tau_n < T\} > \varepsilon, \quad \forall n \geq n_1. \quad (6)$$

For the non-negative function $V(X) = \sum_{i=1}^2 (x_i - 1 - \ln x_i)$, $X = (x_1, x_2)$, $x_i > 0, i = 1, 2$, by the Itô's formula we obtain

$$\begin{aligned} V(X(T \wedge \tau_n)) = V(X_0) + \sum_{i=1}^2 \left\{ \int_0^{T \wedge \tau_n} \left[(x_i(t) - 1) \left(r_i(t) - \frac{b_i(t)x_i(t)}{K_i(t) + x_{3-i}(t)} - c_i(t)x_i(t) \right) + \beta_i(t) + \right. \right. \\ \left. \left. + \int_{\mathbb{R}} \delta_i(t, z)x_i(t)\Pi_2(dz) \right] dt + \int_0^{T \wedge \tau_n} (x_i(t) - 1)\sigma_i(t)dw_i(t) + \int_0^{T \wedge \tau_n} \int_{\mathbb{R}} [\gamma_i(t, z)x_i(t^-) - \ln(1 + \gamma_i(t, z))]\tilde{\nu}_1(dt, dz) + \right. \\ \left. + \int_0^{T \wedge \tau_n} \int_{\mathbb{R}} [\delta_i(t, z)x_i(t^-) - \ln(1 + \delta_i(t, z))]\tilde{\nu}_2(dt, dz) \right\}. \end{aligned} \quad (7)$$

Under conditions of the theorem, there is a constant $L > 0$, such that

$$\begin{aligned} \sum_{i=1}^2 \left[(x_i - 1) \left(r_i(t) - \frac{b_i(t)x_i}{K_i(t) + x_{3-i}} - c_i(t)x_i \right) + \beta_i(t) + \int_{\mathbb{R}} \delta_i(t, z)x_i\Pi_2(dz) \right] \leq \\ \leq \sum_{i=1}^2 \left[\left(r_{\max} + c_{\max} + \frac{b_{\max}}{K_{\min}} + \delta_{\max}\Pi_2(\mathbb{R}) \right) x_i - c_{\min}x_i^2 + \beta_{\max} \right] \leq L, \end{aligned} \quad (8)$$

where $\delta_{\max} = \max_{i=1,2} \{\sup_{t \geq 0, z \in \mathbb{R}} \delta_i(t, z)\}$. From (7) and (8) we have

$$\begin{aligned} V(X(T \wedge \tau_n)) \leq V(X_0) + L(T \wedge \tau_n) + \sum_{i=1}^2 \left\{ \int_0^{T \wedge \tau_n} (x_i(t) - 1)\sigma_i(t)dw_i(t) + \right. \\ \left. + \int_0^{T \wedge \tau_n} \int_{\mathbb{R}} [\gamma_i(t, z)x_i(t^-) - \ln(1 + \gamma_i(t, z))]\tilde{\nu}_1(dt, dz) + \int_0^{T \wedge \tau_n} \int_{\mathbb{R}} [\delta_i(t, z)x_i(t^-) - \ln(1 + \delta_i(t, z))]\tilde{\nu}_2(dt, dz) \right\}. \end{aligned}$$

Whence taking expectations, yields

$$E[V(X(T \wedge \tau_n))] \leq V(X_0) + LT. \quad (9)$$

Set $\Omega_n = \{\tau_n \leq T\}$ for $n \geq n_1$. Then by (6), $P(\Omega_n) = P\{\tau_n \leq T\} > \varepsilon, \forall n \geq n_1$. Note that for every $\omega \in \Omega_n$ there is some i such that $x_i(\tau_n, \omega)$ equals either n or $1/n$. So $V(X(\tau_n)) \geq \min\{n - 1 - \ln n, \frac{1}{n} - 1 + \ln n\}$. It then follows from (9) that $V(X_0) + LT \geq E[\mathbf{1}_{\Omega_n} V(X(\tau_n))] \geq \varepsilon \min\{n - 1 - \ln n, \frac{1}{n} - 1 + \ln n\}$, where $\mathbf{1}_{\Omega_n}$ is the indicator function of Ω_n . Letting $n \rightarrow \infty$ leads to the contradiction $\infty > V(X_0) + LT = \infty$. This completes the proof of the theorem. $\square$

Lemma 1. Let $X(t) = (x_1(t), x_2(t))$ be a solution to the system (4). Then the density of the population $x_i(t), i = 1, 2$ has a property $\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} \leq 0, i = 1, 2$ a.s.

Proof. By the Itô's formula we have for $i = 1, 2$

$$\begin{aligned} e^t \ln x_i(t) - \ln x_{i0} = \int_0^t e^s \left\{ \ln x_i(s) + r_i(s) - \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} - \right. \\ \left. - c_i(s)x_i(s) - \frac{\sigma_i^2(s)}{2} + \int_{\mathbb{R}} [\ln(1 + \gamma_i(s, z)) - \gamma_i(s, z)]\Pi_1(dz) \right\} ds + \psi_i(t), \end{aligned} \quad (10)$$

where

$$\psi_i(t) = \int_0^t e^s \sigma_i(s)dw_i(s) + \int_0^t \int_{\mathbb{R}} e^s \ln(1 + \gamma_i(s, z))\tilde{\nu}_1(ds, dz) + \int_0^t \int_{\mathbb{R}} e^s \ln(1 + \delta_i(s, z))\nu_2(ds, dz), \quad i = 1, 2.$$

By virtue of the exponential inequality ((Borysenko, & Borysenko, 2019), Lemma 2.2) we have

$$P \left\{ \sup_{0 \leq t \leq T} \zeta_i(\mu, t) > \beta \right\} \leq e^{-\mu\beta}, \quad \forall 0 < \mu \leq 1, \beta > 0, i = 1, 2,$$

where

$$\begin{aligned} \zeta_i(\mu, t) = & \psi_i(t) - \frac{\mu}{2} \int_0^t e^{2s} \sigma_i^2(s) ds - \frac{1}{\mu} \int_0^t \int_{\mathbb{R}} \left[(1 + \gamma_i(s, z))^{\mu e^s} - 1 - \right. \\ & \left. - \mu e^s \ln(1 + \gamma_i(s, z)) \right] \Pi_1(dz) ds - \frac{1}{\mu} \int_0^t \int_{\mathbb{R}} \left[(1 + \delta_i(s, z))^{\mu e^s} - 1 \right] \Pi_2(dz) ds, \quad i = 1, 2. \end{aligned}$$

Choosing $T = k\tau, k \in \mathbb{N}, \tau > 0, \mu = e^{-k\tau}, \beta = \theta e^{k\tau} \ln k, \theta > 1$ we get

$$P \left\{ \sup_{0 \leq t \leq k\tau} \zeta_i(\mu, t) > \theta e^{k\tau} \ln k \right\} \leq \frac{1}{k^\theta}, \quad i = 1, 2.$$

By the Borel-Cantelli lemma for almost all $\omega \in \Omega$, there is a random integer $k_0(\omega)$, such that for $\forall k \geq k_0(\omega)$ and $0 \leq t \leq k\tau$ we have

$$\begin{aligned} \psi_i(t) \leq & \frac{1}{2e^{k\tau}} \int_0^t e^{2s} \sigma_i^2(s) ds + e^{k\tau} \int_0^t \int_{\mathbb{R}} \left[(1 + \gamma_i(s, z))^{e^{s-k\tau}} - 1 - e^{s-k\tau} \ln(1 + \gamma_i(s, z)) \right] \Pi_1(dz) ds + \\ & + e^{k\tau} \int_0^t \int_{\mathbb{R}} \left[(1 + \delta_i(s, z))^{e^{s-k\tau}} - 1 \right] \Pi_2(dz) ds + \theta e^{k\tau} \ln k, \quad i = 1, 2. \end{aligned} \quad (11)$$

Applying the inequality $x^r \leq 1 + r(x - 1)$, valid for all $x \geq 0, 0 \leq r \leq 1$, first for $x = 1 + \gamma_i(s, z), r = e^{s-k\tau}$, then for $x = 1 + \delta_i(s, z), r = e^{s-k\tau}$, we derive from (11) the estimates

$$\psi_i(t) \leq \frac{1}{2e^{k\tau}} \int_0^t e^{2s} \sigma_i^2(s) ds + \int_0^t \int_{\mathbb{R}} e^s [\gamma_i(s, z) - \ln(1 + \gamma_i(s, z))] \Pi_1(dz) ds + \int_0^t \int_{\mathbb{R}} e^s \delta_i(s, z) \Pi_2(dz) ds + \theta e^{k\tau} \ln k, \quad i = 1, 2. \quad (12)$$

So from (10) and (12), we get for $i = 1, 2$

$$\begin{aligned} e^t \ln x_i(t) \leq & \ln x_{i0} + \int_0^t e^s \left\{ \ln x_i(s) + r_i(s) - \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} - c_i(s)x_i(s) - \frac{\sigma_i^2(s)}{2} (1 - e^{s-k\tau}) + \right. \\ & \left. + \int_{\mathbb{R}} \delta_i(s, z) \Pi_2(dz) \right\} ds + \theta e^{k\tau} \ln k \leq \ln x_{i0} + \int_0^t e^s [\ln x_i(s) - c_i \inf x_i(s) + L_i] ds + \theta e^{k\tau} \ln k \leq \\ & \leq \ln x_{i0} + L(e^t - 1) + \theta e^{k\tau} \ln k, \quad \forall k \geq k_0(\omega), 0 \leq t \leq k\tau, \end{aligned}$$

for some constant $L > 0$, where $L_i = r_i \sup + \sup_{t \geq 0} \int_{\mathbb{R}} \delta_2(t, z) \Pi_2(dz)$. So for any $(k-1)\tau \leq t \leq k\tau, \forall k \geq k_0(\omega)$ we have

$$\frac{\ln x_i(t)}{\ln t} \leq e^{-t} \frac{\ln x_{i0}}{\ln t} + \frac{L}{\ln t} (1 - e^{-t}) + \frac{\theta e^{k\tau} \ln k}{e^{(k-1)\tau} \ln(k-1)\tau}, \quad i = 1, 2 \quad \text{a.s.}$$

Therefore

$$\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{\ln t} \leq \theta e^\tau, \quad i = 1, 2, \quad \forall \theta > 1, \quad \forall \tau > 0, \quad \text{a.s.}$$

If $\theta \downarrow 1, \tau \downarrow 0$, then we obtain

$$\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{\ln t} \leq 1, \quad i = 1, 2 \quad \text{a.s.}$$

So $\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} \leq 0, i = 1, 2 \quad \text{a.s.}$ □

Lemma 2. Let $p > 0$. Then for any initial value $x_{i0} > 0, i = 1, 2$ we have

$$\limsup_{t \rightarrow \infty} E[x_i^p(t)] \leq L_i(p), \quad i = 1, 2,$$

where $L_i(p) > 0, i = 1, 2$ are some constants depending on p .

Proof. Let τ_n be the stopping time defined in Theorem 1. Applying the Itô's formula to the process $V(t, x_i(t)) = e^t x_i^p(t)$,

$i = 1, 2, p > 0$, we obtain for $i = 1, 2$

$$\begin{aligned} V(t \wedge \tau_n, x_i(t \wedge \tau_n)) &= x_{i0}^p + \int_0^{t \wedge \tau_n} e^s x_i^p(s) \left\{ 1 + p \left(r_i(s) - \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} - c_i(s)x_i(s) \right) + \frac{p(p-1)\sigma_i^2(s)}{2} + \right. \\ &+ \int_{\mathbb{R}} [(1 + \gamma_i(s, z))^p - 1 - p\gamma_i(s, z)] \Pi_1(dz) + \int_{\mathbb{R}} [(1 + \delta_i(s, z))^p - 1] \Pi_2(dz) \Big\} ds + \int_0^{t \wedge \tau_n} p e^s x_i^p(s) \sigma_i(s) dw_i(s) + \\ &+ \int_0^{t \wedge \tau_n} \int_{\mathbb{R}} e^s x_i^p(s^-) [(1 + \gamma_i(s, z))^p - 1] \tilde{\nu}_1(ds, dz) + \int_0^{t \wedge \tau_n} \int_{\mathbb{R}} e^s x_i^p(s^-) [(1 + \delta_i(s, z))^p - 1] \tilde{\nu}_2(ds, dz). \end{aligned} \quad (13)$$

Under Assumption 1 there are constants $L_i(p) > 0, i = 1, 2$, such that

$$\begin{aligned} e^s x_i^p(s) \left\{ 1 + p \left(r_i(s) - \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} - c_i(s)x_i(s) \right) + \frac{p(p-1)\sigma_i^2(s)}{2} + \right. \\ \left. + \int_{\mathbb{R}} [(1 + \gamma_i(s, z))^p - 1 - p\gamma_i(s, z)] \Pi_1(dz) + \int_{\mathbb{R}} [(1 + \delta_i(s, z))^p - 1] \Pi_2(dz) \right\} \leq \\ \leq e^s \left[x_i^p(s) \left(1 + p r_i \sup + \frac{p|p-1|}{2} \sigma_{i \sup}^2 + J_i \sup \right) - p c_{i \inf} x_i^{p+1}(s) \right] \leq e^s L_i(p), \end{aligned} \quad (14)$$

where

$$J_i \sup = \sup_{s \geq 0} \left\{ \int_{\mathbb{R}} [(1 + \gamma_i(s, z))^p - 1 - p\gamma_i(s, z)] \Pi_1(dz) + \int_{\mathbb{R}} [(1 + \delta_i(s, z))^p - 1] \Pi_2(dz) \right\}.$$

From (13) and (14), taking the expectation, we obtain

$$\mathbb{E}[V(t \wedge \tau_n, x_i(t \wedge \tau_n))] \leq x_{i0}^p + L_i(p)e^t, \quad i = 1, 2.$$

If $n \rightarrow \infty$, then we get

$$e^t \mathbb{E}[x_i^p(t)] \leq x_{i0}^p + L_i(p)e^t, \quad i = 1, 2.$$

Hence $\limsup_{t \rightarrow \infty} \mathbb{E}[x_i^p(t)] \leq L_i(p), \quad i = 1, 2.$

□

Lemma 3. If $\min\{(r_1 - \beta_1)_{\inf}, (r_2 - \beta_2)_{\inf}\} > 0$, then for any initial value $x_{i0} > 0, i = 1, 2$ we have

$$\limsup_{t \rightarrow \infty} \mathbb{E} \left[\left(\frac{1}{x_i(t)} \right)^\theta \right] \leq L_i(\theta), \quad 0 < \theta < 1, \quad i = 1, 2, \quad (15)$$

where $L_i(\theta) > 0, i = 1, 2$ are some constants depending on $0 < \theta < 1$.

Proof. For the process $U_i(t) = 1/x_i(t), i = 1, 2$ by the Itô's formula we have

$$\begin{aligned} U_i(t) &= U_i(0) + \int_0^t U_i(s) \left[-r_i(s) + \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} + c_i(s)x_i(s) + \sigma_i^2(s) + \int_{\mathbb{R}} \frac{\gamma_i^2(s, z)}{1 + \gamma_i(s, z)} \Pi_1(dz) \right] ds - \\ &- \int_0^t U_i(s) \sigma_i(s) dw_i(s) - \int_0^t \int_{\mathbb{R}} U_i(s^-) \frac{\gamma_i(s, z)}{1 + \gamma_i(s, z)} \tilde{\nu}_1(ds, dz) - \int_0^t \int_{\mathbb{R}} U_i(s^-) \frac{\delta_i(s, z)}{1 + \delta_i(s, z)} \nu_2(ds, dz). \end{aligned}$$

Then by the Itô's formula we derive for $0 < \theta < 1, i = 1, 2$

$$\begin{aligned} (1 + U_i(t))^\theta &= (1 + U_i(0))^\theta + \int_0^t \theta (1 + U_i(s))^{\theta-2} \left\{ (1 + U_i(s)) U_i(s) \left[-r_i(s) + \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} + c_i(s)x_i(s) + \right. \right. \\ &+ \sigma_i^2(s) + \int_{\mathbb{R}} \frac{\gamma_i^2(s, z)}{1 + \gamma_i(s, z)} \Pi_1(dz) \Big] + \frac{\theta-1}{2} U_i^2(s) \sigma_i^2(s) + \frac{1}{\theta} \int_{\mathbb{R}} \left[(1 + U_i(s))^2 \left(\frac{1 + U_i(s) + \gamma_i(s, z)}{(1 + \gamma_i(s, z))(1 + U_i(s))} \right)^\theta - 1 \right] + \\ &+ \theta (1 + U_i(s)) \frac{U_i(s) \gamma_i(s, z)}{1 + \gamma_i(s, z)} \Big] \Pi_1(dz) + \frac{1}{\theta} \int_{\mathbb{R}} (1 + U_i(s))^2 \left[\left(\frac{1 + U_i(s) + \delta_i(s, z)}{(1 + \delta_i(s, z))(1 + U_i(s))} \right)^\theta - 1 \right] \Pi_2(dz) \Big\} ds - \\ &- \int_0^t \theta (1 + U_i(s))^{\theta-1} U_i(s) \sigma_i(s) dw_i(s) + \int_0^t \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(s^-)}{1 + \gamma_i(s, z)} \right)^\theta - (1 + U_i(s^-))^\theta \right] \tilde{\nu}_1(ds, dz) + \end{aligned} \quad (16)$$

$$+\int_0^t \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(s^-)}{1+\delta_i(s,z)} \right)^\theta - (1+U_i(s^-))^\theta \right] \tilde{\nu}_2(ds, dz) = (1+U_i(0))^\theta + \int_0^t \theta (1+U_i(s))^{\theta-2} J_i(s) ds + \sum_{j=1}^3 I_{j, stoch}(t),$$

where $I_{j, stoch}(t), j = \overline{1, 3}$ are corresponding stochastic integrals in (16). Under Assumption 1 there exist constants $|L_{i,1}(\theta)| < \infty, |L_{i,2}(\theta)| < \infty, i = 1, 2$ such, that for the process $J_i(t)$ we have the estimate

$$\begin{aligned} J_i(t) \leq & (1+U_i(t))U_i(t) \left[-r_i(t) + U_i^{-1}(t) \left(\frac{b_i(t)}{K_i(t)} + c_i(t) \right) + \sigma_i^2(t) + \int \gamma_i(t, z) \Pi_1(dz) \right] + \\ & + \frac{\theta-1}{2} U_i^2(t) \sigma_i^2(t) + \frac{1}{\theta} \int_{\mathbb{R}} (1+U_i(t))^2 \left[\left(\frac{1}{1+\gamma_i(t, z)} + \frac{1}{1+U_i(t)} \right)^\theta - 1 \right] \Pi_1(dz) + \\ & + \frac{1}{\theta} \int_{\mathbb{R}} (1+U_i(t))^2 \left[\left(\frac{1}{1+\delta_i(t, z)} + \frac{1}{1+U_i(t)} \right)^\theta - 1 \right] \Pi_2(dz) \leq U_i^2(t) \left[-r_i(t) + \frac{\sigma_i^2(t)}{2} + \right. \\ & \left. + \int_{\mathbb{R}} \gamma_i(t, z) \Pi_1(dz) + \frac{\theta}{2} \sigma_i^2(t) + \frac{1}{\theta} \int_{\mathbb{R}} [(1+\gamma_i(t, z))^{-\theta} - 1] \Pi_1(dz) + \frac{1}{\theta} \int_{\mathbb{R}} [(1+\delta_i(t, z))^{-\theta} - 1] \Pi_2(dz) \right] + \\ & + L_{i,1}(\theta)U_i(t) + L_{i,2}(\theta) = -L_{i,0}(t, \theta)U_i^2(t) + L_{i,1}(\theta)U_i(t) + L_{i,2}(\theta). \end{aligned}$$

Here we used the inequality $(x+y)^\theta \leq x^\theta + \theta x^{\theta-1}y$, valid for all $0 < \theta < 1, x, y > 0$, in the first integral term for $x = (1+\gamma_i(t, z))^{-1}, y = (1+U_i(t))^{-1}$, and then in the second integral term for $x = (1+\delta_i(t, z))^{-1}, y = (1+U_i(t))^{-1}$. Due to

$$\begin{aligned} \lim_{\theta \rightarrow 0+} \left[\frac{\theta}{2} \sigma_i^2(t) + \frac{1}{\theta} \int_{\mathbb{R}} [(1+\gamma_i(t, z))^{-\theta} - 1] \Pi_1(dz) + \frac{1}{\theta} \int_{\mathbb{R}} [(1+\delta_i(t, z))^{-\theta} - 1] \Pi_2(dz) + \right. \\ \left. + \int_{\mathbb{R}} \ln(1+\gamma_i(t, z)) \Pi_1(dz) + \int_{\mathbb{R}} \ln(1+\delta_i(t, z)) \Pi_2(dz) \right] = \lim_{\theta \rightarrow 0+} \Delta_i(\theta, t) = 0, \end{aligned}$$

and the condition $\min_{i=1,2} (r_i - \beta_i)_{\inf} > 0$ we can choose a sufficiently small $0 < \theta < 1$ so that

$$L_{i,0}(\theta) = \inf_{t \geq 0} L_{i,0}(t, \theta) = \inf_{t \geq 0} [r_i(t) - \beta_i(t) - \Delta_i(\theta, t)] > 0$$

is satisfied. So from (16) and the estimate for $J_i(t)$ we derive

$$\begin{aligned} d \left[(1+U_i(t))^\theta \right] \leq & \theta (1+U_i(t))^{\theta-2} [-L_{i,0}(\theta)U_i^2(t) + L_{i,1}(\theta)U_i(t) + L_{i,2}(\theta)] dt - \theta (1+U_i(t))^{\theta-1} U_i(t) \sigma_i(t) dw_i(t) + \\ & + \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(t^-)}{1+\gamma_i(t, z)} \right)^\theta - (1+U_i(t^-))^\theta \right] \tilde{\nu}_1(dt, dz) + \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(t^-)}{1+\delta_i(t, z)} \right)^\theta - (1+U_i(t^-))^\theta \right] \tilde{\nu}_2(dt, dz). \end{aligned} \quad (17)$$

By the Itô's formula and (17) we have

$$\begin{aligned} d \left[e^{\lambda t} (1+U_i(t))^\theta \right] = & \lambda e^{\lambda t} (1+U_i(t))^\theta dt + e^{\lambda t} d \left[(1+U_i(t))^\theta \right] \leq e^{\lambda t} \theta (1+U_i(t))^{\theta-2} \left[- \left(L_{i,0}(\theta) - \frac{\lambda}{\theta} \right) U_i^2(t) + \right. \\ & + \left(L_{i,1}(\theta) + \frac{2\lambda}{\theta} \right) U_i(t) + L_{i,2}(\theta) + \frac{\lambda}{\theta} \right] dt - \theta e^{\lambda t} (1+U_i(t))^{\theta-1} U_i(t) \sigma_i(t) dw_i(t) + \\ & + e^{\lambda t} \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(t^-)}{1+\gamma_i(t, z)} \right)^\theta - (1+U_i(t^-))^\theta \right] \tilde{\nu}_1(dt, dz) + e^{\lambda t} \int_{\mathbb{R}} \left[\left(1 + \frac{U_i(t^-)}{1+\delta_i(t, z)} \right)^\theta - (1+U_i(t^-))^\theta \right] \tilde{\nu}_2(dt, dz). \end{aligned} \quad (18)$$

Let us choose $\lambda = \lambda(\theta) > 0$ such that $\min_{i=1,2} L_{i,0}(\theta) - \lambda/\theta > 0$. Then there is a constants $L_i(\theta) > 0, i = 1, 2$, such that for $i = 1, 2$ we have

$$(1+U_i(t))^{\theta-2} \left[- \left(L_{i,0}(\theta) - \frac{\lambda}{\theta} \right) U_i^2(t) + \left(L_{i,1}(\theta) + \frac{2\lambda}{\theta} \right) U_i(t) + L_{i,2}(\theta) + \frac{\lambda}{\theta} \right] \leq L_i(\theta). \quad (19)$$

Let τ_n be the stopping time defined in the Theorem 1. Then by integrating (18) from 0 to $t \wedge \tau_n$, using (19) and taking the expectation we obtain

$$\mathbb{E} \left[e^{\lambda(t \wedge \tau_n)} (1+U_i(t \wedge \tau_n))^\theta \right] \leq \left(1 + \frac{1}{x_{i0}} \right)^\theta + \frac{\theta}{\lambda} L_i(\theta) (e^{\lambda t} - 1).$$

Letting $n \rightarrow \infty$ leads to the estimate

$$e^t \mathbb{E} \left[(1+U_i(t))^\theta \right] \leq \left(1 + \frac{1}{x_{i0}} \right)^\theta + \frac{\theta}{\lambda} L_i(\theta) (e^{\lambda t} - 1). \quad (20)$$

From (20) we obtain

$$\limsup_{t \rightarrow \infty} E \left[\left(\frac{1}{x_i(t)} \right)^\theta \right] = \limsup_{t \rightarrow \infty} E \left[U_i^\theta(t) \right] \leq \limsup_{t \rightarrow \infty} E \left[(1 + U_i(t))^\theta \right] \leq \frac{\theta}{\lambda(\theta)} L_i(\theta),$$

and this implies (15). $\square$

2. The long time behaviour

Definition 1. ((Li, & Mao, 2009)) The solution $X(t)$ to the system (4) will be said stochastically ultimately bounded, if for any $\varepsilon \in (0, 1)$, there is a positive constant $\chi = \chi(\varepsilon) > 0$, such that for any initial value $X_0 \in \mathbb{R}_+^2$, the solution to the system (4) has the property that

$$\limsup_{t \rightarrow \infty} P \{ |X(t)| > \chi \} < \varepsilon.$$

In what follows in this section we will assume that Assumption 1 holds.

Theorem 2. The solution $X(t)$ to the system (4) is stochastically ultimately bounded for any initial value $X_0 \in \mathbb{R}_+^2$.

Proof. From Lemma 2 we have an estimate

$$\limsup_{t \rightarrow \infty} E[x_i(t)] \leq L_i, \quad i = 1, 2. \quad (21)$$

For $X = (x_1, x_2) \in \mathbb{R}_+^2$ we have $|X| \leq x_1 + x_2$, therefore, from (21) $\limsup_{t \rightarrow \infty} E[|X(t)|] \leq L = L_1 + L_2$. Let $\chi > L/\varepsilon$, $\forall \varepsilon \in (0, 1)$. Then applying the Chebyshev inequality yields

$$\limsup_{t \rightarrow \infty} P\{|X(t)| > \chi\} \leq \frac{1}{\chi} \limsup_{t \rightarrow \infty} E|X(t)| \leq \frac{L}{\chi} < \varepsilon.$$

$\square$

Definition 2. ((Li, Gao, & Wang, 2016)) The solution $X(t)$ to the equation (4) is said to be stochastically permanent if for any $\varepsilon > 0$, there are positive constants $H = H(\varepsilon)$, $h = h(\varepsilon)$ such that for $i = 1, 2$

$$\liminf_{t \rightarrow \infty} P\{x_i(t) \leq H\} \geq 1 - \varepsilon, \quad \liminf_{t \rightarrow \infty} P\{x_i(t) \geq h\} \geq 1 - \varepsilon,$$

for any initial value $X_0 \in \mathbb{R}_+^2$.

Theorem 3. Under the conditions of Lemma 3, the solution $X(t)$ to the system (4) is stochastically permanent for any initial value $X_0 \in \mathbb{R}_+^2$.

Proof. From Lemma 2 we have an estimate

$$\limsup_{t \rightarrow \infty} E[x_i(t)] \leq L, \quad i = 1, 2.$$

Thus for any given $\varepsilon > 0$, let $H = L/\varepsilon$, by Chebyshev's inequality, we can derive that

$$\limsup_{t \rightarrow \infty} P\{x_i(t) > H\} \leq \frac{1}{H} \limsup_{t \rightarrow \infty} E[x_i(t)] \leq \varepsilon.$$

Consequently $\liminf_{t \rightarrow \infty} P\{x_i(t) \leq H\} \geq 1 - \varepsilon$, $i = 1, 2$.

From Lemma 3 we have the estimate

$$\limsup_{t \rightarrow \infty} E \left[\left(\frac{1}{x_i(t)} \right)^\theta \right] \leq L(\theta), \quad 0 < \theta < 1.$$

For any given $\varepsilon > 0$, let $h = (\varepsilon/L(\theta))^{1/\theta}$, then by Chebyshev's inequality, we have

$$\limsup_{t \rightarrow \infty} P\{x_i(t) < h\} \leq \limsup_{t \rightarrow \infty} P \left\{ \left(\frac{1}{x_i(t)} \right)^\theta > h^{-\theta} \right\} \leq h^\theta \limsup_{t \rightarrow \infty} E \left[\left(\frac{1}{x_i(t)} \right)^\theta \right] \leq \varepsilon.$$

Consequently $\liminf_{t \rightarrow \infty} P\{x_i(t) \geq h\} \geq 1 - \varepsilon$. $\square$

Theorem 4. If for $i = 1, 2$ we have

$$\bar{p}_i^* = \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t p_i(s) ds < 0, \quad \text{where } p_i(s) = r_i(s) - \beta_i(s),$$

then the solution $X(t)$ to the system (4) with initial condition $X_0 \in \mathbb{R}_+^2$ will be extinct: $\lim_{t \rightarrow \infty} x_i(t) = 0$, $i = 1, 2$ a.s.

Proof. By the Itô's formula, we have for $i = 1, 2$

$$\ln x_i(t) = \ln x_{i0} + \int_0^t \left[r_i(s) - \frac{b_i(s)x_i(s)}{K_i(s) + x_{3-i}(s)} - c_i(s)x_i(s) - \beta_i(s) \right] ds + M_i(t) \leq \ln x_{i0} + \int_0^t p_i(s)ds + M_i(t), \quad (22)$$

where the martingale

$$M_i(t) = \int_0^t \sigma_i(s)dw_i(s) + \int_0^t \int_{\mathbb{R}} \ln(1 + \gamma_i(s, z))\tilde{\nu}_1(ds, dz) + \int_0^t \int_{\mathbb{R}} \ln(1 + \delta_i(s, z))\tilde{\nu}_2(ds, dz), \quad i = 1, 2, \quad (23)$$

has quadratic variation (Meyer's angle bracket process)

$$\langle M_i, M_i \rangle(t) = \int_0^t \sigma_i^2(s)ds + \int_0^t \int_{\mathbb{R}} \ln^2(1 + \gamma_i(s, z))\Pi_1(dz)ds + \int_0^t \int_{\mathbb{R}} \ln^2(1 + \delta_i(s, z))\Pi_2(dz)ds \leq Kt, \quad i = 1, 2.$$

Then the strong law of large numbers for local martingales (Lipster, 1980) yields $\lim_{t \rightarrow \infty} M_i(t)/t = 0, i = 1, 2$ a.s. Therefore, from (22) we obtain

$$\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} \leq \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t p_i(s)ds < 0, \quad \text{a.s.}$$

So $\lim_{t \rightarrow \infty} x_i(t) = 0, i = 1, 2$ a.s. □

Theorem 5. If $\bar{p}_i^* = 0, i = 1, 2$, then the solution $X(t)$ to the system (4) with initial condition $X_0 \in \mathbb{R}_+^2$ will be non-persistent in the mean:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s)ds = 0 \quad \text{a.s., } i = 1, 2.$$

Proof. From the first equality in (22) we have

$$\ln x_i(t) \leq \ln x_{i0} + \int_0^t p_i(s)ds - c_i \inf \int_0^t x_i(s)ds + M_i(t), \quad i = 1, 2, \quad (24)$$

where martingale $M_i(t)$ is defined in (23). From the definition of $\bar{p}_i^*$ and the strong law of large numbers for $M_i(t)$ it follows, that $\forall \varepsilon > 0, \exists t_0 \geq 0$, and $\exists \Omega_\varepsilon \subset \Omega, P(\Omega_\varepsilon) \geq 1 - \varepsilon$ such that

$$\frac{1}{t} \int_0^t p_i(s)ds \leq \bar{p}_i^* + \frac{\varepsilon}{2}, \quad \frac{M_i(t)}{t} \leq \frac{\varepsilon}{2}, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

So, from (24) and condition $\bar{p}_i^* = 0, i = 1, 2$ we derive

$$\ln x_i(t) - \ln x_{i0} \leq t(\bar{p}_i^* + \varepsilon) - c_i \inf \int_0^t x_i(s)ds = t\varepsilon - c_i \inf \int_0^t x_i(s)ds, \quad i = 1, 2, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon. \quad (25)$$

Let $y_i(t) = \int_0^t x_i(s)ds, i = 1, 2$, then from (25) we have

$$\ln \left(\frac{dy_i(t)}{dt} \right) \leq \varepsilon t - c_i \inf y_i(t) + \ln x_{i0}, \quad i = 1, 2, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

Therefore

$$e^{c_i \inf y_i(t)} \frac{dy_i(t)}{dt} \leq x_{i0} e^{\varepsilon t}, \quad i = 1, 2, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

By integrating the last inequality from t_0 to t we obtain

$$e^{c_i \inf y_i(t)} \leq \frac{c_i \inf x_{i0}}{\varepsilon} (e^{\varepsilon t} - e^{\varepsilon t_0}) + e^{c_i \inf y_i(t_0)}, \quad i = 1, 2, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

So

$$y_i(t) \leq \frac{1}{c_i \inf} \ln \left[e^{c_i \inf y_i(t_0)} + \frac{c_i \inf x_{i0}}{\varepsilon} (e^{\varepsilon t} - e^{\varepsilon t_0}) \right], \quad i = 1, 2, \quad \forall t \geq t_0, \omega \in \Omega_\varepsilon,$$

and therefore

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s)ds \leq \frac{\varepsilon}{c_i \inf}, \quad i = 1, 2, \quad \forall \omega \in \Omega_\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary and $x_i(t) > 0, i = 1, 2$ a.s., we have $\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s) ds = 0, i = 1, 2$ a.s. □

Theorem 6. *If $\bar{p}_i^* > 0, i = 1, 2$, then the solution $X(t)$ to the system (4) with initial condition $X_0 \in \mathbb{R}_+^2$ will be weakly persistent in the mean:*

$$\bar{x}_i^* = \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s) ds > 0, i = 1, 2, \quad \text{a.s.}$$

Proof. If the assertion of the theorem is not true, then $P\{\omega \in \Omega \mid \bar{x}_2^*(\omega) = 0\} > 0$. From the first equality in (22) we get

$$\frac{\ln x_i(t)}{t} \geq \frac{\ln x_{i0}}{t} + \frac{1}{t} \int_0^t p_i(s) ds - \left(\frac{b_i \sup}{K_i \inf} + c_i \sup \right) \frac{1}{t} \int_0^t x_i(s) ds + \frac{M_i(t)}{t}, i = 1, 2, \quad (26)$$

where martingale $M_i(t), i = 1, 2$ is defined in (23). For $\forall \omega \in \{\omega \in \Omega \mid \bar{x}_2^* = 0\}$ in virtue of the strong law of large numbers for martingale $M_i(t), i = 1, 2$ we have

$$\limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} \geq \bar{p}_i^* > 0, i = 1, 2.$$

Therefore

$$P\left\{\omega \in \Omega : \limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} > 0\right\} > 0, i = 1, 2.$$

But from Lemma 1 we have

$$P\left\{\omega \in \Omega : \limsup_{t \rightarrow \infty} \frac{\ln x_i(t)}{t} \leq 0\right\} = 1, i = 1, 2.$$

So we obtain a contradiction. □

Theorem 7. *If $\bar{p}_{i*} = \liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t p_i(s) ds > 0, i = 1, 2$, where $p_i(t) = r_i(t) - \beta_i(t), i = 1, 2$, then*

$$\bar{x}_{i*} = \liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s) ds \geq \frac{\bar{p}_{i*}}{m_i}, \text{ a.s., where } m_i = \frac{b_i \inf}{K_i \inf} + c_i \sup, i = 1, 2,$$

Therefore, the solution $X(t) = (x_1(t), x_2(t))$ of system (4) with initial value $X_0 \in \mathbb{R}_+^2$ will be strongly persistent in the mean: $\bar{x}_{i} > 0$ almost surely, $i = 1, 2$.*

Proof. From (26) we get

$$\ln x_i(t) \geq \ln x_{i0} + \int_0^t p_i(s) ds - m_i \int_0^t x_i(s) ds + M_i(t), i = 1, 2, \quad (27)$$

where martingales $M_i(t), i = 1, 2$ are defined in (23). From the definition of $\bar{p}_{i*}, i = 1, 2$ and the strong law of large numbers for $M_i(t), i = 1, 2$ it follows that $\forall \varepsilon > 0, \exists t_0 \geq 0$, and $\exists \Omega_\varepsilon \subset \Omega$, such that $P(\Omega_\varepsilon) \geq 1 - \varepsilon$,

$$\frac{1}{t} \int_0^t p_i(s) ds \geq \bar{p}_{i*} - \frac{\varepsilon}{2}, \frac{M_i(t)}{t} \geq -\frac{\varepsilon}{2}, i = 1, 2, \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

Hence, from (27) we have

$$\ln x_i(t) \geq \ln x_{i0} + t(\bar{p}_{i*} - \varepsilon) - m_i \int_0^t x_i(s) ds, i = 1, 2, \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

Applying the same arguments for the processes $y_i(t) = \int_0^t x_i(s) ds, i = 1, 2$ as in the proof of Theorem 5, we derive the estimate

$$y_i(t) \geq \frac{1}{m_i} \ln \left[e^{m_i y_i(t_0)} + \frac{m_i x_{i0}}{\bar{p}_{i*} - \varepsilon} \left(e^{(\bar{p}_{i*} - \varepsilon)t} - e^{(\bar{p}_{i*} - \varepsilon)t_0} \right) \right], i = 1, 2, \forall t \geq t_0, \omega \in \Omega_\varepsilon.$$

So

$$\liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t x_i(s) ds \geq \frac{\bar{p}_{i*} - \varepsilon}{m_i}, i = 1, 2, \omega \in \Omega_\varepsilon.$$

Using the arbitrariness of $\varepsilon > 0$, we complete the proof of the theorem. □

Discussion and conclusions

In this paper, we consider the non-autonomous stochastic mutualism model driven by the system of stochastic differential equations with white noise, centered and non-centered Poisson noises. So, we take into account not only "small" jumps,

corresponding to the centered Poisson measure, but also the “large” jumps, corresponding to the non-centered Poisson measure. For the considered system, sufficient criteria for the existence and uniqueness of a global positive solution are obtained. We derive sufficient conditions of stochastic ultimate boundedness, stochastic permanence, non-persistence in the mean, weak and strong persistence in the mean, and extinction of the population densities in the considered stochastic mutualism model. The results strictly generalize existing results, so they are meaningful and important.

Authors' contribution: Oleksandr Borysenko – conceptualization, writing, and editing; Olga Borysenko – methodology, writing, viewing, and editing.

Acknowledgments. The authors would like to thank the anonymous referee who provided useful and detailed comments on an earlier version of the manuscript.

Sources of funding. The study was conducted mostly at the authors' own expense. Some financial support was partially provided by Taras Shevchenko National University of Kyiv.

References

- Borysenko, O., & Borysenko, D. (2019). Persistence and extinction in stochastic nonautonomous logistic model of population dynamics. *Theory of Probability and Mathematical Statistics*, 99, 67–75. <https://doi.org/10.1090/tpms/1080>
- Borysenko, O., & Borysenko, V. (2023). Extinction and persistence in stochastic predator population density-dependent predator-prey model with jumps. *Bulletin of Taras Shevchenko National University of Kyiv Series: Physics & Mathematics*, 1, 30–36. <https://doi.org/10.17721/1812-5409.2023/1>
- Gikhman, I. I., & Skorokhod, A. V. (1982). *Stochastic differential equations and its applications*. Naukova Dumka [in Russian]. [Гихман, И. И., & Скороход, А. В. (1982). *Стохастические дифференциальные уравнения и их приложения*. Наукова думка].
- Li, M., Gao, H., Sun, C., & Gong, Y. (2015). Analysis of a mutualism model with stochastic perturbations. *International Journal of Biomathematics*, 08(06), 1550072. <https://doi.org/10.1142/S1793524515500710>
- Li, M., Gao, H., & Wang, B. (2016). Analysis of a non-autonomous mutualism model driven by lévy jumps. *Discrete & Continuous Dynamical Systems - B*, 21(4), 1189–1202. <https://doi.org/10.3934/dcdsb.2016.21.1189>
- Li, X., & Mao, X. (2009). Population dynamical behavior of non-autonomous lotka-volterra competitive system with random perturbation. *Discrete & Continuous Dynamical Systems*, 24(2), 523–545. <https://doi.org/10.3934/dcds.2009.24.523>
- Lipster, R. (1980). A strong law of large numbers for local martingales. *Stochastics*, 3, 217–228. <https://doi.org/10.1080/17442508008833146>
- May, R. M. (1976). Models of two interacting populations. In May (Ed.), *Theoretical ecology: Principles and application* (pp. 78–104). PA: Saunders.

Отримано редакцією журналу / Received: 29.03.25

Прорецензовано / Revised: 20.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Олександр БОРИСЕНКО, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-6670-8605

e-mail: oleksandr_borysenko@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

Ольга БОРИСЕНКО, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-2099-2911

e-mail: oborisenko1373@gmail.com

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

ДИНАМІКА НЕАВТОНОМНОЇ СТОХАСТИЧНОЇ МОДЕЛІ МУТУАЛІЗМУ ЗІ СТРИБКАМИ

У природі ми можемо знайти багато прикладів взаємодії двох або більшої кількості популяцій, за яких вигоди одержує кожна із популяцій. Моделі мутуалізму виникають тоді, коли одна популяція надає певні вигоди іншій в обмін на вигоди для себе. Прикладом можуть слугувати популяція мурах і популяція попелюхи. Математичні детерміновані моделі широко використовуються під час вивчення динаміки таких популяційних систем. У реальному житті на популяційні системи часто впливають випадкові збурення, тому доцільно використовувати стохастичні моделі популяційної динаміки. Якщо ми хочемо враховувати такі раптові збурення середовища, як епідемії, пожежі, землетруси, тоді нам доцільно ввести центрований і нецентрований пуассонівські шуми у популяційну модель. Отже, ми будемо враховувати не тільки “малі” стрибки, що відповідають центрованому пуассонівському шуму, але і “великі” стрибки, що відповідають нецентрованому пуассонівському шуму. У роботі доведено існування і єдиність глобального додатного розв’язку системи стохастичних диференціальних рівнянь, що описує неавтономну модель мутуалізму, збурену “білим” шумом, центрованим і нецентрованим пуассонівськими шумами. Одержано достатні умови граничної стохастичної обмеженості, стохастичної перманентності, вимирання, неживання у середньому, слабкого і сильного виживання у середньому для розв’язку системи стохастичних диференціальних рівнянь, що описує відповідну модель мутуалізму.

Ключові слова: стохастична модель мутуалізму, глобальний розв’язок, гранична стохастична обмеженість, стохастична перманентність, вимирання, неживання у середньому, слабке і сильне виживання у середньому.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.4>

Olga VASYLYK, DSc (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-0880-3751
e-mail: vasylyk@matan.kpi.ua

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

Yehor OKUNIEV, Master's Student
ORCID ID: 0009-0009-3695-490X
e-mail: egorky96@gmail.com

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv School of Economics, Kyiv, Ukraine

MODELLING HIV/AIDS EPIDEMIC IN UKRAINE AS A CONTINUOUS-TIME MARKOV PROCESS

Ukraine remains among the countries with the highest rates of HIV/AIDS spread globally. In this study, we fit a continuous-time Markov model to describe the progression of HIV to AIDS and ultimately to death for an individual and so obtain the means to model the epidemic on the global level. Our analysis reconstructs the information lost due to left- and interval-censorship of empirical data. Our methodology integrates survival analysis, maximum likelihood estimation, Monte Carlo simulation, and the construction of confidence intervals. The results provide insights into the dynamics of the epidemic in Ukraine and can further be used in actuarial research of HIV/AIDS as an insurance risk.

Key words: survival models, multiple decrements, Markov models, simulation, stochastic process.

AMS 2020 classification: 60G15, 60G22, 60G51, 68U20.

Introduction

The HIV/AIDS epidemic remains one of the most pressing challenges facing Ukraine's healthcare system (Kruglov et al., 2008; Barnett et al., 2000). According to international reports, Ukraine is among the countries with the highest rates of HIV/AIDS transmission (Miranda et al., 2021; Parczewski et al., 2025). To ensure effective resource allocation, strategic planning, and informed healthcare policy-making, a formalized approach to modeling the epidemic at the individual level is essential.

The use of discrete multi-state models to describe the progression of chronic diseases has long been a standard practice, with Markov models being particularly prominent (Johnson, Kemp, & Kotz, 2005). In this study, we first fit a discrete-time Markov chain to represent the progression of HIV/AIDS at the individual level and then extend this framework to a continuous-time Markov process.

A major challenge in the modeling process is the presence of left- and interval-censoring in the empirical data (Cox, 1972). After fitting the continuous-time Markov process, we address how the information lost due to censoring is reconstructed by the model's design.

The objective of this paper is to develop a statistical model of the HIV/AIDS epidemic that captures both local and broader dynamics, taking into account the epidemiological context of Ukraine. Furthermore, we conduct simulation using *msm* package in R (Jackson, 2007) and explore the application of the derived model in the actuarial analysis of HIV/AIDS as an insurance-related risk.

1. Data sources introduction and preparation

This study uses aggregated region-level data from the State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine (HIV/AIDS.Data, 2025), covering the period from January 2014 up until the present time. The dataset includes the following information for each region and month:

- number of newly reported cases of HIV,
- number of newly reported cases of AIDS,
- number of AIDS-related deaths,
- total number of people with HIV/AIDS under medical supervision at month-end,
- corresponding rates per 100 000 population.

Due to the nature of data collection, these data are left- and interval-censored. The presence of statistics calculated per 100 000 people and the lack of explicit population data imply that these values were estimated. Additionally, we used data on general mortality rates at the region level, sourced from the website of the State Statistics Service of Ukraine, covering the period from January 2015 to January 2022 (State Statistics Service of Ukraine, 2022). Information for later dates is unavailable due to Russia's full-scale invasion of Ukraine.

The empirical data from the sources were stored in a format unsuitable for direct import into Microsoft Excel, so Optical Character Recognition (OCR) technology was employed. For this task, ABBYY software was used, as it specializes in OCR of documents resembling Excel spreadsheets. However, this approach had several limitations. Notably, the software occasionally misinterpreted the digits 0 and 1 as the letters "o" and "i", respectively; introduced rare and sporadic data corruption; and interpreted center-aligned symbols as entries containing multiple spaces.

In order to restore data lost due to shortcomings of OCR, we manually removed extra spaces and applied global character substitution of the form $o \rightarrow 0$.

Afterward, where possible, we derived the number of people who died due to causes different from AIDS (CDA). Later, we made a reasonable assumption that the number of people who die due to CDA within a certain group (e.g., people with HIV) is proportional to the number of people who died due to CDA in the general population (P).

This allowed us to estimate the number of people who died due to CDA in the following groups: people who are undiagnosed (U), people living with HIV (H), and people living with AIDS (A). It should be emphasized that undiagnosed includes individuals with HIV or AIDS who are unaware of their condition, which is realistic, given that HIV may be undetectable for up to half a year after contraction and asymptomatic for years thereafter.

Further, we will use the following right subscripts:

- $_$ - denoting the total number of people in a certain group,
- N - denoting the increment in the size of a certain group,
- D - denoting the number of people who died in a certain group,
- A - denoting death due to AIDS,
- CDA - denoting death due to CDA,
- $.100k$ - calculated per 100 000 people.

Additionally, we use the left subscript “ t ” to specify the moment in time. For example, the total number of people who died due to CDA at time point 4 can be written as ${}_4P_{D,CDA}$. Thus, our estimate for the number of people who died due to CDA within an arbitrary group G can be expressed as:

$${}_tG_{D,CDA} = {}_tP_{D,CDA} \cdot \frac{{}_tG}{{}_tP}.$$

As previously mentioned, statistics per 100 000 people were estimated and thus implicitly carry information about estimated population sizes. This allows us to recover estimates for group sizes using the formula:

$${}_tG_{_,100k} = \frac{{}_tG_{_,100k}}{{}_tP} \cdot 100\,000,$$

where $_$ stands for any statistic where the calculation of the corresponding statistic per 100 000 people is present.

Lastly, to address sporadic data corruption, we applied the following formulae to reconstruct the data:

$$\begin{aligned} {}_tH &= {}_{t-1}H + {}_tH_N - {}_tH_{D,CDA}; \\ {}_tA &= {}_{t-1}A + {}_tA_N - {}_tA_{D,CDA} - {}_tA_{D,A}. \end{aligned}$$

It should be noted that the number of progressions and regressions between HIV and AIDS is not accounted for, since currently, all people with HIV/AIDS under medical supervision receive antiretroviral therapy (ART), which eventually reverts AIDS back to HIV and puts HIV into remission. This makes such transitions unlikely and impossible to estimate with the given empirical data.

2. Designing the model

Natural choice of states would be Undiagnosed (U), has HIV (H), has AIDS (A), and Dead (D). Intuitively, one might want to connect these states in the way the virus progresses naturally, i.e., $U \rightarrow H \rightarrow A$, with all of them connected to D . However, that model is not realistic, due to several reasons:

- The widespread use of ART treatment allows transition $A \rightarrow H$,
- People under medical supervision almost never transition from H to A (Raboud et al., 2002),
- Most people in state A are “late presenters,” i.e., individuals who are diagnosed in the late stages of the infection (Miranda et al., 2021; Parczewski et al., 2025).

So, a more realistic picture (Fig. 1) of the Markov model would be:

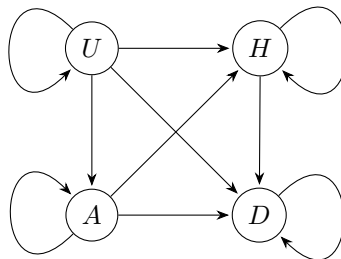


Fig. 1. Detailed model (inestimable)

The only problem with that model is that it is not possible to estimate the transition probability for $A \rightarrow H$ from the empirical data. Therefore, for the sake of further research, we introduce a new state H/A , for which all the transition probabilities can be estimated from empirical data. In this model, Markov chain (Fig. 2) and the monthly transition probability matrix take the following forms:

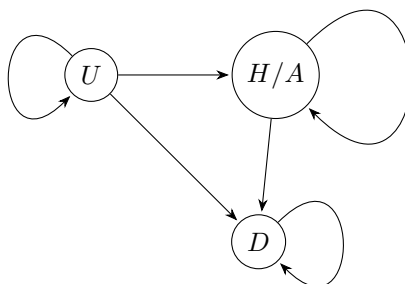


Fig. 2. Simplified model (estimable)

$$\begin{bmatrix} 1 - p_{U,H/A} - p_{U,D} & p_{U,H/A} & p_{U,D} \\ 0 & 1 - p_{H/A,D} & p_{H/A,D} \\ 0 & 0 & 1 \end{bmatrix}.$$

It should be noted, that reversions $A \rightarrow D$ are not represented in ${}_tH_N$.

Lastly, even though state U includes people with HIV/AIDS, it does not contribute to the number of deaths caused by AIDS, since most AIDS-related deaths occur among individuals who are diagnosed with AIDS at a critical stage and die before their immune system can be restored through ART.

With a discrete time-homogeneous Markov model in mind, we consider that the time until a certain transition is geometrically distributed, and the total number of transitions during a month is binomially distributed (with the same p , and different n for each month). It is intuitive to derive ML-estimator for p :

$$\begin{aligned} L(p) &= \prod_{i=1}^n C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i-x_i}, \\ \mathcal{L}(p) &= \ln L(p) = \sum_{i=1}^n \ln C_{n_i}^{x_i} + x_i \ln p + (n_i - x_i) \ln(1-p), \\ \mathcal{L}'(p) &= \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (n_i - x_i) = \frac{1}{p(1-p)} \left(\sum_{i=1}^n x_i - pn_i \right) = 0, \\ \therefore p &= \frac{\sum x_i}{\sum n_i}. \end{aligned}$$

Thus, estimators for transition probabilities of interest are:

$$\begin{aligned} p_{U,H/A} &= \frac{\sum {}_tH_N + \sum {}_tA_N}{\sum {}_tU}, \\ p_{U,D} &= \frac{\sum {}_tU_{D.CDA}}{\sum {}_tU}, \\ p_{H/A,D} &= \frac{\sum {}_tH_{D.CDA}}{\sum {}_tH} + \frac{\sum {}_tA_{D.CDA} + \sum {}_tA_{D.A}}{\sum {}_tA}. \end{aligned}$$

With the fitted model, we can address the problem of censorship. Firstly, it is easy to find a continuous-time Markov chain that, when discretized by months, turns into the discrete time-homogeneous Markov model we have just fitted to the empirical data. In order to obtain the transition matrix for the discrete time-homogeneous Markov chain from a continuous-time Markov chain with generator Q , we use the formula $P = e^{tQ}$; thus, to reconstruct the generator function we use $tQ = \ln P$.

With the interval censorship removed from the distribution, the best way to address left-censorship is to shift all the information by a period of 6 months to account for the undetectability of HIV in the early stages. With exponential sojourn times, accounting for the starting age of observation does not yield any improvements, due to the lack of memory property of the exponential distribution.

With the individual epidemic Markov model derived, it is straightforward and described in (Johnson, Kemp, & Kotz, 2005) how to obtain the population epidemic Markov model, given the individual model generator Q and the initial population by groups, the mean vector and covariance matrix are expressed as follows:

$$x(t) = x(0)e^{tQ}, \quad (1)$$

$$\mathcal{C}(t) = \text{Cov}[x(t)] = \text{Diag}(x(0)e^{tQ}) - e^{tQ^\top} \text{Diag}(x(0))e^{tQ}. \quad (2)$$

3. Simulation testing

Using results from the previous section, we obtain the following estimate for the discrete-time transition matrix P :

$$\begin{bmatrix} 0.998931 & 3.361732e-05 & 0.001035343 \\ 0 & 9.450082e-01 & 0.054991825 \\ 0 & 0 & 1 \end{bmatrix}.$$

Estimate for the continuous counterpart of the same individual-level process generator Q is:

$$\begin{bmatrix} -0.001069532 & 3.459568e-05 & 0.001034936 \\ 0 & -0.056561701 & 0.056561701 \\ 0 & 0 & 0 \end{bmatrix}.$$

The estimated initial population vector is:

$$\hat{x}(0) = (43192218, 171418, _).$$

The number of dead individuals is set to zero, since we are interested in modeling only the number of people living with HIV or AIDS.

Expected vector after 1 month of observation:

$$\hat{x}(1) = (43145390, 153817, _).$$

Absolute percentage differences from observed data at $t = 1$:

$$\delta x(1) = (0.02018511, -0.1086147, _).$$

Expected vector after 3 months of observation:

$$\hat{x}(3) = (43051887, 124446, _).$$

Absolute percentage differences from observed data at $t = 3$:

$$\delta x(3) = (-0.03802216, -0.2782137, _).$$

It can be seen that forecasts based on the initial time of observation tend to underestimate the total number of individuals living with HIV/AIDS. This is caused by two factors: the assumption of exponential sojourn times, and the inability to account for the starting age at the beginning of the observation period. Estimate for the population process covariance matrix $\mathcal{C}(3)$:

$$\begin{bmatrix} 1795865.4 & -119254.9170 & -179489.9249 \\ -119254.9 & 124101.3380 & -518.8347 \\ -179489.9 & -518.8347 & 186522.1861 \end{bmatrix}.$$

We used the *msm* library for R (Jackson, 2007) to simulate the population process. To achieve this, we conducted multiple individual-level simulations and aggregated the resulting data into a simulated population-level process. This allowed us to obtain confidence intervals for the population-level values at each moment in time. By performing multiple simulations, we derived empirical confidence intervals for the population process at each time point. The results are shown on the Fig. 3-5

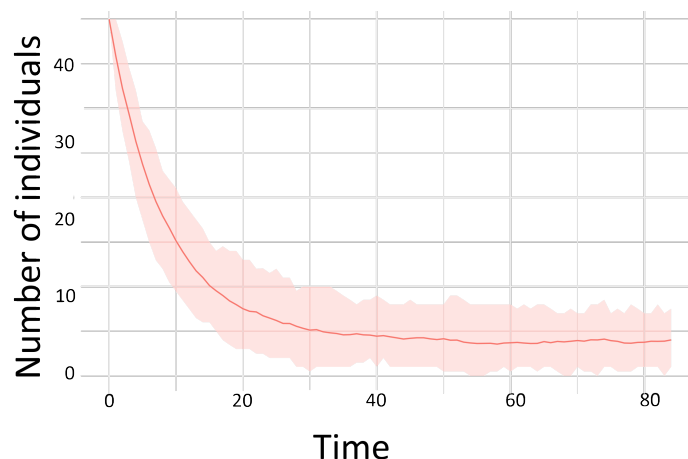


Fig. 3. CI for the number of people with HIV/AIDS with population 10000 and proportional counts

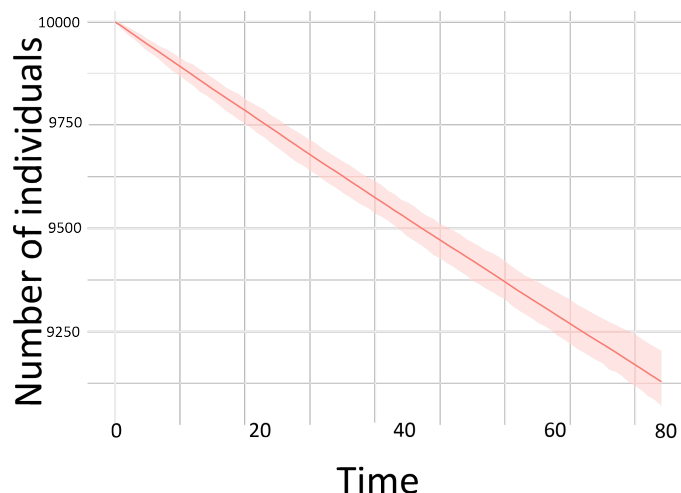


Fig. 4. CI for the number of undiagnosed people with or without HIV/AIDS with population 10000 and proportional counts

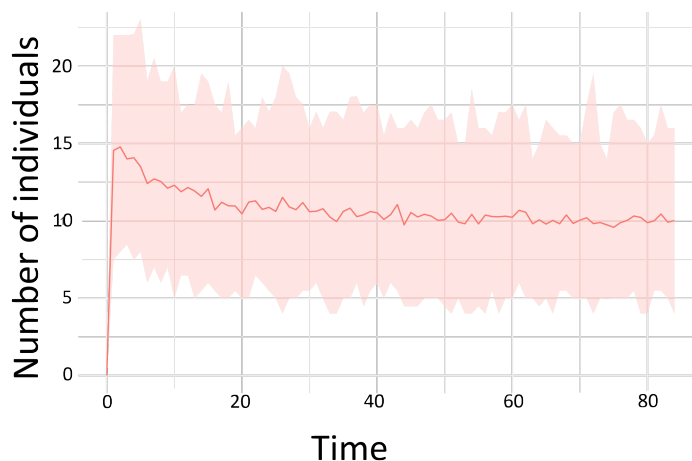


Fig. 5. CI for the number of dead people with or without HIV/AIDS with population 10000 and proportional counts

Within this model, we did not account for population growth, since most people who tested positive for HIV or AIDS were aged 15 or older. The number of parenteral sources of infection, as can be seen in (HIV/AIDS.Data, 2025), is negligible. Additionally, as stated in (Fowler, Flynn, & Aizire, 2018), children born to women living with HIV/AIDS undergo prophylaxis, and less than 1 % of such children contract HIV; thus, accounting for newborns or other parenterally infected individuals (Kleinman, & Busch, 2000) is unnecessary during the observation period.

Discussions and conclusions

This study explored the dynamics of the HIV/AIDS epidemic in Ukraine using a continuous-time Markov process model fitted to the empirical data. The model assumes that the time to transition between states U, H/A, and D is exponentially distributed, which carries an inherent assumption about memoryless progression. While analytically tractable, this assumption introduces certain limitations in capturing temporal variability in transition behavior.

The forecasts generated by the model tend to underestimate the future number of individuals living with HIV/AIDS. This bias is attributable to the non-constant nature of transition intensities over time, as shown in the Fig. 6:

In particular, the observed variation suggests that the assumption of homogeneous transition rates is too restrictive for capturing the evolving epidemiological landscape-driven by factors such as changes in testing practices, treatment accessibility, and behavioral patterns.

A natural methodological improvement would involve changing the model to model the individual level as a semi-Markov process, in which sojourn times follow more flexible distributions, such as the Weibull distribution. This extension would relax the memoryless assumption and allow for modeling duration-dependent transitions. Importantly, it would also enable more accurate inference about long-term epidemic trajectories.

The semi-Markov approach has a discrete-time counterpart in time-inhomogeneous Markov chains, which could be used to estimate a new model's parameters using LSE estimation on available data. Moreover, such models offer a framework for incorporating the starting age of individuals, thereby addressing the left-censoring issue present in the current analysis. Integrating age-structured entry would allow the model to more faithfully represent the natural history of HIV progression in a population with varied entry times into the observation window.

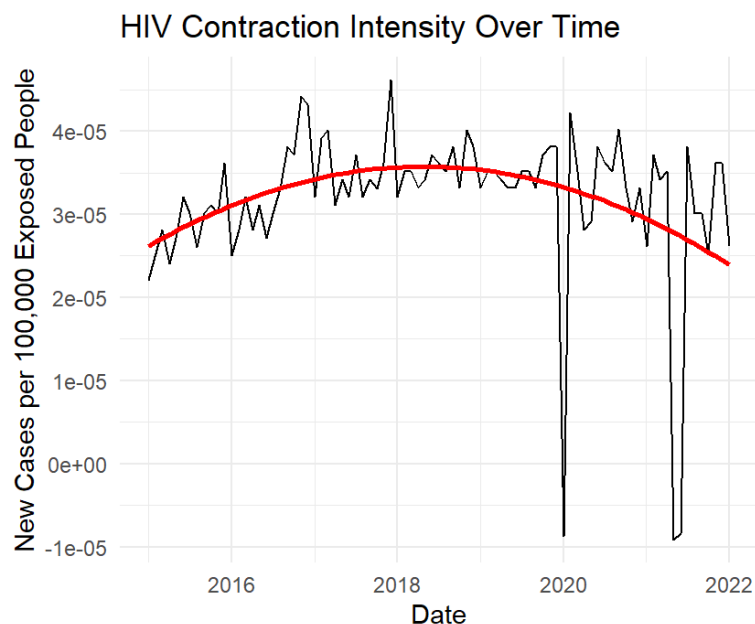


Fig. 6. HIV contraction intensity over time

Another critical extension involves incorporating reverse transitions from AIDS to HIV, enabled by ART. Estimating parameters for $A \rightarrow H$ transition in a 4-state Markov model would more realistically reflect disease dynamics and improve the model's descriptive and predictive accuracy. However, this extension requires access to detailed medical data on patient-level health outcomes and ART-adherence information, which are currently limited.

Finally, improving the model's demographic realism could be achieved by estimating the inflow of newly observed individuals aged 15 and above, which constitutes the age threshold for inclusion in most HIV surveillance programs. Modeling this influx explicitly would support a more robust population-level framework, accommodating both epidemiological transitions and demographic processes.

In summary, the proposed refinements—semi-Markov modeling, inclusion of reverse transitions, and dynamic population entry—represent promising directions for improving the accuracy and applicability of stochastic models of HIV/AIDS in Ukraine. These extensions would allow for more nuanced analysis of epidemic trajectories and provide more informative guidance for public health policy and intervention planning.

In this article, we constructed and justified a reasonable model for the process of HIV/AIDS diagnostics. We fitted continuous-time individual-level and population-level Markov processes to describe the HIV/AIDS epidemic in Ukraine, estimating the discrete-time transition matrix P and its continuous-time counterpart – the generator matrix Q . We managed to retrieve information about distributions lost due to interval censoring, obtained the mean vector of the population Markov process $x(t)$, and its covariance matrix $C(t)$. Finally, we used Monte Carlo simulations to build confidence intervals for the counting process of the total number of people living with HIV/AIDS.

Authors' contribution: Olga Vasylyk - conceptualization and methodology; Yehor Okuniey - methodology and research, software, simulation, and analysis of the results, and preparation of the literature review.

Acknowledgments. The authors would like to thank the anonymous referees for their useful and detailed comments on an earlier version of the manuscript.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Barnett, T., Whiteside, A., Khodakevich, L., Kruglov, Y., & Steshenko, V. (2000). The HIV/AIDS epidemic in Ukraine: its potential social and economic impact. *Social science & medicine*, 51(9), 1387–1403. [https://doi.org/10.1016/s0277-9536\(00\)00104-0](https://doi.org/10.1016/s0277-9536(00)00104-0)
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. https://doi.org/10.1007/978-1-4612-4380-9_37
- Fowler, M., Flynn, P., & Aizire, J. (2018). What is new in perinatal HIV prevention? *Current opinion in pediatrics*, 30(1), 144–151. <https://doi.org/10.1097/mop.0000000000000579>
- HIV/AIDS.Data. (2025). *Statistics on HIV/AIDS in Ukraine collected by State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine*. <https://phc.org.ua/kontrol-zakhvoryuvan/vilsnid/statistika-z-vilsnidu>
- Jackson, C. (2007). *Multi-state modelling with R: the msm package*. Cambridge. <https://doi.org/10.32614/CRAN.package.msm>
- Johnson, N., Kemp, A., & Kotz, S. (2005). *Univariate discrete distributions*. John Wiley and Sons. <https://doi.org/10.1002/0471715816>
- Kleinman, S., & Busch, M. (2000). The risks of transfusion-transmitted infection: direct estimation and mathematical modelling. *Best Practice and Research Clinical Haematology*, 13(4), 631–649. <https://doi.org/10.1053/beha.2000.0104>
- Kruglov, Y., Kobyshecha, Y., Salyuk, T., Varetska, O., Shakarishvili, A., & Saldanha, V. (2008). The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007. *Sexually transmitted infections*, 84(1), 37–41. <https://doi.org/10.1136/sti.2008.031195>

- Miranda, M., Pingarilho, M., Pimentel, V., Martins, M. R. O., Vandamme, A.-M., Bobkova, M., Böhm, M., Seguin-Devaux, C., Paredes, R., Rubio, R., Zazzi, M., Incardona, F., & Abecasis, A. (2021). Determinants of hiv-1 late presentation in patients followed in europe. *Pathogens*, 10(7), 835. <https://doi.org/10.3390/pathogens10070835>
- Parczewski, M., Gökengin, D., Sullivan, A., de Amo, J., Cairns, G., Bivol, S., Kuchukhidze, G., Vasylyev, M., & Rockstroh, J. K. (2025). Control of HIV across the WHO European region: progress and remaining challenges. *The Lancet Regional Health–Europe*, 52, 1–14. <https://doi.org/10.1016/j.lanepe.2025.101243>
- Raboud, J., Rae, S., Woods, R., Harris, M., Montaner, J., & INCAS and AVANTI Study Groups. (2002). Consecutive rebounds in plasma viral load are associated with virological failure at 52 weeks among HIV-infected patients. *Aids*, 16(12), 1627–1632. <https://doi.org/10.1097/00002030-200208160-00008>
- State Statistics Service of Ukraine. (2022). *Mortality rates by regions of Ukraine, 2015–2022 [data set]*. https://www.ukrstat.gov.ua/operativ/operativ2022/ds/kjp_reg/arh_kjp_reg2022_u.html

Отримано редакцією журналу / Received: 08.08.25
Прорецензовано / Revised: 27.11.25
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Ольга ВАСИЛИК, д-р фіз.-мат. наук, доц.

ORCID ID: 0000-0002-0880-3751

e-mail: vasylyk@matan.kpi.ua

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна

Єгор ОКУНЄВ, магістр

ORCID ID: 0009-0009-3695-490X

e-mail: egorky96@gmail.com

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”,
Київська школа економіки, Київ, Україна

МОДЕЛЮВАННЯ ЕПІДЕМІЇ ВІЛ/СНІД В УКРАЇНІ ЯК МАРКОВСЬКИЙ ПРОЦЕС ІЗ НЕПЕРЕРВНИМ ЧАСОМ

Упродовж тривалого часу Україна входить до переліку країн з найвищими темпами епідемії ВІЛ/СНІД. У представленій роботі оцінено параметри марковського ланцюга з неперервним часом, щоб описати процес до діагностування ВІЛ/СНІД та, врешті-решт, смерть індивіда, і в такий спосіб отримати модель епідемії на рівні всієї популяції. Завдяки аналізу здійснено спробу відновити інформацію, втрачену через ліве та інтервальне цензурування, наявне в емпіричних даних. Запропонована методологія поєднує аналіз виживання, оцінювання максимальної правдоподібності, моделювання методом Монте-Карло та побудову довірчих інтервалів. Отримані результати дають змогу краще зрозуміти динаміку епідемії ВІЛ/СНІД в Україні та надалі можуть бути використані для актуарних досліджень епідемії, розглядаючи зараження ВІЛ як страховий ризик.

Ключові слова: моделі виживання, кратні декременти, моделі Маркова, симуляція, випадкові процеси.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.21:519.23

DOI: <https://doi.org/10.17721/1812-5409.2026/1.5>

Oleksandr VOLKOV, Master's Student
 ORCID ID: 0009-0004-0247-9921
 e-mail: oleksandr_volkov@berkeley.edu
 University of California, Berkeley, Berkeley, USA

Yurii VOLKOV, DSc (Phys. & Math.), Prof.
 ORCID ID: 0000-0002-2270-3407
 e-mail: yuriiivolkov38@gmail.com
 Volodymyr Vynnychenko Central Ukrainian State University, Kropyvnytskyi, Ukraine

Nataliia VOINALOVYCH, PhD, Assoc. Prof.
 ORCID ID: 0000-0002-0523-7889
 e-mail: voinalovychn@gmail.com
 Volodymyr Vynnychenko Central Ukrainian State University, Kropyvnytskyi, Ukraine

ON THE CLASS OF EXPONENTIAL STATISTICAL STRUCTURES OF TYPE B

The article is devoted to the study of exponential statistical structures of type B, which constitute an important subclass of exponential families of probability distributions. This class is characterized by a number of analytical and probabilistic properties that make it a convenient tool for solving both theoretical and applied problems in mathematical statistics. The relevance of this research lies in the need to generalize known classes of distributions and to develop a unified framework for their analysis, which is essential for applications in stochastic modeling, machine learning, and financial mathematics.

The paper proposes a formal definition of type B distributions based on the Laplace transform of dominating measures and a system of functional-differential equations describing their structure. Necessary and sufficient conditions for a statistical structure to belong to class B are established, and it is proved that such structures can be represented through a dominating measure with an explicit Laplace transform. The obtained results make it possible to describe a wide range of well-known one-dimensional and multivariate distributions, including the Binomial, Poisson, Normal, Gamma, Polynomial, and Logarithmic distributions, as well as specific cases such as the Borel–Tanner and Random Walk distributions.

Particular attention is given to the proof of structural theorems that determine the stability of class B under linear transformations and the addition of independent random vectors. It is shown that if a distribution belongs to class B, its linear transformations and sums also belong to this class with the corresponding parameters. Recursive relations for initial and central moments as well as for semi-invariants are obtained, providing an efficient analytical and computational framework for their evaluation.

Furthermore, the tails of type B distributions are investigated using the properties of the Laplace transform. As a result, new exponential inequalities for estimating the probabilities of large deviations are derived, which extend classical approaches to the analysis of statistical distributions. The obtained results can be applied in theoretical studies and in practical problems of stochastic modeling.

Key words: exponential statistical structures, class B, probability distributions, Laplace transform, stochastic modeling, linear operators.

AMS 2020 classification: 62E10, 62H12, 62F10.

Introduction

The concept of an exponential statistical structure was formed in the mid-20th century within the framework of the theory of exponential families of distributions. The first systematic constructions appeared in the works of J.-R. Barra (Barra, 1981), who developed the analytical foundations of mathematical statistics and introduced the concept of natural parameters of the exponential family. Classical discrete distributions – Binomial, Poisson, and Negative Binomial – were generalized in the monograph by Johnson and Kotz (Johnson, & Kotz, 1969). Further development of the theory was achieved through the works of A. Noack (Noack, 1950), who proposed a new class of discrete distributions with the property of closure with respect to the addition operation, and C. Morris (Morris, 1982), who systematized Natural Exponential Families with Quadratic Variance Functions (NEF–QVF). The classical works of Kendall and Stuart (Kendall, & Stuart, 1969) presented the foundations of the statistical theory of moments and cumulants, which became the basis for further analytical developments.

At the same time, starting from the 2000s, exponential structures have been considered not only as an analytical tool but also as part of the broader framework of information geometry. In particular, in the works of S. Amari (Amari, 2016), Ay and Jost (Ay, Jost, & Lê, 2017), as well as Kass and Vos (Kass, & Vos, 2019), it was shown that exponential families can be described through the Riemannian structure of the parameter space, where the Fisher–Rao metric is defined as $g_{ij}(\theta) = \mathbb{E}_{\theta} \left[\frac{\partial \ln f(X; \theta)}{\partial \theta_i} \frac{\partial \ln f(X; \theta)}{\partial \theta_j} \right]$. This approach made it possible to establish a connection between classical exponential distributions, Bayesian models, and the geometric properties of statistical manifolds.

Furthermore, generalizations of the classical NEF–QVF constructions have emerged. In particular, in the works of Letac and Mora (Letac, & Mora, 1990), natural exponential families with cubic variance functions (NEF–CVF) were studied, which are described by the relation $V(m) = am^3 + bm^2 + cm + d$, where $V(m)$ is the variance function, and m is the mean.

In recent years, the theory of exponential families has been further extended in both theoretical and applied directions. In particular, Shushi (Shushi, 2018) introduced a generalized exponential family of distributions, providing new insights into their structural properties and extending classical characterizations beyond standard natural exponential families. These developments highlight the flexibility of exponential-type models in capturing complex stochastic behavior.

At the same time, Efron (Efron, 2022) presented a comprehensive modern treatment of exponential families, emphasizing their central role in contemporary statistical theory and applications, including Bayesian inference, large-scale data analysis, and machine learning. This work unifies classical results with modern perspectives, reinforcing the importance of exponential structures as a fundamental tool in statistics.

These advances demonstrate the continued relevance of exponential families and motivate the study of new subclasses and generalizations, such as the exponential statistical structures of type B considered in this paper.

Recently, Laplace transform methods have been reconsidered as a powerful tool for analyzing the asymptotic properties of exponential models. Katsevich et al. (Katsevich, 2024) developed a rigorous approach to the multidimensional Laplace expansion: $I(\lambda) = \int_{\mathbb{R}^n} e^{-\lambda\varphi(x)} a(x) dx \sim e^{-\lambda\varphi(x_0)} \left(\frac{2\pi}{\lambda}\right)^{n/2} \frac{a(x_0)}{\sqrt{\det \varphi''(x_0)}}, \lambda \rightarrow \infty$, where x_0 is the point of minimum of $\varphi(x)$, which provides precise applicability conditions in high dimensions.

Despite the existing results within the NEF–QVF and NEF–CVF frameworks, the class of exponential structures defined through the Laplace transform of dominating measures that satisfy a functional-differential equation linking parameter variables and moment characteristics has not yet been systematically described. This gap is filled by the present study.

The object of research is exponential statistical structures of type B, which form a subclass of exponential families of distributions.

The aim and objectives of the research are to establish the necessary and sufficient conditions for a statistical structure to belong to class B, to construct the corresponding analytical apparatus through the Laplace transform of dominating measures, and to prove structural theorems regarding the closure of this class with respect to linear transformations and convolutions.

1. Exponential statistical structures of type B

Among exponential statistical structures (Noack, 1950, p. 177–190), one can distinguish a subclass characterized by a number of important analytical properties.

Definition 1. We say that a statistical structure $(Q(x), x \in X)$, where $X \subseteq \mathbb{R}^m$ is an open set representing the parameter space, belongs to class B if there exists a function $\varphi(z, x)$, which is the Laplace transform of the measure $Q(x)$, and this function satisfies the equation:

$$\frac{\partial \varphi}{\partial z} + \frac{\partial \varphi}{\partial x} V(x) + x\varphi = 0, \quad \varphi(0, x) = 1 \quad \forall x \in X. \quad (1)$$

The distributions $Q(x)$ themselves will be called distributions of type B.

Theorem 1. For a statistical structure $\{Q(x), x \in X\}$ to belong to class B, it is necessary and sufficient that there exists a measure μ dominating the distributions $Q(x)$ such that

$$Q(x)(dt) = (u(s(x)))^{-1} \exp(-s(x)t') \mu(dt),$$

where $u(z)$ is the Laplace transform of the measure μ , and $s(x)$ is the s -characteristic of the measure μ .

Proof. Let us prove sufficiency. Suppose that there exists a measure μ such that

$$Q(x)(dt) = (u(s(x)))^{-1} \exp(-s(x)t') \mu(dt),$$

where $u(z)$ is the Laplace transform of μ .

Then the Laplace transform of $Q(x)$ is given by

$$\varphi(z, x) = \int_{\mathbb{R}^m} e^{-zt'} Q(x)(dt) = (u(s(x)))^{-1} \int_{\mathbb{R}^m} e^{-(z+s(x))t'} \mu(dt) = \frac{u(z+s(x))}{u(s(x))}.$$

Differentiating $\varphi(z, x)$ with respect to z and x , and using the properties of the Laplace transform, one verifies by direct substitution that $\varphi(z, x)$ satisfies equation (1). Hence, $Q(x)$ belongs to class B.

Now, let us prove necessity. Since $\varphi(0, x) = 1$, the function $\varphi(z, x)$ is the unique solution of equation (1).

Let us express it differently. To do this, fix an arbitrary point (z, x) and denote by $s(x)$ a particular solution of the equation $ds/dx = -W(x) = -(V(x))^{-1}$. Since for all $x \in X$ the matrix $W(x)$ is positive definite, there exists a neighborhood δ of the point x in which the mapping $s(x)$ has an inverse. Denote it by $x(s)$, $s \in \delta$. Next, consider the equation $db/dx = x(s)$. Since $dx_i/dx_j = dx_j/dx_i$, $i, j = \overline{1, m}$ (as a consequence of the symmetry of the matrix $W(x)$), this equation has at least one solution. Denote by $b(s)$ the analytic continuation of this solution to the strip $T_\delta = \{\zeta \in \mathbb{C}^m : \operatorname{Re} \zeta \in \delta\}$. Then, by direct substitution into equation (1), it can be verified that

$$\varphi(z, x) = \exp(\langle b(s(x)) - b(s(x) + z) \rangle).$$

Consider the measure

$$\mu(dt) = \exp(s(x)t' - b(s(x))) Q(x)(dt).$$

This measure does not depend on x , since its Laplace transform is the function $u(z) = \exp(-b(z))$, $z \in T_\delta$. Let us prove that the measure μ is the desired one. Indeed, consider the measure

$$q(x)(dt) = (u(s(x)))^{-1} \exp(-s(x)t') \mu(dt).$$

The Laplace transform of the measure $q(x)$ is the function $\exp(\langle b(s(x)) - b(s(x) + z) \rangle)$, i.e., $\varphi(z, x)$. Due to the one-to-one correspondence between measures and their Laplace transforms, this means that $q(x) = Q(x)$, which completes the proof. $\square$

In practice, the membership of a family of distributions $\{Q(x), x \in X\}$ in the class B is established by checking the relation (1).

We symbolically denote that the family $\{Q(x), x \in X\}$ belongs to the class B as:

$$Q(x) \in B(x; V(x)).$$

If $\xi \sim Q(x)$, we also write: $\xi \in B(x; V(x))$.

Many well-known families of distributions belong to the class B .

Example 1. The Binomial distribution $B_p^n \in B(np; np(1-p))$.

Example 2. The Poisson distribution $P_\lambda \in B(\lambda; \lambda)$.

Example 3. The Negative Binomial distribution

$$NB_p^n \in B(n(1-p)p^{-1}; np^{-2}(1-p)).$$

Example 4. The Normal distribution $\Phi_{\alpha, \sigma^2} \in B(\alpha; \sigma^2)$.

Example 5. The Gamma distribution $\Gamma_{\alpha, \lambda} \in B(\lambda/\alpha; \lambda/\alpha^2)$, (parameter λ is fixed). In particular, the exponential distribution $\Gamma_{\alpha, 1} \in B(\alpha^{-1}; \alpha^{-2})$.

Example 6. The Multivariate Normal distribution $\Phi_{\alpha, \sigma^2} \in B(\alpha; \sigma^2)$, where $\alpha = (\alpha_1, \dots, \alpha_m)$, $\sigma = (\sigma_{ij}), i, j = \overline{1, m}$.

Example 7. The Multinomial distribution

$$B_{(p_1, \dots, p_m)}^n \in B(n(p_1, \dots, p_m); np_i(\delta_{ij} - p_j)), i, j = \overline{1, m}.$$

Example 8. The Negative Multinomial distribution (Johnson, & Kotz, 1969, p. 292)

$$NB_{(p_1, \dots, p_m)}^n \in B(n(p_1, \dots, p_m); np_i(\delta_{ij} + p_j)), i, j = \overline{1, m}.$$

Example 9. Consider the Multivariate Logarithmic distribution $\Lambda_{(\theta_1, \dots, \theta_m)}$ (Johnson, & Kotz, 1969, p. 303), defined as follows: let $k = (k_1, \dots, k_m) \in \mathbb{N}^m, k_1 + \dots + k_m \geq 1$. Then

$$\begin{aligned} \xi \in \Lambda_{\{\theta_1, \dots, \theta_m\}} &\iff \\ \iff P\{\xi = (k_1, \dots, k_m)\} &= -(k_1 + \dots + k_m - 1)(k_1! \dots k_m!)^{-1} \theta_1^{k_1} \dots \theta_m^{k_m} \log(1 - \theta_1 - \dots - \theta_m)^{-1}, \\ 0 < \theta_i < 1, \quad \theta_1 + \dots + \theta_m &< 1. \end{aligned}$$

This family of distributions belongs to the class B :

$$\Lambda_{(\theta_1, \dots, \theta_m)} \in B\left(-\frac{(\theta_1, \dots, \theta_m)}{(1 - |\theta|) \log(1 - |\theta|)}; -\frac{\theta_i}{(1 - |\theta|) \log(1 - |\theta|)}\left(\delta_{ij} + \frac{\theta_j(1 + \log(1 - |\theta|))}{(1 - |\theta|) \log(1 - |\theta|)}\right)\right),$$

where $(i, j = \overline{1, m}, \quad |\theta| = \theta_1 + \dots + \theta_m)$.

In the univariate case:

$$\Lambda_\theta \in B\left(-\frac{\theta}{(1 - \theta) \log(1 - \theta)}, -\frac{\theta}{(1 - \theta)^2 \log(1 - \theta)}\left(1 + \frac{\theta}{\log(1 - \theta)}\right)\right).$$

Example 10. Define the Multivariate Random Walk distribution $B_{(p_1, \dots, p_m)}^n$.

Let

$$k \in \mathbb{N}^m, \quad |k| = k_1 + \dots + k_m, \quad 0.5 < p_i < 1, \quad i = \overline{1, m}, \quad p = \frac{1}{2} \left(1 + \left(\sum_{i=1}^m (2p_i - 1)^{-1}\right)^{-1}\right).$$

Then

$$\xi \in B_{(p_1, \dots, p_m)}^n \iff P\{\xi = k\} = \frac{n(|k| - 1)}{|k|!} \left(\frac{(|k| + n)/2}{|k|}\right) p^{|k|+n} (1-p)^{|k|-n} \prod_{i=1}^m \left(\frac{2p_i - 1}{2p_i - 1}\right)^{k_i}.$$

We have for $i, j = \overline{1, m}$:

$$B_{(p_1, \dots, p_m)}^n \in B\left(n((2p_1 - 1)^{-1}, \dots, n(2p_m - 1)^{-1}); n(2p_i - 1)^{-1}(\delta_{ij} - (2p_j - 1)^{-1})(2(2p - 1) - (2p - 1)^{-1})\right).$$

In the univariate case:

$$B_p^n \in (n(2p-1)^{-1}; n(2p-1)^{-1}((2p-1)^{-2}-1)).$$

Example 11. Define the Multivariate Borel–Tanner distribution $BT_{(\alpha_1, \dots, \alpha_m)}^n \in B$.

Let

$$k \in \mathbb{N}^m, \quad 0 < \alpha_j < 1, \quad i = \overline{1, m}, \quad \alpha = 1 - \left(\sum_{i=1}^m (1 - \alpha_i)^{-1} \right)^{-1}.$$

Then

$$\xi \in BT_{(\alpha_1, \dots, \alpha_m)}^n \iff P\{\xi = k\} = \frac{|k|!}{k! (|k| - n)!} n |k|^{n-1} \prod_{i=1}^m \left(\frac{1 - \alpha_i}{1 - \alpha_j} \right)^{k_i} \exp(-|k| \alpha) \alpha^{|k| - n}.$$

We have for $i, j = \overline{1, m}$:

$$BT_{(\alpha_1, \dots, \alpha_m)}^n \in B \left(n((1 - \alpha_1)^{-1}, \dots, (1 - \alpha_m)^{-1}); n(1 - \alpha_j)^{-1} (\delta_{ij} - (1 - \alpha_j)^{-1} (2 - ((1 - \alpha)^{-1} - \alpha))) \right).$$

In the univariate case:

$$BT_{\alpha}^n \in B \left(\frac{n}{1 - \alpha}; \frac{n\alpha}{(1 - \alpha)^3} \right).$$

The class B also includes the Noack distribution family [3] and the so-called natural exponential families with quadratic variance function (NEF-QVF distributions) (Morris, 1982).

Theorem 2. Let $a = (a_{ij}), i, j = \overline{1, m}$ be an arbitrary non-singular matrix, and $b = (b_1, \dots, b_m)$ an arbitrary vector. Then, if $\xi \in B(x, V(x))$, we have

$$\eta = a\xi + b \in B(ax + b; aV(x)a').$$

Proof. To prove the statement, it is sufficient to verify the equality

$$\frac{\partial \varphi_{\eta}}{\partial z} + \frac{\partial \varphi_{\eta}}{\partial y} V_{\eta}(y) - y \varphi_{\eta} = 0,$$

with $y = ax + b$, where $\varphi_{\eta} = \varphi_{\eta}(z, y)$ is the Laplace transform of the distribution of the random vector η , and $V_{\eta}(y) = \text{cov } \eta$. The verification of the given equality is carried out in coordinate form.

Note that $M_{\eta} = y = ax + b$, hence

$$y_j = \sum_{k=1}^m a_{jk} x_k + b_j, \quad j = \overline{1, m}, \quad \text{cov } \eta = \text{cov}(a\xi + b) = aV(x)a',$$

therefore

$$V_{\eta} = \left(\sum_{k,r=1}^m a_{ir} a_{jr} \nu_{rk}(x) \right), \quad i, j = \overline{1, m}, \quad \varphi_{\eta} = M \exp(-z(a\xi + b)') = \exp(-zb') \varphi(az, x).$$

Let $\omega = az$, $\omega = (\omega_1, \dots, \omega_m)$. Then

$$\begin{aligned} & \frac{\partial \varphi_{\eta}}{\partial z_i} + \sum_{j=1}^m \left(\sum_{k,r=1}^m a_{ir} a_{jr} \nu_{rk}(x) \right) \frac{\partial \varphi_{\eta}}{\partial y_j} + y_j \varphi_{\eta} = \\ & = \exp(-zb') \left(\sum_{r=1}^m a_{ri} \frac{\partial \varphi(\omega, x)}{\partial \omega_r} + \sum_{q=1}^m \frac{\partial \varphi(\omega, x)}{\partial x_q} \sum_{k,r=1}^m \left(\sum_{j=1}^m a_{qj}^{(-1)} a_{jk} \right) a_{ir} \nu_{rk}(x) + \sum_{r=1}^m a_{ir} x_r \varphi(\omega, x) \right), \end{aligned}$$

where $a_{qj}^{(-1)}$ are the elements of the matrix a^{-1} .

Since

$$\sum_{j=1}^m a_{qj}^{(-1)} a_{jk} = \delta_{qk},$$

we obtain

$$\exp(-zb') \sum_{r=1}^m a_{ir} \left(\frac{\partial \varphi(\omega, x)}{\partial \omega_r} + \sum_{q=1}^m \frac{\partial \varphi(\omega, x)}{\partial x_q} \nu_{rq}(x) + x_r \varphi(\omega, x) \right) = 0,$$

which completes the proof. $\square$

Theorem 3. Let $\xi_1, \dots, \xi_n, \dots$ be a sequence of independent and identically distributed random vectors, with $\xi_1 \in B(x; V(x))$. Then

$$\eta_n = \xi_1 + \dots + \xi_n \in B(nx; nV(x)).$$

The validity of this statement follows from the fact that

$$\varphi_{\eta_n}(z, x) = (\varphi(z, x))^n.$$

Corollary 1. $\zeta_n := (\xi_1 + \dots + \xi_n)/n \in B(x; n^{-1}V(x))$.

Theorem 4. If the distribution $Q(x) \in B(x; V(x))$, then $I(x) = V^{-1}(x)$, where $I(x)$ is the Fisher information matrix of the family $\{Q(x), x \in X\}$.

Proof. Let the measure μ dominate the distributions $Q(x)$. Then

$$(dQ(x)/d\mu)(t) = p(t, x) = \exp(-s(x)t' - \ln u(s(x))).$$

Let $q_{ij}(x), i, j = \overline{1, m}$, denote the elements of the matrix $I(x)$, and let $\xi \sim Q(x)$. Then

$$q_{ij}(x) = M \left(\frac{\partial \ln p(\xi, x)}{\partial x_i} \cdot \frac{\partial \ln p(\xi, x)}{\partial x_{i,j}} \right) = M \left((p(\xi, x))^{-2} \frac{\partial p(\xi, x)}{\partial x_i} \cdot \frac{\partial p(\xi, x)}{\partial x_{i,j}} \right),$$

and if we denote $\omega := s(x)t' + \ln u(s(x))$, then

$$q_{ij}(x) = L_1 \left(e^{2\omega} (e^{-\omega})_x^{(e_i)} (e^{-\omega})_x^{(e_j)}; x \right) = \omega_{ij}(x) = W(x) = V^{-1}(x),$$

which completes the proof. $\square$

Theorem 5. Let $Q(x) \in B(x; V(x))$, $a_k = a_k(x)$ be the raw moments of order $k = (k_1, \dots, k_m)$ of the distribution $Q(x)$, $\beta_k = \beta_k(x)$ the central moments, and $\sigma_k = \sigma_k(x)$ the semivariances. Then for all $k \in \mathbb{N}^m$ the following recurrence relations hold:

$$a_{k+e_i} = \sum_{j=1}^m \nu_{ij}(x) \frac{\partial a_k}{\partial x_j} + x_i a_k, \quad a_{e_i} = x_i, \quad i = \overline{1, m}; \quad (2)$$

$$\beta_{k+e_i} = \sum_{j=1}^m \nu_{ij}(x) \left(\frac{\partial \beta_k}{\partial x_j} + k_j \beta_{k-e_j} \right), \quad \beta_0 = 1, \quad \beta_{e_i} = 0, \quad i = \overline{1, m}; \quad (3)$$

$$\sigma_{k+e_i} = \sum_{j=1}^m \nu_{ij}(x) \frac{\partial \sigma_k}{\partial x_j}, \quad \sigma_{e_i} = x_i, \quad i = \overline{1, m}. \quad (4)$$

Proof. Rewrite relation (1) in coordinate form:

$$\frac{\partial \varphi}{\partial z_i} + \sum_{j=1}^m \nu_{ij}(x) \frac{\partial \varphi}{\partial x_j} + x_i \varphi = 0, \quad i = \overline{1, m}. \quad (5)$$

Since $a_k = (-1)^k \varphi^{(k)}(0, x)$, differentiating equality (5) k times and setting $z = 0$ yields (2).

Let $\beta_k(x)$ be the central moments related to the initial moments $a_k(x)$ by the relation

$$\beta_k(x) = \sum_{p \leq k} (-1)^{|k-p|} C_k^p a_p(x) a_1(x)^{k-p}.$$

Differentiating with respect to x_j and taking into account (2), we have

$$\frac{\partial \beta_k}{\partial x_j} = \sum_{p \leq k} \frac{\partial \beta_k}{\partial a_p} (a_{p+e_j} - x_j a_p) = \sum_{p \leq k} \frac{\partial \beta_k}{\partial a_p} a_{p+e_j} - k_j \beta_{k-e_j}.$$

Substituting this into (2), we obtain (3).

To obtain relations (4), note that $\sigma_k = (-1)^k (\ln \varphi(z, x))_{z=0}^{(k)}$. Moreover, from relations (1) it follows that the function $\psi = \ln \varphi(z, x)$ satisfies the equation

$$\frac{\partial \psi}{\partial z} + V(x) \frac{\partial \psi}{\partial x} + x = 0,$$

or in coordinate form

$$\frac{\partial \psi}{\partial z_i} + \sum_{j=1}^m \nu_{ij}(x) \frac{\partial \psi}{\partial x_j} + x_i = 0, \quad i = \overline{1, m},$$

from which (4) follows. $\square$

Corollary 2. In the one-dimensional case ($m = 1$) the following relations hold:

$$a_{k+1} = V(x) \frac{\partial a_k}{\partial x} + x a_k, \quad a_1 = x, \quad k = 1, 2, \dots$$

$$\beta_{k+1} = V(x) \left(\frac{\partial \beta_k}{\partial x} + k \beta_{k-1} \right), \quad \beta_0 = 1, \quad \beta_1 = 0, \quad k = 1, 2, \dots$$

$$\sigma_{k+1} = V(x) \frac{\partial \sigma_k}{\partial x}, \quad \sigma_1 = x, \quad k = 1, 2, \dots \quad (6)$$

The obtained recurrence relations for moments and semivariances are quite convenient for their computation. In some cases, all moments and semivariances can be derived explicitly from these relations.

We limit ourselves to some examples for finding semivariances, since there exist formulas expressing moments in terms of semivariances (Kendall, & Stuart, 1969).

Example 12. Let $V(x) = x(ax + b)$, where a, b are arbitrary constants. Then the solution of equation (6) (can be verified directly) is the function

$$\sigma_{k+1}(x) = x(ax + b) \sum_{m=1}^k m! b^{k-m} S(k, m) (ax)^{m-1}, \quad k = 1, 2, \dots \quad (7)$$

where $S(k, m)$ are the Stirling numbers of the second kind, i.e.,

$$S(k, m) = \frac{1}{m!} \sum_{j=0}^m \binom{m}{j} (-1)^j (m-j)^k.$$

In particular, setting $a = -1$, $b = 1$, $x = np$ in (7), we obtain explicit formulas for the semivariances of the binomial distribution B_p^n :

$$\sigma_{k+1} = np(1-p) \sum_{m=1}^k m! S(k, m) p^{m-1}, \quad k = 1, 2, \dots$$

setting $a = 1$, $b = 1$, $x = n(1-p)p^{-1}$, we obtain explicit formulas for the semivariances of the negative binomial distribution NB_p^n :

$$\sigma_{k+1} = np^{-2}(1-p) \sum_{m=1}^k m! S(k, m) p^{1-m} (1-p)^{m-1}, \quad k = 1, 2, \dots$$

setting $a = 0$, $b = 1$, $x = \lambda$, we obtain the semivariances of the Poisson distribution P_λ :

$$\sigma_k = \lambda, \quad k = 1, 2, \dots \quad (8)$$

setting $a = \lambda^{-1}$, $b = 0$, $x = \lambda a^{-1}$, we obtain the semivariances of the gamma distribution $\Gamma_{\lambda, a}$:

$$\sigma_k = (k-1)! \lambda a^{-k}. \quad (9)$$

Relations (8) and (9) are known (Kendall, & Stuart, 1969, p. 108–109).

Example 13. Consider Bernoulli random walks on a line: a particle moves along the axis t at times $t = 1, 2, \dots$, it moves left with probability $p > 1/2$ and right with probability $q = 1 - p$ by one step. Suppose initially the particle is at position $t = n$. The walk ends if the particle reaches the origin. Let ξ be the random variable representing the time of the walk. This random variable can take values $n, n+2, n+4, \dots$ with probabilities (Volkov, 1992, p. 152)

$$\omega_{nk} = \frac{n}{k} \binom{k}{(k+n)/2} p^{(k+n)/2} q^{(k-n)/2}, \quad k = n, n+2, \dots$$

Since the generating function of this distribution is $u(s) = \left(1 - \sqrt{1 - 4pqs^2}\right)^n (2qs)^{-n}$, it can be easily verified that

$$\xi \in B \left(\frac{n}{2p-1}, \left(\frac{n}{2p-1} \right)^3 - \frac{n}{2p-1} \right),$$

thus the semivariances of the random variable ξ satisfy the equation:

$$\sigma_{k+1}(x) = (x^3 n^{-2} - x) \sigma'_k(x), \quad \sigma_1(x) = x, \quad x = \frac{n}{2p-1}, \quad k = 1, 2, \dots$$

It can be directly verified that the solution of this equation is the function

$$\sigma_{k+2} = (x^3 n^{-2} - x) \sum_{r=0}^m (-1)^{k+r} (2r+1)! c_{k,r} (x n^{-1})^{2r},$$

where

$$c_{k,r} = \sum_{m=r}^k \binom{k}{m} 2^{m-r} S(m, r), \quad k = 0, 1, 2, \dots$$

Next, we will consider the estimation of the tails of distributions of type B .

Theorem 6. Let $Q(x) \in B(x; V(x))$, $x \in X$, and $s(x)$ be the s -characteristic of the measure μ dominating the distributions $Q(x)$. Then for any y such that $s(x) \geq s(y)$, the following inequality holds:

$$Q(x)(\{t : t \geq y\}) \leq \exp \left(-(y-x) \left(\int_0^1 (1-t) V'(x+t(y-x)) dt \right) (y-x)' \right). \quad (10)$$

Remark 1. If $t = (t_1, \dots, t_m)$, $y = (y_1, \dots, y_m)$, then

$$t \geq y \Leftrightarrow t_1 \geq y_1, \dots, t_m \geq y_m;$$

$$s(x) \geq s(y) \Leftrightarrow s_1(x_1, \dots, x_m) \geq s_1(y_1, \dots, y_m), \dots, s_m(x_1, \dots, x_m) \geq s_m(y_1, \dots, y_m).$$

Proof. Consider the operator

$$L_1(f; x) = \int_{\mathbb{R}^m} f(t) Q(x)(dt).$$

Let μ dominate the distributions $Q(x)$, $s(x)$ be the s -characteristic of μ , and $u(z)$ be the Laplace transform of μ . Then, if $\eta(t)$ is the Heaviside unit function, for any vector $a = (a_1, \dots, a_m)$ with non-negative coordinates, $\eta(t_1) \dots \eta(t_m) \leq \exp(at')$, and hence for any $y \in \mathbb{R}^m$

$$L_1(\eta(t_1 - y_1) \dots \eta(t_m - y_m); x) \leq L_1(\exp(a(t - y)'); x) = \exp \left(- \left(ay' - \ln \frac{u(s(x) - a)}{u(s(x))} \right) \right).$$

The last inequality is equivalent to

$$Q(x)\{t : t \geq y\} \leq \exp \left(- \left(ay' - \ln \frac{u(s(x) - a)}{u(s(x))} \right) \right). \quad (11)$$

Set $\beta(a) = ay' - \ln \frac{u(s(x) - a)}{u(s(x))}$, and using properties of the Laplace transform of μ , we obtain

$$\frac{d\beta}{da} = y - x(s(x) - a), \quad x(s) = -d \ln u(s)/ds$$

– the mapping inverse to the mapping

$$s(x) : \frac{d^2\beta}{da^2} = -V(x(s(x) - a)).$$

Since the matrix V is positive definite, the function $\beta(a)$ is convex, and hence at the point $a(y) = (a_1(y), \dots, a_m(y))$ defined by the system

$$y = x(s(x) - a).$$

it attains a maximum. Therefore,

$$a(y) = s(x) - s(y).$$

Set

$$A(y) = \beta(a(y)) = a(y)y' - \ln \frac{u(s(x) - a(y))}{u(s(x))},$$

which is the maximum value of the function $\beta(a)$. The expression for $A(y)$ can be simplified. Note that $A(x) = 0$. Moreover,

$$\frac{dA}{dy} = \left(\frac{da_1}{dy}, \dots, \frac{da_m}{dy} \right) y' + a(y) - x(s(y)) \left(\frac{ds_1}{dy}, \dots, \frac{ds_m}{dy} \right) = a(y).$$

Hence, $\frac{dA}{dy} \Big|_{y=x=0}$. Further,

$$\frac{d^2A}{dy^2} = \frac{da}{dy} = -\frac{ds(y)}{dy} = -V^{-1}(y).$$

Then, representing the function $A(y)$ by its Taylor expansion around the point x with the remainder in integral form, we obtain

$$A(y) = (y-x) \left(\int_0^1 (1-t) V^{-1}(x+t(y-x)) dt \right) (y-x)'.$$

Since inequality (11) holds for any a with non-negative coordinates, it also holds for $a = a(y)$, if $s(x) \geq s(y)$ (for example, if all elements of V^{-1} are non-negative, then $a(y) \geq 0$ provided $x \leq y$). Hence, $Q(x)\{t : t \geq y\} \leq \exp(-A(y))$, which completes the proof. $\square$

Discussion and conclusions

The relations (2)–(4) obtained in this work generalize the classical results of Kendall and Stuart (Kendall, & Stuart, 1969) regarding moment equations, but unlike them, they are derived within a unified approach to exponential structures defined through the Laplace transform. This allows for the derivation of recurrent formulas for raw moments, central moments, and semi-invariants not only for individual distributions but also for a wide class of measure-dominated families that satisfy the

functional-differential equation (1). Such an approach provides a unified algorithm for computing analytical characteristics for various distributions, including Binomial, Poisson, Normal, Gamma, and Lognormal, which turned out to be special cases of the class $B(x; V(x))$.

Of particular importance are the obtained structural theorems (2)–(4), which prove the closure of the class B with respect to linear transformations and convolutions. This means that when independent random vectors are added or affine transformations of the distribution parameters are applied, the distribution remains within the same class. This property is fundamental for constructing multivariate stochastic models, particularly in problems of risk aggregation and modeling of cumulative random effects in financial processes.

An important result of the research is also the proof of exponential inequalities (10), which provide estimates of large deviation probabilities. This opens up the possibility of applying the methods of class B in problems of risk control, system reliability assessment, and the analysis of the tail behavior of distributions in statistical models.

The practical significance of the obtained results lies in the fact that structures of type B form a theoretical basis for constructing generalized exponential models capable of describing both symmetric and asymmetric stochastic processes with exponential tails. This makes them promising for modeling loss distributions in insurance, the behavior of financial assets, as well as for training machine learning models that take into account nonlinear variance dependencies.

Further research should be directed towards extending the theory of exponential structures of type B to non-stationary processes, analyzing multivariate parametric models with dependent components, and developing efficient numerical algorithms for computing moments and semi-invariants for high-dimensional distributions. A separate promising direction is the application of the obtained results to problems of optimal statistical inference, parameter estimation in complex stochastic systems, and the construction of adaptive models in machine learning and financial analytics.

Authors' contribution: Oleksandr Volkov – conceptualization, methodology, writing-viewing, and editing; Yurii Volkov – conceptualization, methodology; Nataliia Voinalovych – analysis and preparation of the literature, writing-viewing and editing.

Sources of funding. The research was conducted using the authors' own funds.

References

- Amari, S. (2016). *Information geometry and its applications*. Springer.
- Ay, N., Jost, J., & L  , H. V. (2017). *Information geometry*. Springer.
- Barra, J.-R. (1981). *Mathematical basis of statistics*. Academic Press, New York.
- Efron, B. (2022). *Exponential families in theory and practice* (Vol. 16). Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108773157>
- Johnson, N. L., & Kotz, S. (1969). *Distributions in statistics: Discrete distributions*. Houghton Mifflin, Boston.
- Kass, R. E., & Vos, P. W. (2019). *Geometrical foundations of asymptotic inference*. Springer.
- Katsevich, A. (2024). *The Laplace asymptotic expansion in high dimensions*. arXiv. <https://doi.org/10.48550/arXiv.2406.12706>
- Kendall, M., & Stuart, A. (1969). *The advanced theory of statistics. vol. 1: Distribution theory*. Charles Griffin, London.
- Letac, G., & Mora, M. (1990). Natural real exponential families with cubic variance functions. *The Annals of Statistics*, 18(1), 1–37. <https://doi.org/10.1214/aos/1176347491>
- Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Annals of Mathematical Statistics*, 10(2), 65–80.
- Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, 21(1), 127–132. <https://doi.org/10.1214/aoms/1177729894>
- Shushi, T. (2018). The generalized exponential family of distributions and its characteristics. *Communications in Statistics - Theory and Methods*, 47(10), 2520–2526. 10.1080/03610926.2017.1337730
- Volkov, Y. (1992). *Positive operators. approximation. probability*. NMC VO [in Ukrainian]. [Волков, Ю. (1992). *Додатні оператори. Наближення. Імовірність*. НМК ВО].

Отримано редакцією журналу / Received: 29.10.25
Прорецензовано / Revised: 09.04.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Олександр ВОЛКОВ, магістр
ORCID ID: 0009-0004-0247-9921
e-mail: oleksandr_volkov@berkeley.edu
Університет Каліфорнії, Берклі, Берклі, США

Юрій ВОЛКОВ, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-2270-3407
e-mail: yuriiivolkov38@gmail.com
Центральноукраїнський державний університет імені Володимира Винниченка, Кропивницький, Україна

Наталія ВОЙНАЛОВИЧ, канд. пед. наук, доц.
ORCID ID: 0000-0002-0523-7889
e-mail: vojnalovychn@gmail.com
Центральноукраїнський державний університет імені Володимира Винниченка, Кропивницький, Україна

КЛАС ЕКСПОНЕНЦІАЛЬНИХ СТАТИСТИЧНИХ СТРУКТУР ТИПУ В

Статтю присвячено дослідженню експоненціальних статистичних структур типу В, що формують важливий підклас експоненціальних сімей розподілів. Цей клас вирізняється низкою аналітичних і ймовірнісних властивостей, що робить його зручним інструментом для теоретичних і прикладних задач математичної статистики. Актуальність теми зумовлена потребою в узагальненні відомих класів розподілів і побудові єдиного апарату для їх аналізу, що має практичне значення у стохастичному моделюванні, машинному навчанні та фінансовій математиці.

У роботі запропоновано формальне означення розподілів типу В на основі перетворення Лапласа домінуючих мір і системи функціонально-диференціальних рівнянь, що описують їхню структуру. Встановлено необхідні й достатні умови належності статистичної структури до класу В, доведено, що такі структури можуть бути подані через домінуючу міру з явним перетворенням Лапласа. Отримані результати дають змогу описати широкий спектр відомих розподілів, серед яких біноміальний, пуассонівський, нормальний, гамма-розподіл, поліноміальний, логарифмічний, а також специфічні випадки: розподіл Бореля – Таннера та розподіли випадкових блукань.

Особливу увагу приділено доведенню структурних теорем, що визначають стійкість класу В відносно лінійних перетворень та операції додавання незалежних випадкових векторів. Показано, що коли розподіл належить до класу В, то його лінійні перетворення та суми також залишаються в цьому класі. Отримано рекурентні співвідношення для початкових і центральних моментів, а також для семінваріантів, що забезпечують ефективний апарат для їхнього аналітичного та чисельного обчислення.

Крім того, досліджено властивості "хвостів" розподілів типу В за допомогою характеристик перетворення Лапласа. У підсумку виведено нові експоненціальні нерівності для оцінювання ймовірностей великих відхилень, що розширюють класичні підходи до аналізу статистичних розподілів. Отримані результати можуть бути застосовані у теоретичних дослідженнях та в задачах прикладного стохастичного моделювання.

Ключові слова: експоненціальні статистичні структури, клас В, ймовірнісні розподіли, перетворення Лапласа, лінійні оператори, стохастичне моделювання.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.6>

Ievgen KARNAUKH, PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0001-5885-7347

e-mail: karnauh@mmf.dnu.edu.ua

Oles Honchar Dnipro National University, Dnipro, Ukraine

BOUNDS SIMULATION FOR FLUID FLOW MODEL OF THE SECOND ORDER

This paper presents an approximation formula for the conditional cumulative distribution function of the maximum of a Brownian bridge with a given minimum. The formula enables us to develop an algorithm that generates the minimum and the upper bound of the maximum for a Brownian bridge trajectory. Thus, we can determine the simulation procedure for the upper and lower limits of a second-order fluid flow model. The demonstration of the results of the algorithm for a specific set of parameters is given. The extrema empirical distributions obtained using the sequential discretisation method and the approximate quantile method are compared. The possibility of sampling points in the middle of intervals, as well as choosing one of the subintervals at which the minimum is achieved, is considered.

Key words: strong simulation, acceptance-rejection sampling with squeezing, quantile transformation.

AMS 2020 classification: 60G51, 65C05, 91G60.

Introduction

Fluid models can be considered as a special class of Markov additive processes that allow the study of inhomogeneous dynamics. The parameters are assumed to depend on the state of the external environment, which can be described in terms of a finite Markov chain. Two main types of fluid models are recognized: first- and second-order. The first-order models imply piecewise linear paths that change direction at the transition moments of the environment, or regime-switching moments. A detailed list of references on the applications of these models in various fields, as well as an overview of known extensions, can be found in (Latouche, & Nguyen, 2018). Such models are appropriate if the system characteristics can be completely determined by first moments. The second-order model emerges when controlling the first and second moments of the rates of change is necessary (cf. (Karandikar, & Kulkarni, 1995)). This model evolve like a Brownian motion, but the drift and volatility depend on the environment state. It is also referred to as Markov-modulated Brownian motion (see, for example, (Latouche, 2025)).

Fluid models are versatile enough to represent a variety of phenomena while remaining analytically tractable. A matrix-analytical framework has been developed for them. Within this framework, many probabilistic characteristics (at least in terms of some integral transform like the Laplace transform) can be expressed in explicit form as a function of solutions to a certain quadratic matrix equation (see, for example, (Ahn, & Meini, 2019)). The framework stays relevant for some extensions like models with phase-type jumps (see, e.g., (Breuer, 2010a; D'Auria et al., 2010)), models with barriers ((Ivanovs, 2010; Breuer, 2010b; Horváth, & Telek, 2017), to name just a few), and with the level dependent evolution (see (Almoussa, Horváth, & Telek, 2020) and the references therein).

Including jumps in the model whose distribution differs from the phase-type distribution (or, more generally, from a distribution with a rational characteristic function) significantly complicates the possibility of obtaining an explicit solution. The problem of establishing effective numerical procedures becomes apparent in this case. One direct approach is to use the Monte Carlo method. Rather than trying to derive analytical expressions for probabilistic characteristics, we can generate a large enough number of paths to get the empirical distribution of the characteristic of interest. Under the assumption that the jump part is independent of the diffusion part, the main issue will be related to the latter (sample paths of it are infinite dimensional random elements). For a general overview of the issue see, e.g., (Pollock, 2013). Using a discretisation scheme (simulating from the finite dimensional distributions) introduces the bias to the Monte Carlo estimator. To overcome this issue, (Beskos, Peluchetti, & Roberts, 2012) presented the ε -strong algorithm, a rejection sampling method for constructing upper and lower Brownian path boundaries. This method allows for unbiased Monte Carlo estimators over path expectations. For an extension to the case of jump diffusion, see (Pollock, Johansen, & Roberts, 2016). For some alternative approaches, see, for instance, (Gribaudo, Iacono, & Manini, 2021), (D'Onofrio, & Lanteri, 2023), (Herrmann, & Massin, 2023), and the references therein.

The ε -strong algorithm involves iterative use of the rejection method at all steps, including the initialization one. At this step, it was proposed to select from an increasing sequence of nested space intervals the one (the so-called Bessel layer) that contains the entire trajectory at a given time interval (for more details, see (Pollock, Johansen, & Roberts, 2016, Section 7.1)). In this paper, we present and justify an algorithm for modeling upper and lower bounds using the inversion method. Specifically, we propose to simulate the endpoints of a path over each time interval with the minimum and an upper bound of the maximum (cf. (Pollock, Johansen, & Roberts, 2016, Section 7.2)). The possibility of midpoint values simulation with subsequent inclusion of refinement and bisection steps of the ε -strong algorithm for certain subintervals is also discussed. This can provide enhanced control over path behavior on the corresponding interval. Thus, the proposed algorithm complements the existing range of diffusion modeling methods.

This paper is structured as follows. In Section 1, we provide the definition of the second-order fluid model. Section 2 contains an in-depth discussion of the modeling methods employed. In Section 3, we present a monotonic sequence that approximates the conditional cdf of the Brownian bridge maximum with given minimum from below. We use this sequence to justify the algorithm for modeling the minimum and an upper bound of the maximum. An example for comparison with the results obtained by the standard discretization method is provided. Section 4 is devoted to discussing the possibility

of modeling midpoint values in order to clarify the behavior of the trajectory on subintervals. Finally, Section 5 describes the simulation procedure that combines the steps of constructing a path envelope with endpoints and midpoints, taking regime-switching into account.

1. The model

Let $J = \{J_t, t \geq 0\}$ be a homogeneous irreducible Markov chain with right continuous sample paths and finite state space $E = \{1, \dots, N\}$. Denote by $Q = \|Q_{ij}\|_{i,j=1}^N$ the rate transition matrix and by $\pi = (\pi_1 \dots \pi_N)$ the invariant distribution. For certainty, we assume that π is also the initial distribution of the chain. J is usually referred to as a phase or regime process, and J_t can be interpreted as the environment state at the moment t . The environment controls the flow parameters: $a_i \in \mathbb{R}$ (drift) and $\sigma_i^2 \geq 0$ (volatility), $i \in E$. The model will be of the second order if $\exists i \in E: \sigma_i^2 > 0$, and if $\forall i \in E: \sigma_i^2 = 0$, we get the first order model.

Let $W = \{W(t), t \geq 0\}$ be a Wiener process. We assume that J and W are mutually independent. The fluid level process is defined as

$$X_t = X_0 + \int_0^t a_{J_s} ds + \int_0^t \sigma_{J_s}^2 dW_s,$$

where X_0 is the initial level. That is, X follows a Brownian motion in between any two jumps of J . The process $\{X, J\}$ is called a fluid flow process, and it is a special case of Markov additive processes (see, for example, (Çinlar, 1972)).

The simulation of J is straightforward. We begin with a randomly selected initial state according to the initial distribution, say $J_0 = j$. The time J spends in the state is exponentially distributed with rate $-Q_{jj}$. Then, sample the next state from the distribution $\frac{Q_{ji}}{-Q_{jj}}$, $i \neq j$, and generate independently the next inter-transition time, etc.

In between transitions, X is a pure diffusion. We can split the time interval, say, into 2^m parts, $m \in \mathbb{N}$, and simulate the sample path at each point of partition iteratively. This discretization scheme does not sufficiently characterize the simulated sample paths. A path's entry into a region cannot be accurately determined simply by examining its position at a limited number of points. To gain more control over the path behavior within the partition interval, it can be helpful to get upper and lower bounds. Since we fix the initial and terminal values, the Brownian motion is actually a Brownian bridge. Although there is an analytical expression for the joint distribution of the minimum and maximum, it is represented in terms of an infinite series, which complicates the simulation. In the next section, we consider some versions of two basic techniques to simulate a random variable, namely, quantile transformation and the accept-reject method (see, e.g., (Brandimarte, 2014, Chapter 5) or (Owen, 2013, Chapter 4)), which can be useful for overcoming complications.

2. Simulation methods

Let $\tilde{g}(x)$ be a nonnegative function with $\int_{-\infty}^{\infty} \tilde{g}(x) dx = c_g < \infty$. Assume that we can simulate a random variable $\zeta \sim g(x) = \frac{1}{c_g} \tilde{g}(x)$, but we need to simulate a random variable $\xi \sim f(x)$ with

$$f(x) \propto p(x) \tilde{g}(x),$$

where $p(x)$ is the probability of an event which we only know up to the alternate sequence:

$$p_2(x) \leq p_4(x) \leq \dots \leq p_{2n-2}(x) \leq p_{2n}(x) \leq \dots \leq p(x) \leq \dots \leq p_{2n+1}(x) \leq p_{2n-1}(x) \leq \dots \leq p_1(x),$$

$$\lim_{n \rightarrow \infty} p_{2n}(x) = p(x) = \lim_{n \rightarrow \infty} p_{2n-1}(x), \quad \forall x \in \mathbb{R}.$$

Assume $p_{-1}(x) \equiv 1$ and $p_0(x) \equiv 0$. In this case we can use the following accept-reject algorithm with squeezing (see, for instance, (Pollock, 2013, Algorithm 2.6.1)).

1. Set $n = 1$. Simulate independently $\eta \sim Uniform(0, 1)$ and $\zeta \sim g(x)$.
2. If $\eta > p_n(\zeta)$, then back to Step 1 (reject ζ). If not, go further.
3. If $\eta < p_{n+1}(\zeta)$, then put $\xi = \zeta$ and stop (accept ζ). If not, increase n by 1 and go to Step 2.

Proposition 1. Random variable ξ obtained by the accept-reject algorithm with squeezing, has probability density function f .

Let ν be the number of steps we need to simulate a sample from f of size one. Then for $n \geq 2$

$$\Pr\{\nu = n\} = \sum_{\substack{q, k_1, \dots, k_{q+1} \in \mathbb{N}: \\ q \leq n, n - q \text{ is even,} \\ \sum_{i=1}^{q+1} k_i = (n - q)/2}} \prod_{i=1}^q R_{2k_i-1} A_{2k_{q+1}}, \quad (1)$$

where

$$\begin{aligned} R_{2k-1} &= \frac{1}{c_g} \int_{-\infty}^{\infty} (p_{2k-3}(z) \wedge 1 - p_{2k-1}(z) \wedge 1) \tilde{g}(z) dz, \\ A_{2k} &= \frac{1}{c_g} \int_{-\infty}^{\infty} (p_{2k}(z) \vee 0 - p_{2k-2}(z) \vee 0) \tilde{g}(z) dz, \\ x \vee y &:= \max\{x, y\}, \quad x \wedge y := \min\{x, y\}. \end{aligned}$$

Proof. Denote $\tilde{f}(x) = p(x)\tilde{g}(x)$ and $c_f = \int_{-\infty}^{\infty} \tilde{f}(x) dx$. Using the total probability law and independence of η and ζ we get

$$\begin{aligned} \Pr\{\xi < x\} &= \Pr\{\xi < x | \zeta \text{ is accepted}\} \Pr\{\zeta \text{ is accepted}\} + \Pr\{\xi < x | \zeta \text{ is rejected}\} \Pr\{\zeta \text{ is rejected}\} = \\ &= \Pr\{\xi < x | \zeta \text{ is accepted}\} \Pr\{\zeta \text{ is accepted}\} + \Pr\{\xi < x\} \Pr\{\zeta \text{ is rejected}\} = \\ &= \Pr\{\zeta < x, \zeta \text{ is accepted}\} + \Pr\{\xi < x\} \Pr\{\zeta \text{ is rejected}\}. \end{aligned} \quad (2)$$

From the algorithm

$$\begin{aligned} \Pr\{\zeta \text{ is rejected}\} &= \Pr\{\exists k \in \mathbb{N} : \eta > p_{2k-1}(\zeta)\} = \Pr\{\eta > p(\zeta)\} = \Pr\left\{\eta > \frac{\tilde{f}(\zeta)}{\tilde{g}(\zeta)}\right\} = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\left\{\eta > \frac{\tilde{f}(\zeta)}{\tilde{g}(\zeta)}\right\}} \mid \zeta\right]\right] = \\ &= \mathbb{E}\left[1 - \frac{\tilde{f}(\zeta)}{\tilde{g}(\zeta)}\right] = 1 - \int_{-\infty}^{\infty} \frac{\tilde{f}(z)}{\tilde{g}(z)} g(z) dz = 1 - \frac{c_f}{c_g} \end{aligned}$$

and

$$\Pr\{\zeta < x, \zeta \text{ is accepted}\} = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\zeta < x\}} \mathbb{1}_{\left\{\eta < \frac{\tilde{f}(\zeta)}{\tilde{g}(\zeta)}\right\}} \mid \zeta\right]\right] = \mathbb{E}\left[\mathbb{1}_{\{\zeta < x\}} \frac{\tilde{f}(\zeta)}{\tilde{g}(\zeta)}\right] = \int_{-\infty}^x \frac{\tilde{f}(z)}{\tilde{g}(z)} g(z) dz = \frac{c_f}{c_g} \int_{-\infty}^x f(z) dz.$$

Substitution in (2) yields $\Pr\{\xi < x\} = \int_{-\infty}^x f(z) dz$. Thus, $\xi \sim f(x)$.

To get (1) just note that

$$\begin{aligned} A_{2k} &= \Pr\{\zeta \text{ is accepted in } 2k \text{ steps}\} = \Pr\{p_{2k-2}(\zeta) < \eta < p_{2k}(\zeta)\}, \\ R_{2k-1} &= \Pr\{\zeta \text{ is rejected in } 2k-1 \text{ steps}\} = \Pr\{p_{2k-1}(\zeta) < \eta < p_{2k-3}(\zeta)\}. \end{aligned}$$

□

From a practical point of view, the use of formula (1) to estimate the complexity of the algorithm may require considerable analytical effort. In accordance with the algorithm, we will check within the loop whether a proposal falls within the acceptance or rejection region. One way to control the execution time of the loop is to impose a maximum number of steps. Consequently, it would be suitable to consider the probability that the check of the proposal will require no more steps than the pre-established limit.

Fix $n_{max} \in \mathbb{N}$ and denote the events

$$H = \{\text{Check of the proposal takes more than } n_{max} \text{ steps}\}, S = \bar{H},$$

$$A = \{\text{The proposal is accepted}\}, R = \bar{A}.$$

For concise notation, we will use P_E as the probability of event E and P_{AB} as $\Pr(A \cap B)$.

Proposition 2. Let $\tilde{\nu}$ be the number of proposal checks that take more than n_{max} steps. Then

$$\Pr\{\tilde{\nu} = 0\} = \frac{P_{AS}}{1 - P_{RS}}, \Pr\{\tilde{\nu} = m\} = \left(\frac{P_{RH}}{1 - P_{RS}}\right)^{m-1} \frac{P_A P_H}{(1 - P_{RS})^2}, m \in \mathbb{N}. \quad (3)$$

Proof. Event $\{\tilde{\nu} = 0\}$ corresponds to the case when all checks take less than or equal to n_{max} steps. So,

$$\Pr\{\tilde{\nu} = 0\} = P_{AS} \sum_{k=0}^{\infty} P_{RS}^k = \frac{P_{AS}}{1 - P_{RS}}.$$

Conditioning on whether the last step of acceptance take more than n_{max} steps and on the number of preceding rejections, we have

$$\Pr\{\tilde{\nu} = m\} = P_{AH} \sum_{k=m-1}^{\infty} C_k^{m-1} P_{RS}^{k-m+1} P_{RH}^{m-1} + P_{AS} \sum_{k=m}^{\infty} C_k^m P_{RS}^{k-m} P_{RH}^m.$$

Since $\frac{1}{(1 - P_{RS})^{m+1}} = \sum_{k=m}^{\infty} C_k^m P_{RS}^{k-m}$ we get

$$\Pr\{\tilde{\nu} = m\} = \frac{P_{AH} P_{RH}^{m-1}}{(1 - P_{RS})^m} + \frac{P_{AS} P_{RH}^m}{(1 - P_{RS})^{m+1}} = \left(\frac{P_{RH}}{1 - P_{RS}}\right)^{m-1} \left(\frac{P_{AH} (1 - P_{RS}) + P_{AS} P_{RH}}{(1 - P_{RS})^2}\right).$$

The observation that $P_{AH} (1 - P_{RS}) + P_{AS} P_{RH} = P_A P_H$ completes the proof.

□

Remark 1. From Proposition 2 we have that the distribution of $\tilde{\nu}$ is the mixture (with parameter $\frac{P_{AS}}{1 - P_{RS}}$) of the distribution concentrated at 0 and the geometric distribution $Geom\left(1 - \frac{P_{RH}}{1 - P_{RS}}\right)$. This is to be expected as we wait for a particular event

to occur in successive independent experiments and count only those that meet the given criteria. From (3) we deduce

$$E \tilde{\nu} = \left(\frac{1}{1 - \frac{P_{RH}}{1 - P_{RS}}} \right)^{-2} \frac{P_A P_H}{(1 - P_{RS})^2} = P_H \frac{1}{P_A}.$$

This result can be obtained as a consequence of Wald's equality and the representation $\tilde{\nu} = \sum_{i=1}^{\kappa} X_i$, where κ is the number of proposals we need to generate $\xi \sim f(x)$ and $\{X_i\}_{i=1}^{\infty}$ are i.i.d. indicators of event A . From the proof of Proposition 1 the probability of acceptance is $\frac{c_f}{c_g}$. The number of proposals that we must make before one is accepted has the distribution $Geom\left(\frac{c_f}{c_g}\right)$ and the expected number of proposals is $\frac{c_g}{c_f}$. We count only the proposals that require complex verification which can happen with the probability

$$P_H = \frac{1}{c_g} \int_{-\infty}^{\infty} d_{n_{max}}(z) \tilde{g}(z) dz,$$

where

$$d_n(z) = \begin{cases} \mathbb{I}_{\{p_n(z) > 1\}} + p_n(z) \mathbb{I}_{\{p_n(z) < 1\}} - p_{n-1}(z) \mathbb{I}_{\{p_{n-1}(z) > 0\}}, & n \text{ is odd,} \\ \mathbb{I}_{\{p_{n-1}(z) > 1\}} + p_{n-1}(z) \mathbb{I}_{\{p_{n-1}(z) < 1\}} - p_n(z) \mathbb{I}_{\{p_n(z) > 0\}}, & n \text{ is even.} \end{cases}$$

Thus, the expected number of such proposals is

$$E \tilde{\nu} = \frac{\int_{-\infty}^{\infty} d_{n_{max}}(z) \tilde{g}(z) dz}{\int_{-\infty}^{\infty} p(z) \tilde{g}(z) dz}.$$

Assuming that n_{max} is even and large enough (such that $p_{n_{max}}(z) \geq 0$ and $p_{n_{max}-1}(z) \leq 1$) we get the upper estimate

$$E \tilde{\nu} \leq \frac{\int_{-\infty}^{\infty} p_{n_{max}-1}(z) \tilde{g}(z) dz}{\int_{-\infty}^{\infty} p_{n_{max}}(z) \tilde{g}(z) dz} - 1. \quad (4)$$

Now, consider the following modification of the quantile transformation method. Let ξ be a random variable with strictly monotonic continuous cdf $F(x) = \Pr\{\xi \leq x\}$. The exact formula for $F(x)$ is not known. We can only approximate it from below by $F_n(x)$: $F_n(x) \leq F(x)$, $F_n(x)$ is strictly monotone and continuous on the set of nonnegative values of F_n , and $\lim_{n \rightarrow \infty} F_n(x) = F(x)$, $\forall x \in \mathbb{R}$. Define $\tilde{F}_n(x) = F_n(x) \mathbb{I}_{\{F_n(x) \geq 0\}}(x)$ and $\tilde{F}_n^{-1}(x) = \inf\{y : \tilde{F}_n(y) \geq x\}$, then $\forall x \in \mathbb{R}$: $\tilde{F}_n(x) \uparrow F(x)$ and $\forall y > 0$: $\tilde{F}_n^{-1}(y) \downarrow F^{-1}(y)$ (see Fig. 1). Therefore, the following statement holds.

Lemma 1. Let $\eta \sim Uniform(0, 1)$ and put $\xi_n = \tilde{F}_n^{-1}(\eta)$. Then the sequence $\{\xi_n\}_{n=1}^{\infty}$ converges almost surely to a random variable with cdf F .

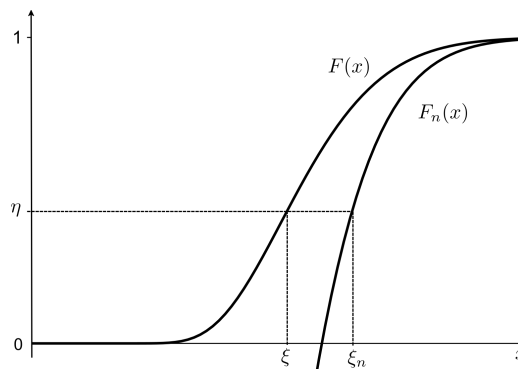


Fig. 1. Upper bound simulation for ξ

Lemma 1 justifies the quantile method of modeling an upper bound that monotonically approaches a random variable ξ in the case when its cumulative distribution function is known to within an approximating sequence. The exact representation for $\tilde{F}_n^{-1}(x)$ is the exception rather than the rule. To find $\xi_n = \tilde{F}_n^{-1}(\eta)$, we will use a numeric procedure for solving the nonlinear equation (e.g., the van Wijngaarden-Dekker-Brent method (Forsythe, Malcolm, & Moler, 1977, Section 7.2)). In this case, we need to be careful that we are trying to approach the solution from above.

3. Simulation of the minimum and an upper bound of the maximum for a Brownian bridge

Let $\{B_{x,l,y}(t), t \in [0, l]\}$ be the Brownian bridge process with value $x \in \mathbb{R}$ at 0 and value $y \in \mathbb{R}$ at $l > 0$. Since

$$B_{x,l,y}(t) \sim x + \frac{t}{l}(y - x) + \frac{l-t}{\sqrt{l}} W_{\frac{t}{l}}, 0 \leq t \leq l,$$

we have that the event that the Brownian bridge stays in an interval corresponds to the event that a Wiener process does not cross linear boundaries (for more details see (Casella, & Roberts, 2008, Appendix 1)):

$$\Pr\{u < B_{x,l,y}(t) < v, \forall t \in [0, l]\} = \Pr\left\{-\frac{(y-u)t + (x-u)}{\sqrt{l}} < W_t < \frac{(v-y)t + (v-x)}{\sqrt{l}}, \forall t \in [0, l]\right\}.$$

Set $B_{x,l,y}^- = \inf_{t \in [0, l]} B_{x,l,y}(t)$ and $B_{x,l,y}^+ = \sup_{t \in [0, l]} B_{x,l,y}(t)$. Then from (Beskos, Peluchetti, & Roberts, 2012, (4.3)) we get

$$\Pr\{B_{x,l,y}^- > u, B_{x,l,y}^+ < v\} = \left(1 - \sum_{n=1}^{\infty} H_n(l, v-u, x-u, y-u)\right) \mathbb{I}_{\{u < x, y < v\}}, \quad (5)$$

where

$$H_n(l, \Delta, \Delta_x, \Delta_y) = e^{-q_1^+} + e^{-q_2^+} - e^{-q_1^-} - e^{-q_2^-}, \quad \Delta \geq \Delta_x, \Delta_y \geq 0,$$

with

$$q_1^+ = \frac{2}{l}(n\Delta - \Delta_x)(n\Delta - \Delta_y), q_2^+ = \frac{2}{l}((n-1)\Delta + \Delta_x)((n-1)\Delta + \Delta_y),$$

$$q_1^- = \frac{2n\Delta}{l}(n\Delta + (\Delta_x - \Delta_y)), q_2^- = \frac{2n\Delta}{l}(n\Delta + (\Delta_y - \Delta_x)).$$

As a consequence, we can get the marginal cdfs (see also (Doob, 1949, (4.2)))

$$F_{B_{x,l,y}^-}(u) = \begin{cases} e^{-\frac{2}{l}(x-u)(y-u)}, & u \leq x, y, \\ 1 & \text{otherwise,} \end{cases} \quad F_{B_{x,l,y}^+}(u) = \begin{cases} 1 - e^{-\frac{2}{l}(v-x)(v-y)}, & v \geq x, y, \\ 0 & \text{otherwise,} \end{cases}$$

and expectations

$$\mathbb{E} B_{x,l,y}^- = x \wedge y - \sqrt{\frac{\pi l}{8}} e^{\frac{(x-y)^2}{2l}} \operatorname{erfc}\left(\frac{x \vee y - x \wedge y}{\sqrt{2l}}\right),$$

$$\mathbb{E} B_{x,l,y}^+ = x \vee y + \sqrt{\frac{\pi l}{8}} e^{\frac{(x-y)^2}{2l}} \operatorname{erfc}\left(\frac{x \vee y - x \wedge y}{\sqrt{2l}}\right),$$

where $\operatorname{erfc}(x) = 1 - \sqrt{\frac{2}{\pi}} \int_0^x e^{-t^2} dt$ is the complementary error function. Each addend of $H_n(l, v-u, x-u, y-u)$ is monotone in u , and from (Jones, 2001, Ch. 16, Sec C) we can use term-wise differentiation of the series in (4), which gives the following statement.

Proposition 3. *The conditional cdf of the Brownian bridge maximum with given minimum $u \leq x$ can be represented as the series*

$$F_{B_{x,l,y}^+ | B_{x,l,y}^- = u}(v) = \begin{cases} \frac{1}{f_{B_{x,l,y}^-}(u)} \sum_{n=1}^{\infty} G_n(l, v-u, x-u, y-u), & v \geq x \vee y, \\ 0 & v < x \vee y, \end{cases}$$

where $f_{B_{x,l,y}^-}(u)$ is the pdf of minimum, and

$$G_n(l, \Delta, \Delta_x, \Delta_y) = e^{-q_1^+} p_1^+ + e^{-q_2^+} p_2^+ - e^{-q_1^-} p_1^- - e^{-q_2^-} p_2^-, \quad \Delta \geq \Delta_x, \Delta_y \geq 0,$$

with

$$p_1^+ = \frac{2(n-1)}{l}(2n\Delta - \Delta_x - \Delta_y), p_2^+ = \frac{2n}{l}(2(n-1)\Delta + \Delta_x + \Delta_y),$$

$$p_1^- = \frac{2n}{l}(2n\Delta + \Delta_x - \Delta_y), p_2^- = \frac{2n}{l}(2n\Delta + \Delta_y - \Delta_x).$$

Theorem 1. *Function $\tilde{F}_m(v) = \tilde{F}_m(l, x, y, u, v) = \frac{1}{f_{B_{x,l,y}^-}(u)} \sum_{n=1}^m G_n(l, v-u, x-u, y-u) \mathbb{I}_{\{v \geq x \vee y\}}$ is continuous and strictly monotone on the set on which it is non-negative. Moreover, $\tilde{F}_m(v) \uparrow F_{B_{x,l,y}^+ | B_{x,l,y}^- = u}(v)$ as $m \rightarrow \infty$.*

Proof. Fix $n \geq 2, l > 0, \Delta_x \geq 0, \Delta_y \geq 0$, then the function G_n can be considered as a linear combination of functions of the following type $f(\Delta) = e^{-q(\Delta)} p(\Delta)$ with $q(\Delta) = (a\Delta + b_1)(a\Delta + b_2)$, $p(\Delta) = k\Delta + m$, $a > 0, k > 0, \Delta \geq \tilde{\Delta}$, where $\tilde{\Delta} := \max\{0, -\frac{b_1}{a}, -\frac{b_2}{a}, -\frac{m}{k}\}$. For the derivative, we have the representation

$$f'(\Delta) = e^{-q(\Delta)} (-q'(\Delta) p(\Delta) + p'(\Delta)) = e^{-q(\Delta)} (-a(a\Delta + b_1 + a\Delta + b_2)(k\Delta + m) + k).$$

Since $0 \leq (a\Delta + b_1 + a\Delta + b_2)(k\Delta + m) \uparrow +\infty$ as $\Delta \uparrow +\infty$, we have only the two possible cases for behavior of f . If $f'(\tilde{\Delta}) > 0$, then f first increases to some point $\Delta^* > \tilde{\Delta}$ (the corresponding values of Δ^* for each term of G_n are $\frac{\sqrt{l} + \Delta_x + \Delta_y}{2n}$, $\frac{\sqrt{l} - \Delta_x - \Delta_y}{2(n-1)}$, $\frac{\sqrt{l} - \Delta_x + \Delta_y}{2n}$, and $\frac{\sqrt{l} + \Delta_x - \Delta_y}{2n}$), and then monotonically approaches 0. If $f'(\tilde{\Delta}) \leq 0$, then f monotonically approaches 0 from the starting point (see Fig. 2).

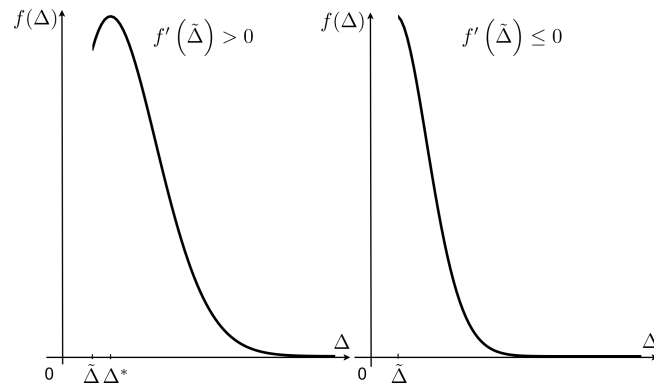


Fig. 2. Cases of possible behavior of $e^{-q(\Delta)}p(\Delta)$

Since $q_1^+ \leq q_1^-$ and $q_2^+ \leq q_2^-$, $e^{-q_1^+}p_1^+ + e^{-q_2^+}p_2^+$ tends to 0 more slowly than $e^{-q_1^-}p_1^- + e^{-q_2^-}p_2^-$. Hence, G_n starting from some $\bar{\Delta}_n$ will be positive monotonically decreasing to 0 (see the top right graph in Fig. 3), and this $\bar{\Delta}_n$ will decrease monotonically in n (the bottom left graph in Fig. 3).

If $n = 1$, then $e^{-q_1^-}p_1^-$ and $e^{-q_2^-}p_2^-$ show the similar behavior as shown in Fig. 2. While $e^{-q_1^+}p_1^+$ and $e^{-q_2^+}p_2^+$ are constant in Δ . More precisely, $e^{-q_1^+}p_1^+ + e^{-q_2^+}p_2^+|_{n=1, \Delta=v-u, \Delta_x=x-u, \Delta_y=y-u} = f_{B_{x,l,y}^-}(u)$. The term G_1 as function of Δ starts from the non-positive value (for certainty, let $\Delta_x \leq \Delta_y$, then $G_1(l, \Delta_y, \Delta_x, \Delta_y) = -e^{-\frac{2\Delta_y}{l}(2\Delta_y - \Delta_x)} \frac{2}{l} (3\Delta_y - \Delta_x) \leq 0$), reaches zero at some point $\bar{\Delta}_1$, and then monotonically approaches $f_{B_{x,l,y}^-}(u)$. So, the sum $\sum_{i=1}^n G_i$ is continuous and, starting from some point, strictly monotone in Δ (the bottom right graph in Fig. 3).

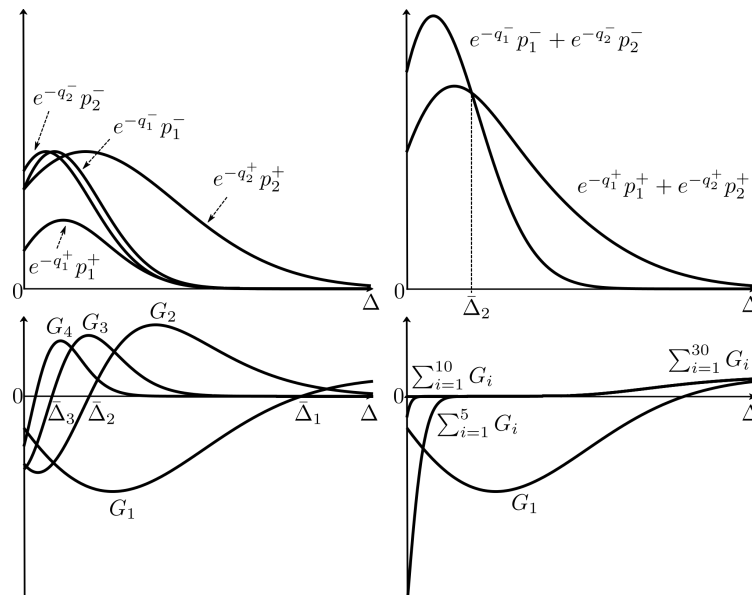


Fig. 3. Behavior of $\sum_{i=1}^m G_i$ with $l = 2, \Delta_x = 0.1, \Delta_y = 0.2$

□

Combining Lemma 1 with Theorem 1 gives us the following modified quantile transformation method for modeling the minimum and an upper bound of the maximum. Let $\tilde{F}_m^{-1}$ be the inverse function of $\tilde{F}_m$ restricted to the set $\{v : \tilde{F}_m(v) \geq 0\}$ and ζ_1, ζ_2 are independent $Unifom(0, 1)$ -distributed random variables. Define $\xi = F_{B_{x,l,y}^-}^{-1}(\zeta_1)$ and $\eta_m = \tilde{F}_m^{-1}(\zeta_2)|_{u=\xi}$, then $\eta_m \geq \eta$ and

$$\Pr\{\xi < r, \eta_m < q\} = E \mathbb{I}_{\left\{\zeta_1 < F_{B_{x,l,y}^-}^{-1}(r)\right\}} \tilde{F}_m(q)|_{u=F_{B_{x,l,y}^-}^{-1}(\zeta_1)} = \int_{-\infty}^r \tilde{F}_m(q)|_{u=z} f_{B_{x,l,y}^-}(z) dz.$$

By Lebesgue's dominated convergence theorem

$$\Pr\{\xi < r, \eta_m < q\} \xrightarrow{m \rightarrow \infty} \int_{-\infty}^r F_{B_{x,l,y}^+|B_{x,l,y}^-}^{-1}(q) f_{B_{x,l,y}^-}(z) dz = \Pr\{B_{x,l,y}^- < r, B_{x,l,y}^+ < q\}.$$

Example. Just to illustrate, we fixed $x = 2.1$, $y = 2.2$, $l = 1$, and simulated 100000 of Brownian bridge trajectories using standard discrete-time approximation. Namely, we simulate a Wiener process path $\{W_{t_i}\}_{i=1}^{2^m}$ at points $t_i = \frac{il}{2^m}$, $m = 12$, and calculate $B_{x,l,y}(t_i) = x + \frac{t_i}{l}(y - x) + W_{t_i} - \frac{t_i}{l}W_l$, $i = \overline{1, m}$. The points $(\min_{1 \leq i \leq 2^m} B_{x,l,y}(t_i), \max_{1 \leq i \leq 2^m} B_{x,l,y}(t_i))$ determined for each trajectory are presented in Fig. 4 (on the left). The contours in Fig. 4 are constructed based on the approximation of the joint density of the minimum and maximum, obtained as the sum of the derivatives with respect to u and v of the first 30 terms in (4). The graph on the right in Fig. 4 shows the results of 100000 simulations using the modified quantile transformation. To find $\tilde{F}_m^{-1}(\zeta_2)|_{u=\xi}$, we used the van Wijngaarden-Dekker-Brent method, iteratively increasing m to no more than 50, until the difference between successive solutions became less than 10^{-8} . The sample and theoretical values of the expectations, together with the 0.025 and 0.975 quantiles, rounded to 6 significant figures, are shown in Tab. 1.

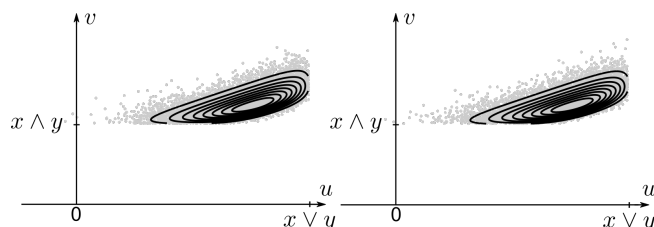


Fig. 4. Generated minimum and maximum of the Brownian bridge using sequential discrete-time simulation (left) and the approximate quantile method (right)

Table 1

Estimates and exact values of the mean and quantiles for extrema				
	Minimum $B_{x,l,y}^-$		Maximum $B_{x,l,y}^+$	
Theoretical Mean	1.52037		2.77963	
Theoretical Quantiles	(0.790978, 2.02688)		(2.27312, 3.50902)	
	Discrete-time approximation	Quantile transform approximation	Discrete-time approximation	Quantile transform approximation
Sample Mean	1.52782	1.52104	2.76925	2.78048
Sample Quantiles	(0.795669, 2.03636)	(0.799451, 2.02684)	(2.26324, 3.50005)	(2.2754, 3.50600)

In general, the discrete-time approximation gives us an overestimate for the minimum and underestimate for the maximum. To demonstrate, we simulated 1000 Brownian bridge paths at points $\{t_i\}_{i=1}^{2^m}$, found the average values for the minimum and maximum, and repeated this process 150 times. Similar steps were performed for the quantile transformation. The smooth kernel density estimates for the resulting data are shown in Fig. 5.

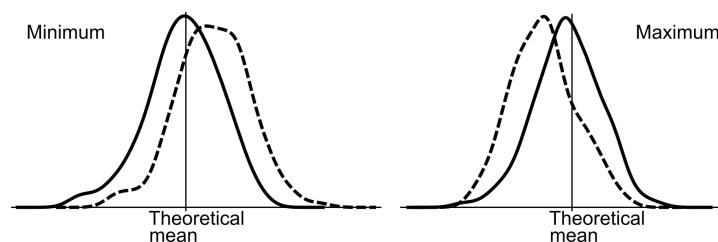


Fig. 5. Kernel-smoothed empirical marginal distributions for the average of minimum (left) and maximum (right) using sequential discrete-time simulation (dashed lines) and the approximate quantile method (solid lines)

Remark 2. It is possible to apply the considered algorithm to model Brownian motion with drift a and volatility σ^2 . To do so, substitute $l\sigma^2$ instead of l in the obtained relations (see, e.g. (Metwally, & Atiya, 2002, Section II)). According to Girsanov's theorem, the value of the drift is not essential (Beskos, Peluchetti, & Roberts, 2012, Section 6.1).

Example. Let $a = -1$, $\sigma = \sqrt{5}$. We split the interval $[0, 1]$ into 80 subintervals of the same length and start from $x = 2$. Fig. 6 shows sampled points of a Brownian motion path with corresponding lower and upper bounds.

The first passage time to the level 0 has an inverse Gaussian distribution (see, e.g., (Karatzas, & Shreve, 1998, Ch. 3, (5.12))) and the probability of crossing this level on the interval $[0, 1]$ up to 6 digits is 0.527339. We did a series of 1000 runs. In each run, we simulated 500 times the points and bounds on the interval $[0, 1]$, which was divided uniformly into 200 subintervals. We then determined the average number of times a path crossed the zero level. The empirical distribution of these averages is shown in Fig. 7, (a). The total average for all runs was 0.527874 and the empirical 0.025, 0.5, and 0.975 quantiles were 0.482, 0.528, and 0.568, respectively.

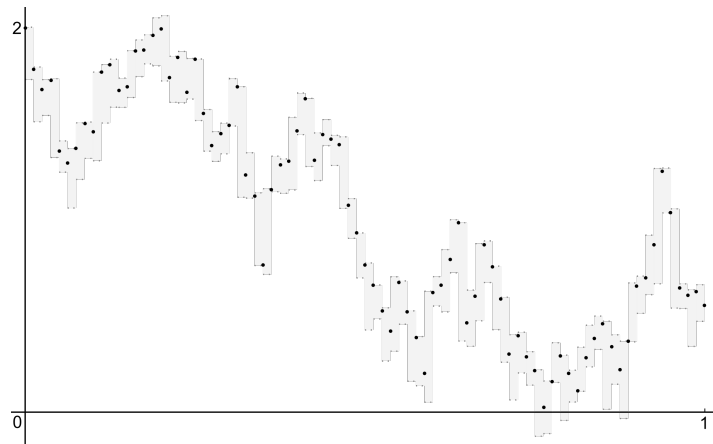


Fig. 6. Simulated path points of the Brownian motion with drift $\alpha = -1$ and volatility $\sigma = \sqrt{5}$

For a series of 50000 simulations, we recorded the value of the left boundary of the subinterval on which the crossing happened. Fig. 7, (b) shows the corresponding histogram and the conditional pdf of the first passage time through level 0, given that the crossing occurred within $[0, 1]$. Given the specified parameter values, and despite the potential for bias, the estimated outcomes appear to be reasonable. For a method of direct and exact simulation of the first-passage time, the reader can refer to (Herrmann, & Zucca, 2019). This method avoids simulating the full path, introduces no approximation error (only statistical error), and does not restrict the first passage time to be within a fixed time interval.

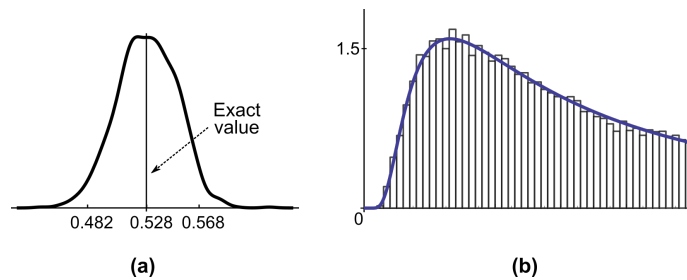


Fig. 7. (a) Kernel-smoothed empirical distribution for the average number of times trajectories cross zero level. (b) Histogram of an approximation of the first passage time and its conditional pdf (solid line)

4. Midpoint simulation

To further detail the trajectory behavior in the interval, we can adapt the idea of the ε -strong algorithm from (Beskos, Peluchetti, & Roberts, 2012). After defining the boundaries, the next step is to determine the position at the middle of the interval. The modeling routine can be justified using the following consequence of (Beskos, Peluchetti, & Roberts, 2012, Proposition 5.1).

Lemma 2. The conditional pdf $f_{mid}(z)$ of $B_{x,l,y}(\frac{l}{2})$ given $B_{x,l,y}^- = u$ and $B_{x,l,y}^+ < v$ is proportional to $p^\pm(z) \phi(z)$, where

$$p^\pm(z) = p^\pm(l, x, y, z, u, v) = -\frac{\partial}{\partial u} \left(\Pr \left\{ B_{x,\frac{l}{2},z}^- > u, B_{x,\frac{l}{2},z}^+ < v \right\} \Pr \left\{ B_{z,\frac{l}{2},y}^- > u, B_{z,\frac{l}{2},y}^+ < v \right\} \right),$$

$$\phi(z) = e^{-\frac{2}{l} \left(z - \frac{x+y}{2} \right)^2}.$$

For sampling from $f_{mid}(z)$ we can use the accept-reject algorithm with squeezing. First, note that

$$\max_{z \geq u} f_{B_{z,l,x}}^-(u) = \max_{z \geq u} f_{B_{x,l,z}}^-(u) = \begin{cases} \frac{e^{\frac{2(x-u)^2-l}{l}}}{x-u}, & x-u \leq \sqrt{\frac{l}{2}}, \\ \frac{2(x-u)}{l}, & \text{otherwise.} \end{cases}$$

Second, the monotonicity of probabilities gives $0 \leq -\frac{\partial}{\partial u} \Pr \left\{ B_{x,\frac{l}{2},z}^- > u, B_{x,\frac{l}{2},z}^+ < v \right\} \leq f_{B_{x,\frac{l}{2},z}}^-(u)$. Hence,

$$\begin{aligned} 0 \leq p^\pm(z) &= \left(-\frac{\partial}{\partial u} \Pr \left\{ B_{x,\frac{l}{2},z}^- > u, B_{x,\frac{l}{2},z}^+ < v \right\} \right) \Pr \left\{ B_{z,\frac{l}{2},y}^- > u, B_{z,\frac{l}{2},y}^+ < v \right\} + \\ &+ \left(-\frac{\partial}{\partial u} \Pr \left\{ B_{z,\frac{l}{2},y}^- > u, B_{z,\frac{l}{2},y}^+ < v \right\} \right) \Pr \left\{ B_{x,\frac{l}{2},z}^- > u, B_{x,\frac{l}{2},z}^+ < v \right\} \leq \\ &\leq f_{B_{x,\frac{l}{2},z}}^-(u) + f_{B_{z,\frac{l}{2},y}}^-(u) \leq \max_{z \geq u} f_{B_{x,\frac{l}{2},z}}^-(u) + \max_{z \geq u} f_{B_{z,\frac{l}{2},y}}^-(u) =: C_{x,l,y,u}. \end{aligned} \quad (6)$$

The algorithm requires $0 \leq p^\pm(z) \leq 1$, which can be satisfied using $\frac{1}{C_{x,l,y,u}} p^\pm$ instead of $p^\pm$. Third, the function $\phi(z)$ up to a constant is the density of normal distribution $N\left(\frac{x+y}{2}, \frac{l}{4}\right)$. Finally, without explicitly specifying the dependency on the corresponding arguments, the alternating sequence for $\Pr\{B^- > u, B^+ < v\}$ is $\{h_n\}_{n=1}^\infty$ (for details see (Beskos, Peluchetti, & Roberts, 2012, Section 4)) and for $\left(-\frac{\partial}{\partial u} \Pr\{B^- > u, B^+ < v\}\right)$ is $\{g_n\}_{n=1}^\infty$ where

$$h_{2n-1} = 1 - \sum_{i=1}^n H_i, \quad h_{2n} = \max\left\{1 - \sum_{i=1}^{n-1} H_i - \left(e^{-q_1^+} + e^{-q_2^+}\right)|_{i=n}, 0\right\},$$

$$g_{2n-1} = \begin{cases} \sum_{i=1}^n G_i + \left(e^{-q_1^+} p_1^+ + e^{-q_2^+} p_2^+\right)|_{i=n+1}, & \sum_{i=1}^n G_i \geq 0, \\ f_{B^-}(u), & \text{otherwise,} \end{cases} \quad g_{2n} = \sum_{i=1}^n G_i \mathbb{1}_{\{\sum_{i=1}^n G_i \geq 0\}}.$$

The expression for the lower bound g_{2n} follows directly from Theorem 1. For the upper bound g_{2n-1} we can use arguments similar to those in the proof of the theorem. In case $\sum_{i=1}^n G_i < 0$, instead of the upper bound $f_{B^-}(u)$, a more accurate estimate would be $\left(\sum_{i=1}^n G_i + \left(e^{-q_1^+} + e^{-q_2^+}\right)|_{i=n+1}\right)$ at the root of the equation $\sum_{i=1}^n G_i = 0$, but finding the corresponding solution at each iteration step can be computationally expensive. Thus, the alternating sequence for $p^\pm$ is defined by $\frac{g_n^{(1)} h_n^{(2)} + g_n^{(2)} h_n^{(1)}}{C_{x,l,y,u}}$ (see Fig. 8), where the number in the superscript indicates the set of terms arguments in (6), $(x, \frac{l}{2}, z, u, v)$ or $(z, \frac{l}{2}, y, u, v)$, respectively.

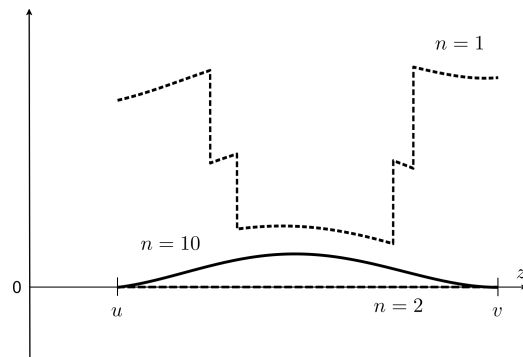


Fig. 8. 1st, 2nd, and 10th elements of the alternating sequence for $p^\pm$ with $l = 1, x = 2.1, y = 2.2, u = 1.8, v = 2.5$

For illustrative purposes, we have included in Fig. 9 the histogram of a sample of 5000 points from $f_{mid}(z)$ with $l = 1, x = 2.1, y = 2.2, u = 1.8$, and $v = 2.5$. The solid curve corresponds to the graph of $\frac{g_n^{(1)} h_n^{(2)} + g_n^{(2)} h_n^{(1)}}{C_{x,l,y,u}}$ with $n = 11$ (a further increase in n does not result in a noticeable change in the graph).

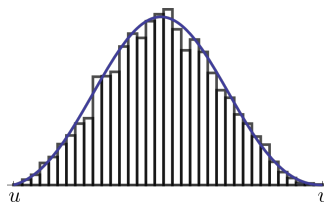


Fig. 9. Histogram of simulated midpoints with $l = 1, x = 2.1, y = 2.2, u = 1.8, v = 2.5$

Remark 3. After bipartition, we take into account that the minimum is reached at least in one of the subintervals. To determine on which one, we simulate the value of the indicator $\mathbb{1}\{\eta < r\}$, where $\eta \sim \text{Uniform}[0, 1]$,

$$r = \frac{\left(-\frac{\partial}{\partial u} \Pr\left\{B_{x, \frac{l}{2}, z}^- > u, B_{x, \frac{l}{2}, z}^+ < v\right\}\right) \Pr\left\{B_{z, \frac{l}{2}, y}^- \geq u, B_{z, \frac{l}{2}, y}^+ < v\right\}}{p^\pm(z)},$$

using the technique from (Beskos, Peluchetti, & Roberts, 2012, Section 4.2). That is, we increase n until $\frac{g_{2n-1}^{(1)} h_{2n-1}^{(2)}}{g_{2n}^{(1)} h_{2n}^{(2)} + g_{2n-1}^{(2)} h_{2n-1}^{(1)}} < \eta$ (and assume that the minimum was reached in the first half of the interval) or $\frac{g_{2n}^{(1)} h_{2n}^{(2)}}{g_{2n-1}^{(1)} h_{2n-1}^{(2)} + g_{2n}^{(2)} h_{2n}^{(1)}} > \eta$ (respectively, in the second). If the denominator is equal to 0, the fraction is defined as 1. For the interval with the given minimum, we can continue to define the value at the midpoint using Lemma 2. For another interval, we can employ the refinement with consequent bisection as described in (Beskos, Peluchetti, & Roberts, 2012, Section 5).

5. Algorithm description

The following basic modeling steps can be obtained by collecting the considered facts together. Let d be the maximum length of subintervals. The upper limits of some loop indices have not been explicitly specified. They can be defined in

accordance with a particular condition. For example, the loop by k can be executed until T_k exceeds the predetermined value. We may continue to execute the loop by m until the accuracy constraints are met.

1. Set $T_0 = 0$, $X_{T_0} = X_0$. Generate J_{T_0} from the initial distribution.

2. Loop by $k \geq 1$.

(a) Draw J_{T_k} from the discrete distribution $\left\{ \frac{Q_{J_{T_{k-1}}j}}{-Q_{J_{T_{k-1}}J_{T_{k-1}}}}, j \neq J_{T_{k-1}} \right\}$.

(b) Generate the inter-jump time $\Delta T_k \sim \text{Exponential}(-Q_{J_{T_{k-1}}J_{T_{k-1}}})$. Set $T_k = T_{k-1} + \Delta T_k$.

(c) Put $t_{k-1}^0 = T_{k-1}$, $n_k = \left\lceil \frac{\Delta T_k}{d} \right\rceil + 1$, $l_k = \frac{\Delta T_k}{n_k}$, where $[\cdot]$ is the integer part.

(d) Loop by $i = \overline{1, n_k}$.

i. Set $t_{k-1}^i = t_{k-1}^{i-1} + l_k$.

ii. Generate

$$X_{t_{k-1}^i} = X_{t_{k-1}^{i-1}} + a_{J_{T_{k-1}}} l_k + \sigma_{J_{T_{k-1}}} \sqrt{l_k} \zeta, \zeta \sim N(0, 1).$$

iii. Generate the minimum $B_{k,i}^-$ and the upper bound of maximum $\check{B}_{k,i}^+$ for Brownian bridge $B_{X_{t_{k-1}^{i-1}}, \sigma_{J_{T_{k-1}}}^2 l_k, X_{t_{k-1}^i}}$ (see Section 3).

iv. Set $s_L = t_{k-1}^{i-1}$, $s_R = t_{k-1}^i$.

v. Loop by $m \geq 2$.

A. Put $s_m = \frac{s_L + s_R}{2}$. Given $B_{k,i}^-$ and $\check{B}_{k,i}^+$ generate midpoint value X_{s_m} using Lemma 2.

B. Choose between intervals $[s_L, s_m]$ and $[s_m, s_R]$ with minimum using Remark 3. Set $s_R = s_m$ in the former case and $s_L = s_m$ in the latter.

C. For the chosen interval, we continue the loop, and for the other interval, we can apply ε -strong algorithm from (Beskos, Peluchetti, & Roberts, 2012).

Remark 4. According to the proposed algorithm, the lower bound for the maximum and the upper bound for the minimum required in the ε -strong algorithm need to be refined only for one of the intervals obtained after midpoint simulation.

Example. Let's choose the following parameters for Z . Assume that for J the state space is $E = \{1, 2\}$, the intensity matrix is $Q = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ 1 & -1 \end{pmatrix}$, and the initial distribution is $\pi = (\frac{2}{3}, \frac{1}{3})$. For X the drift parameters are $a_1 = \frac{1}{2}$, $a_2 = -\frac{3}{2}$, and the volatility parameters are $\sigma_1 = \sqrt{5}$, $\sigma_2 = \sqrt{2}$. Fig. 10 shows the results of applying the algorithm without the ε -strong algorithm steps. Specifically, it displays the simulated trajectory points at the moments of state transitions, the minimum and the upper bound of the maximum on the corresponding intervals. It also displays the midpoints. For the purposes of simplicity, we took $d = \infty$, indicating that $[T_{k-1}, T_k]$ were not divided into subintervals.

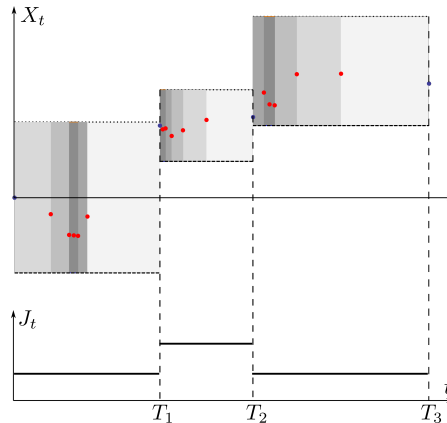


Fig. 10. Regime-switching moments (vertical dashed lines), values of X at those moments (blue points), values of midpoints (red points), the minimum of X (horizontal dotted lines), and the upper bound of maximum (horizontal dashed lines) between regime-switching moments. The gray bars indicate the interval selection (on Step B.). The darker color indicates a later selection

Discussion and conclusions

We propose an algorithm that permits us to generate lower and upper bounds for a fluid flow model of the second order. The algorithm has a direct generalization for models with jumps (under the assumption that the jump part is independent of the diffusion part) and allows for reducing the problem of estimation bias due to discretization. It is based on the ε -strong algorithm incorporating a combination of rejection sampling and quantile transformation.

The algorithm has some pros and cons compared to the ε -strong algorithm. The main advantage is that in the initialization step, we simulate the precise value (up to machine precision) of the minimum and the upper bound of the maximum. By performing additional calculations, it is possible to ascertain a lower bound for the maximum, thereby ensuring the precision of the path upper boundary. The specification of the path behavior at intermediate points with a given value of the minimum (or maximum) necessitates the use of a distinct type of process. For instance, conditional on a Brownian bridge sample path minimum and the time instance when it is achieved, the law of the path parts is that of Bessel bridge (see, e.g., (Pollock, 2013, Section 2.8)). As the value of d decreases, the necessity for intermediate point modeling also decreases; at the same time, the number of calculations increases, which may result in redundancy. The determination of the optimal rule for d depending on the functional under study and the required level of accuracy may be of interest for further research.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Ahn, S., & Meini, B. (2019). Matrix equations in Markov modulated Brownian motion: theoretical properties and numerical solution. *Stochastic Models*, 36(2), 251–284. <https://doi.org/10.1080/15326349.2019.1704785>
- Almousa, S. A.-D., Horváth, G., & Telek, M. (2020). Transient analysis of piecewise homogeneous Markov fluid models. *Annals of Operations Research*, 310(2), 333–353. <http://dx.doi.org/10.1007/s10479-020-03831-1>
- Beskos, A., Peluchetti, S., & Roberts, G. (2012). ε -strong simulation of the Brownian path. *Bernoulli*, 18(4), 1223–1248. <http://dx.doi.org/10.3150/11-BEJ383>
- Brandimarte, P. (2014). *Handbook in Monte Carlo simulation. Applications in financial engineering, risk management, and economics*. John Wiley & Sons. <http://dx.doi.org/10.1002/9781118593264>
- Breuer, L. (2010a). A quintuple law for Markov additive processes with phase-type jumps. *Journal of Applied Probability*, 47(2), 441–458. <http://dx.doi.org/10.1017/s0021900200006744>
- Breuer, L. (2010b). The total overflow during a busy cycle in a Markov-additive finite buffer system. In Müller-Clostermann, E., & Rathgeb (Eds.), *Measurement, modelling, and evaluation of computing systems and dependability and fault tolerance* (pp. 198–211). Springer. http://dx.doi.org/10.1007/978-3-642-12104-3_16
- Casella, B., & Roberts, G. O. (2008). Exact Monte Carlo simulation of killed diffusions. *Advances in Applied Probability*, 40(1), 273–291. <http://dx.doi.org/10.1239/aap/1208358896>
- Çınlar, E. (1972). Markov additive processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(2), 85–93. <http://dx.doi.org/10.1007/bf00532536>
- D'Auria, B., Ivanovs, J., Kella, O., & Mandjes, M. (2010). First passage of a Markov additive process and generalized Jordan chains. *Journal of Applied Probability*, 47(4), 1048–1057. <http://dx.doi.org/10.1017/s002190020000735x>
- D'Onofrio, G., & Lanteri, A. (2023). Approximating the first passage time density of diffusion processes with state-dependent jumps. *Fractal and Fractional*, 7(1), 30. <http://dx.doi.org/10.3390/fractalfract7010030>
- Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 20(3), 393–403. <http://dx.doi.org/10.1214/aoms/117729991>
- Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). *Computer methods for mathematical computations*. Prentice Hall.
- Gribaudo, M., Iacono, M., & Manini, D. (2021). Simulation of n -dimensional second-order fluid models with different absorbing, reflecting and mixed barriers. In *Quantitative evaluation of systems* (pp. 276–292). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-85172-9_15
- Herrmann, S., & Massin, N. (2023). Exact simulation of the first passage time through a given level of jump diffusions. *Mathematics and Computers in Simulation*, 203, 553–576. <http://dx.doi.org/10.1016/j.matcom.2022.07.007>
- Herrmann, S., & Zucca, C. (2019). Exact simulation of the first-passage time of diffusions. *Journal of Scientific Computing*, 79(3), 1477–1504. <http://dx.doi.org/10.1007/s10915-018-00900-3>
- Horváth, G., & Telek, M. (2017). Matrix-analytic solution of infinite, finite and level-dependent second-order fluid models. *Queueing Systems*, 87(3–4), 325–343. <http://dx.doi.org/10.1007/s11134-017-9544-z>
- Ivanovs, J. (2010). Markov-modulated Brownian motion with two reflecting barriers. *Journal of Applied Probability*, 47(4), 1034–1047. <http://dx.doi.org/10.1017/s0021900200007348>
- Jones, F. (2001). *Lebesgue integration on euclidean space*. Jones and Bartlett Publishers.
- Karandikar, R. L., & Kulkarni, V. G. (1995). Second-order fluid flow models: Reflected Brownian motion in a random environment. *Operations Research*, 43(1), 77–88. <http://dx.doi.org/10.1287/opre.43.1.77>
- Karatzas, I., & Shreve, S. E. (1998). *Brownian motion and stochastic calculus* (2nd ed.). Springer. <http://dx.doi.org/10.1007/978-1-4612-0949-2>
- Latouche, G. (2025). *Markov-modulated Brownian motion. The matrix-analytic perspective*. Springer. <https://link.springer.com/book/9783031985751>
- Latouche, G., & Nguyen, G. T. (2018). Analysis of fluid flow models. *Queueing Models and Service Management*, 1(2), 001-029. https://qmsm.pu.edu.tw/papers/papers/2018_Issue_2/paper_1_Guy_Latouche_p_1_29.pdf
- Metwally, S. A. K., & Atiya, A. (2002). Using Brownian bridge for fast simulation of jump-diffusion processes and barrier options. *The Journal of Derivatives*, 10(1), 43–54. <http://dx.doi.org/10.3905/jod.2002.319189>
- Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. (Accessed: 2025-07-03) <https://artowen.su.domains/mc/>
- Pollock, M. (2013). *Some Monte Carlo methods for jump diffusions* (phdthesis, University of Warwick). <https://wrap.warwick.ac.uk/id/eprint/60602/>
- Pollock, M., Johansen, A. M., & Roberts, G. O. (2016). On the exact and ε -strong simulation of (jump) diffusions. *Bernoulli*, 22(2), 794–856. <http://dx.doi.org/10.3150/14-BEJ676>

Отримано редакцією журналу / Received: 03.08.25
Прорецензовано / Revised: 14.03.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Євген КАРНАУХ, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-5885-7347
e-mail: karnauh@mmf.dnu.edu.ua
Дніпровський національний університет імені Олеся Гончара, Дніпро, Україна

МОДЕЛЮВАННЯ МЕЖ ДЛЯ МОДЕЛІ ПОТОКУ РІДИНИ ДРУГОГО ПОРЯДКУ

Отримано наближену формулу для умовної функції розподілу максимуму броунівського мосту із заданим мінімумом. Формула дає змогу розробити алгоритм моделювання мінімуму та верхньої межі максимуму для траєкторії броунівського мосту. Це, зі свого боку, формує можливість визначити процедуру моделювання для верхньої та нижньої границі для моделі потоку рідини другого порядку. Показано результати роботи алгоритму для певної множини параметрів. Проведено порівняння емпіричних розподілів екстремумів, отриманих за допомогою методу послідовної дискретизації та методу наближених квантилів. Розглянуто можливість моделювання точок посередині інтервалів, а також вибір одного з підінтервалів, на якому досягається мінімум.

Ключові слова: *сильне моделювання, метод відсіювання зі стисканням, квантильне перетворення.*

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.7>

Volodymyr KUDIN, DSc (Engin.), Prof.

ORCID ID: 0009-0005-9071-1213

e-mail: vol.kudin@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Viacheslav ONOTSKYI, PhD (Phys. & Math.), Assist.

ORCID ID: 0000-0002-1920-0905

e-mail: onotskyi@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

ANALYSIS OF LINEAR SYSTEMS WITH COMPLEX ELEMENTS BY THE BASIS MATRIX METHOD

This work proposes the application of the mathematical framework of complex variable function theory (specifically, for linear systems with complex elements) to describe the mechanism of interaction between components in a system ("model" – "researcher") during the process of mathematical modeling. The method of basis matrices is extended to linear systems with complex elements. An algorithm and computational procedure are developed for analyzing the properties and solving such systems, taking into account the interaction mechanism between the real and imaginary components in the representation of the mathematical model. A computational experiment is conducted to verify the correctness of the proposed algorithm and computational procedure.

The object of the study is processes of various natures in the context of solving the problem of mathematical modeling. The subject of the study is methods, algorithms, and computational procedures of mathematical modeling, particularly the mathematical apparatus of complex variable function theory and linear algebra.

Research Objective: to develop algorithms and computational procedures for analyzing the properties of two-component mathematical models of processes (with complex elements), capable of accounting for the researcher's activity in the problem-solving process.

Research Tasks: to analyze known approaches to solving computational problems while considering the "activity" of the human analyst, operations researcher, or decision-maker; to identify basic mathematical methods in the class of linear systems that incorporate the ability to account for (human) activity in the problem-solving process; to develop computational procedures and algorithms adapted to modeling tasks (at various stages) for comprehensive analysis of linear systems (two-component systems), with a component accounting for human activity; to conduct a computational experiment based on the proposed method, algorithm, and procedure to verify their correctness on model examples; to analyze and summarize the experimental results and outline prospects for further research.

The mathematical modeling procedure involves a series of sequential stages (with feedback). Unfortunately, the unavoidable issue of the adequacy of the mathematical model in representing the studied process remains a challenge in solving applied problems. It is well known that modeling problems are expressed using mathematical tools of varying complexity and are often simplified (reduced) to linear models. This includes the discretization of continuous problems and the linearization of nonlinear ones.

It should be noted that the modeling process does not explicitly account for the influence of the researcher's experience, i. e., the human factor, on the quality of the modeling process at all stages. In fact, the two-component structure of the problem ("explicit" component – the mathematical model, and "implicit" – the researcher within the modeling loop) can be "traced".

This underscores the relevance of developing new mathematical problem formulations for modeling with complex elements (two-component systems), in which mechanisms are embedded to account for the researcher's experience in achieving greater adequacy of the model to the studied process at various problem-solving stages.

Key words: system of linear algebraic equations, basis matrix method, complex analysis.

AMS 2020 classification: 15A06, 65F05, 30A99.

Introduction

In recent decades, a scientific direction has gained intensive development (in practical tasks), which involves taking into account the presence of a person (in various guises) in the problem-solving contour of a model when presenting it. For example, dialogic (Dul et al., 2012), human-machine (Mrachkovskiy, 2023; Tan, Feng, & Shen, 2017), active (Strauch, 2017), ergatic (Korobiichuk, Tokar, & Danik, 2019), expert decision-making systems, and others. In practice, they include various interpretations of two-component systems – taking into account a person (analyst, researcher, expert, operations researcher). That is, the mechanisms of human interaction with the system (influences on a certain mathematical model). In the mathematical aspect (as a generalization), the nature of the category of a complex number (in the theory of functions of complex variables) is based on two-componentity (real and imaginary parts). And the modulus of a complex number can be interpreted as the resulting interaction of two components: real and imaginary.

In general, there is an implicit tendency towards an increasingly widespread involvement of the mathematical apparatus of complex (and hypercomplex) numbers in the description of complex processes (mathematical representation) (Demmel, 1997; Blackford et al., 2002).

The use of a more complex formal mathematical apparatus, on the one hand, "leads" to a more complicated representation of the model, and on the other hand, contributes to a more adequate representation of processes of a complex nature. Accordingly, the use of the mathematical apparatus of linear systems with complex elements can be a "golden mean", that is, a parity "measure of balance" between the complexity of the study and the adequacy of the representation.

Further study of the properties of linear systems with complex elements seems important in the field of practical application in the description of applied problems.

The object of research is processes of various nature in the context of solving the problem of mathematical modeling.

The subject of the research is methods, algorithms and computational procedures of mathematical modeling. In particular, the mathematical apparatus of the theory of functions of a complex variable, linear algebra.

© Kudin Volodymyr, Onotskyi Viacheslav, 2026

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

The purpose of the research: to develop algorithms and computational procedures for analyzing the properties of two-component mathematical models of processes (with complex elements) with the ability to take into account the activity of the researcher in the process of solving the problem.

Research objectives:

- to analyze known approaches to solving computational problems taking into account the "activity" of a person, such as an analyst, operations researcher, decision maker, etc.,
- to identify basic mathematical methods in the class of linear systems, which include the ability to take into account (and activity) of a person in the process of solving a problem, i.e. two-component systems,
- to develop computational procedures and algorithms for the complex analysis of linear systems (two-component systems) adapted to modeling tasks (at different stages), with a component of taking into account human activity,
- to conduct a computational experiment based on the proposed method, algorithm and procedure to verify the correctness of their "work" on model examples,
- to analyze and form conclusions based on the results of the experiments and outline prospects for further research.

1. Overview and relevance

It is known that mathematical models of processes are presented in the class of nonlinear problems, often optimization problems. When studying them, simplification procedures are used, for example, by transitioning from continuous problems to discrete ones (discretization of continuity), replacing nonlinear complex dependencies with simpler ones (linearization of nonlinearity in the neighborhood), etc. As a result of such transformations, the initial problem is formalized in the class of linear systems, in particular, equations, inequalities, etc.

It is natural (desirable) that the study of a linear system with complex variables can also be reduced to a system of linear algebraic equations with real elements (generally, of higher dimension). This indicates a wide range of possible practical applications of linear systems.

Of course, when analyzing solvability (unsolvability), the appropriate basic theoretical background can be applied (Demmel, 1997). It should be noted that the issue of solvability of systems of linear algebraic equations (SLAE) was embedded in classical theoretical results: the fundamental theorem of linear algebra, Kronecker – Capella, Mynkivsky – Farkash. A number of studies on the analysis of the solvability of linear systems and the influence of the parameters of the organization of calculations on the quality of the mathematical model were conducted, in particular, in (Granlund, 2023; Fromberger, 2005). The practice of performing calculations revealed differences between the mathematical (theoretical or exact) and machine (approximate computer) representation of the model, which was described in (Demmel, 1997; Blackford et al., 2002). These differences (as irreducible errors) are an additional source of inadequacy of the model and the process (for example, the incorrectness of the problem).

It has also been established that when using computational methods for solving SLAE, it is important to estimate and take into account the rank value and its completeness (it depends on the accuracy of the model representation). It has been established that for SLAE with an arbitrary ratio of the number of rows and columns, the Gaussian method can be considered the most universal (Demmel, 1997). This method allows solving the problem without restrictions on the structure of the constraint matrix and the dimension of the corresponding matrix. The importance (for correct calculations) of controlling the rank values (and the properties of its completeness) (Kudin et al., 2019a), conditionality (Kudin et al., 2019b) has been established. It is also important for analyzing the structure of the system (Kudin, Onyshchenko, A., & Onyshchenko? I., 2019) to find the corresponding rank matrix (basis), fulfill the conditions of support, and non-degeneracy of the basis matrix (Kudin et al., 2007). In particular, in the work (Kudin et al., 2019a) the influence of the accuracy of the representation of model elements (mantissa length, order value, machine zero threshold and overflow) and the options for organizing calculations on the properties of solutions was investigated. The corresponding software implementation was developed for calculations using the basis matrix methods (BMM) (Kudin et al., 2019a). Long arithmetic for models with rational elements was used here. In the works (Kudin et al., 2019b), algorithms and computer implementation of Gauss-type methods and artificial basis matrices (a variant of the basis matrix method) (Kudin et al., 2007) in the Matlab and Visual C++ environments were proposed using the technology of exact calculations of method elements, primarily for ill-conditioned systems of different dimensions (Kudin et al., 2019b).

It has been confirmed by experiments that computational procedures can be developed and applied in different ways (subjectively). This is also observed when choosing: a computational algorithm, the accuracy of calculations, the representation of numbers (the length of the mantissa), the tool for performing calculations (the threshold value of the machine zero), and the degree of taking into account the conditionality of the system. This ultimately forms a "different vision and assessment" of the course of calculations. That is, the influence of such subjectivity in organizing calculations can be significant. And in the future (due to the established value of the rank of the constraint matrix) to affect the solvability, the dimensionality of the solution space (in the case of non-uniformity). This indicates the importance of developing a mechanism to take into account the possible influence of the researcher (his implicit activity) during the construction of the model at the stage of developing the computational method and algorithm. In general, through the quantitative selection of the parameters of the organization of calculations, it is possible to actively influence (i.e., interact with the system) the quality of the mathematical model at the stage of organizing calculations. Such activity can be quantitatively assessed, i.e., acquire its meaning, by introducing into consideration the coefficient of influence of the subjective on the mathematical model. The use of the mathematical apparatus of complex numbers makes it possible to quantitatively take into account the experience of the researcher (through the coefficient of influence activity) on the real component (mathematical model) when solving the problem.

2. Main definitions and results

Let us investigate the properties of a linear system with complex elements. In particular, the conditions of solvability and non-solvability in the context of using the method of basis matrices (artificial basis matrices) (Kudin et al., 2007).

We have a SLAE with complex elements

$$A^{(CO)}u = C^{(CO)}, \quad (1)$$

where $A^{(CO)} = A^{(R)} + iA^{(S)}$, $C^{(CO)} = C^{(R)} + iC^{(S)}$ – real and imaginary components ($i^2 = -1$), i. e. matrices and constraint vectors of a complex linear system, $u = (u_1, u_2, \dots, u_m)^T$ – vector of variables, $A^{(R)} = (a_{ij}^{(R)})_{i,j=1}^{m,m}$, $A^{(S)} = (a_{ij}^{(S)})_{i,j=1}^{m,m}$ – matrices (and elements)

of real and imaginary components (components), $C^{(R)} = (c_i^{(R)})_{i=1}^m$, $C^{(S)} = (c_i^{(S)})_{i=1}^m$ – respectively, the vectors of the right-hand sides of the constraints of the component of the complex system. In particular, $a_{ij}^{(CO)} = a_{ij}^{(R)} + i a_{ij}^{(S)}$, $c_i^{(CO)} = c_i^{(R)} + i c_i^{(S)}$, $i = \overline{1, m}$, $j = \overline{1, m}$ – complex elements of a linear system.

Let us consider a system of the form:

$$A^{(Mod)} u = C^{(Mod)}, \quad (2)$$

whose elements are formed by the corresponding modules of complex numbers of the elements of the model (1), where

$$A^{(Mod)} = (a_{ij}^{(Mod)})_{i,j=1}^{m,m}, \quad C^{(Mod)} = (c_i^{(Mod)})_{i=1}^m, \quad a_{ij}^{(Mod)} = \sqrt{a_{ij}^{(R)2} + a_{ij}^{(S)2}}, \quad c_i^{(Mod)} = \sqrt{c_i^{(R)2} + c_i^{(S)2}}, \quad i = \overline{1, m}, \quad j = \overline{1, m}.$$

That is, the classical mechanism of (element-wise) interaction of real and imaginary components of a complex linear system.

Without loss of generality, system (1) can be written in equivalent form:

$$\begin{cases} A^{(R)} u = C^{(R)} \\ A^{(S)} u = C^{(S)} \end{cases} \quad (3)$$

Let us rewrite the linear system (3) in the form :

$$A u = C, \quad (4)$$

where $A = \begin{pmatrix} A^{(R)} \\ A^{(S)} \end{pmatrix}$, $C = \begin{pmatrix} C^{(R)} \\ C^{(S)} \end{pmatrix}$, accordingly, the matrix A of dimension $2m \times m$, which will be called the main or generating matrix

in the future, C is the constraint vector of dimension $2m$.

When studying system (4), we will also use a system of the form:

$$E^{(m)} u = I^{(m)}, \quad (5)$$

in which $E^{(m)} = (e_{ij}^{(m)})_{i,j=1}^{m,m} = \delta_{ij}$, $e_{ij}^{(m)} = 1, i = j$, $e_{ij}^{(m)} = 0, i \neq j$ is a trivial square matrix of dimension $(m \times m)$,

$I^{(m)} = (i_i^{(m)})_{i=1}^m$, $i_i^{(m)} = 1, i = \overline{1, m}$ is a constraint vector of dimension m , with a known solution and inverse matrix. In the following, we will call it an auxiliary system.

The basis of the proposed method of artificial basis matrices (ABM) and its corresponding algorithms is the idea of a basis matrix formed by linearly independent rows. The process of iterative "incrementing" (replacing) the constraints of the auxiliary system (5) with the relaxed constraints of the main system (4) is carried out. During the iterations (based on the formed artificial basis matrices), the magnitude of the rank of the matrices of the formed intermediate systems is analyzed and the main elements of the method are found.

Definition 1. The submatrix A_b of the matrix A, composed of m linearly independent normals $I_b = (i_1, i_2, \dots, i_m) \subset I = \{1, 2, \dots, 2m\}$ of constraints (4), will be called artificial basis, and the solution $u_0 = (u_{01}, u_{02}, \dots, u_{0m})^T$ of the corresponding system of equations

$$A_b u_0 = C_b,$$

where $C_b = (c_{i_1}, c_{i_2}, \dots, c_{i_m})^T$ is the subvector C, artificial basis, $\Delta_r = a_r u_0 - c_r$, $r \in I$ is the residuals of the constraints.

Let e_{ri} be elements of the inverse matrix $A_b^{-1} = \|e_{ij}\| (i = \overline{1, m}; j = \overline{1, m})$ composed of m linearly independent column vectors $(j_1, j_2, \dots, j_m)$ $(e_{.j} = (e_{1j}, e_{2j}, \dots, e_{mj}), j = 1, 2, \dots, m)$ inverse to A_b ; $\alpha_r = (\alpha_{r1}, \alpha_{r2}, \dots, \alpha_{rm})$ – vector of expansion of the constraint normal vector $a_r u = c_r$ along the rows of the basis matrix A_b ($\alpha_r = a_r \times A_b^{-1}$); $\alpha_{lk} \neq 0$ – leading element of the transformation method, where l is the number of the normal-row of the constraint matrix that is introduced, and k is the number of the normal-row of the constraint matrix that is extracted from the artificial basis matrix. It has been established that between the coefficients of expansion of the constraint normals (4) along the rows of the artificial basis matrix, elements of the inverse matrices, artificial basis solutions, and constraints (4) in two adjacent artificial basis matrices, the relations given in (Kudin et al., 2007) take place.

Let us establish the properties of solutions of a linear system with a square matrix of constraints. Let system (5) have a square matrix of constraints, for definiteness of dimension m ($r = \text{rang}(A) = m$).

Proposition 1. For the existence of a unique solution (4), it is necessary and sufficient that $\alpha_{lk}^{(i)} \neq 0, i = \overline{1, m}$, where $\alpha_{lk}^{(i)}$ are the leading elements of the simplex iteration of the method of artificial basis matrices by replacing the rows of the basis matrix formed by (5) with the normals (4).

Proof. Let the simplex iterations of replacing the rows of the matrix (5) with the normals of the constraints (4) (Kudin et al., 2007) satisfy the condition $\alpha_{lk}^{(i)} \neq 0, i = \overline{1, m}$. That is, at all m iterations, the condition of support of the intermediate artificial basis matrix (linear independence of the rows of the matrix) and the nondegeneracy of the matrix (4) are satisfied. This means that there is a solution to the system (4) and, unique. It is not difficult to verify that the "reverse" condition is also satisfied. $\square$

Proposition 2. The matrix A of the main system is full rank (non-degenerate for this case) if $\alpha_{lk}^{(i)} \neq 0, i = \overline{1, m}$, where i is the number of iterations for replacing the normal vectors of the auxiliary system (5) with the vectors of the main system (4).

Proof. The validity of the statement follows from the fulfillment of the support condition (linear independence of the rows of the basis matrix) when performing the substitution iteration and its non-degeneracy.

Corollary 1. *The rank of system (4) is determined by the number of correct substitutions (fulfillment of the support condition – linear independence of rows) based on simplex iterations.*

Let $r = \text{rang}(A) < m$. In general, the solvability analysis of the SLAE (4) is carried out in two steps. In the first step, we find a separate solution of the relaxed SLAE (4) with the normal constraints forming the basis matrix. In the second step, we analyze the solvability of (4) as a whole, i.e., we establish the admissibility of the found solution to the basic system. It should be noted that the solvability conditions of the above statement can be extended to the general case of the structural properties of the constraint matrix and the rank value, i. e., $r = \text{rang}(A) < \min(n, m)$.

Proposition 3. *A necessary and sufficient condition for the undecidability of system (4) for which $r = \text{rang}(A) = m$ is the violation of the relation for at least one of the expansions of the row (the normal vector $\alpha_i, i \in I$ of the constraints)*

$$c_i \neq \bar{c}_i = \alpha_i C_b, i \in I / I_b, \quad (6)$$

i. e., $\exists i_0 \in I / I_b, c_{i_0} \neq \bar{c}_{i_0} = \alpha_{i_0} C_b$.

Proof. Necessity. Proof by contradiction. Let system (4) be solvable. Taking into account the solvability property for $\forall u_0 \in U$ (in particular, our intermediate) in the assumption of the value of rank r , we have

$$a_r u_0 = c_r, i \in I$$

in particular, the system $a_{i_0} u_0 = c_{i_0}$ is equivalent $\alpha_{i_0} A_{(b)} u_0 = \alpha_{i_0} C_b$ as a linear combination of basis constraints, then $\alpha_{i_0} A_{(b)} = a_{i_0}, \alpha_{i_0} C_b = c_{i_0} = \bar{c}_{i_0}$. We have arrived at a contradiction, since by the condition must $\exists i_0, c_{i_0} \neq \bar{c}_{i_0} = \alpha_{i_0} C_b$.

Sufficiency. Proof also from the converse. Let us have the fulfillment of condition (6), i.e. among $\alpha_i, i \in I$ – series expansions of normals $\exists i_0, c_{i_0} \neq \bar{c}_{i_0} = \alpha_{i_0} C_b$, with solvability (4). For $\forall u_0 \in U$ the solution of system (4), $a_r u_0 = c_r, i \in I$. In particular, under the assumption of solvability, it follows that $\alpha_{i_0} A_{(b)} u_0 = \alpha_{i_0} C_b, \alpha_{i_0} A_{(b)} = a_{i_0}, \alpha_{i_0} C_b = c_{i_0} = \bar{c}_{i_0}$. Due to the assumption of (6), we have reached a contradiction. □

Proposition 4. *A necessary and sufficient condition for the solvability of system (4) for which $r = \text{rang}(A) = m$ is the fulfillment of the relation for vectors $\alpha_i, i \in I_H$, row expansions of normals (4):*

$$c_i = \bar{c}_i = \alpha_i C_b, i \in I_H. \quad (7)$$

The proof of the statement is carried out according to the scheme of the previous one (statement 3). Taking into account the solvability of (4) we obtain the validity of (7). On the other hand, the assumption of the validity of (7) for $\forall i_0 \in I_H : a_{i_0} u_0 = \alpha_{i_0} A_{(b)} u_0 = \alpha_{i_0} C_b = \bar{c}_{i_0} = c_{i_0}$, will determine the solvability of (4). □

It should be noted that the conditions of the statement correspond to the provisions of the well-known Kronecker-Capelli theorem on the solvability of a system of equations.

Let us consider the main stages of the relaxation scheme of sequential inclusion of relaxed constraints (4) into system (5) with known elements of the method. At the last iteration of the inclusion, the rank value, the solution of the relaxed ("intermediate") system of algebraic equations (4) with constant elements for a non-degenerate system, will be established.

Algorithm 1

Preparatory step. We assume that the solution u_0 and the inverse matrix A_b^{-1} of the SLAE of type (5) of the auxiliary system are initially known.

Step 1. We perform simplex iterations by replacing the rows of the basic matrix of the system (5) with the normals of the constraints of the system (4), according to the relations (Kudin et al., 2007).

We find the corresponding elements of the method: row expansion of the basic matrices with constraints from (4), the inverse basic matrix, artificial basic solutions $u_0^{(k)}$, where k is the iteration number.

Step 2. We check the number of iterations r of replacing the rows of the auxiliary system (5) with the rows of the main system (4) for which the conditions are met: $\alpha_{lk} \neq 0$ is the number determines the rank of the main system.

Step 3. If the number of iterations for which $\alpha_{lk}^{(i)} \neq 0$ is equal to m , then we proceed to the next step. Otherwise, to the penultimate step.

Step 4. We find the unique solution of the relaxed system (4) $A_b^{-1} \cdot C_b = u^0$.

Step 5. Fulfilling the condition $r < m$ means violating the condition of completeness of rank and uniqueness of the solution and requires further analysis of the solvability of the problem.

The last step. Formation of initial information based on the results of the analysis (5).

3. Computational experiment

A specific model problem was constructed ($m = 5$ selected), for which:

- 1) The pseudo-random number sensor generated a solution u_0 with components from the range $[0, 99]$;
- 2) the off-diagonal elements of the coefficient matrix $A^{(CO)}$ are complex numbers $a_{ij}^{(CO)}$ with real and imaginary parts – numbers from the range $[0, 1]$;

3) diagonal elements $a_{ii}^{(CO)}$ are calculated by the formula:

$$a_{ii}^{(CO)} = \sum_{j=1, j \neq i}^m a_{ij}^{(R)} + i \sum_{j=1, j \neq i}^m a_{ij}^{(S)} = a_{ii}^{(R)} + i a_{ii}^{(S)},$$

where m is the dimension of the coefficient matrix.

4) the vector of the right-hand sides is calculated by the formula $C^{(CO)} = A^{(CO)}u_0 = (A^{(R)} + iA^{(S)})u_0 = C^{(R)} + iC^{(S)}$, i. e. the components of the system with complex elements (3) are found;

5) system (4) is formed and "normalized" (each row of constraints is divided into the corresponding diagonal element of the matrix).

As a result (de facto) we obtain a SLAE of type (4) with a diagonal advantage of the component constraint matrices for both the real and imaginary parts ($A^{(R)}$, $A^{(S)}$) of the form:

$$A^{(R)} = \begin{pmatrix} 1 & 0.205882 & 0.169118 & 0.257353 & 0.367647 \\ 0.472973 & 1 & 0.027027 & 0.135135 & 0.364865 \\ 0 & 0.330189 & 1 & 0.5 & 0.169811 \\ 0.25283 & 0.350943 & 0.283019 & 1 & 0.113208 \\ 0.334601 & 0.114068 & 0.368821 & 0.18251 & 1 \end{pmatrix},$$

$$A^{(S)} = \begin{pmatrix} 1 & 0.285714 & 0.107143 & 0 & 0.607143 \\ 0.175115 & 1 & 0.299539 & 0.442396 & 0.0829493 \\ 0.436893 & 0.0485437 & 1 & 0.038835 & 0.475728 \\ 0.235632 & 0 & 0.195402 & 1 & 0.568966 \\ 0.608696 & 0.236025 & 0.0807453 & 0.0745342 & 1 \end{pmatrix}, \quad (8)$$

$$C^{(R)} = (120.199 \quad 83.4189 \quad 165.047 \quad 149.238 \quad 154.365)^T,$$

$$C^{(S)} = (111.071 \quad 102.23 \quad 162.087 \quad 172.546 \quad 130.161)^T, \quad (9)$$

where $A^{(CO)} = A^{(R)} + iA^{(S)}$, $C^{(CO)} = C^{(R)} + iC^{(S)}$, $i^2 = -1$ - imaginary one.

The exact solution of such a generated system is $u_0 = (45 \ 16 \ 98 \ 95 \ 84)^T$.

For testing, a computer program was developed in the C++ programming language that implements the ABM with calculations within the *long double* type (for real, imaginary times, solution and intermediate values).

The accuracy parameter of the computational experiment was chosen as $\text{eps} = 1e-12$ – the threshold value for comparison with zero (Kudin et al., 2019a).

Experiment 1. Solving the SLAE (4) with the coefficient matrix (8) and the right-hand side (9) using the ABM.

As a result of the 5 iterations, a solution $u = (45 \ 16 \ 98 \ 95 \ 84)^T$ with an error of 10^{-13} was obtained, which is due to the error of the input data, the error of calculations and the machine "epsilon". During each iteration, the fulfillment of the condition $\alpha_{ik} \neq 0$ was observed. In addition, the solvability conditions of the redefined system (4) were checked by calculating and evaluating the residuals of the solution $\max_{r \in I} |\Delta_r| \leq 4.3 \cdot 10^{-14}$, which are within the computational error and confirm the correctness of the solution of the redefined system (4).

The content of the iteration results is given in Appendix A.

Experiment 2. Numerical modeling of the mechanism of interaction between real and imaginary components of the system – for a complex two-component model.

The activity coefficient is introduced into consideration, a function of the influence of the imaginary component of a complex system on the real one. For example, in the form of $A = \alpha(k) = 1/k$, with parameter $k \in [0.5, 2]$.

In this computational experiment, using the ABM, solutions of the SLAE were found, the elements of the model of which (constraint matrix) depend on the activity function $\alpha(k)$ and are formed by the relations: $\bar{a}_{ij} = \min\{a_{ij}^{(I)}, \alpha a_{ij}^{(S)}\}$, where $a_{ij}^{(I)}$ and $a_{ij}^{(S)}$ are the real and imaginary parts of the elements of the matrix $A^{(CO)}$.

Figures 1–5 shows graphs of the dependence of the components of the SLAE solution on the parameter $k \in [0.5, 2]$, which varies with a step of 0.15. The horizontal lines drawn in black correspond to the components of the solution of the initial SLAE (1) with coefficients (8), (9).

The graphs above (Fig. 1–5) show the dynamics of changes in the solution components (5 of them) when the activity coefficient changes (within $[0.5, 2]$). An increase is observed for all solution components except the first component when increasing, and for the first component from the 8th iteration.

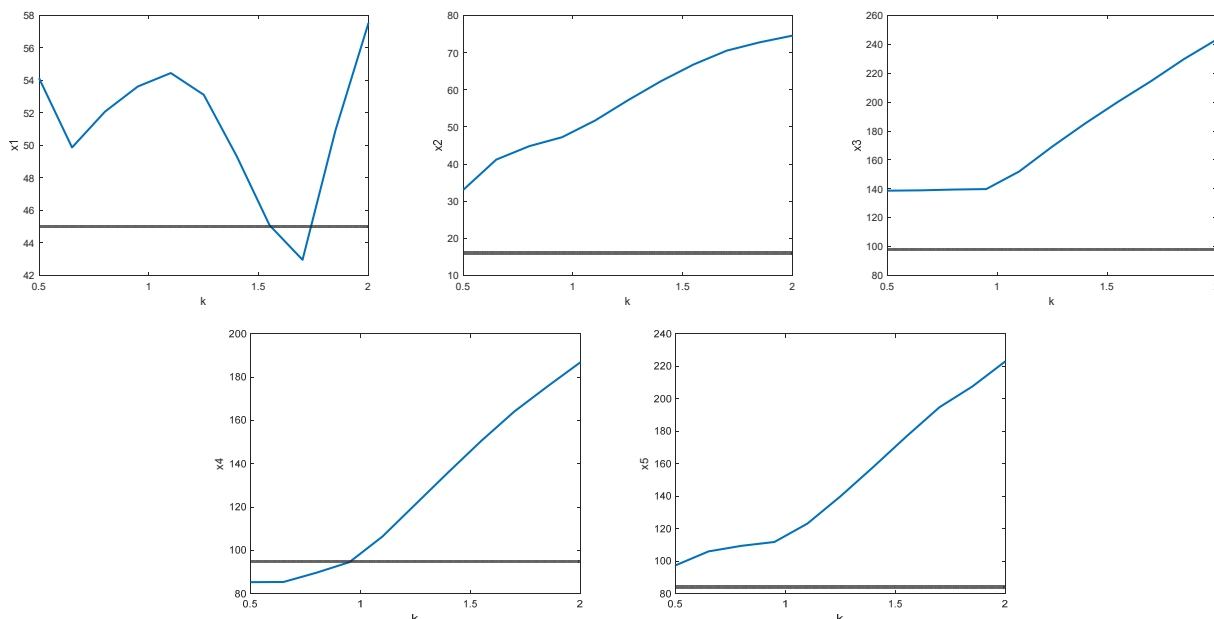


Fig. 1–5. Evolution of solution components depending on k ($k = 0.5 \dots 2$) – abscissa axis

Discussion and conclusions

According to the research objectives, an analysis of known approaches to solving computational problems, taking into account human "activity" was conducted. Basic mathematical methods (in the class of linear systems) were identified, which include the ability to take into account the activity of a person in the process of solving a problem. The computational procedures for the complex analysis of linear systems (two-component systems) were developed (adapted to modeling tasks at different stages), with a component of taking into account human activity. A computational experiment was conducted to verify the correctness of the "work" of the procedures on model examples.

The results of the experiments (obtained residuals of order 10^{-13} and Figures 1-5) confirm the correctness of using ABM in numerical modeling of problems related to solving SLAE with complex elements. Prospects for further research are outlined, in particular, the "complete" complex linear system, and the development of a generalization of the developed computational procedures for the analysis and solution of complex systems based on classical methods.

Appendix A

Program output when solving SLAE (1) with coefficient matrix (8) and right-hand side (9). Here tek_alfa_l_k is the leading element of the transformation (α_{lk}).

Table A.1

(A C)					
1	0.205882	0.169118	0.257353	0.367647	120.199
0.472973	1	0.027027	0.135135	0.364865	83.4189
0	0.330189	1	0.5	0.169811	165.047
0.25283	0.350943	0.283019	1	0.113208	149.238
0.334601	0.114068	0.368821	0.18251	1	154.365
1	0.285714	0.107143	0	0.607143	111.071
0.175115	1	0.299539	0.442396	0.0829493	102.23
0.436893	0.0485437	1	0.038835	0.475728	162.087
0.235632	0	0.195402	1	0.568966	172.546
0.608696	0.236025	0.0807453	0.0745342	1	130.161

Table A.2

Iteration 1				
tek_alfa_l_k=1;				
The inverse of A is A ⁽⁻¹⁾ :				
1	-0.205882	-0.169118	-0.257353	-0.367647
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1
Права частина c:				
120.199	1	1	1	1

Table A.3

Iteration 2				
tek alfa k=0.902623;				
The inverse of A is A ⁽⁻¹⁾ :				
1.10788	-0.228093	-0.181198	-0.254293	-0.324086
-0.523998	1.10788	0.0586746	-0.0148613	-0.211581
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1
c:				
120.199	83.4189	1	1	1

Table A.4

Iteration 3				
tek alfa k = 1.01937;				
The inverse of A is A ⁽⁻¹⁾ :				
1.07713	-0.163069	-0.177754	-0.166289	-0.30632
-0.514039	1.08683	0.0575595	-0.0433586	-0.217334
0.16973	-0.358858	0.980995	-0.485683	-0.0980501
0	0	0	1	0
0	0	0	0	1
c:				
120.199	83.4189	165.047	1	1

Table A.5

Iteration 4				
tek alfa k=0.805283;				
The inverse of A is A ⁽⁻¹⁾ :				
1.10603	-0.113794	-0.125531	-0.206497	-0.320416
-0.506503	1.09967	0.0711762	-0.0538426	-0.221009
0.254148	-0.21494	1.13352	-0.603121	-0.13922
-0.173813	-0.296321	-0.314049	1.2418	0.0847668
0	0	0	0	1
c:				
120.199	83.4189	165.047	149.238	1

Table A.6

Iteration 5				
tek alfa k=0.831702;				
The inverse of A is A ⁽⁻¹⁾ :				
1.25024	-0.131513	0.000395104	-0.233865	-0.385253
-0.407036	1.08745	0.158035	-0.0727201	-0.265731
0.316805	-0.222638	1.18824	-0.615013	-0.167392
-0.211963	-0.291633	-0.347363	1.24904	0.10192
-0.450059	0.0553001	-0.393009	0.0854148	1.20235
c:				
120.199	83.4189	165.047	149.238	154.365

Table A.7

Program completion									
Residuals vector:									
- 4.26e-14	- 1.42e-14	2.84e-14	0	0	- 4.26e-14	0	0	0	- 2.84e-14
The system is correct									
Solution u:									
45	16	98	95	84					
Execution time: 0.034 seconds									

Authors' contribution: Volodymyr Kudin – concept, problem statement, method, algorithm, source analysis, preparation of literature review or theoretical foundations of the study; Vyacheslav Onotsky – computational procedures, software, experiment (collection and verification of empirical data, research), source analysis, preparation of literature review or theoretical foundations of the study.

Sources of funding. Funding is partially provided by Taras Shevchenko National University of Kyiv.

References

- BigInt. (2013). *Large Integer Package*. <https://github.com/luckyabib/BigInt>
- Blackford, L., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., & Whaley, R. (2002). An Updated Set of Basic Linear Algebra Subprograms (BLAS). *ACM Transactions on Mathematical Software*, 28(2), 135–151. <https://doi.org/10.1145/567806.567807>
- Demmel, J. (1997). *Applied Numerical Linear Algebra*. SIAM.
- Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W., Wilson, J., & van der Doelen B. (2012). A strategy for human factors/ergonomics: developing the discipline and profession. *Ergonomics*, 55(4), 377–395. <https://doi.org/10.1080/00140139.2012.661087>

- Fromberger, M. (2005). *Multiple Precision Integer Library (MPI)*. <https://github.com/servo/nss/tree/master/lib/freebl/mpl>
- Granlund, T. (2023). *The GNU Multiple Precision Arithmetic Library. Edition 6.3.0*. <https://gmplib.org/list-archives/gmp-announce/2023-July/000050.html>
- Korobiichuk, I., Tokar, A., & Danik, Y. (2019). Evaluation methods for the ergatic system reliability operator. *Automation 2019: progress in automation, robotics and measurement techniques*, 920, 560–570. http://dx.doi.org/10.1007/978-3-030-13273-6_52
- Kudin, V., Lyashko, S., Khritonenko, N., & Yatsenko, Y. (2007). Analysis of the properties of a linear system using the method of artificial basis matrices. *Cybernetics and Systems Analysis*, 43(4), 563–570. <https://doi.org/10.1007/s10559-007-0081-3>
- Kudin, V., Onotskiy, V., Al-Ammouri, A., & Shkvarchuk, L. (2019a). Advancement of a long arithmetic technology in the construction of algorithms for studying linear systems. *Eastern-European Journal of Enterprise Technologies*, 1(4), 14–21. <https://doi.org/10.15587/1729-4061.2019.157521>
- Kudin, V., Onotskiy, V., Onyshchenko, A., & Stupak, Y. (2019b). Forming a methodology of basic matrices in the study of poorly conditioned linear systems. *Eastern-European Journal of Enterprise Technologies*, 6(4), 57–67. <https://doi.org/10.15587/1729-4061.2019.184530>
- Kudin, V., Onyshchenko, A., & Onyshchenko, I. (2019). Algorithmizing the methods of basis matrices in the study of balance intersectoral ecological and economic models. *Eastern-European Journal of Enterprise Technologies*, 3(4), 45–55. <https://doi.org/10.15587/1729-4061.2019.170516>
- Mrachkovskiy, D. (2023). Development of a mathematical model of the operator in a human-machine control system for a quarry excavator. *Electromechanical Systems, Optimization Methods and Modeling*, 62(3), 41–48. <https://doi.org/10.32782/2072-2052.2023.3.62.4>
- Strauch, B. (2017). *Investigating Human Error: Incidents, Accidents, and Complex Systems*. Taylor & Francis Group.
- Tan, Y., Feng, D., & Shen, H. (2017). Research for Unmanned Aerial Vehicle components reliability evaluation model considering the influences of human factors. In B. Xu, & Y. Chen (Eds.), *3rd International Conference on Mechanical, Electronic and Information Technology Engineering*, Vol. 139. EDP Sciences. <https://doi.org/10.1051/mateconf/201713900221>

Отримано редакцією журналу / Received: 08.05.25
Прорецензовано / Revised: 12.11.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Володимир КУДІН, д-р техн. наук, проф.
ORCID ID: 0009-0005-9071-1213
e-mail: vol.kudin@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

В'ячеслав ОНОЦЬКИЙ, канд. фіз.-мат. наук, асист.
ORCID ID: 0000-0002-1920-0905
e-mail: onotskiy@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

АНАЛІЗ ЛІНІЙНИХ СИСТЕМ ІЗ КОМПЛЕКСНИМИ ЕЛЕМЕНТАМИ МЕТОДОМ БАЗИСНИХ МАТРИЦЬ

Запропоновано застосування математичного апарату теорії функції комплексної змінної (в частині лінійних систем із комплексними елементами) для опису механізму взаємодії компонент у системі ("модель" – "дослідник") у процесі математичного моделювання. Розвинуто метод базисних матриць до лінійних систем із комплексними елементами. Розроблено алгоритм й обчислювальну процедуру аналізу властивостей і розв'язання з урахуванням механізму взаємодії дійсної та уявної складових компонент у поданні математичної моделі. Проведено обчислювальний експеримент з перевірки коректності роботи запропонованого алгоритму й обчислювальної процедури.

Об'єктом дослідження є процеси різної природи в контексті розв'язання задачі математичного моделювання.
Предмет дослідження – методи, алгоритми й обчислювальні процедури математичного моделювання, зокрема, математичний апарат теорії функції комплексної змінної, лінійної алгебри.

Мета дослідження: розробити алгоритми й обчислювальні процедури аналізу властивостей двох компонентних математичних моделей процесів (з комплексними елементами) зі здатністю враховувати активність дослідника в процесі розв'язання задачі.

Завдання дослідження: провести аналіз відомих підходів до розв'язання обчислювальних задач з урахуванням "активності" людини як аналітика, дослідника операцій, особи, яка ухвалює рішення тощо; виявити базові математичні методи в класі лінійних систем, у яких закладено здатність урахувати (й активність) людини в процесі розв'язання задачі; розробити адаптовані під задачі моделювання (на різних етапах) обчислювальні процедури й алгоритми комплексного аналізу лінійних систем (двокомпонентних систем), зі складовою врахування активності людини; провести обчислювальний експеримент на основі запропонованого методу, алгоритму та процедури з перевірки їхньої коректності "роботи" на модельних прикладах; зробити аналіз і сформулювати висновки за результатами експериментів та окреслити перспективи подальших досліджень.

У процедурі математичного моделювання закладено проведення ряду послідовних етапів (з оберненим зв'язком). І, на жаль, неусуєна проблема адекватності математичної моделі подання досліджуваному процесу є супутником розв'язання прикладних задач. Як відомо, задачі моделювання подаються математичним апаратом різної складності і часто "приводяться" (спрощуються) зведенням до лінійних моделей, зокрема, здійснюється дискретизація неперервних і лінеаризація нелінійних задач.

Варто зазначити, що в процесі моделювання не закладено у явному вигляді врахування впливу досвіду дослідника, тобто людського фактору, на якість процесу моделювання на всіх етапах. Де-факто, двокомпонентність структури задачі ("явна" складова – математична модель та "неявна" – дослідник у контурі моделювання) "простежується".

Це обумовлює актуальність розроблення нових математичних постановок задач моделювання з комплексними елементами (двокомпонентних систем), у яких закладено механізми врахування досвіду дослідника щодо досягнення більшої адекватності моделі досліджуваному процесу під час розв'язання задачі на різних етапах.

Ключові слова: система лінійних алгебраїчних рівнянь, метод базисних матриць, комплексний аналіз.

Один з авторів (В. Кудін) є членом редколегії видання, тому він не брав участі у рецензуванні та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішеннях про публікацію результатів.

One of the authors (V. Kudin) is a member of the journal editorial board; he did not take part in the peer-review process or in the decision to publish this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.21

DOI: <https://doi.org/10.17721/1812-5409.2026/1.8>

Yana MOKANU, PhD Student, Assist.

ORCID ID: 0009-0006-0752-6584

e-mail: yana.mokanu3075@gmail.com

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

ON ERGODIC PROPERTY OF SOME LÉVY-TYPE PROCESSES IN $\mathbb{R}^d$

In this paper, we investigate the ergodicity in total variation of the process X , related to some integro-differential operator with unbounded coefficients, and describe the speed of convergence to the respective invariant measure. Roughly speaking, a Markov process is said to be ergodic if there exists an invariant probability measure to which the transition probability of the considered process converges. In this case, the speed of convergence is also called the ergodic rate, and it plays an important role in further investigations of asymptotic properties, in particular, in simulating the processes, calculating their functionals, and also in limit theorems.

In the proofs, we use the Lyapunov-type criterion, which resembles the well-known from the theory of ordinary differential equations Lyapunov's second method. Some examples are provided.

Key words: ergodicity, Lévy-type process, Lyapunov criterion, Lyapunov function.

AMS 2020 classification: 60G51, 37A50, 60G17.

Introduction

Consider an integro-differential operator

$$Lf(x) = \ell(x) \cdot \nabla f(x) + \int_{\mathbb{R}^d \setminus \{0\}} (f(x+u) - f(x) - \nabla f(x) \cdot u \mathbb{1}_{|u|<1}) \nu(x, du), \quad (1)$$

defined on the set $C_0^2(\mathbb{R}^d)$ of twice continuously differentiable functions with compact support, where $\ell(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$, and the Lévy-type kernel $\nu(x, du)$ satisfies

$$\int_{\mathbb{R}^d \setminus \{0\}} (1 \wedge |u|^2) \nu(x, du) < \infty \quad \text{for any } x \in \mathbb{R}^d.$$

To simplify the notation, we assume that the kernel $\nu(x, du)$ does not have an atom at 0 and write $\int_{\mathbb{R}^d}$ instead of $\int_{\mathbb{R}^d \setminus \{0\}}$.

Under certain conditions (see Section 1 below) one can associate with $(L, C_0^2(\mathbb{R}^d))$ a Markov process X , whose transition probability kernel $P_t(x, dy)$ possesses certain regularity properties. The process X is often called a Lévy-type process, because locally its structure resembles the structure of a Lévy process, e.g. a process with independent stationary increments.

The aim of this paper is to derive sufficient conditions for ergodicity of the process X related to (1) and find the ergodic rates.

This work further develops our previous studies on the ergodicity of Lévy-type processes: we began our investigation in (Knopova, & Mokanu, 2024) with the one-dimensional case and bounded Lévy-type kernel; continued in (Knopova, & Mokanu, 2025), where a Lévy-driven stochastic differential equation (SDE) with unbounded coefficients was considered in dimension d .

The long-time behaviour of Markov processes has been extensively investigated. Skorokhod in (Skorokhod, 1987) discussed the long-time behaviour of general Markov processes and solutions to SDEs, as well as stabilisation properties of solutions to systems of SDEs. Masuda in (Masuda, 2004), (Masuda, 2007) investigated a Lévy-driven Ornstein-Uhlenbeck process, in particular, the existence of a smooth density, ergodicity, and convergence rates. Sufficient conditions for recurrence, positive recurrence, and exponential ergodicity of one-dimensional Lévy-type processes generated by the integro-differential operator were derived in (Wang, 2008). In (Wang, 2013) sufficient conditions for exponential and strong ergodicity of SDEs driven by symmetric α -stable processes were obtained. Next, exponential ergodicity of solutions to certain SDEs (in particular, SDEs of McKean-Vlasov type), damping Hamiltonian dynamics with state-dependent and non-local collisions, and Lévy driven Langevin dynamics with singular potentials was studied in (Liang, Majka, & Wang, 2021), (Bao, & Wang, 2023) and (Bao, Fang, & Wang, 2024), respectively. Ergodicity of a solution to the jump-diffusion SDE was studied by Wee in (Wee, 1999). In (Fort, & Roberts, 2005) and (Douc, Fort, & Guillin, 2009) sufficient conditions for ergodicity were given in the form of so-called (f, r) -modulated moments. In (Kulik, & Scheutzow, 2018) the problem of the uniqueness of an invariant probability measure of a Markov process was investigated; here the authors also derived sufficient conditions for the weak convergence of the transition probability to the invariant measure. Kulik in (Kulik, 2018) discussed the asymptotic behaviour of a wide range of Markov processes from simple Markov chains with finite state spaces to more sophisticated models (including solutions to Lévy driven SDEs), in particular, the author examined different types of ergodic rates and their further application in limit theorems; the question of weak ergodicity was addressed as well. Sandrić in (Sandrić, 2013) gave sufficient conditions for recurrence, transience, and ergodicity of stable-like processes in terms of their coefficients. Further, sufficient conditions for recurrence, transience, and ergodicity of stable-like Markov chains under certain assumptions on their density functions and drift were derived in (Sandrić, 2014). In (Sandrić, 2016) the author discussed recurrence, transience, ergodicity, strong, (sub)-exponential ergodicity, and mixing properties of Lévy-type processes driven by the infinitesimal generator.

The structure of this paper is as follows. In Section 1, we recall the notion of ergodicity, discuss the assumptions, and formulate our results. Proofs are given in Section 2 and examples in Section 3.

1. Main definitions and results

A Markov process $(X_t)_{t \geq 0}$ on $\mathbb{R}^d$ with a transition probability kernel $P_t(x, dy)$ is called (strongly) *ergodic* (see (Sandrić, 2016)) if there exists an invariant probability measure $\pi(\cdot)$ such that

$$\lim_{t \rightarrow \infty} \|P_t(x, \cdot) - \pi(\cdot)\|_{TV} = 0$$

for any $x \in \mathbb{R}^d$. Here $\|\cdot\|_{TV}$ is the total variation norm. Moreover, one can describe the speed of convergence to the invariant measure, if certain assumptions are satisfied. The process is called (non-uniformly) ψ -ergodic (see (Sandrić, 2016) for polynomial and exponential rates), if there exists a function $\psi \geq 0$, $\psi(t) \rightarrow 0$ as $t \rightarrow \infty$, such that

$$\|P_t(x, \cdot) - \pi(\cdot)\|_{TV} \leq C(x)\psi(t), \quad t \geq 1, \quad (2)$$

for some $C: \mathbb{R}^d \rightarrow [0, \infty)$.

There are several possible ways to show ergodicity of a Markov process (see (Kulik, 2018, § 3.5)). We use the skeleton chain approach, described in (Kulik, 2018), which allows us to reduce the continuous-time setting to the discrete one by constructing a skeleton chain $X^h := \{X_{nh}, n \in \mathbb{Z}_+\}$ and apply ergodic theorems, developed specifically for Markov chains. Under certain irreducibility assumptions (e.g., Dobrushin condition), ergodicity of the skeleton chain implies ergodicity of the initial process with the same convergence rate (cf. (Kulik, 2018, § 3.2), (Sandrić, 2016)).

To see if the chain is ergodic, one can check the (local) Dobrushin (or other forms of irreducibility condition, suitable for a particular task (see the discussion in (Kulik, 2018, p. 64))) and the Lyapunov-type conditions ((Kulik, 2018, Th. 2.8.9, Th. 3.2.3)). Note that the irreducibility-type assumption is crucial for the described approach, but if it fails, it is still possible for the process to be ergodic only in a weaker sense (see the discussion in (Kulik, 2018, § 4)).

A Markov chain X with a transition kernel $\mathbb{P}(x, dy)$ satisfies the *local Dobrushin* condition (cf. (Kulik, 2018, p. 42)) on a measurable set $B \subset \mathbb{R}^d \times \mathbb{R}^d$ if

$$\sup_{x, y \in B} \|\mathbb{P}(x, \cdot) - \mathbb{P}(y, \cdot)\|_{TV} < 2. \quad (3)$$

In practice, it is sufficient and reasonable to check if there exists an absolutely continuous component of $\mathbb{P}(x, dy)$, such that $x \mapsto \mathbb{P}(x, \cdot)$ is continuous in L_1 (see (Kulik, 2018, Prop. 2.9.1)), instead of proving (3) directly. This, in turn, is the case if the initial Markov process possesses a strictly positive transition probability density $p_t(x, \cdot)$, such that $x \mapsto p_t(x, \cdot)$ is continuous in x in L_1 (see (Sandrić, 2016)). It means that the respective semigroup $(P_t)_{t \geq 0}$, $P_t f(x) := \mathbb{E}^x f(X_t)$ possesses the C_b -Feller property, that is, it preserves the space of bounded continuous functions (see (Böttcher, Schilling, & Wang, 2013)). The Dobrushin condition ensures that the process converges to the same invariant measure, regardless of its starting point.

The *Lyapunov-type* condition (see (Kulik, 2018, p. 52)) allows us to establish moment estimates, responsible for (non-uniform) ergodic rates. Suppose that there exist:

- a norm-like function $V: \mathbb{R}^d \rightarrow [1, \infty)$ (i.e. $V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$), bounded on a compact K ;
- a function $f: [1, \infty) \rightarrow (0, \infty)$, that admits a non-negative increasing and concave extension to $[0, \infty)$;
- a constant $C > 0$;

such that the following relations hold:

$$\mathbb{E}^x V(X_1) - V(x) \leq -f(V(x)) + C, \quad x \in \mathbb{R}^d, \quad (4)$$

$$f\left(1 + \inf_{x \notin K} V(x)\right) > 2C.$$

Inequality (4) is called the *Lyapunov inequality*, and such a function V is called the *Lyapunov function*.

Given that we are in the continuous-time settings, it is reasonable to consider the Lyapunov-type condition in terms of the process X itself, namely, in terms of its *full generator* $\mathcal{L}$:

$$\mathcal{L}V(x) \leq -f(V(x)) + C, \quad (5)$$

where f, V and C have the same meaning as above.

The full generator is taken into consideration because Lyapunov functions are often unbounded and the domain of L is too narrow to include them. Recall the “stochastic version” of a *full generator*, see (Böttcher, Schilling, & Wang, 2013, (1.50)). The full generator of a measurable contraction semigroup $(P_t)_{t \geq 0}$ on the space $C_\infty(\mathbb{R}^d)$ of continuous functions vanishing at infinity is

$$\hat{L} := \left\{ (f, g) \in (\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^d)) : \quad M_t^{[f, g]} := f(X_t) - f(x) = \int_0^t g(X_s) ds, \quad x \in \mathbb{R}^d, t \geq 0, \right. \\ \left. \text{is a local } \mathcal{F}_t^X := \sigma(X_s, s \leq t) \text{ martingale} \right\}.$$

Note that in general $\hat{L}$ is not single-valued, see the discussion in (Böttcher, Schilling, & Wang, 2013, p. 36).

Since for $f \in C_0^2(\mathbb{R}^d)$

$$P_t f(x) - f(x) = \int_0^t P_s L f(x) ds \quad \forall t \geq 0, x \in \mathbb{R}^d, \quad (6)$$

see (Ethier, & Kurtz, 1986, Ch. 1, Prop. 1.5), by the strong Markov property and the Dynkin formula $(f, Lf) \subset \hat{L}$ for any $f \in C_0^2(\mathbb{R}^d)$. Further, it is not difficult to check that $Lf(x) < \infty$ for $f \in \mathcal{D}_0$, where

$$\mathcal{D}_0 := \left\{ f \in C^2(\mathbb{R}^d) : \left| \int_{|u| \geq 1} (f(x+u) - f(x)) \nu(x, du) \right| < \infty \quad \forall x \in \mathbb{R}^d \right\}.$$

Using the localization procedure, the Ito formula, and taking the expectation, we derive (6) for any $f \in \mathcal{D}_0$ and, consequently, we get that $(f, Lf) \in \hat{L}$, $f \in \mathcal{D}_0$. We denote this extension of L to $\mathcal{D}_0$ by $\mathcal{L}$.

Summarising, condition (5) for the initial process implies (4) for its skeleton chain (see (Kulik, 2018, Th. 3.2.3)), which, along with the Dobrushin condition, ensures the ergodicity in total variation of this chain and provides the (non-uniform) ergodic rate (cf. (Kulik, 2018, Th. 2.8.6, Th. 2.8.8)). As it was mentioned above, the ergodicity of the chain implies the ergodicity of the initial process with the same ergodic rate.

The speed of convergence to π in (2) as $t \rightarrow \infty$ is (cf. (Kulik, 2018, Th. 2.8.8, Th. 3.2.3))

$$\psi(t) := \left(\frac{1}{f(F^{-1}(\gamma t))} \right)^\delta, \quad F(t) := \int_1^t \frac{dw}{f(w)}, \quad t \in (1, \infty), \quad (7)$$

where $\gamma, \delta \in (0, 1)$. Moreover,

$$\int_{\mathbb{R}^d} f(V(x)) \pi(dx) < \infty.$$

Our standing assumptions in this paper would be

There exists a solution $\{X_t, t > 0\}$ to the martingale problem for $(L, C_0^2(\mathbb{R}^d))$; (A1)

The skeleton chain $X^h := \{X_{nh}, n \in \mathbb{Z}_+\}$ satisfies the Dobrushin condition. (A2)

Condition (A1) assures the existence of the process associated with (1) (which we call the *Markovian selection* related to L), but it looks rather abstract and is not easy to check. The sufficient conditions ensuring (A1) were obtained in (Kühn, 2020) and are the following:

The symbol q is locally bounded, continuous in x , satisfies $q(x, 0) = 0$; (A1')

$\lim_{R \rightarrow \infty} \sup_{|x| \leq R} \sup_{|\xi| \leq R^{-1}} |q(x, \xi)| = 0$. (A1'')

Recall that the symbol $q(x, \xi)$ of the operator L is a function, such that for f from the Schwartz space $S(\mathbb{R}^d)$

$$L f(x) = - \int_{\mathbb{R}^d} e^{i\xi \cdot x} q(x, \xi) \hat{f}(\xi) d\xi,$$

where $\hat{f}$ is the Fourier transform of f , and

$$q(x, \xi) = -i\ell(x) \cdot \xi + \int_{\mathbb{R}^d} \left(1 - e^{i\xi \cdot u} + i\xi \cdot u \mathbb{1}_{|u| \leq 1} \right) \nu(x, du).$$

Under (A1') and (A1''), there exists a solution to the Martingale problem for $(L, C_0^\infty(\mathbb{R}^d))$, which is strongly Markov, see (Kühn, 2020, Th. 1.1, Cor. 4.1).

The uniqueness of the process in the case of unbounded coefficients is harder to prove. Up to our knowledge, at the moment there are no general results establishing the uniqueness of the solution to the Martingale problem for $(L, C_0^\infty(\mathbb{R}^d))$, if the kernel $\nu(x, du)$ is unbounded, unless one can rewrite L (using the change of measure) in the form, usually used in Lévy-driven SDEs, and check the conditions for the existence and uniqueness of the (strong or weak) solution to the respective SDE. We do not pursue this topic in the current article and confine ourselves with finding the conditions that guarantee the ergodicity of *some* Markovian selection, related to $(L, C_0^\infty(\mathbb{R}^d))$.

It is a kind of general knowledge that ergodicity can come from the drift or diffusion parts of the generator. We investigate the conditions on the Lévy-type measure under which (5) is satisfied. Our interest lies both in the drift- and noise-induced ergodicity. Further, we will show that in the second case, when the ergodic behaviour comes from the kernel $\nu(x, \cdot)$ itself, the support of $\nu(x, \cdot)$ must depend on the direction of the vector x .

Denote by $|u|$ the norm of a vector $u \in \mathbb{R}^d$; by $\mathcal{B}(r)$ and $\mathcal{B}^c(r)$ the ball centred at 0 with radius r and its complement, respectively; by $\mathcal{A}(r, R)$ the annulus with the inner radius r and the outer radius R .

As a Lyapunov function we take

$$V(x) = 1 + \phi^p(x), \quad p \in (0, 1), \quad x \in \mathbb{R}^d, \quad (8)$$

where $\phi : \mathbb{R}^d \rightarrow \mathbb{R}_+$, $\phi \in C^2(\mathbb{R}^d)$, such that $\phi(x) = |x|$ for $|x| > 1$ and $\phi(x) \leq |x|$ for $|x| \leq 1$.

We also denote by $e_x = \frac{x}{\|x\|}$ the unit vector of $x \in \mathbb{R}^d$, and by $\gamma_{x,u}$ the cosine of the angle between unit vectors e_x, e_u :

$$\gamma_{x,u} := \cos \angle(e_x, e_u).$$

The aim of the article is to find conditions ensuring the Lyapunov-type inequality (5) for $\mathcal{L}$. Our strategy is to select its negative and positive components to compare their impact on the generator. For the sake of simplicity, we consider separately the “drift” and “kernel” parts:

$$\begin{aligned} \mathcal{L}V(x) &= \nabla V(x) \cdot \left(\ell(x) + \int_{\mathcal{A}(1,|x|)} u \nu(x, du) \right) + \\ &+ \int_{\mathbb{R}^d} (V(x+u) - V(x) - \nabla V(x) \cdot u \mathbb{1}_{|u| \leq |x|}) \nu(x, du) =: \\ &=: \mathcal{L}^{drift}V(x) + \mathcal{L}^{ker}V(x). \end{aligned}$$

Consider the “drift” part. For $|x|$ big enough $V(x) = 1 + |x|^p$, so we can transform

$$\mathcal{L}^{drift}V(x) = p|x|^{p-1} \left(|\ell(x)| \gamma_{x,\ell(x)} + \int_{\mathcal{A}(1,|x|)} \gamma_{x,u} |u| \nu(x, du) \right).$$

Since our ultimate goal is to compare the speed of all generator components, we define the first “growth” function

$$\phi^{drift}(x) := |x|^{p-1} \max \left(|\ell(x)|, \int_{\mathcal{A}(1,|x|)} |u| \nu(x, du) \right).$$

It is not hard to see that there exists $\zeta \in [-2, 2]$, such that

$$\mathcal{L}^{drift}V(x) \leq -p\zeta \phi^{drift}(x). \quad (9)$$

Additionally, fix $\lambda \in (0, 1)$ and split the “kernel” part $\mathcal{L}^{ker}$ into the “ball” and “tail” sub-components, based on the jump size $|u|$:

$$\begin{aligned} \mathcal{L}^{ker}V(x) &= \int_{\mathcal{B}(|x|^\lambda)} (V(x+u) - V(x) - \nabla V(x) \cdot u) \nu(x, du) + \\ &+ \int_{\mathcal{B}^c(|x|^\lambda)} (V(x+u) - V(x) - \nabla V(x) \cdot u \mathbb{1}_{\mathcal{A}(|x|^\lambda, |x|)}) \nu(x, du) =: \\ &=: \mathcal{L}^{ball}V(x) + \mathcal{L}^{tail}V(x). \end{aligned} \quad (10)$$

Further, again according to the jump size $|u|$, split the “tail” part:

$$\begin{aligned} \mathcal{L}^{tail}V(x) &= \int_{\mathcal{A}(|x|^\lambda, |x|)} (V(x+u) - V(x) - \nabla V(x) \cdot u) \nu(x, du) + \\ &+ \int_{\mathcal{B}^c(|x|)} (V(x+u) - V(x)) \nu(x, du) =: \\ &=: \mathcal{L}^{big}V(x) + \mathcal{L}^{large}V(x). \end{aligned}$$

Now we are ready to introduce the rest of the “growth” functions:

$$\phi_+^{ball}(x) := |x|^{p-2} \int_{\mathcal{B}(|x|^\lambda)} |u|^2 \nu(x, du), \quad \phi_-^{ball}(x) := |x|^{p-2} \int_{\mathcal{B}(|x|^\lambda)} |u|^2 \gamma_{x,u}^2 \nu(x, du),$$

and

$$\phi^{big}(x) := \frac{p}{2} |x|^{p-1} \int_{\mathcal{A}(|x|^\lambda, |x|)} |u| \nu(x, du), \quad \phi^{large}(x) := \int_{\mathcal{B}^c(|x|)} |u|^p \nu(x, du).$$

Therefore, we arrive at the following result.

Lemma 1. For V from (8) and $\zeta \in [-2, 2]$:

$$\mathcal{L}V(x) \leq -p\zeta \phi^{drift}(x) + \frac{p}{4} (1 + o(1)) (\phi_+^{ball}(x) + (p-2) \phi_-^{ball}(x)) + \phi^{big}(x) + \phi^{large}(x), \quad |x| \rightarrow \infty. \quad (11)$$

Inequality (11) gives us a first vision on how the final condition may look. Indeed, consider the right-hand side of (11). The last two terms (which relate to “tail”) are positive by construction. Consequently, the negativity of $\mathcal{L}V$ can come from either the “drift” or “ball” components, or a combination thereof. According to this, we introduce several constants that illustrate the interconnection of the corresponding “growth” functions.

First, it can be seen from (11) that ϕ_+^{ball} and ϕ_-^{ball} correspond to the positive and negative parts of “ball” respectively thus, the sign of $\mathcal{L}^{ball}V$ is fully determined by their relation. Define

$$C^{ball} := \limsup_{|x| \rightarrow \infty} \frac{\phi_+^{ball}(x)}{\phi_-^{ball}(x)}.$$

By definition $C^{ball} \geq 1$ and it has to be less than $2 - p$ for $\mathcal{L}^{ball}V$ to be negative. In other words, it means that the measure $\nu(x, \cdot)$ has to be concentrated along the direction, collinear to x .

To compare “drift” with “tail” we introduce

$$C^{big} := \limsup_{|x| \rightarrow \infty} \frac{\phi^{big}(x)}{\phi^{drift}(x)} \quad \text{and} \quad C^{large} := \limsup_{|x| \rightarrow \infty} \frac{\phi^{large}(x)}{\phi^{drift}(x)}.$$

In the same vein, the constant

$$C^{tail} := \limsup_{|x| \rightarrow \infty} \frac{\max(\phi^{big}(x), \phi^{large}(x))}{\phi_-^{ball}(x)}. \quad (12)$$

shows the relation between “ball” and “tail”.

In both cases, finite constants ensure that “tail” does not spoil the negativity of $\mathcal{L}V$, and under additional assumptions, one can get either “drift”- or “ball”-induced ergodicity (respectively).

Remark 1. In (12) we do not consider the constant, related to ϕ_+^{ball} , because in the case of “ball”-induced ergodicity, the strictly negative component of “ball” (that is ϕ_-^{ball}) has to dominate.

Finally, to compare “drift” and “ball”, we define

$$\tilde{C}_+^{drift} := \limsup_{|x| \rightarrow \infty} \frac{\phi_+^{ball}(x)}{\phi^{drift}(x)} \quad \text{and} \quad \tilde{C}_-^{drift} := \liminf_{|x| \rightarrow \infty} \frac{\phi_-^{ball}(x)}{\phi^{drift}(x)}.$$

Note that the functions ϕ^{drift} and ϕ_-^{ball} are not necessarily monotone; thus, in order to formulate our main result in a more “user-friendly” manner, we assume that they can be bounded from below by some power-type functions. Namely, suppose that for $|x|$ big enough there exist some strictly positive K^{drift} , K^{ball} , and γ^{drift} , γ^{ball} , such that the following inequalities hold:

$$\phi^{drift}(x) \geq K^{drift}|x|^{\gamma^{drift}}, \quad \phi_-^{ball}(x) \geq K^{ball}|x|^{\gamma^{ball}}. \quad (13)$$

Now we are ready to formulate the main results of the paper.

Theorem 1. Case 1. (“Drift-induced” ergodicity). Suppose that (9) holds with $\zeta \in (0, 2]$, and $\tilde{C}_+^{drift}$, $\tilde{C}_-^{drift}$, C^{big} , C^{large} are finite. Then (5) holds with $f(u) = Ku^{\gamma^{drift}/p}$, $K > 0$, if

$$-p\zeta + \frac{p}{4} \left(\tilde{C}_+^{drift} + (p-2)\tilde{C}_-^{drift} \right) + C^{big} + C^{large} < 0. \quad (14)$$

Case 2. (“Ball-induced” ergodicity). Suppose that C^{ball} , $C^{tail} < \infty$. Then (5) holds with $f(u) = Ku^{\gamma^{ball}/p}$, $K > 0$, if either

a) $\tilde{C}_-^{drift} = \infty$, $\zeta \in [-2, 2]$ and

$$\frac{p}{4}(C^{ball} + p - 2) + 2C^{tail} < 0. \quad (15)$$

or

b) $\tilde{C}_-^{drift} \in (0, \infty)$, $\zeta \in [-2, 0]$ and

$$-\frac{p\zeta}{\tilde{C}_-^{drift}} + \frac{p}{4}(C^{ball} + p - 2) + 2C^{tail} < 0. \quad (16)$$

Remark 2. In Case 2 we have purely “ball”-induced ergodicity: in Case 2a) “drift” does not have any impact compared to “ball” and in Case 2b) “drift” is positive and thus works even against. In Case 1 ergodicity comes from “drift” only, if $\tilde{C}_-^{drift} = 0$; otherwise both components contribute.

2. Proofs

Proof of Lemma 1. We already have the estimate of $\mathcal{L}^{drift}V$ (see (9)). Consider now $\mathcal{L}^{ball}V$. Since $|u| \in \mathcal{B}(|x|^\lambda)$, then for $|x| \gg 1$:

$$|x + u| \geq |x| \cdot |1 - |x|^{\lambda-1}| > 1,$$

and $V(x+u) = 1 + |x+u|^p$. Thus, we can apply Taylor's formula to the integrand function of the first term of (10) and get

$$V(x+u) - V(x) - \nabla V(x) \cdot u = \frac{1}{2} \int_0^1 (1-\theta) u' D^2 V(x+\theta u) u d\theta,$$

where

$$D^2 V(x) = \{\partial_{ij}^2 V(x)\}_{i,j=1}^d = p|x|^{p-2} \begin{pmatrix} 1 + (p-2) \frac{x_1^2}{|x|^2} & (p-2) \frac{x_1 x_2}{|x|^2} & \cdots & (p-2) \frac{x_1 x_d}{|x|^2} \\ (p-2) \frac{x_1 x_2}{|x|^2} & 1 + (p-2) \frac{x_2^2}{|x|^2} & \cdots & (p-2) \frac{x_2 x_d}{|x|^2} \\ \cdots & \cdots & \ddots & \cdots \\ (p-2) \frac{x_1 x_d}{|x|^2} & (p-2) \frac{x_2 x_d}{|x|^2} & \cdots & 1 + (p-2) \frac{x_d^2}{|x|^2} \end{pmatrix}.$$

Then

$$\begin{aligned} \mathcal{L}^{ball} V(x) &= \frac{1}{2} \int_{\mathbb{B}(|x|^\lambda)} \int_0^1 (1-\theta) u' \cdot D^2 V(x+\theta u) \cdot u \nu(x, du) d\theta = \\ &= \frac{p}{2} \int_{\mathbb{B}(|x|^\lambda)} \int_0^1 (1-\theta) |x+\theta u|^{p-2} \left(\sum_{i=1}^d \left(1 + (p-2) \frac{(x_i + \theta u_i)^2}{|x+\theta u|^2} \right) u_i^2 + \right. \\ &\quad \left. + (p-2) \sum_{i,j=1, i \neq j}^d u_i u_j \frac{(x_i + \theta u_i)(x_j + \theta u_j)}{|x+\theta u|^2} \right) d\theta \nu(x, du). \end{aligned}$$

Regroup the sums in the inner integral and apply the sum of squares formula to $u_i(x_i + \theta u_i)$:

$$\begin{aligned} \mathcal{L}^{ball} V(x) &= \frac{p}{2} \int_{\mathbb{B}(|x|^\lambda)} \int_0^1 (1-\theta) |x+\theta u|^{p-2} \left(\sum_{i=1}^d u_i^2 + \frac{p-2}{|x+\theta u|^2} \left(\sum_{i=1}^d u_i(x_i + \theta u_i) \right)^2 \right) d\theta \nu(x, du) = \\ &= \frac{p}{2} \int_{\mathbb{B}(|x|^\lambda)} \int_0^1 (1-\theta) |x+\theta u|^{p-2} |u|^2 \left(1 + \underbrace{(p-2) (e_u \cdot e_{x+\theta u})^2}_{\text{negative}} \right) d\theta \nu(x, du). \end{aligned}$$

Consider the squared scalar product

$$\begin{aligned} (e_u \cdot e_{x+\theta u})^2 &= \left(\gamma_{x,u} \frac{|x|}{|x+\theta u|} + \frac{\theta |u|}{|x+\theta u|} \right)^2 \geq \gamma_{x,u}^2 \left(\frac{|x|}{|x|+|u|} \right)^2 - \frac{2\theta |x| \cdot |u|}{(|x|+|u|)^2} \geq \\ &\geq \frac{\gamma_{x,u}^2}{(1+|x|^{\lambda-1})^2} - \frac{2|x|^\lambda}{|x|(1-|x|^{\lambda-1})^2}. \end{aligned}$$

Estimating the inner integral of $\mathcal{L}^{ball} V$ depending upon its sign (again by triangle inequalities), we get

$$\frac{1}{2} (|x| \cdot |1 + |x|^{\lambda-1}|)^{p-2} \leq \int_0^1 (1-\theta) |x+\theta u|^{p-2} d\theta \leq \frac{1}{2} (|x| \cdot |1 - |x|^{\lambda-1}|)^{p-2}.$$

Summarising, we derive the final estimate of the "ball" part:

$$\begin{aligned} \mathcal{L}^{ball} V(x) &\leq \frac{p}{4} |x|^{p-2} \int_{\mathbb{B}(|x|^\lambda)} |u|^2 (C_1(x) + (p-2) C_2(x) \gamma_{x,u}^2) \nu(x, du) = \\ &= \frac{p}{4} C_1(x) \phi_+^{ball}(x) + \frac{p(p-2)}{4} C_2(x) \phi_-^{ball}(x), \end{aligned}$$

where

$$C_1(x) = \frac{1}{(1-|x|^{\lambda-1})^{2-p}} + \frac{2(2-p)|x|^{\lambda-1}}{(1+|x|^{\lambda-1})^{2-p}(1-|x|^{\lambda-1})^2}, \quad C_2(x) = \frac{1}{(1+|x|^{\lambda-1})^{4-p}}.$$

Both $C_1(x)$ and $C_2(x)$ tend to 1 as $|x|$ tends to infinity; therefore for $|x|$ large enough:

$$\mathcal{L}^{ball} V(x) \leq \frac{p}{4} (1+o(1)) (\phi_+^{ball}(x) + (p-2) \phi_-^{ball}(x)). \quad (17)$$

Next, consider the "tail" part, beginning with the big jumps. Since by definition $V(x)$ is either $1 + |x|^p$ or $1 + \phi^p(x)$

(depending on the value of $|x|$), and $\phi(x) \leq |x|$, then $V(x+u) \leq 1 + |x+u|^p$, and

$$\begin{aligned} \mathcal{L}^{big} V(x) &\leq \int_{\mathcal{A}(|x|^\lambda, |x|)} (|x+u|^p - |x|^p - p|x|^{p-1} \cdot |u|\gamma_{x,u}) \nu(x, du) = \\ &= |x|^p \int_{\mathcal{A}(|x|^\lambda, |x|)} \left(\left(1 + 2\frac{|u|}{|x|} \gamma_{x,u} + \left(\frac{|u|}{|x|} \right)^2 \right)^{\frac{p}{2}} - 1 \right) \nu(x, du) - \\ &- p|x|^{p-1} \int_{\mathcal{A}(|x|^\lambda, |x|)} |u|\gamma_{x,u} \nu(x, du) =: J_1(x) + J_2(x). \end{aligned}$$

Now using the fact that $u \in \mathcal{A}(|x|^\lambda, |x|)$, we derive

$$\begin{aligned} J_1(x) &\leq |x|^p \int_{\mathcal{A}(|x|^\lambda, |x|)} \left(\left(1 + (1 + 2\gamma_{x,u}) \frac{|u|}{|x|} \right)^{\frac{p}{2}} - 1 \right) \nu(x, du) \leq \\ &\leq \frac{p}{2} |x|^{p-1} \int_{\mathcal{A}(|x|^\lambda, |x|)} (1 + 2\gamma_{x,u}) |u| \nu(x, du). \end{aligned}$$

Summing $J_1(x)$ and $J_2(x)$, we obtain the final estimate of the big jumps part:

$$\mathcal{L}^{big} V(x) \leq \frac{p}{2} |x|^{p-1} \int_{\mathcal{A}(|x|^\lambda, |x|)} |u| \nu(x, du) = \phi^{big}(x). \quad (18)$$

In the case of large jumps we again use that by definition $V(x+u) \leq 1 + |x+u|^p$. Then

$$\mathcal{L}^{large} V(x) \leq \int_{\mathcal{B}^c(|x|)} (|x+u|^p - |x|^p) \nu(x, du) \leq \int_{\mathcal{B}^c(|x|)} |u|^p \nu(x, du) = \phi^{large}(x). \quad (19)$$

Together estimates (9), (17), (18) and (19) yield the statement of the lemma, namely,

$$\mathcal{L} V(x) \leq -p\zeta \phi^{drift}(x) + \frac{p}{4} (1 + o(1)) (\phi_+^{ball}(x) + (p-2)\phi_-^{ball}(x)) + \phi^{big}(x) + \phi^{large}(x), \quad |x| \rightarrow \infty.$$

□

Now, using the results of the proven lemma, we can easily prove Theorem 1.

Proof of Theorem 1. Case 1. "Drift-induced" ergodicity. Suppose that (9) holds with $\zeta \in (0, 2]$. Dividing both sides of (11) by ϕ^{drift} and taking $\limsup$ as $|x| \rightarrow \infty$, we get

$$\limsup_{|x| \rightarrow \infty} \frac{\mathcal{L} V(x)}{\phi^{drift}(x)} \leq -p\zeta + \frac{p}{4} \left(\tilde{C}_+^{drift} + (p-2)\tilde{C}_-^{drift} \right) + C^{big} + C^{large}. \quad (20)$$

Then it is clear that the Lyapunov-type condition (5) is satisfied if $\tilde{C}_+^{drift}$, $\tilde{C}_-^{drift}$, C^{big} , C^{large} are finite, and moreover the right-hand side of (20) is strictly negative.

Case 2. "Ball-induced" ergodicity.

a) Assume that $\tilde{C}_-^{drift} = \infty$ (that is, "ball" is way faster than "drift") and $\zeta \in [-2, 2]$. Dividing this time both sides of (11) by ϕ_-^{ball} and taking $\limsup$ as $|x| \rightarrow \infty$, we get

$$\limsup_{|x| \rightarrow \infty} \frac{\mathcal{L} V(x)}{\phi_-^{ball}(x)} \leq \frac{p}{4} (C^{ball} + p-2) + 2C^{tail}. \quad (21)$$

If $C^{ball}, C^{tail} < \infty$, and the right-hand side of (21) is strictly negative, condition (5) is satisfied.

b) Assume $\tilde{C}_-^{drift} \in (0, \infty)$ and $\zeta \in [-2, 0]$. Following the same approach as above we get

$$\limsup_{|x| \rightarrow \infty} \frac{\mathcal{L} V(x)}{\phi_-^{ball}(x)} \leq -\frac{p\zeta}{\tilde{C}_-^{drift}} + \frac{p}{4} (C^{ball} + p-2) + 2C^{tail}. \quad (22)$$

If $C^{ball}, C^{tail} < \infty$, and the right-hand side of (22) is strictly negative, condition (5) holds.

In each case, the form of f from (5) (and thus of the ergodic rates ψ from (7)) is determined by (13).

□

3. Examples

Example. Consider a one-dimensional symmetric bounded $\nu(\cdot, \cdot)$ and suppose that the drift part $\ell(x)$ is absent. Then ϕ^{drift} , ϕ^{big} , and ϕ^{large} are all equal to zero for all $x \in \mathbb{R}$, and we are in the setting of Case 2a) with $C^{tail} = 0$. Moreover, given that we are in the one-dimensional setting, $|\gamma_{x,u}| = 1$, thus, $\phi_+^{ball} = \phi_-^{ball}$ and $C^{ball} = 1$. Then (15) is always satisfied, yielding the Lyapunov-type condition (5).

Example. Consider now an α -stable Lévy-type measure $\nu(x, du) = |x|^\beta |u|^{-d-\alpha} du$, $x \in \mathbb{R}^d$, with $\beta \in (1-p, 1)$ and $\alpha \in (0, 2)$. Take $\ell(x) = |x|^\eta$, $\eta \in (0, \beta)$, and assume that $\zeta \in [1/2, 1]$.

First, estimate

$$\phi_-^{ball}(x) \leq \phi_+^{ball}(x) = \frac{S(d-1)}{2-\alpha} |x|^{p-2+\beta+\lambda(2-\alpha)},$$

where $S(d-1)$ is the surface area of the unit sphere of dimension $d-1$.

Next, to calculate the rest of the “growth” functions we define

$$\phi_\tau^{ann}(x) := |x|^{p-1+\beta} \int_{\mathcal{A}(|x|^\tau, |x|)} |u|^{1-d-\alpha} du, \quad \tau \in \{0, \lambda\}.$$

Simple calculations lead us to

$$\phi_\tau^{ann}(x) \asymp S(d-1) \begin{cases} (\alpha-1)^{-1} |x|^{p-1+\beta+(1-\alpha)\tau}, & \alpha > 1, \\ (1-\alpha)^{-1} |x|^{p+\beta-\alpha}, & \alpha < 1, \\ (1-\tau) |x|^{p-1+\beta} \ln |x|, & \alpha = 1. \end{cases}$$

Then (remember that $\eta \in (0, \beta)$)

$$\phi^{drift}(x) = \phi_0^{ann}(x), \quad \phi^{big}(x) = \frac{p}{2} \phi_\lambda^{ann}(x).$$

Finally,

$$\phi^{large}(x) = \frac{S(d-1)}{\alpha-p} |x|^{p+\beta-\alpha}.$$

We can easily see that $\tilde{C}_+^{drift}$ (hence and $\tilde{C}_-^{drift}$) are equal to zero for every $\alpha \in (0, 2)$. Furthermore, given that $\zeta \in [1/2, 1]$, inequality (14) is automatically satisfied for $\alpha \geq 1$, thus, Lyapunov-type condition (5) holds with $f(u) = Cu^{\frac{p-1+\beta}{p}}$, $C \leq p\zeta$, for $\alpha > 1$, and with $f(u) = Cu^{\frac{p-1+\beta}{p}} \ln u$, $C \leq \frac{p}{2}(2\zeta-1+\lambda)$, for $\alpha = 1$. For $\alpha < 1$, (5) holds with $f(u) = Cu^{\frac{p+\beta-\alpha}{p}}$, $C \leq p(\zeta - \frac{1}{2}) - \frac{1-\alpha}{\alpha-p}$, if p satisfies the inequality

$$p^2(1-2\zeta) - p\alpha(1-2\zeta) - 2(1-\alpha) > 0.$$

In each case, C was derived directly from (14).

Discussion and conclusions

In this paper, we investigated the long-time behaviour of the process X , related to some integro-differential operator with unbounded coefficients. Using the Lyapunov-type criterion, we found the conditions on the Lévy-type measure of the process X ensuring its ergodicity in total variation norm; in particular, we showed that in the case of noise-induced ergodicity the Lévy-type measure must be concentrated along the direction, collinear to the spatial coordinate x . The conditions were given in terms of the coefficients of the process’s generator, making the results easily applicable in practice.

Acknowledgments. The author expresses deepest gratitude to her supervisor, Dr. Victoria Knopova, for her mentorship and constant encouragement. The author sincerely thanks Prof. Dr. René Schilling for the hospitality and valuable, insightful comments and hints. The author is also grateful to the anonymous reviewers for their comments and constructive suggestions, which helped improve the quality and clarity of this manuscript.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Bao, J., Fang, R., & Wang, J. (2024). Exponential ergodicity of Lévy driven Langevin dynamics with singular potentials. *Stochastic Processes and their Applications*, 172, 104341. <https://doi.org/10.1016/j.spa.2024.104341>
- Bao, J., & Wang, J. (2023). Exponential ergodicity for damping Hamiltonian dynamics with state-dependent and non-local collisions. *Bernoulli*, 29(3), 2442–2465. <https://doi.org/10.3150/22-BEJ1548>
- Böttcher, B., Schilling, R., & Wang, J. (2013). *Lévy matters. III*. Springer, Cham. <https://doi.org/10.1007/978-3-319-02684-8>
- Douc, R., Fort, G., & Guillin, A. (2009). Subgeometric rates of convergence of f -ergodic strong Markov processes. *Stochastic Processes and their Applications*, 119(3), 897–923. <https://doi.org/10.1016/j.spa.2008.03.007>
- Ethier, S. N., & Kurtz, T. G. (1986). *Markov processes. Characterization and convergence*. John Wiley & Sons. <https://doi.org/10.1002/9780470316658>
- Fort, G., & Roberts, G. O. (2005). Subgeometric ergodicity of strong Markov processes. *The Annals of Applied Probability*, 15(2), 1565–1589. <https://doi.org/10.1214/105051605000000115>
- Knopova, V., & Mokanu, Y. (2024). On ergodic properties of some Lévy-type processes. *Journal of Theoretical Probability*, 37(1), 582–602. <https://doi.org/10.1007/s10959-023-01252-x>
- Knopova, V., & Mokanu, Y. (2025). On ergodic property of the solution to a Lévy-driven SDE. *Cybernetics and Systems Analysis*, 61(5), 98–109. <https://doi.org/10.1007/s10559-025-00830-7>

- Kühn, F. (2020). Existence of (Markovian) solutions to martingale problems associated with Lévy-type operators. *Electronic Journal of Probability*, 25, 1–26. <https://doi.org/10.1214/20-ejp424>
- Kulik, A. (2018). *Ergodic behavior of Markov processes: With applications to limit theorems*. De Gruyter. <https://doi.org/10.1515/9783110458930>
- Kulik, A., & Scheutzow, M. (2018). Generalized couplings and convergence of transition probabilities. *Probability Theory and Related Fields*, 171, 333–376. <https://doi.org/10.1007/s00440-017-0779-8>
- Liang, M., Majka, M., & Wang, J. (2021). Exponential ergodicity for SDEs and McKean–Vlasov processes with Lévy noise. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(3), 1665–1701. <https://doi.org/10.1214/20-AIHP1123>
- Masuda, H. (2004). On multidimensional Ornstein–Uhlenbeck processes driven by a general Lévy process. *Bernoulli*, 10(1), 97–120. <https://doi.org/10.3150/bj/1077544605>
- Masuda, H. (2007). Ergodicity and exponential β -mixing bounds for multidimensional diffusions with jumps. *Stochastic Processes and their Applications*, 117(1), 35–56. <https://doi.org/10.1016/j.spa.2006.04.010>
- Sandrić, N. (2013). Long-time behavior of stable-like processes. *Stochastic Processes and their Applications*, 123(4), 1276–1300. <https://doi.org/10.1016/j.spa.2012.12.004>
- Sandrić, N. (2014). Ergodic property of stable-like Markov chains. *Journal of Theoretical Probability*, 29(2), 459–490. <https://doi.org/10.1007/s10959-014-0586-4>
- Sandrić, N. (2016). Ergodicity of Lévy-type processes. *ESAIM. Probability and Statistics*, 20, 154–177. <https://doi.org/10.1051/ps/2016009>
- Skorokhod, A. (1987). *Asymptotic methods of the theory of stochastic differential equations*. Naukova Dumka [in Russian]. [Скороход, А. (1987). *Асимптотические методы теории стохастических дифференциальных уравнений*. Наукова думка].
- Wang, J. (2008). Criteria for ergodicity of Lévy type operators in dimension one. *Stochastic Processes and their Applications*, 118(10), 1909–1928. <https://doi.org/10.1016/j.spa.2007.11.003>
- Wang, J. (2013). Exponential ergodicity and strong ergodicity for SDEs driven by symmetric α -stable processes. *Applied Mathematics Letters*, 26, 654–658. <https://doi.org/10.1016/j.aml.2013.01.004>
- Wee, I.-S. (1999). Stability for multidimensional jump-diffusion processes. *Stochastic Processes and their Applications*, 80(2), 193–209. [https://doi.org/10.1016/S0304-4149\(98\)00078-7](https://doi.org/10.1016/S0304-4149(98)00078-7)

Отримано редакцією журналу / Received: 13.09.25

Прорецензовано / Revised: 12.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Яна МОКАНУ, асп., асист.

ORCID ID: 0009-0006-0752-6584

e-mail: yana.mokanu3075@gmail.com

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна

ЕРГОДИЧНІСТЬ ДЕЯКИХ ПРОЦЕСІВ ТИПУ ЛЕВІ В $\mathbb{R}^d$

Досліджено ергодичність у повній варіації процесу X , пов'язаного з деяким інтегро-диференціальним оператором з необмеженими коефіцієнтами, та описано швидкість збіжності до відповідної інваріантної міри. Іншими словами, процес Маркова є ергодичним, якщо існує інваріантна ймовірнісна міра, до якої збігається ядро ймовірності переходу відповідного процесу. У такому разі швидкість збіжності називається коефіцієнтом ергодичності; вона відіграє важливу роль у подальшому вивченні асимптотичних властивостей, зокрема в моделюванні процесів, обчисленні їхніх функціоналів, а також у граничних теоремах.

Для доведення застосовано критерій типу Ляпунова, що нагадує відомий з теорії звичайних диференціальних рівнянь другий метод Ляпунова. Теоретичні результати проілюстровано прикладами.

Ключові слова: ергодичність, процес типу Леві, критерій Ляпунова, функція Ляпунова.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 517.52

DOI: <https://doi.org/10.17721/1812-5409.2026/1.9>

Mykola MOROZ, PhD, Junior Researcher

ORCID ID: 0000-0001-6658-4924

e-mail: moroznik22@gmail.com

Institute of Mathematics of NAS of Ukraine, Kyiv, Ukraine

THE CONVERGENCE OF SEQUENCES IN TERMS OF POSITIVE AND ALTERNATING PERRON EXPANSIONS

In this paper we study conditions for the convergence of sequences in terms of positive and alternating Perron expansions of real numbers (P -representation and P^- -representation). These conditions make it possible to verify the continuity of functions defined via the P -representation or P^- -representation.

The main results of this work are as follows. First, we systematically consider all possible cases and, by taking into account the geometric structure of Perron expansions, significantly reduce the number of distinct cases that require separate analysis. Second, for each case, we formulate and rigorously prove a convergence condition that is both necessary and sufficient. Third, the results are obtained not for specific expansions (such as the Engel or Pierce expansions), but for the Perron expansions, which include many classical representations of real numbers as special cases.

These results provide a foundation for further analysis of fractal functions defined through Perron expansions or their special cases (the Lüroth, Engel, Pierce, and Sylvester expansions).

Keywords: convergence sequences, Perron expansions, Lüroth expansions, Engel expansions, Pierce expansions.

AMS 2020 classification: 40A05, 11A67, 11K55.

Introduction

Many expansions of real numbers serve as effective tools for defining and analyzing functions that exhibit locally complex behavior and fractal properties, including nowhere monotonic, nowhere differentiable, and singular functions. For instance, various number expansions can be employed to define singular functions such as the Minkowski question-mark function, the Salem function, and the Cantor function, as well as continuous functions that are nowhere monotonic and nowhere differentiable, such as the Tribin function (Prats'ovtyi, Baranovs'kyi, & Maslova, 2021).

In recent years, the Engel and Pierce expansions have been increasingly utilized for such purposes, enabling the modeling of fractal functions with structures more intricate than self-similar ones (see, e.g., (Albeverio et al., 2013; Baranovskiy, & Pratsiovytyi, 2023; Moroz, 2017, 2024)). Verifying the continuity of such functions is often essential, yet it is not always a trivial task. However, a general approach to verifying the continuity of such functions was not proposed in these works. In this paper, we aim to lay the foundations of such a general approach. Continuity is understood here in the Heine sense – namely, in terms of the convergence of sequences. Accordingly, our primary focus is on establishing conditions for the convergence of sequences whose elements are given in terms of such expansions.

We formulate and prove all propositions for the positive and alternating Perron expansions of real numbers (P -representations and P^- -representations) (Moroz, 2024, 2025). These expansions generalize a number of classical representations, including the positive and alternating Lüroth expansions (Zhykharyeva, & Pratsiovytyi, 2012; Kalpazidou, Knopfmacher, & Knopfmacher, 1990), Engel expansions (Erdős, Rényi, & Szusz, 1958), Sylvester expansions (Erdős, Rényi, & Szusz, 1958), DKB-expansions (Galambos, 1998), Pierce expansions (Shallit, 1986), and second Ostrogradsky expansions (Torbin, & Pratsyovtya, 2010). They are also closely related to Oppenheim expansions (Galambos, 1976; Oppenheim, 1972; Sydoruk, & Torbin, 2017).

1. General information about Perron expansions and main notations

Consider functions $\varphi_n(x_1, \dots, x_n): \mathbb{N}^n \rightarrow \mathbb{N}$ for all $n \in \mathbb{N}$, with $\varphi_0 \equiv \text{const} \in \mathbb{N}$. A fixed sequence of such functions $(\varphi_n)_{n=0}^\infty$ is denoted by P . Each sequence P defines two types of Perron expansions of real numbers: the P -representation (positive Perron expansion) and the P^- -representation (alternating Perron expansion).

The P -representation of a number $x \in (0, 1]$ is defined as its expansion into a positive Perron series:

$$x = \sum_{n=0}^{\infty} \frac{r_0 \cdots r_n}{(p_1 - 1)p_1 \cdots (p_n - 1)p_n p_{n+1}}, \quad (1)$$

where $r_0 = \varphi_0$, $r_n = \varphi_n(p_1, \dots, p_n)$, and $p_n \geq r_{n-1} + 1$ for all $n \in \mathbb{N}$. The number x and its P -representation (1) are denoted briefly by $\Delta_{p_1 p_2 \dots}^P$. For each sequence P of functions φ_n , every $x \in (0, 1]$ has a unique P -representation (Moroz, 2024, Theorem 1). The number $p_n = p_n(x)$ is called the n th P -digit of x .

The P -cylinder of rank k with base $c_1 \dots c_k$ is defined as the set $\Delta_{c_1 \dots c_k}^P$ consisting of all numbers $x \in (0, 1]$ such that $p_1(x) = c_1, \dots, p_k(x) = c_k$. It is known (Moroz, 2024, Lemma 3, Corollary 3) that the P -cylinder $\Delta_{c_1 \dots c_k}^P$ is an interval of the form $(a, b]$. The infimum, supremum, and diameter of the P -cylinder $\Delta_{c_1 \dots c_k}^P$ are given by the following formulas:

$$\inf \Delta_{c_1 \dots c_k}^P = \sum_{n=0}^{k-1} \frac{r_0 \cdots r_n}{(c_1 - 1)c_1 \cdots (c_n - 1)c_n c_{n+1}}, \quad (2)$$

$$\sup \Delta_{c_1 \dots c_k}^P = \inf \Delta_{c_1 \dots c_k}^P + \frac{r_0 \cdots r_{k-1}}{(c_1 - 1)c_1 \cdots (c_k - 1)c_k}, \quad (3)$$

$$|\Delta_{c_1 \dots c_k}^P| = \frac{r_0 \dots r_{k-1}}{(c_1 - 1)c_1 \dots (c_k - 1)c_k}, \quad (4)$$

where $r_0 = \varphi_0$ and $r_n = \varphi_n(c_1, \dots, c_n)$ for all $n = 1, \dots, k-1$.

The geometry of the P -representation was studied in greater detail in our previous work (Moroz, 2024). Figures 1 and 2 illustrate the relative positions of P -cylinders and provide insight into the geometric structure of the P -representation.

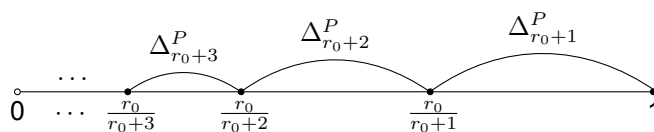


Fig. 1. P -cylinders of rank 1 inside $(0, 1]$, $r_0 = \varphi_0$

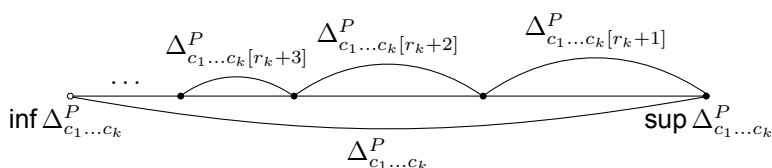


Fig. 2. P -cylinders of rank $k+1$ inside the P -cylinder $\Delta_{c_1 \dots c_k}^P$ of rank k , $r_k = \varphi_k(c_1, \dots, c_k)$

The P^- -representation of a number $x \in (0, 1]$ is defined as its expansion into an alternating Perron series:

$$x = \sum_{n=0}^{\infty} \frac{(-1)^n r_0 \dots r_n}{(q_1 - 1)q_1 \dots (q_n - 1)q_n(q_{n+1} - 1)}, \quad (5)$$

where $r_0 = \varphi_0$, $r_n = \varphi_n(q_1, \dots, q_n)$, and $q_n \geq r_{n-1} + 1$ for all $n \in \mathbb{N}$. The number x and its P^- -representation (5) are denoted briefly by $\Delta_{q_1 q_2 \dots}^{P^-}$. For each sequence P of functions φ_n , there exists a countably infinite set $IS^{P^-} \subset (0, 1]$ whose elements do not have its P^- -representation. However, every $x \in (0, 1) \setminus IS^{P^-}$ has a unique P^- -representation (Moroz, 2025, Theorem 3.12). For $x \in (0, 1) \setminus IS^{P^-}$, the number $q_n = q_n(x)$ is called the n th P^- -digit of x .

The P^- -cylinder of rank k with base $c_1 \dots c_k$ is defined as the set $\Delta_{c_1 \dots c_k}^{P^-}$ of all numbers $x \in (0, 1) \setminus IS^{P^-}$ such that $q_1(x) = c_1, \dots, q_k(x) = c_k$. It is known (Moroz, 2025, Corollary 3.14) that the P^- -cylinder $\Delta_{c_1 \dots c_k}^{P^-}$ is a set of the type $(a, b) \setminus IS^{P^-}$. The infimum, supremum, and diameter of P^- -cylinder $\Delta_{c_1 \dots c_k}^{P^-}$ are given by the following formulas:

- if k is odd, then

$$\sup \Delta_{c_1 \dots c_k}^{P^-} = \sum_{n=0}^{k-1} \frac{(-1)^n r_0 \dots r_n}{(c_1 - 1)c_1 \dots (c_n - 1)c_n(c_{n+1} - 1)}, \quad (6)$$

$$\inf \Delta_{c_1 \dots c_k}^{P^-} = \sup \Delta_{c_1 \dots c_k}^{P^-} - \frac{r_0 \dots r_{k-1}}{(c_1 - 1)c_1 \dots (c_k - 1)c_k}, \quad (7)$$

- if k is even, then

$$\inf \Delta_{c_1 \dots c_k}^{P^-} = \sum_{n=0}^{k-1} \frac{(-1)^n r_0 \dots r_n}{(c_1 - 1)c_1 \dots (c_n - 1)c_n(c_{n+1} - 1)}, \quad (8)$$

$$\sup \Delta_{c_1 \dots c_k}^{P^-} = \inf \Delta_{c_1 \dots c_k}^{P^-} + \frac{r_0 \dots r_{k-1}}{(c_1 - 1)c_1 \dots (c_k - 1)c_k}, \quad (9)$$

- in both cases

$$|\Delta_{c_1 \dots c_k}^{P^-}| = \frac{r_0 \dots r_{k-1}}{(c_1 - 1)c_1 \dots (c_k - 1)c_k}, \quad (10)$$

where $r_0 = \varphi_0$ and $r_n = \varphi_n(c_1, \dots, c_n)$ for all $n = 1, \dots, k-1$. It is important that IS^{P^-} is the set of infima and suprema of all P^- -cylinders.

The geometry of the P^- -representation was studied in greater detail in our previous work (Moroz, 2025). Figures 1, 2, and 3 in (Moroz, 2025) illustrate the relative positions of P^- -cylinders and provide insight into the geometric structure of the P^- -representation. Note that the geometry of the P^- -representation is similar to that of continued fractions.

2. The convergence of sequences in terms of P -representation

We now prove several propositions that establish a connection between the convergence of a numerical sequence and the P -digits of its elements.

Let $x_0 = \Delta_{c_1 c_2 \dots}^P \in (0, 1]$, and let $(x_n)_{n=1}^\infty$ be a sequence of numbers from $(0, 1] \setminus \{x_0\}$. For each $n \in \mathbb{N}$, let k_n denote the smallest index such that $p_{k_n}(x_n) \neq p_{k_n}(x_0)$.

Proposition 1. *If $\lim_{n \rightarrow \infty} k_n = +\infty$, then $\lim_{n \rightarrow \infty} x_n = x_0$.*

Proof. Assume, without loss of generality, that $k_n \geq 2$. Then x_0 and x_n are contained in the same P -cylinder $\Delta_{c_1 \dots c_{k_n-1}}^P$ of rank $k_n - 1$, and we have

$$0 < |x_0 - x_n| < \left| \Delta_{c_1 \dots c_{k_n-1}}^P \right|.$$

Since $k_n \rightarrow +\infty$, according to (Moroz, 2024, Lemma 4) we have $\left| \Delta_{c_1 \dots c_{k_n-1}}^P \right| \rightarrow 0$ as $n \rightarrow \infty$. It follows that $x_n \rightarrow x_0$. $\square$

The condition $\lim_{n \rightarrow \infty} k_n = \infty$ is sufficient for the convergence of the sequence $(x_n)_{n=1}^\infty$ to x_0 . However, it is generally not necessary. There are several fundamental cases, each characterized by its own necessary and sufficient conditions for convergence.

Proposition 2. *If x_0 is not the endpoint of some P -cylinder, then*

$$\lim_{n \rightarrow \infty} x_n = x_0 \iff \lim_{n \rightarrow \infty} k_n = +\infty.$$

Proof. Given the Proposition 1, it suffices to prove that

$$\lim_{n \rightarrow \infty} x_n = x_0 \implies \lim_{n \rightarrow \infty} k_n = +\infty.$$

Suppose to the contrary that $\lim_{n \rightarrow \infty} x_n = x_0$ and $\lim_{n \rightarrow \infty} k_n \neq +\infty$. Then there exists a constant M and an infinite sequence $(n_t)_{t=1}^\infty$ such that $k_{n_t} < M$. Since x_0 is neither the infimum nor the supremum of any P -cylinder, we have

$$\inf \Delta_{c_1 \dots c_M}^P < x_0 = \Delta_{c_1 c_2 \dots}^P < \sup \Delta_{c_1 \dots c_M}^P.$$

However, it follows that $x_{n_t} \notin \Delta_{c_1 \dots c_M}^P$, so

$$|x_0 - x_{n_t}| \geq \min \left\{ x_0 - \inf \Delta_{c_1 \dots c_M}^P, \sup \Delta_{c_1 \dots c_M}^P - x_0 \right\} = \text{const} > 0.$$

Since this inequality holds for infinitely many indices n_t , the sequence $(x_n)_{n=1}^\infty$ cannot converge to x_0 . This contradicts our assumption. Hence, $k_n \rightarrow +\infty$. $\square$

It follows from the geometry of the P -representation that the infimum of any P -cylinder coincides with the supremum of the adjacent left P -cylinder of the same rank. Moreover, the supremum of any P -cylinder, except for $\Delta_{[r_0+1]}^P$, coincides with the infimum of another P -cylinder, though not necessarily of the same rank.

The following proposition provides necessary and sufficient conditions for the sequence $(x_n)_{n=1}^\infty$ to converge to x_0 from the left, where x_0 is the supremum of a P -cylinder.

Proposition 3. *Let $x_0 \in (0, 1]$ be the supremum of a P -cylinder, and let $(x_n)_{n=1}^\infty$ be a sequence such that $0 < x_n < x_0$ for all $n \in \mathbb{N}$. Then*

$$\lim_{n \rightarrow \infty} x_n = x_0 \iff \lim_{n \rightarrow \infty} k_n = +\infty.$$

Proof. The proof is similar to that of Proposition 2. $\square$

The following proposition provides necessary and sufficient conditions for the sequence $(x_n)_{n=1}^\infty$ to converge to x_0 from the right, where x_0 is the infimum of a P -cylinder.

Proposition 4. *Let $x_0 = \inf \Delta_{c_1 \dots c_k}^P$, and let $(x_n)_{n=1}^\infty$ be a sequence such that $x_0 < x_n \leq 1$ for all $n \in \mathbb{N}$. Then $\lim_{n \rightarrow \infty} x_n = x_0$ if and only if there exists an index n_0 such that*

$$\begin{cases} p_i(x_n) = c_i \text{ for all } i \leq k \text{ and } n \geq n_0, \\ \lim_{n \rightarrow \infty} p_{k+1}(x_n) = +\infty. \end{cases} \quad (11)$$

Proof. Sufficiency. It follows from (11) that

$$x_n \in \Delta_{c_1 \dots c_k p_{k+1}(x_n)}^P \subseteq \Delta_{c_1 \dots c_k}^P$$

for all $n \geq n_0$. Thus, $\inf \Delta_{c_1 \dots c_k}^P < x_n \leq \sup \Delta_{c_1 \dots c_k p_{k+1}(x_n)}^P$. Using (Moroz, 2024, Corollary 1), which follows from (2) and (3), we obtain

$$0 < x_n - x_0 \leq \sup \Delta_{c_1 \dots c_k p_{k+1}(x_n)}^P - \inf \Delta_{c_1 \dots c_k}^P = \frac{r_0 r_1 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k (p_{k+1}(x_n) - 1)} \rightarrow 0$$

as $p_{k+1}(x_n) \rightarrow \infty$, where $r_0 = \varphi_0$ and $r_i = \varphi_i(c_1, \dots, c_i)$ for all $i \leq k$. Hence, $x_n \rightarrow x_0$.

Necessity. Assume, for the sake of contradiction, that (11) fails to hold for any n_0 . We distinguish two cases to show that in both $x_n \not\rightarrow x_0$.

Case 1: There exists a sequence $(n_t)_{t=1}^\infty$ and indices $i_t \leq k$ such that $p_{i_t}(x_{n_t}) \neq c_{i_t}$ for all $t \in \mathbb{N}$. In this case, $x_{n_t} \notin \Delta_{c_1 \dots c_k}^P$, so

$$x_0 = \inf \Delta_{c_1 \dots c_k}^P < \sup \Delta_{c_1 \dots c_k}^P < x_{n_t},$$

and therefore

$$x_{n_t} - x_0 > \left| \Delta_{c_1 \dots c_k}^P \right| = \text{const} > 0.$$

Since this inequality holds for infinitely many indices n_t , the sequence $(x_n)_{n=1}^\infty$ does not converge to x_0 .

Case 2: Suppose that $p_i(x_n) = c_i$ for all $i \leq k$ and $n \geq n_0$, but there exist a constant M and a sequence $(n_t)_{t=1}^\infty$ such that $p_{k+1}(x_{n_t}) < M$. Assuming without loss of generality that $n_t \geq n_0$, from (Moroz, 2024, Corollary 1) we have

$$\begin{aligned} x_{n_t} - x_0 &\geq \inf \Delta_{c_1 \dots c_k p_{k+1}(x_{n_t})}^P - \inf \Delta_{c_1 \dots c_k}^P = \frac{r_0 r_1 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k p_{k+1}(x_{n_t})} > \\ &> \frac{r_0 r_1 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k M} = \text{const} > 0, \end{aligned}$$

where $r_0 = \varphi_0$ and $r_i = \varphi_i(c_1, \dots, c_i)$ for all $i \leq k$. Since this inequality holds for infinitely many indices n_t , the sequence $(x_n)_{n=1}^\infty$ does not converge to x_0 .

Thus, condition (11) is both necessary and sufficient for the convergence of $(x_n)_{n=1}^\infty$ to x_0 . $\square$

Since the number 0 does not have a P -representation, we will consider the convergence of the sequence $(x_n)_{n=1}^\infty$ to 0 separately.

Proposition 5. $\lim_{n \rightarrow \infty} x_n = 0 \iff \lim_{n \rightarrow \infty} p_1(x_n) = +\infty$.

Proof. This proposition follows from the inequality $0 < \frac{r_0}{p_1(x_n)} < x_n \leq \frac{r_0}{p_1(x_n) - 1}$, where $r_0 = \varphi_0$. $\square$

3. The convergence of sequences in terms of P^- -representation

Not every number $x \in (0, 1)$ has a P^- -representation. Therefore, we will assume that all elements of the sequence $(x_n)_{n=1}^\infty$ belong to $(0, 1) \setminus IS^{P^-}$. However, the limit of such a sequence can be any number from $[0, 1]$. Consequently, several cases need to be considered.

Let $x_0 = \Delta_{c_1 c_2 \dots}^{P^-} \in (0, 1) \setminus IS^{P^-}$, and let $(x_n)_{n=1}^\infty$ be a sequence of numbers from $(0, 1) \setminus (IS^{P^-} \cup \{x_0\})$. For each $n \in \mathbb{N}$, denote by k_n the smallest index for which $q_{k_n}(x_n) \neq q_{k_n}(x_0)$.

Proposition 6. $\lim_{n \rightarrow \infty} x_n = x_0 \iff \lim_{n \rightarrow \infty} k_n = +\infty$.

Proof. The proof is similar to that of Proposition 1 and Proposition 2. $\square$

Consider the cases in which the sequence $(x_n)_{n=1}^\infty$ converges to a number that does not have a P^- -representation.

Proposition 7. $\lim_{n \rightarrow \infty} x_n = 0 \iff \lim_{n \rightarrow \infty} q_1(x_n) = +\infty$.

Proof. The proof is similar to that of Proposition 5. $\square$

Lemma 1. Every number from IS^{P^-} is the supremum of some P^- -cylinder of odd rank. In particular,

- $\inf \Delta_{c_1 \dots c_k}^{P^-} = \sup \Delta_{c_1 \dots c_{k-1} [c_k+1]}^{P^-}$ if k is odd,
- $\sup \Delta_{c_1 \dots c_k}^{P^-} = \sup \Delta_{c_1 \dots c_k [r_k+1]}^{P^-}$ if k is even, where $r_k = \varphi_k(c_1, \dots, c_k)$,
- $\inf \Delta_{c_1 \dots c_k}^{P^-} = \inf \Delta_{c_1 \dots c_{k-1}}^{P^-} = \sup \Delta_{c_1 \dots [c_{k-1}+1]}^{P^-}$ if k is even and $c_k = \varphi_{k-1}(c_1, \dots, c_{k-1}) + 1$,
- $\inf \Delta_{c_1 \dots c_k}^{P^-} = \sup \Delta_{c_1 \dots c_{k-1} [c_k-1]}^{P^-} = \sup \Delta_{c_1 \dots c_{k-1} [r_k-1] [r_k+1]}^{P^-}$ if k is even and $c_k > \varphi_{k-1}(c_1, \dots, c_{k-1}) + 1$, where $r_k = \varphi_k(c_1, \dots, c_{k-1}, c_k - 1)$.

This lemma follows directly from the geometry of the P^- -representation and from formulas (6), (7), (8), and (9).

Proposition 8. Let $x_0 = \sup \Delta_{c_1 \dots c_k}^{P^-}$, where k is odd, and let $(x_n)_{n=1}^\infty$ be a sequence such that $0 < x_n < x_0$ for all $n \in \mathbb{N}$. Then $\lim_{n \rightarrow \infty} x_n = x_0$ if and only if there exists an index n_0 such that

$$\begin{cases} q_i(x_n) = c_i \text{ for all } i \leq k \text{ and } n \geq n_0, \\ \lim_{n \rightarrow \infty} q_{k+1}(x_n) = +\infty. \end{cases} \quad (12)$$

Proof. Sufficiency. It follows from (12) that

$$x_n \in \Delta_{c_1 \dots c_k q_{k+1}(x_n)}^{P^-} \subseteq \Delta_{c_1 \dots c_k}^{P^-}$$

for all $n \geq n_0$. Then $\inf \Delta_{c_1 \dots c_k q_{k+1}(x_n)}^{P^-} < x_n < \sup \Delta_{c_1 \dots c_k}^{P^-}$. Using (Moroz, 2025, Corollary 3.8), which follows from formulas (6), (7), (8), and (9), we obtain

$$0 < x_0 - x_n < \sup \Delta_{c_1 \dots c_k}^{P^-} - \inf \Delta_{c_1 \dots c_k q_{k+1}(x_n)}^{P^-} = \frac{r_0 r_1 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k (q_{k+1}(x_n) - 1)} \rightarrow 0$$

as $q_{k+1}(x_n) \rightarrow \infty$, where $r_0 = \varphi_0$ and $r_i = \varphi_i(c_1, \dots, c_i)$ for all $i \leq k$. Hence, $x_n \rightarrow x_0$.

Necessity. Assume, for the sake contradiction, that (12) fails to hold for any n_0 . We consider two possible cases and show that in each case $x_n \not\rightarrow x_0$.

Case 1: There exists a sequence $(n_t)_{t=1}^\infty$ and indices $i_t \leq k$ such that $q_{i_t}(x_{n_t}) \neq c_{i_t}$ for all $t \in \mathbb{N}$. In this case, $x_{n_t} \notin \Delta_{c_1 \dots c_k}^{P^-}$, so

$$x_{n_t} < \inf \Delta_{c_1 \dots c_k}^{P^-} < \sup \Delta_{c_1 \dots c_k}^{P^-} = x_0,$$

and therefore

$$x_0 - x_{n_t} > \left| \Delta_{c_1 \dots c_k}^{P^-} \right| = \text{const} > 0.$$

Since this inequality holds for infinitely many indices n_t , the sequence $(x_n)_{n=1}^\infty$ does not converge to x_0 .

Case 2: Suppose that $q_i(x_n) = c_i$ for all $i \leq k$ and $n \geq n_0$, but there exist a constant M and a sequence $(n_t)_{t=1}^\infty$ such that $q_{k+1}(x_{n_t}) < M$. Assuming without loss of generality that $n_t \geq n_0$, from (Moroz, 2025, Corollary 3.8) we have

$$\begin{aligned} x_0 - x_{n_t} &> \sup \Delta_{c_1 \dots c_k}^{P^-} - \sup \Delta_{c_1 \dots c_k q_{k+1}(x_{n_t})}^{P^-} = \frac{r_0 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k q_{k+1}(x_{n_t})} > \\ &> \frac{r_0 \dots r_k}{(c_1 - 1) c_1 \dots (c_k - 1) c_k M} = \text{const} > 0, \end{aligned}$$

where $r_0 = \varphi_0$ and $r_i = \varphi_i(c_1, \dots, c_i)$ for all $i \leq k$. Since this inequality holds for infinitely many indices n_t , the sequence $(x_n)_{n=1}^\infty$ does not converge to x_0 .

Thus, condition (12) is both necessary and sufficient for the convergence of $(x_n)_{n=1}^\infty$ to x_0 . $\square$

Lemma 2. Every number from IS^{P^-} , except for 1, is the infimum of some P^- -cylinder of even rank. In particular,

- $\sup \Delta_{c_1 \dots c_k}^{P^-} = \inf \Delta_{c_1 \dots c_{k-1} [c_k + 1]}^{P^-}$ if k is even,
- $\inf \Delta_{c_1 \dots c_k}^{P^-} = \inf \Delta_{c_1 \dots c_k [r_k + 1]}^{P^-}$ if k is odd, where $r_k = \varphi_k(c_1, \dots, c_k)$,
- $\sup \Delta_{c_1 \dots c_k}^{P^-} = \sup \Delta_{c_1 \dots c_{k-1}}^{P^-} = \inf \Delta_{c_1 \dots [c_{k-1} + 1]}^{P^-}$ if k is odd and $c_k = \varphi_{k-1}(c_1, \dots, c_{k-1}) + 1$,
- $\sup \Delta_{c_1 \dots c_k}^{P^-} = \inf \Delta_{c_1 \dots c_{k-1} [c_k - 1]}^{P^-} = \inf \Delta_{c_1 \dots c_{k-1} [c_k - 1] [r_k + 1]}^{P^-}$ if k is odd and $c_k > \varphi_{k-1}(c_1, \dots, c_{k-1}) + 1$, where $r_k = \varphi_k(c_1, \dots, c_{k-1}, c_k - 1)$.

This lemma follows directly from the geometry of the P^- -representation and from formulas (6), (7), (8), and (9).

Proposition 9. Let $x_0 = \inf \Delta_{c_1 \dots c_k}^{P^-}$, where k is even, and let $(x_n)_{n=1}^\infty$ be a sequence such that $x_0 < x_n < 1$ for all $n \in \mathbb{N}$. Then $\lim_{n \rightarrow \infty} x_n = x_0$ if and only if there exists an index n_0 such that

$$\begin{cases} q_i(x_n) = c_i \text{ for all } i \leq k \text{ and } n \geq n_0, \\ \lim_{n \rightarrow \infty} q_{k+1}(x_n) = +\infty. \end{cases} \quad (13)$$

Proof. The proof is similar to that of Proposition 8. $\square$

Discussion and conclusions

The above propositions provide necessary and sufficient conditions for verifying the convergence of any sequence defined in terms of positive or alternating Perron expansions. If a sequence contains infinitely many elements on both sides of the limit point, which is the endpoint of some P -cylinder or is an element of IS^{P^-} , it suffices to consider two corresponding subsequences and apply the relevant proposition to each of them. Moreover, to establish the continuity of a function at a given point, the propositions concerning one-sided limits are fully sufficient.

Sources of funding. This work was supported by a grant from the Simons Foundation (1290607, M.M.).

References

- Albeverio, S., Baranovskyi, O., Kondratiev, Y., & Pratsiovytyi, M. (2013). On one class of functions related to Ostrogradsky series and containing singular and nowhere monotonic functions. *Naukovyj chasopys NPU imeni M. P. Drahomanova. Seriya 1. Fizyko-matematychni nauky*, 15, 35–55.
- Baranovskyi, O., & Pratsiovytyi, M. (2023). One class of continuous functions with complicated local properties related to Engel series. *Functiones et Approximatio, Commentarii Mathematici*, 68(2), 143–162. <https://doi.org/10.7169/facm/1963>

- Erdős, P., Rényi, A., & Szusz, P. (1958). On Engel's and Sylvester's series. *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae, Sectio Mathematica*, 1, 7–32.
- Galambos, J. (1976). *Representations of Real Numbers by Infinite Series*. Springer. <https://doi.org/10.1007/BFb0081642>
- Galambos, J. (1998). Further metric results on series expansions. *Publicationes Mathematicae Debrecen*, 52(3–4), 377–384. <http://dx.doi.org/10.5486/PMD.1998.2037>
- Kalpazidou, S., Knopfmacher, A., & Knopfmacher, J. (1990). Lüroth-type alternating series representations for real numbers. *Acta Arithmetica*, 55(4), 311–322. <https://doi.org/10.4064/aa-55-4-311-322>
- Moroz, M. (2017). Projector of the Δ^O -representation of numbers into the Δ^E -representation. *Transactions of Institute of Mathematics of NAS of Ukraine*, 14(4), 49–64 [in Ukrainian]. [Мороз, М. (2017). Проектор Δ^O -зображення чисел в Δ^E -зображення. *Збірник праць Інституту математики НАН України*, 14(4), 49–64].
- Moroz, M. (2024). Representation of Real Numbers by Perron Series, Their Geometry, and Some Applications. *Journal of Mathematical Sciences*, 279(3), 384–399. <https://doi.org/10.1007/s10958-024-07020-4>
- Moroz, M. (2025). Representation of real numbers by alternating Perron series and their geometry. *Expositiones Mathematicae*, 43(1), 1–18. <https://doi.org/10.1016/j.exmath.2024.125635>
- Oppenheim, A. (1972). The representation of real numbers by infinite series of rationals. *Acta Arithmetica*, 21(1), 391–398. <https://doi.org/10.4064/aa-21-1-391-398>
- Prats'ovtyi, M., Baranovs'kyi, O., & Maslova, Y. (2021). Generalization of the Tribin Function. *Journal of Mathematical Sciences*, 253(2), 276–288. <https://doi.org/10.1007/s10958-021-05227-3>
- Shallit, J. O. (1986). Metric theory of Pierce expansions. *The Fibonacci Quarterly*, 24(1), 22–40. <https://doi.org/10.1080/00150517.1986.12429786>
- Sydroruk, L., & Torbin, G. (2017). On singularity of distribution of random variables with independent symbols of Oppenheim expansions. *Modern Stochastics: Theory and Applications*, 4(3), 273–283. <https://doi.org/10.15559/17-VMSTA87>
- Torbin, G., & Pratsyovtya, I. (2010). Singularity of the second Ostrogradskii random series. *Theory of Probability and Mathematical Statistics*, 81, 187–195. <https://doi.org/10.1090/S0094-9000-2011-00819-8>
- Zhykharyeva, Y., & Pratsyovtyi, M. (2012). Expansions of numbers in positive Lüroth series and their applications to metric, probabilistic and fractal theories of numbers. *Algebra and Discrete Mathematics*, 14(1), 145–160.

Отримано редакцією журналу / Received: 31.07.25

Прорецензовано / Revised: 30.09.25

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Микола МОРОЗ, д-р філософії, мол. наук. співроб.

ORCID ID: 0000-0001-6658-4924

e-mail: mykola.moroz@imath.kiev.ua

Інститут математики НАН України, Київ, Україна

ЗБІЖНІСТЬ ЧИСЛОВИХ ПОСЛІДОВНОСТЕЙ У ТЕРМІНАХ ДОДАТНИХ І ЗНАКОЗМІННИХ РОЗКЛАДІВ ПЕРРОНА

Розглянуто умови збіжності числових послідовностей, елементи яких подано у вигляді їхніх додатних і знакозмінних розкладів Перрона (P -представлення та P^- -представлення дійсних чисел відповідно). Ці умови дають можливість дослідити на неперервність функції, що визначені за допомогою P -представлення чи P^- -представлення чисел.

Основні результати дослідження полягають у такому. По-перше, ми систематично розглядаємо всі можливі випадки, та, враховуючи геометричні властивості розкладів Перрона, значно зменшуємо кількість випадків, що потребують окремого розгляду. По-друге, для кожного випадку ми знаходимо та доводимо умову, що є одночасно необхідною та достатньою. По-третє, наші результати отримані не для конкретних розкладів (таких як розклади Енгеля чи Пірса), а для розкладів Перрона загального виду, що охоплюють багато класичних розкладів дійсних чисел (як спеціальні випадки).

Зазначені результати закладають основу для аналізу фрактальних функцій, що визначені в термінах розкладів Перрона чи їхніх часткових випадків (розклади Люрота, Енгеля, Пірса, Сильвестера тощо).

Ключові слова: збіжні послідовності, розклади Перрона, розклади Люрота, розклади Енгеля, розклади Пірса.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.10>

Denis PLATONOV, PhD Student

ORCID ID: 0009-0005-3660-5516

e-mail: dplatonov@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

WEAK CONVERGENCE ANALYSIS OF DYNAMIC CUTTING FOR LÉVY-TYPE PROCESSES

We derive weak error bounds for Euler – Maruyama schemes of the one-dimensional Lévy-driven SDE, where small jumps are truncated in a time-dependent way using a dynamic cutting technique. Large jumps are simulated exactly, while small jumps are either (i) omitted or (ii) replaced by a Gaussian term with matching variance. Under standard Lipschitz-growth and smoothness conditions on the coefficients and Lévy measure, we prove that the weak error of both schemes is of order $O(n^{-1})$. This rate is achieved by choosing the scaling hyperparameter as $h = n^{-\alpha/(\varepsilon(2-\alpha))}$ for (i) and $h = n^{-\alpha/(\varepsilon(3-\alpha))}$ for (ii) respectively.

Key words: dynamic cutting, Lévy-driven SDEs, weak convergence, Euler – Maruyama scheme.

AMS 2020 classification: 60H35, 65C30, 60G51.

Introduction

We consider a Lévy-driven SDE of the type

$$X_t = X_0 + \int_0^t a(s, X_s) ds + \int_0^t b(s, X_s) dB_s + \int_0^t \int_{-\infty}^{\infty} c(s, X_{s-}, z) \tilde{N}(ds, dz), \quad (1)$$

where $\tilde{N}(ds, dz)$ is the compensated Poisson random measure corresponding to a Lévy process Z_t , and $a(s, x)$, $b(s, x)$, $c(s, x, z)$ are some measurable functions, conditions on which we specify in Section 2.

The analysis of weak approximation errors for Lévy-driven SDEs began with the work of (Protter, & Talay, 1997), who established that the error of the Euler scheme can be expanded in powers of $1/n$. Building on this, (Jacod, 2004) established functional limit theorems for the normalized error process.

A main practical obstacle is the infinite activity of many Lévy processes, making the exact simulation of small jumps impossible. To tackle this (Asmussen, & Rosiński, 2001) introduced a truncation method where jumps below a fixed threshold ε are removed and their effect is approximated by a Gaussian term with suitable variance. Recently, (Bossy, & Maurer, 2025) revisited the fixed-cutoff for the time-inhomogeneous SDEs, demonstrating that the weak convergence rate of $1/n + \varepsilon^{3-\beta}$ can be achieved by choosing threshold ε optimally.

Classical approaches rely on a *fixed* truncation level. In contrast, time-dependent threshold for simulating Lévy processes was introduced by (Ivanenko, Knopova, & Platonov, 2025). Later, the same *dynamic cutting* approach was applied to analyze strong error bounds of the time-inhomogeneous SDEs (Knopova, & Platonov, 2025). In this paper, we transfer this *dynamic cutting* framework to weak error analysis.

1. Dynamic cutting setup

We assume that the driving jump process Z admits the Lévy-Itô representation

$$Z_t = \int_0^t \int_{\mathbb{R}} z (N(ds, dz) - \nu(dz)ds) = \int_0^t \int_{\mathbb{R}} z \tilde{N}(ds, dz), \quad (2)$$

where $N((0, t], B)$ counts the jumps of Z that fall in B , $\tilde{N}(ds, dz) = N(ds, dz) - \nu(dz)ds$ denotes the compensated Poisson random measure, and ν is the associated Lévy measure.

Let $\nu^+ := \nu|_{(0, +\infty)}$ and $\nu^- := \nu|_{(-\infty, 0)}$. Denote for $r > 0$

$$N^+(r) := \nu^+(r, +\infty), \quad N^-(r) := \nu^-(-\infty, -r),$$

and their generalized inverses at t^{-1} for $t > 0$ by

$$\tau^\pm(t) := \sup\{r \geq 0 : N^\pm(r) \geq t^{-1}\}, \quad (3)$$

$$\tau(t) := \max(\tau^+(t), \tau^-(t)). \quad (4)$$

For $s > 0$, define

$$B(s, h, \varepsilon) := \{z : -\tau^-((sh)^\varepsilon) < z < \tau^+((sh)^\varepsilon)\}.$$

Here, $h \in (0, 1)$ and $\varepsilon \in (0, 1)$ are the dynamic cutting parameters introduced in (Ivanenko, Knopova, & Platonov, 2025). They are responsible for controlling the shape and scaling of the time-dependent threshold.

Consider two types of Euler-Maruyama schemes:

$$X_{t_{i+1}}^{1,\tau} := X_{t_i}^{1,\tau} + a(t_i, X_{t_i}^{1,\tau})\Delta_i + b(t_i, X_{t_i}^{1,\tau})\sqrt{\Delta_i}\xi_i + C_{t_i}(X_{t_i}^{1,\tau}), \quad (5)$$

$$X_{t_{i+1}}^{2,\tau} := X_{t_i}^{2,\tau} + a(t_i, X_{t_i}^{2,\tau})\Delta_i + b(t_i, X_{t_i}^{2,\tau})\sqrt{\Delta_i}\xi_i + C_{t_i}(X_{t_i}^{2,\tau}) + \sigma_i(X_{t_i}^{2,\tau})\zeta_i, \quad (6)$$

where:

- ξ_i, ζ_i – sequences of i.i.d. standard Gaussian random variables;
- $C_{t_i}(x)$ is the compensated frozen large-jump increment

$$C_{t_i}(x) := \int_{t_i}^{t_{i+1}} \int_{\mathbb{R} \setminus B(t_i, h, \varepsilon)} c(t_i, x, z) \tilde{N}(ds, dz).$$

- $\sigma_i^2(x)$ is the variance of the Gaussian correction and uses the frozen left-endpoint value

$$\sigma_i^2(x) := \Delta_i \int_{B(t_i, h, \varepsilon)} c^2(t_i, x, z) \nu(dz).$$

For simplicity, we write X_t^τ to denote $X_t^{1,\tau}$ or $X_t^{2,\tau}$ when the distinction between the schemes is not relevant.

Let $0 = t_0 < t_1 < \dots < t_n = T$ be the grid formed by combining equally spaced points and large jumps (positive and negative). This construction guarantees that large jumps are included as grid points. In the weak-error proof below, we use the standard frozen-coefficient continuous Euler interpolation associated with these grid values and, for notational simplicity, denote this interpolation again by X_t^τ .

In addition, let us define the lower Pruitt functions as

$$\psi^{L,\pm}(\xi) := \int_{\{u \neq 0, |u\xi| \leq 1\}} |u\xi|^2 \nu^\pm(du), \quad \psi^L := \psi^{L,+} + \psi^{L,-}. \quad (7)$$

2. Assumptions

Throughout this section we write $\lesssim$ to indicate that particular constants are not important for the bounds.

Assumption 1 (Regularity of coefficients). *For all $s, t \in [0, T]$ and $x, y, z \in \mathbb{R}$:*

1. *The coefficients are globally Lipschitz in the space-time variables:*

$$|a(t, x) - a(s, y)| + |b(t, x) - b(s, y)| \lesssim (|t - s| + |x - y|),$$

$$|c(t, x, z) - c(s, y, z)| \lesssim (|t - s| + |x - y|) (1 \wedge |z|).$$

2. *The coefficients satisfy a linear growth condition in the state variable:*

$$|a(t, x)| + |b(t, x)| \lesssim (1 + |x|),$$

$$|c(t, x, z)| \lesssim (1 + |x|) (1 \wedge |z|).$$

Assumption 2 (Smoothness). *Fix an integer $m \geq 3$.*

1. *For all $t \in [0, T]$, $x \in \mathbb{R}$, $z \neq 0$ and integers $k, \ell \geq 0$ with $1 \leq k + \ell \leq m$:*

$$|\partial_x^k a(t, x)| + |\partial_x^k b(t, x)| + |\partial_x^k \partial_z^\ell c(t, x, z)| \lesssim 1.$$

2. *There exists a measurable function $q \in \bigcap_{p \geq 2} L^p(\nu)$ such that for every $k = 1, \dots, m$:*

$$|\partial_x^k c(t, x, z)| \lesssim q(z), \quad \text{for all } x \in \mathbb{R}, z \neq 0.$$

Assumption 3 (Kernel assumptions). *Recall the definition of the lower Pruitt functions (7). Then*

1. *$\nu(du)$ is absolutely continuous.*
2. *$N^\pm(r) \asymp \psi^{L,\pm}(1/r)$.*

Remark 1. Under Assumption 3.1 $N^\pm$ are strictly decreasing and $N^\pm(\tau^\pm(t)) = \frac{1}{t}$. Moreover, under Assumption 3.2 one can show that there exists $\alpha^\pm \in (0, 2)$ such that $\tau^\pm(t) \lesssim t^{1/\alpha^\pm}$. We also define

$$\alpha := \max(\alpha^-, \alpha^+). \quad (8)$$

3. Well-posedness and some implications

We define the operator $\mathcal{L}_t$ on $f \in C_b^2(\mathbb{R})$ by

$$\mathcal{L}_t f(x) = a(t, x) f'(x) + \frac{1}{2} b(t, x)^2 f''(x) + \int_{\mathbb{R}} [f(x + c(t, x, z)) - f(x) - c(t, x, z) f'(x)] \nu(dz). \quad (9)$$

Proposition 1. *Let $u(t, x) := \mathbb{E}[\Phi(X_T^{t,x})]$, where $\Phi \in C_b^3(\mathbb{R})$ and $X^{t,x}$ is the strong solution of the SDE (1). Then under Assumptions 1 and 2, the following statements hold:*

(i) $u \in C^{1,3}((0, T) \times \mathbb{R})$ and its derivatives satisfy the polynomial growth condition for some $p \geq 0$:

$$|\partial_x^k u(t, x)| \lesssim (1 + |x|^p). \quad (10)$$

(ii) The function u is the unique classical solution of the backward Kolmogorov PDE:

$$\partial_t u(t, x) + \mathcal{L}_t u(t, x) = 0, \quad u(T, x) = \Phi(x). \quad (11)$$

Proof. Under Assumption 1, the SDE (1) has a unique strong solution (Situ, 2005, Thm. 117), and standard moment estimates give the polynomial bound for the solution itself. By Assumption 2 and (Breton, & Privault, 2020, Prop. 4.1), the solution flow $x \mapsto X_s^{t,x}$ is three-times differentiable in x . Moreover, (Breton, & Privault, 2020, Thm. 5.1) gives the corresponding L^q bounds for the flow derivatives. Thus, for every $q \geq 2$,

$$\sup_{s \in [t, T]} \mathbb{E}[|X_s^{t,x}|^q] \lesssim 1 + |x|^q, \quad \sup_{s \in [t, T]} \mathbb{E}[|\partial_x^k X_s^{t,x}|^q] < \infty, \quad k = 1, 2, 3. \quad (12)$$

The moment bounds in (12) combined with the boundedness of $\Phi^{(k)}$ allow us to apply the dominated convergence theorem to justify differentiating under the expectation, as the moment bounds provide the necessary integrable majorant.

For $k = 1, 2, 3$, differentiation of the stochastic flow using the chain rule yields

$$\partial_x^k u(t, x) = \mathbb{E} \left[\sum_{m=1}^k \Phi^{(m)}(X_T^{t,x}) B_{k,m}(\partial_x X_T^{t,x}, \dots, \partial_x^{k-m+1} X_T^{t,x}) \right],$$

where $B_{k,m}$ are the exponential Bell polynomials given by the Faà di Bruno formula (Comtet, 1974, Chap. III, Thm. A). The boundedness of the derivatives of Φ , together with the moment bounds for the derivatives of the stochastic flow, gives the required polynomial growth estimates for $\partial_x^k u$ and confirms that $u(t, \cdot) \in C^3(\mathbb{R})$.

Having established the spatial regularity of u , we now prove its differentiability in time and show that it solves the backward Kolmogorov PDE using the semigroup approach. Let $(P_{s,t})_{0 \leq s \leq t \leq T}$ be the time-inhomogeneous two-parameter semigroup defined by $P_{s,t} f(x) = \mathbb{E}[f(X_t^{s,x})]$. Then we can then write $u(t, x) = P_{t,T} \Phi(x)$. By the Markov property, for a small $h > 0$ such that $t + h < T$, we have $P_{t,T} = P_{t,t+h} P_{t+h,T}$. This allows us to write:

$$u(t, x) = P_{t,t+h} P_{t+h,T} \Phi(x) = P_{t,t+h} u(t + h, x).$$

Rearranging this and dividing by h gives us an expression:

$$\frac{u(t + h, x) - u(t, x)}{h} = - \left(\frac{P_{t,t+h} - I}{h} \right) u(t + h, x). \quad (13)$$

Let us examine the r.h.s. of the above equation. Let $f \in C^2(\mathbb{R})$. By Itô's formula on $[t, t + h]$,

$$f(X_{t+h}^{t,x}) - f(X_t^{t,x}) = \int_t^{t+h} \mathcal{L}_s f(X_s^{t,x}) ds + M_{t,t+h},$$

where $M_{t,t+h}$ is a local martingale part with $\mathbb{E}[M_{t,t+h}] = 0$. Taking expectations and using $X_t = x$ gives us

$$P_{t,t+h} f(x) - f(x) = \mathbb{E} \left[\int_t^{t+h} \mathcal{L}_s f(X_s^{t,x}) ds \right].$$

If $f(s, \cdot)$ and its derivatives have polynomial growth, and the coefficients satisfy the growth and integrability conditions in Assumption 1, the operator $\mathcal{L}_t$ is bounded by a polynomial: $|\mathcal{L}_t f(x)| \lesssim (1 + |x|^q)$ for some $q \geq 0$. By applying the moment bounds from (12), we can use Fubini's theorem to interchange the expectation and the integral.

$$\frac{P_{t,t+h} f(x) - f(x)}{h} = \frac{1}{h} \int_t^{t+h} \mathbb{E}[\mathcal{L}_s f(X_s^{t,x})] ds. \quad (14)$$

Hence, by the dominated convergence and Lebesgue differentiation theorems, noting that $X_t^{t,x} = x$, we get

$$\frac{1}{h} \int_t^{t+h} \mathbb{E}[\mathcal{L}_s f(X_s^{t,x})] ds \xrightarrow{h \rightarrow 0} \mathbb{E}[\mathcal{L}_t f(X_t^{t,x})] = \mathcal{L}_t f(x),$$

Consequently,

$$\lim_{h \rightarrow 0} \frac{P_{t,t+h} f(x) - f(x)}{h} = \mathcal{L}_t f(x).$$

Using the limit results, together with $u(t + h, \cdot) \rightarrow u(t, \cdot)$, we obtain the backward Kolmogorov equation:

$$\partial_t u(t, x) = -\mathcal{L}_t u(t, x). \quad (15)$$

For uniqueness, let $v \in C^{1,3}$ be another classical solution and set $w := v - u$. By linearity, w solves $\partial_t w + \mathcal{L}_t w = 0$ with $w(T, x) = 0$. Applying Itô's formula to $w(s, X_s^{t,x})$ on $[t, T]$ gives a sum of an ordinary integral and a martingale term. Taking expectations eliminates the martingale term (as it has zero mean) and gives:

$$\mathbb{E}[w(T, X_T^{t,x})] - \mathbb{E}[w(t, X_t^{t,x})] = \mathbb{E}\left[\int_t^T (\partial_s w + \mathcal{L}_s w)(s, X_s^{t,x}) ds\right] = 0.$$

Using the terminal condition $w(T, \cdot) = 0$ and the initial condition $X_t^{t,x} = x$, this simplifies to $0 - w(t, x) = 0$. Thus, $w(t, x) = 0$ for all (t, x) , which implies $v \equiv u$. $\square$

Corollary 2. *If we additionally assume higher smoothness, namely $u \in C^{1,5}$, then under the hypotheses of Proposition 1, for every integer $0 \leq \ell \leq 3$ and for some $p \geq 0$, the mixed derivatives satisfy the growth condition*

$$|\partial_t \partial_x^\ell u(t, x)| \lesssim (1 + |x|^p).$$

Proof. From Proposition 1, we have the identity $\partial_t u = -\mathcal{L}_t u$. The regularity $u \in C^{1,5}$ allows us to differentiate with respect to x up to order ℓ :

$$\partial_t \partial_x^\ell u(t, x) = -\partial_x^\ell (\mathcal{L}_t u).$$

We bound the term on the right by analyzing its drift, diffusion and jump components separately.

The drift and diffusion components of the operator's derivative are

$$I_1^\ell := \partial_x^\ell [a(t, x) \partial_x u(t, x)], \quad I_2^\ell := \frac{1}{2} \partial_x^\ell [b^2(t, x) \partial_x^2 u(t, x)].$$

Applying the Leibniz rule gives the expansions

$$I_1^\ell = \sum_{k=0}^{\ell} C_\ell^k (\partial_x^k a) (\partial_x^{\ell-k+1} u), \quad I_2^\ell = \frac{1}{2} \sum_{k=0}^{\ell} C_\ell^k (\partial_x^k b^2) (\partial_x^{\ell-k+2} u),$$

where the term $\partial_x^k b^2$ expands to $\sum_{m=0}^k C_k^m (\partial_x^m b) (\partial_x^{k-m} b)$.

Proposition 1 yields polynomial growth $|\partial_x^m u| \lesssim (1 + |x|)^{p_0}$ for some $p_0 \geq 0$. For $j \geq 1$, Assumption 2 gives uniform bounds on the derivatives of the coefficients, $|\partial_x^j a| \leq C$ and $|\partial_x^j b| \leq C$; for $j = 0$, Assumption 1 gives $|a| + |b| \lesssim (1 + |x|)$. Hence, each term in the Leibniz sums is either bounded by $(1 + |x|)$ (from a) or $(1 + |x|)^2$ (from b^2). Consequently, for some $p_1 \geq 0$

$$|I_1^\ell(t, x)| + |I_2^\ell(t, x)| \lesssim (1 + |x|)^{p_1}.$$

Let $c := c(t, x, z)$ and $I_3^\ell(t, x) := \partial_x^\ell \int_{\mathbb{R}} J(t, x, z) \nu(dz)$, where by the Taylor theorem,

$$J(t, x, z) := u(t, x + c) - u(t, x) - c \partial_x u(t, x) = \frac{1}{2} c^2 \int_0^1 (1-s) \partial_x^2 u(t, x + sc) ds.$$

We need to move ∂_x^ℓ inside the z -integral. To justify this step, we build an integrable majorant for $|\partial_x^\ell J|$. Using the Leibniz rule, we obtain

$$\partial_x^\ell J(t, x, z) = \frac{1}{2} \sum_{k=0}^{\ell} C_\ell^k (\partial_x^k c^2) \int_0^1 (1-s) \partial_x^{\ell-k+2} u(t, x + sc) ds.$$

For $j = 0$, Assumption 1 provides the bound $|\partial_x^j c| \lesssim (1 + |x|)(1 \wedge |z|)$, while for $j = 1, 2, 3$, Assumption 2 gives $|\partial_x^j c| \lesssim q(z)$. Since each $\partial_x^j c^2$ is a product of at most two such entries, we have

$$|\partial_x^k c^2(t, x, z)| \lesssim (1 + |x|)^2 \left[(1 \wedge |z|)^2 + (1 \wedge |z|) q(z) + q(z)^2 \right].$$

Proposition 1 says that for $m \leq 5$ for some $p_2 \geq 0$

$$|\partial_y^m u(t, y)| \lesssim (1 + |y|)^{p_2}, \quad t \in [0, T], \quad y \in \mathbb{R}.$$

Because $|c| \lesssim (1 + |x|)$, replacing y by $x + sc$, $s \in [0, 1]$ shows that the s -integral is bounded by $(1 + |x|)^{p_3}$ for some $p_3 \geq 0$.

Combining the previous two steps, for some $p_4 \geq 0$ we obtain

$$|\partial_x^\ell J(t, x, z)| \lesssim (1 + |x|)^{p_4} G(z), \quad G(z) := (1 \wedge |z|)^2 + (1 \wedge |z|) q(z) + q(z)^2.$$

Next, we show that $G \in L^1(\nu)$. Indeed,

- (1) $\int_{\mathbb{R}} (1 \wedge |z|)^2 \nu(dz) < \infty$ based on the Lévy measure definition.
- (2) $\int_{\mathbb{R}} q(z)^2 \nu(dz) < \infty$ by Assumption 2.
- (3) Using the Cauchy-Schwarz inequality and letting $q \in L^2(\nu)$,

$$\int_{\mathbb{R}} (1 \wedge |z|) q(z) \nu(dz) \leq \left(\int_{\mathbb{R}} (1 \wedge |z|)^2 \nu(dz) \right)^{1/2} \left(\int_{\mathbb{R}} q(z)^2 \nu(dz) \right)^{1/2} < \infty.$$

Since $|\partial_x^\ell J| \leq (1 + |x|)^{p_4} G(z)$ with $G \in L^1(\nu)$, dominated convergence theorem allows us to move ∂_x^ℓ inside the integral in the definition of the $I_3^\ell(t, x)$, giving

$$|I_3^\ell(t, x)| \lesssim (1 + |x|)^{p_4}.$$

Adding the polynomial bounds for the drift, diffusion, and jump parts together yields the result. $\square$

Lemma 1. *Under Assumption 1, for any $p \geq 1$, the grid-point moments of the schemes are finite:*

$$\mathbb{E} \left[\max_{0 \leq i \leq n} |X_{t_i}^\tau|^p \right] < \infty.$$

Proof. Let the grid be $0 = t_0 < t_1 < \dots < t_n = T$. It is enough to control the moments at the grid points. Indeed,

$$\mathbb{E} \left[\max_{0 \leq i \leq n} |X_{t_i}^\tau|^p \right] \leq \sum_{i=0}^n \mathbb{E} [|X_{t_i}^\tau|^p].$$

The number of points n is finite. Thus, it is sufficient to show that $\mathbb{E} [|X_{t_i}^\tau|^p] < \infty$ for all i . We do this by induction. We have the base case $\mathbb{E} [|X_0|^p] < \infty$. For the inductive step, we derive a recursive bound for $A_i := \mathbb{E} [|X_{t_i}^\tau|^p]$.

Fix i and consider the scheme's update on $[t_i, t_{i+1}]$ with coefficients frozen at $(t_i, X_{t_i}^\tau)$:

$$X_{t_{i+1}}^\tau = X_{t_i}^\tau + a_i \Delta_i + b_i (B_{t_{i+1}} - B_{t_i}) + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R} \setminus B(t_i, h, \varepsilon)} c_i(z) \tilde{N}(ds, dz) + \mathbb{I}_{\{\text{scheme 2}\}} \sigma_i \xi_i,$$

where the coefficients a_i, b_i, c_i are defined as frozen at time t_i .

Next, our goal is to bound $\mathbb{E} [|X_{t_{i+1}}^\tau|^p | \mathcal{F}_{t_i}]$. We can do this by applying the inequality

$$|x_1 + \dots + x_k|^p \leq C_p (|x_1|^p + \dots + |x_k|^p)$$

and considering each term from the right-hand side separately.

Using Assumption 1 to bound coefficients, the moment bounds for the Brownian terms, and the Novikov's theorem 6 for the compensated Poisson integral, we obtain the one-step conditional bound:

$$\mathbb{E} [|X_{t_{i+1}}^\tau|^p | \mathcal{F}_{t_i}] \leq |X_{t_i}^\tau|^p + C \Delta_i^{\beta_p} (1 + |X_{t_i}^\tau|^p), \quad \beta_p := \min(1, p/2).$$

Taking full expectations gives the following recursion:

$$A_{i+1} \leq (1 + C \Delta_i) A_i + C \Delta_i, \quad A_i := \mathbb{E} [|X_{t_i}^\tau|^p].$$

Since $A_0 < \infty$, by induction we have that $A_i < \infty$ for all i . Therefore,

$$\mathbb{E} \left[\max_{0 \leq i \leq n} |X_{t_i}^\tau|^p \right] < \infty.$$

$\square$

Throughout the following theorem and its proof we write $\mathbb{E}[\cdot]$ for $\mathbb{E}^x[\cdot]$ with non-random initial condition $X_0 = x$.

Theorem 3. *Let Assumptions 1–3 be satisfied and $\Phi \in C_b^3(\mathbb{R})$. Then the weak error between the true solution X_T of SDE (1) and the approximations X_T^τ from the Euler schemes (5) and (6) is of order n^{-1} :*

$$|\mathbb{E}[\Phi(X_T)] - \mathbb{E}[\Phi(X_T^\tau)]| \lesssim n^{-1}.$$

This rate is achieved when h is chosen as $h = n^{-\frac{\alpha}{\varepsilon(2-\alpha)}}$ for scheme (5), and $h = n^{-\frac{\alpha}{\varepsilon(3-\alpha)}}$ for scheme (6).

Proof. In the proof we rely on the approach outlined in (Bossy, & Maurer, 2025).

Our goal is to find an upper bound for

$$\left| \mathbb{E}[\Phi(X_t)] - \mathbb{E}[\Phi(X_t^\tau)] \right|, \quad \Phi \in C_b^3(\mathbb{R}). \quad (16)$$

Applying Itô's formula from Theorem 4 to $u(t, X_t)$, we obtain

$$u(t, X_t) = u(0, X_0) + \int_0^t \underbrace{(\partial_s u + \mathcal{L}_s u)}_{=0}(s, X_s) ds + M_t,$$

where M_t is the martingale part with zero expected value defined by

$$M_t := \int_0^t b(s, X_s) \partial_x u(s, X_s) dB_s + \int_0^t \int_{\mathbb{R}} \left[u(s, X_{s-} + c(s, X_{s-}, z)) - u(s, X_{s-}) \right] \tilde{N}(ds, dz).$$

Therefore, $\mathbb{E}[u(t, X_t)] = u(0, X_0)$.

Let $\eta(s) := \sup\{t_k \geq 0 \mid t_k \leq s\}$. Then, using Kolmogorov PDE (11), for DC schemes (5) and (6) we obtain

$$u(t, X_t^\tau) = u(0, X_0^\tau) + \int_0^t (\partial_s u + \mathcal{L}_s^\tau u)(s, X_s^\tau) ds + M_t^\tau = u(0, X_0^\tau) + \int_0^t (-\mathcal{L}_s u + \mathcal{L}_s^\tau u)(s, X_s^\tau) ds + M_t^\tau,$$

where M_t^τ is the martingale part with zero expected value defined below for both schemes respectively

$$M_t^{\tau,1} := \int_0^t \int_{\mathbb{R} \setminus B(\eta(s), h, \varepsilon)} \left[u(s, X_{s-}^{\tau,1} + c(\eta(s), X_{\eta(s)}^{\tau,1}, z)) - u(s, X_{s-}^{\tau,1}) \right] \tilde{N}(ds, dz) + \\ + \int_0^t b(\eta(s), X_{\eta(s)}^{\tau,1}) \partial_x u(s, X_s^{\tau,1}) dB_s,$$

$$M_t^{\tau,2} := \int_0^t \int_{\mathbb{R} \setminus B(\eta(s), h, \varepsilon)} \left[u(s, X_{s-}^{\tau,2} + c(\eta(s), X_{\eta(s)}^{\tau,2}, z)) - u(s, X_{s-}^{\tau,2}) \right] \tilde{N}(ds, dz) + \\ + \int_0^t b(\eta(s), X_{\eta(s)}^{\tau,2}) \partial_x u(s, X_s^{\tau,2}) dB_s + \int_0^t d(\eta(s), X_{\eta(s)}^{\tau,2}) \partial_x u(s, X_s^{\tau,2}) dW_s,$$

$$d^2(s, x) := \int_{B(s, h, \varepsilon)} c^2(s, x, z) \nu(dz).$$

Thus, on $[t_i, t_{i+1})$, the Gaussian interpolation uses $d(t_i, X_{t_i}^\tau)$, with the variance over the interval $\sigma_i^2(X_{t_i}^\tau) = \Delta_i d^2(t_i, X_{t_i}^\tau)$.

$$\mathcal{L}_t^{1,\tau} f(x) = a_i f'(x) + \frac{1}{2} b_i^2 f''(x) + \int_{\mathbb{R} \setminus B(t_i, h, \varepsilon)} [f(x + c_i(z)) - f(x) - c_i(z) f'(x)] \nu(dz), \quad (17)$$

$$\mathcal{L}_t^{2,\tau} f(x) = a_i f'(x) + \frac{1}{2} b_i^2 f''(x) + \int_{\mathbb{R} \setminus B(t_i, h, \varepsilon)} [f(x + c_i(z)) - f(x) - c_i(z) f'(x)] \nu(dz) + \quad (18)$$

$$+ \frac{1}{2} f''(x) \int_{B(t_i, h, \varepsilon)} c_i^2(z) \nu(dz), \quad (19)$$

where $a_i := a(t_i, X_{t_i}^\tau)$, $b_i := b(t_i, X_{t_i}^\tau)$, $c_i(z) := c(t_i, X_{t_i}^\tau, z)$.

Note that since $X_0^\tau = X_0 = x$, we have $u(0, X_0^\tau) = u(0, X_0)$. Then, by taking expectations, we get

$$\mathbb{E}[\Phi(X_t)] - \mathbb{E}[\Phi(X_t^\tau)] = \int_0^T \mathbb{E}[(\mathcal{L}_t - \mathcal{L}_t^\tau)u(t, X_t^\tau)] dt. \quad (20)$$

Since the coefficients of the frozen Euler interpolation are frozen on $[t_i, t_{i+1})$ at the left endpoint value, we analyze the operator difference at the fixed state $x = X_{t_i}^\tau$.

Setting $f(x) := u(t, x)$, this difference can be decomposed as

$$(\mathcal{L}_t - \mathcal{L}_t^\tau)f(x) = (A) + (B) + (C) + (D), \quad (21)$$

where for $t \in [t_i, t_{i+1})$

$$(A) := [a(t, x) - a(t_i, x)] f'(x),$$

$$(B) := \frac{1}{2} [b^2(t, x) - b^2(t_i, x)] f''(x),$$

$$(C) := \int_{\mathbb{R} \setminus B(t_i, h, \varepsilon)} [f(x + c(t, x, z)) - f(x + c(t_i, x, z)) - (c(t, x, z) - c(t_i, x, z)) f'(x)] \nu(dz),$$

$$(D_1) := \int_{B(t_i, h, \varepsilon)} [f(x + c(t, x, z)) - f(x) - c(t, x, z) f'(x)] \nu(dz). \quad (\text{Scheme 1})$$

For Scheme 2, write $(D_2) = (D_{2,1}) + (D_{2,2})$, where

$$(D_{2,1}) := \int_{B(t_i, h, \varepsilon)} \left[f(x + c(t, x, z)) - f(x) - c(t, x, z) f'(x) - \frac{1}{2} c^2(t, x, z) f''(x) \right] \nu(dz),$$

$$(D_{2,2}) := \frac{1}{2} f''(x) \int_{B(t_i, h, \varepsilon)} [c^2(t, x, z) - c^2(t_i, x, z)] \nu(dz).$$

Bound (A). Using Assumption 1 and Proposition 1, we obtain

$$|(A)| \lesssim |t - t_i| |f'(x)| \lesssim \Delta_i (1 + |x|^{p_a}), \quad \text{for some } p_a \geq 0. \quad (22)$$

Bound (B). By Assumption 1 we have

$$|b^2(t, x) - b^2(s, x)| = (b(t, x) - b(s, x))(b(t, x) + b(s, x)) \lesssim |t - s|(1 + |x|).$$

Thus, adding Proposition 1, we obtain

$$|(B)| \lesssim (t - t_i)(1 + |x|) |f''(x)| \lesssim \Delta_i (1 + |x|)(1 + |x|^p) \lesssim \Delta_i (1 + |x|^{p_b}), \quad \text{for some } p_b \geq 0. \quad (23)$$

Bound (C). Let

$$c := c(t, x, z), \quad c_i := c(t_i, x, z), \quad \delta := c - c_i.$$

Assumptions 1–2 give us the following bounds

$$|\delta| \lesssim |t - t_i| (1 \wedge |z|), \quad |c_i| \lesssim (1 + |x|) (1 \wedge |z|).$$

The first order Taylor expansion of $g(u) := f(x + u)$ at $u = c_i$ yields:

$$g(c) = g(c_i) + \delta g'(c_i) + \frac{1}{2} \delta^2 g''(c_i + \theta \delta), \quad \theta \in (0, 1).$$

Hence,

$$g(c) - g(c_i) - \delta g'(c_i) = \frac{1}{2} \delta^2 g''(c_i + \theta \delta).$$

Applying the mean-value theorem and Proposition 1, we get:

$$|g'(c_i) - g'(0)| \lesssim (1 + |x|^{p_1}) |c_i|.$$

The second term is bounded by Proposition 1. After that the argument $x + c_i + \theta \delta$ grows at most linearly with $|x|$ by Assumption 1. Consequently, we have the following polynomial bound:

$$|g''(c_i + \theta \delta)| = |\partial_x^2 u(t, x + c_i + \theta \delta)| \lesssim (1 + |x + c_i + \theta \delta|^{p_2}) \lesssim (1 + |x|^{p_2}).$$

Then, for $f(x) = u(t, x)$ we have

$$\begin{aligned} |f(x + c) - f(x + c_i) - \delta f'(x)| &\lesssim (1 + |x|^{p_1}) |\delta| |c_i| + \frac{1}{2} |\delta|^2 (1 + |x|^{p_2}) \lesssim (1 + |x|^{p_c}) (|\delta| |c_i| + |\delta|^2) \lesssim \\ &\lesssim (1 + |x|^{p_c}) (|t - t_i| (1 \wedge |z|^2) + |t - t_i|^2 (1 \wedge |z|^2)) \lesssim (1 + |x|^{p_c}) \Delta_i (1 \wedge |z|^2). \end{aligned}$$

Finally, by definition of the Lévy measure we get

$$|(C)| \lesssim \Delta_i (1 + |x|^{p_c}), \quad \text{for some } p_c \geq 0. \quad (24)$$

Bound (D). For each z there exists $\theta_1 = \theta_1(z) \in (0, 1)$ and $\theta_2 = \theta_2(z) \in (0, 1)$ such that the Taylor expansion for both Euler schemes can be written as following:

$$f(x + c) - f(x) - cf'(x) = \frac{1}{2} c^2 f''(x + \theta_1 c), \quad (25)$$

$$f(x + c) - f(x) - cf'(x) - \frac{1}{2} c^2 f''(x) = \frac{1}{6} c^3 f'''(x + \theta_2 c). \quad (26)$$

Using Assumption 1 and Proposition 1,

$$|f(x + c) - f(x) - cf'(x)| \lesssim (1 + |x|^{p+2}) |z|^2,$$

$$|f(x + c) - f(x) - cf'(x) - \frac{1}{2} c^2 f''(x)| \lesssim (1 + |x|^{p+3}) |z|^3.$$

Here, we used the fact that $(1 \wedge |z|) = |z|$, since we are in the small-jump region.

For Scheme 2, the term $(D_{2,1})$ is controlled by the cubic Taylor remainder:

$$\left| f(x + c) - f(x) - cf'(x) - \frac{1}{2} c^2 f''(x) \right| \lesssim (1 + |x|^{p+3}) |z|^3.$$

Therefore,

$$|D_{2,1}| \lesssim (1 + |x|^{p+3}) \int_{B(t_i, h, \varepsilon)} |z|^3 \nu(dz).$$

It remains to control $(D_{2,2})$. By Assumption 1,

$$|c^2(t, x, z) - c^2(t_i, x, z)| \leq |c(t, x, z) - c(t_i, x, z)| (|c(t, x, z)| + |c(t_i, x, z)|).$$

Using the Lipschitz and growth bounds for c , we get

$$|c^2(t, x, z) - c^2(t_i, x, z)| \lesssim |t - t_i|(1 + |x|)^2(1 \wedge |z|)^2.$$

Hence, for some $p \geq 0$,

$$|D_{2,2}| \lesssim |t - t_i|(1 + |x|^p) \int_{\mathbb{R}} (1 \wedge |z|)^2 \nu(dz) \lesssim \Delta_i(1 + |x|^p).$$

Completing the proof.

Taking expectations and using the grid-point moment bound from Lemma 1, and noting that $\Delta_i \leq \frac{T}{n} \lesssim n^{-1}$, we get:

$$\int_0^T \mathbb{E}[|(A)| + |(B)| + |(C)|] dt \lesssim n^{-1}.$$

In addition, for the (D) part we apply Theorem 5 and obtain

$$\int_0^T \mathbb{E}[|(D_1)|] dt \lesssim \frac{\tau^2(h^\varepsilon)}{h^\varepsilon},$$

and, for Scheme 2,

$$\int_0^T \mathbb{E}[|(D_{2,1})|] dt \lesssim \frac{\tau^3(h^\varepsilon)}{h^\varepsilon}, \quad \int_0^T \mathbb{E}[|(D_{2,2})|] dt \lesssim n^{-1}.$$

Consequently,

$$\int_0^T \mathbb{E}[|(D_2)|] dt \lesssim \frac{\tau^3(h^\varepsilon)}{h^\varepsilon} + n^{-1}.$$

Using the Remark 1 and taking $h = n^{-\frac{\alpha}{\varepsilon(2-\alpha)}}$ for scheme 1, and $h = n^{-\frac{\alpha}{\varepsilon(3-\alpha)}}$ for scheme 2, we get

$$|\mathbb{E}[\Phi(X_T)] - \mathbb{E}[\Phi(X_T^\tau)]| \lesssim n^{-1}.$$

□

Appendix

Theorem 4 (Itô's formula for a compensated Lévy-type SDE (Applebaum, 2009, Thm. 4.4.7)). *Let X_t be a solution to (1). Then, for any function $f \in C_b^2(\mathbb{R})$, the following holds a.s. for all $t \geq 0$:*

$$\begin{aligned} f(X_t) - f(X_0) &= \int_0^t a(s, X_s) f'(X_s) ds + \frac{1}{2} \int_0^t b(s, X_s)^2 f''(X_s) ds + \int_0^t b(s, X_s) f'(X_s) dB_s + \\ &+ \int_0^t \int_{\mathbb{R}} [f(X_{s-} + c(s, X_{s-}, z)) - f(X_{s-}) - c(s, X_{s-}, z) f'(X_{s-})] \nu(dz) ds + \\ &+ \int_0^t \int_{\mathbb{R}} [f(X_{s-} + c(s, X_{s-}, z)) - f(X_{s-})] \tilde{N}(ds, dz). \end{aligned}$$

We state the next theorem in the form adopted by our framework.

Theorem 5 (Small jumps moments (Knopova, & Platonov, 2025, Lem. A.3)). *Let X_t be a solution to (1). Then for any $p \geq 2$*

$$\mathbb{E} \left[\int_0^t \int_{B(s, h, \varepsilon)} |c(s, X_s, z)|^p \nu(dz) ds \right] \lesssim \frac{\tau^p(h^\varepsilon)}{h^\varepsilon}.$$

Theorem 6. (Kühn, & Schilling, 2023, Th. 4.20) *Let $(X_t)_{t \geq 0}$ be a one-dimensional stochastic process given by*

$$X_t = \int_0^t \int_{y \neq 0} F(s, y) \tilde{N}(dy, ds), \quad t \geq 0,$$

where F is a predictable stochastic process, and $\tilde{N} = N - \hat{N}$ is a compensated Poisson random measure with compensator $\hat{N}(dy, ds) = \nu(dy)ds$, where the measure ν satisfies $\int_{y \neq 0} \min\{1, |y|^2\} \nu(dy) < \infty$.

1. (Novikov, 1975) *If*

$$\mathbb{E} \left[\int_0^T \int_{y \neq 0} |F(s, y)| \nu(dy) ds \right] < \infty,$$

for some $T > 0$, then for any $a \in [1, 2]$ and $p \in [0, a] \subseteq [0, 2]$, we have

$$\mathbb{E} \left(\sup_{t \leq T} |X_t|^p \right) \leq C_{p,a} \mathbb{E} \left[\left(\int_0^T \int_{y \neq 0} |F(s, y)|^a \nu(dy) ds \right)^{p/a} \right]. \quad (27)$$

2. (Kunita, 2004, Th.2.11) For all $p \geq 2$, we have

$$\mathbb{E} \left(\sup_{t \leq T} |X_t|^p \right) \leq c_p \mathbb{E} \left[\left(\int_0^T \int_{y \neq 0} |F(s, y)|^2 \nu(dy) ds \right)^{p/2} \right] + c_p \mathbb{E} \left[\int_0^T \int_{y \neq 0} |F(s, y)|^p \nu(dy) ds \right]. \quad (28)$$

Discussion and conclusions

We studied weak approximation of time-inhomogeneous Lévy-driven SDEs, analyzing two Euler schemes: (i) small jumps are removed completely and (ii) small jump are approximated with a Gaussian term. Under the stated assumptions, we proved $\mathcal{O}(n^{-1})$ weak error when scaling hyperparameter h is chosen optimally.

Sources of funding. The study was partially financially supported by Taras Shevchenko National University of Kyiv.

References

- Applebaum, D. (2009). *Lévy processes and stochastic calculus* (Second ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511809781>
- Asmussen, S., & Rosiński, J. (2001). Approximations of small jumps of Lévy processes with a view towards simulation. *Journal of Applied Probability*, 38(2), 482–493. <https://doi.org/10.1017/s0021900200019987>
- Bossy, M., & Maurer, P. (2025). On the ε -Euler-Maruyama scheme for time inhomogeneous jump-driven SDEs. *Stochastic Processes and their Applications*, 190, 104747. <https://doi.org/10.1016/j.spa.2025.104747>
- Breton, J.-C., & Privault, N. (2020). Integrability and regularity of the flow of stochastic differential equations with jumps. *Theory of Probability and its Applications*, 65(1), 82–101. <https://doi.org/10.1137/S0040585X97T989830>
- Comtet, L. (1974). *Advanced combinatorics. The art of finite and infinite expansions*. Springer Dordrecht. <https://doi.org/10.1007/978-94-010-2196-8>
- Ivanenko, D., Knopova, V., & Platonov, D. (2025). On approximation of some lévy processes. *Austrian Journal of Statistics*, 54(1), 177–199. <https://doi.org/10.17713/ajs.v54i1.1991>
- Jacod, J. (2004). The Euler scheme for Lévy driven stochastic differential equations: limit theorems. *The Annals of Probability*, 32(3A), 1830–1872. <https://doi.org/10.1214/009117904000000667>
- Knopova, V., & Platonov, D. (2025). *Strong convergence rates for euler schemes of levy-driven sde using dynamic cutting*. <https://doi.org/10.48550/arXiv.2504.11988>
- Kühn, F., & Schilling, R. L. (2023). Maximal inequalities and some applications. *Probability Surveys*, 20, 382–485. <https://doi.org/10.1214/23-ps17>
- Kunita, H. (2004). Stochastic differential equations based on Lévy processes and stochastic flows of diffeomorphisms. In *Real and stochastic analysis* (pp. 305–373). Birkhäuser Boston.
- Novikov, A. A. (1975). On discontinuous martingales. *Theory of Probability & Its Applications*, 20(1), 11–26. <https://doi.org/10.1137/1120002>
- Protter, P., & Talay, D. (1997). The Euler scheme for Lévy driven stochastic differential equations. *The Annals of Probability*, 25(1), 393–423. <https://doi.org/10.1214/aop/1024404293>
- Situ, R. (2005). *Theory of stochastic differential equations with jumps and applications*. Springer. <https://doi.org/10.1007/b106901>

Отримано редакцією журналу / Received: 12.09.25

Прорецензовано / Revised: 03.12.25

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Денис ПЛАТОНОВ, асп.

ORCID ID: 0009-0005-3660-5516

e-mail: dplatonov@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

АНАЛІЗ СЛАБКОЇ ЗБІЖНОСТІ ДИНАМІЧНОЇ ЗРІЗКИ ДЛЯ ПРОЦЕСІВ ТИПУ ЛЕВІ

Отримано оцінки слабкої збіжності для схеми Ейлера – Маруями одновимірних СДР, керованих процесом Леві, у яких малі стрибки динамічно зрізаються залежно від часу. Великі стрибки моделюються точно, тоді як малі стрибки (i) або повністю ігноруються, або (ii) замінюються нормальною випадковою величиною з відповідною дисперсією. За стандартних умов Ліпшиця та гладкості коефіцієнтів і міри Леві ми доводимо, що слабка похибка обох схем має порядок $\mathcal{O}(n^{-1})$. Цей порядок досягається вибором гіперпараметра масштабування $h = n^{-\alpha/(\varepsilon(2-\alpha))}$ для варіанта (i) та $h = n^{-\alpha/(\varepsilon(3-\alpha))}$ для варіанта (ii).

Ключові слова: динамічна зрізка, процеси типу Леві, схема Ейлера – Маруями.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.11>

Javier M. HUARCA, PhD, Prof.

ORCID ID: 0000-0002-1564-2343

e-mail: jhuarcao@usmp.pe

Universidad de San Martín de Porres, Lima, Peru

THE H PROBABILITY DISTRIBUTIONS

Probability distributions are important models in all scientific fields; however, they are essential in statistics. We provide a new class of mixed joint probability distributions with probability spaces that are sequences of nested piled $2D$ -lattices. The methodology for producing these models involves embedding a discrete variable extended β -Benford's distribution into a continuous variable modified Bernstein polynomial. β is the parameter of this family of distributions and denotes the base of an arbitrary number system. Our purpose was to describe the mathematical structure and statistical features behind the interactions of these probability distributions. On the basis of the theoretical development and its results, we conclude that these models are valuable for numerous applications in various scientific disciplines. A concrete decimal system case is introduced to support our hypothesis.

Key words: extended β -Benford distribution, Bernstein polynomials, number system, universal properties of a probability distribution, homotopic simplicial complex.

AMS 2020 classification: 60E05, 93A30; 62R40.

Introduction

Probability distributions have played an important role in all fields of science, especially statistics, since the sixteenth century, initiated by G. Cardano, followed by P. de Fermat and B. Pascal in the seventeenth century; this tradition is carried on to this day by scientists such as (Franklin, 2015; Debnath, & Basu, 2015). For example, after the discovery of the renowned Benford's distribution in 1938 (Benford, 1938), which originally went back to 1881, as surmised by astronomer (Newcomb, 1881). Since then, many papers and books supporting its theoretical and practical applications have been published in various fields of science. Namely, in statistics, mathematics, in business (accounting, economics, finance, and management), physics, chemistry, social science, computer science, etc. This means that there is rich and abundant literature on Benford's distribution.

Some researchers aimed to find generalizations of Benford's law, achieving fruitful results, such as (Fu, Shi, & Su, 2007), who proposed an interesting, first-digit, generalized Benford's law

$$p(x) = N \log_{10} \left(1 + \frac{1}{s + x^q} \right), \quad x = 1, 2, 3, 4, 5, 6, 7, 8, 9,$$

where N is a normalization factor that makes $p(x)$ to be a probability distribution, and s and q are some model parameters, see (Fu, Shi, & Su, 2007) for details. In addition, (Pietronero et al., 2001) showed that Benford's law can be generalized to a series of numbers generated from more complex systems, such as catalogs of seismic activity. (Barabesi, & Pratelli, 2020) provide some properties of the Generalized Benford Law, a flexible model for the distribution of significant digits, and abundant references in the introduction of their publication. Although there is abundant literature on this topic, it remains insufficient. This definition can be expanded to encompass broader domains. In addition, new probability distributions need to be identified, and their applicability extended beyond the current distributions to both classical and quantum fields. This need was one of the motivations for writing this paper.

In the introduction, we begin with a brief presentation of the well-known standard Benford probability distribution (SBPD) for a discrete random variable W and for any positive integer $\beta \geq 2$, defined by

$$B(w; \beta) = \log_{\beta} \left(1 + \frac{1}{w} \right), \quad w = 1, 2, \dots, \beta - 1, \quad (1)$$

where $\sum_{w=1}^{\beta-1} B(w; \beta) = 1$ and $B(w; \beta)$ is non-negative for all w , see the proof of (5). Note that for the decimal system, $\beta = 10$, w in $\{1, 2, 3, \dots, 9\}$ is a special case that our brain understands the most. We derive the expected value, variance, and moment generating function of this SBPD for any number of systems of base $\beta \geq 2$, and obtain the following:

$$E(w; \beta) = \log_{\beta} \left[\frac{\beta^{(\beta-1)}}{(\beta-1)!} \right], \quad (2)$$

$$Var(w; \beta) = \log_{\beta} \left[\frac{\beta^{(\beta-1)^2}}{\prod_{i=2}^{\beta} (i-1)^{2i-3}} \right] - \left[\log_{\beta} \left(\frac{\beta^{(\beta-1)}}{(\beta-1)!} \right) \right]^2, \quad (3)$$

$$M_W(t) = \log_{\beta} \left[\frac{\beta^{e^{(\beta-1)t}}}{\prod_{i=2}^{\beta} (i-1)^{e^{(i-1)t}} - e^{(i-2)t}} \right], \quad (4)$$

where e denotes the Euler number. We note that these impressive results depend only on the parameter β , which is the basis of an arbitrary number system. Namely, if we consider the binary number system then $\beta = 2$, and $E(w; 2) = 1$, $Var(w; 2) = 0$, $M_W(t) = e^t$ for real t .

In Section 1 we demonstrate the properties of the extended SBPD, defined in Equation (5), as a more general case to clarify the veracity of Equations (2), (3), and (4). To the best of our knowledge, we do not know whether these interesting results have been published in the literature.

Henceforth, we establish that parameter β functions as both the base of the numerical system and the base of the logarithm. Any system and logarithm base chosen are automatically convertible by computers to either a natural logarithm or a decimal logarithm during processing.

In this study, we provide a novel family of mixed joint probability distributions, denoted by $\mathcal{H}(x, w; \beta, j)$, associated with an extended β -Benford distribution, Equation (5), and special polynomials formulated for an arbitrary number system of bases $\beta \geq 2$. We show that each member of this family defines a mixed two-dimensional probability distribution with its own features and characteristics. We then provide a concrete example for $\beta = 10$ in the decimal system to clarify the theory.

In other words, the primary goal of this research is to introduce finite families of mixed probability distributions characterized by their nested-piled 2D-lattices and their features. It is organized as follows: Section 1 provides the proposed extended β -Benford probability distribution and its preliminary features and methodology. The crucial theoretical developments for a generalized $\mathcal{H}_G(x, w; \beta, j)$ distribution are formalized in Section 2 starting with Theorem 2 and its generalization Theorem 3, followed by several propositions. To clarify this theory, we choose the decimal system $\beta = 10$ as an application. All other bases work in the same manner. Section 4 discusses some theoretical and practical implications of the main $\mathcal{H}$ findings and concludes with a brief summary of the contributions of this study.

This study makes four contributions. First, a new generalized β -Benford distribution for all digits of any number system of bases β was formulated. This distribution constitutes one of the two key components of the structure of our models. Second, a family of 2-dimensional mixed joint probability distributions is introduced along with its statistical features. These are the models called the $\mathcal{H}$ distributions. Third, a connection between these $\mathcal{H}$ distributions and Bézier curves of degree $\beta - 1$ with β control points is provided. Fourth, an analytic allocation distribution function is established.

1. Material and Methodology

The Equations (2), (3), and (4) are the fundamental properties of the SBPD. This widely acknowledged SBPD is also referred to as the first non-zero digit law or the distribution of the most significant digits, as it is restricted to sets of data with non-zero first digits. This means that the digit zero of a number system is not an element of its domain. However, there are many sets of data in all fields of science whose leftmost digit begins with zero, namely: SSN, student ID, ORCID, bank accounts and loans, constants in physics, etc. Therefore, to examine the frequency of the first positional digit for this kind of data, we need to extend the domain in (1) of the SBPD to all digits of any number system.

Thus, to remedy these shortcomings, we provide an extended Benford probability distribution (EBPD), which includes all basic symbols of an arbitrary number system of base β , in the following theorem.

Theorem 1. Let W be a discrete random variable. W has an EBPD $\mathcal{H}(w; \beta)$ given by

$$\mathcal{H}(w; \beta) = \frac{1}{\log_{\beta}(\beta + 1)} \log_{\beta}\left(1 + \frac{1}{w + 1}\right), \quad w = 0, 1, 2, \dots, \beta - 1, \quad (5)$$

where $\beta \geq 2$ is a positive integer that plays the role of the base of a number system and, at the same time, is the base of the logarithm function, as we established above.

We clearly observe that this discrete variable W includes all the basic values of a numerical system. For example, if $\beta = 2$ or $\beta = 10$, we have the well-known binary ($w = 0, 1$) and decimal ($w = 0, 1, 2, \dots, 9$) systems and $\log_2(\cdot)$ and $\log_{10}(\cdot)$, respectively. It is understood that any logarithm function of base β is related to the natural logarithm $\ln$. Fig. 1 illustrates the shapes of these probability distributions for distinct positive integer values of $\beta \geq 2$. For example, for $\beta = 2$, $w = 0; 1$ with probability values $\mathcal{H}(0; 2) = 0.63$ and $\mathcal{H}(1; 2) = 0.37$. For $\beta = 3$, $w = 0; 1; 2$ with probabilities $\mathcal{H}(w; 3) = 0.5, 0.29, 0.21$ (or 50, 29, 21 %), respectively. Note that the total probability in these two examples, as well as in all other instances, is 1, as required, and decays exponentially.

For each value of β and $w = 0, 1, 2, \dots, \beta - 1$, gluing together consecutive points, we have topological building blocks called simplicial complexes. A 0-dimensional simplex is a vertex; a 1-dimensional simplex is a line segment with two vertices; a 2-dimensional simplex has three vertices and three edges, and so on. We clarify these topological objects a little more in Section 2. See (Singer, & Thorpe, 1967) for more details.

Proof. It is easy to verify that (5), as well as (1), satisfies the two universal properties of a probability distribution (p.d.). That is, $\mathcal{H}(w; \beta) \geq 0$ for all w and $\sum_{w=0}^{\beta-1} \mathcal{H}(w; \beta) = 1$. For the second property, when executing the summation, all terms cancel out each other except for the last term, $\log_{\beta}(\beta + 1)$, which survives, and then it cancels out with the constant of normality of (5). Also, to prove it, one should use the telescoping series. $\square$

Equation (5) is one of the two fundamental pieces in the construction of the distribution $\mathcal{H}_G(x, w; \beta, j)$. In what follows, we write $\log(\cdot)$ to denote $\log_{\beta}(\cdot)$. We now compute some natural properties of (5).

Proposition 1. Let W be a random variable distributed as in (5), then the expected value and the variance are given by

$$E(W; \beta) = \frac{1}{\log(\beta + 1)} \log\left(\frac{(\beta + 1)^{(\beta-1)}}{\beta!}\right), \quad (6)$$

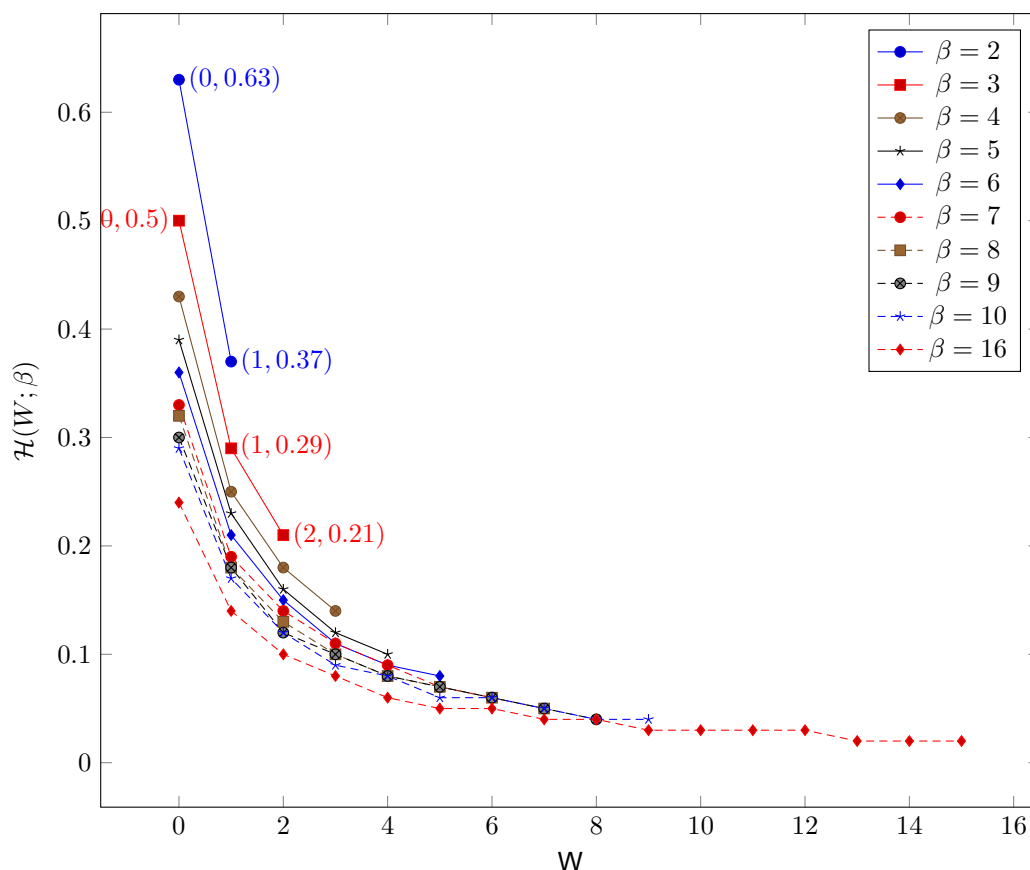


Fig. 1. Graphs of $\mathcal{H}(w; \beta)$ for distinct bases $\beta = 2, 3, \dots, 10, 16$.

$$Var(W; \beta) = \frac{1}{\log(\beta + 1)} \log \left[\frac{(\beta + 1)^{(\beta - 1)^2}}{\prod_{i=1}^{\beta} i^{2i - 3}} \right] - \left[\frac{1}{\log(\beta + 1)} \log \left(\frac{(\beta + 1)^{(\beta - 1)}}{\beta!} \right) \right]^2. \quad (7)$$

Proof. In general, one could show this proposition by finding the k th moment about the origin of the random variable W , usually denoted by μ'_k , the expected value of W^k .

Symbolically, for $\beta \geq 2, w = 0, 1, 2, \dots, \beta - 1$, we have

$$\begin{aligned} E(W^k) &= \frac{1}{\log(\beta + 1)} \sum_{w=0}^{\beta-1} w^k \log \left(1 + \frac{1}{w+1} \right) = \frac{1}{\log(\beta + 1)} \log \left[\prod_{w=0}^{\beta-1} \left(\frac{w+2}{w+1} \right)^{w^k} \right] = \\ &= \frac{1}{\log(\beta + 1)} \log \left[\frac{(\beta + 1)^{(\beta-1)^k}}{\prod_{i=1}^{\beta} i^{(i-1)^k - (i-2)^k}} \right], \end{aligned}$$

where $\mu'_1 = E(W; \beta)$ when $k = 1$ and $Var(W; \beta) = \mu'_2 - (\mu'_1)^2$ when $k = 1, 2$. □

We can notice that as β increases, both the expectation and the variance increase.

Proposition 2. Let W be a random variable whose probability distribution is given in (5), then its moment generating function, for all real values of t is given by

$$M_W(t) = \frac{1}{\log(\beta + 1)} \log \left[\frac{(\beta + 1)^{e^{(\beta-1)t}}}{\prod_{i=1}^{\beta} i^{e^{(i-1)t} - e^{(i-2)t}}} \right], \quad \beta \geq 2, t \in (-\infty, +\infty). \quad (8)$$

Proof. We use the definition of the moment-generating function

$$M_W(t) = E(e^{tW}) = \sum_{w=0}^{\beta-1} e^{tw} \frac{1}{\log(\beta + 1)} \log \left(1 + \frac{1}{w+1} \right) = \frac{1}{\log(\beta + 1)} \log \left[\prod_{w=0}^{\beta-1} \left(\frac{w+2}{w+1} \right)^{e^{tw}} \right].$$

Expanding the product operator, on the right-hand side, associating the exponents of each pair of consecutive factors, and simplifying, yields the proof. □

Of course, from this $M_W(t)$, we can generate all of the moments of W by successively differentiating $M_W(t)$, with respect to t , and then evaluating the result at $t = 0$. This is another way to prove the Proposition 1. We leave this to the reader. Notice that $M_W(0) = 1$, for every $\beta \geq 2$ and $t \in (-\infty, +\infty)$.

2. Main theory and results

In this section, following all notations as above, we formulate the probability distribution $\mathcal{H}(\cdot)$ of two random variables. It is a joint distribution of a mixed random vector (X, W) that is described in the next theorem.

Theorem 2. A random vector (X, W) has a mixed joint probability distribution, denoted by $\mathcal{H}(x, w; \beta)$, with parameter $\beta \geq 2$ if its joint distribution is given by

$$\mathcal{H}(x, w; \beta) = \binom{\beta-1}{w} \frac{x^w (\beta-x)^{\beta-w-1}}{\beta^{\beta-1} \log(\beta+1)} \log\left(1 + \frac{1}{w+1}\right), \quad w = 0, 1, 2, \dots, \beta-1. \quad 0 \leq x \leq \beta. \quad (9)$$

It is a bivariate distribution of a continuous random variable X and a discrete random variable W , sitting on a mesh domain. The foundation of $\mathcal{H}(x, w; \beta)$ is the result of embedding the EBPd defined in (5) with a generalization of β -Bernstein basis polynomials of degree $\beta-1$ over an interval $[0, \beta]$, for any number system of base $\beta \geq 2$. This theorem is the backbone of this research from which abundant theoretical information emerges. This probability distribution has certain properties and characteristics that, traditionally, any probability distribution must have. These include the cumulative distribution, marginals, conditionals, expected values, variance, covariance, correlation coefficient, moment generating function, characteristic function, etc. Our goal in this section is to investigate and characterize the mathematical and statistical structures underlying these probability distributions. Primarily, $\mathcal{H}(x, w; \beta)$ needs to satisfy the universal properties:

$$(i) \quad \mathcal{H}(x, w; \beta) \geq 0 \quad \text{for all values of } (x, w) \quad \text{and} \quad (ii) \quad \sum_{w=0}^{\beta-1} \int_0^{\beta} \mathcal{H}(x, w; \beta) dx = 1.$$

Proof. The property (i) is clear, since all factors of $\mathcal{H}(x, w; \beta)$ are positive for each value of (x, w) on their domains. For (ii) the key step is to solve the integral $\int x^w (\beta-x)^{\beta-w-1} dx$. To solve it, one may use any mathematical software. Here we use (Maple.Soft, 2024), obtaining

$$\beta^{\beta-w-1} \frac{1}{w+1} x^{w+1} \text{hypergeom}([w+1, -\beta+w+1], [2+w], \frac{x}{\beta}),$$

then, evaluating at the limits of the integral $x = \beta$ and $x = 0$, we have $\beta^\beta \Gamma(\beta-w) \Gamma(w+1) / \Gamma(\beta+1)^1$. Now, replacing this result in the Equation (ii) then expanding the binomial coefficient and simplifying it turns out

$$\sum_{w=0}^{\beta-1} \frac{\log(1 + \frac{1}{w+1})}{\log(\beta+1)} = 1,$$

which is the case of Equation (5) and completes the proof. $\square$

Proposition 3. Let (X, W) be a random vector with joint probability distribution established in Theorem 2 for $\beta \geq 2$, then the marginal probability density function of X and the marginal probability mass function of W are, respectively,

$$\mathcal{H}(x; \beta) = \sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \frac{x^w (\beta-x)^{\beta-w-1}}{\beta^{\beta-1} \log(\beta+1)} \log\left(1 + \frac{1}{w+1}\right), \quad 0 \leq x \leq \beta, \quad (10)$$

$$\mathcal{H}(w; \beta) = \frac{1}{\log(\beta+1)} \log\left(1 + \frac{1}{w+1}\right), \quad w = 0, 1, 2, \dots, \beta-1. \quad (11)$$

Proof. Obviously, to be $\mathcal{H}(x; \beta)$ and $\mathcal{H}(w; \beta)$ probability distribution, by themselves, they must satisfy the universal properties (i): non-negativity and (ii): total probability. In both cases the non-negativity is clear. For the (ii) the Equation (11), $\mathcal{H}(w; \beta) = \frac{1}{\log(\beta+1)} \log(1 + \frac{1}{w+1})$, $w = 0, 1, 2, \dots, \beta-1$, was shown in (5) that $\mathcal{H}(w; \beta)$ satisfies the universal property. That is, $\sum_{w=0}^{\beta-1} \mathcal{H}(w; \beta) = 1$. Thus, it qualifies to be a discrete marginal probability distribution of $\mathcal{H}(x, w; \beta)$. To show that (10) is a probability density function (p.d.f) it can be written as

$$\mathcal{H}(x; \beta) = \sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \frac{x^w (\beta-x)^{\beta-w-1}}{\beta^{\beta-1}} \mathcal{H}(w; \beta), \quad 0 \leq x \leq \beta. \quad (12)$$

Now, integrating both sides, we solve the definite integral $\int_0^{\beta} \mathcal{H}(x; \beta) dx$ of Equation (12) as we did in the Theorem 2 and simplifying, we get the result $\int_0^{\beta} \mathcal{H}(x; \beta) dx = 1$. $\square$

¹ $\Gamma(m) = (m-1)!$ is the discrete gamma function.

We show a special case and its applications in the next section. We continue this section showing some standard properties of $\mathcal{H}(x, w; \beta)$.

Proposition 4. Consider all the random variables satisfying assumption of Proposition 1 Then, the means and the variances of X and W are, respectively,

$$E(x; \beta) = \frac{\beta}{(\beta + 1)\log(\beta + 1)} \log\left[\frac{(\beta + 1)^\beta}{\beta!}\right], \quad \beta \geq 2, \quad (13)$$

$$Var(x; \beta) = \frac{\beta^2}{(\beta + 1)(\beta + 2)\log(\beta + 1)} \log\left[\frac{(\beta + 1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}}\right] - \left(\frac{\beta}{(\beta + 1)\log(\beta + 1)} \log\left[\frac{(\beta + 1)^\beta}{\beta!}\right]\right)^2. \quad (14)$$

Proof. The expected value $E(w; \beta)$ and the variance $Var(w; \beta)$ have already been proven in Proposition 1. To demonstrate (13) and (14) we first determine the k th moment about the origin of X . Consider Theorem 2 and follow the strategy that we have done in Proposition 1

$$E(X^k) = \sum_{w=0}^{\beta-1} \int_0^\beta x^k \mathcal{H}(x, w; \beta) dx = \sum_{w=0}^{\beta-1} \int_0^\beta \binom{\beta-1}{w} \frac{x^{w+k} (\beta-x)^{\beta-w-1}}{\beta^{\beta-1} \log(\beta+1)} \log\left(1 + \frac{1}{w+1}\right) dx. \quad (15)$$

Solving the integral in (15), with (Maple.Soft, 2024), and simplifying, we obtain

$$\begin{aligned} E(X^k) &= \sum_{w=0}^{\beta-1} \frac{\beta^{k+1} \Gamma(\beta) \Gamma(w+k+1)}{\Gamma(w+1) \Gamma(\beta+k+1)} \frac{\log(1 + \frac{1}{w+1})}{\log(\beta+1)} = \frac{\beta^k \beta!}{(\beta+k)! \log(\beta+1)} \sum_{w=0}^{\beta-1} \log\left[\frac{w+2}{w+1}\right]^{\frac{(w+k)!}{w!}} = \\ &= \frac{\beta^k \beta!}{(\beta+k)! \log(\beta+1)} \log\left[\prod_{w=0}^{\beta-1} \left[\frac{w+2}{w+1}\right]^{\frac{(w+k)!}{w!}}\right] = \theta_k(\beta) \log\left[\frac{(\beta+1)^{\frac{(k+\beta-1)!}{(\beta-1)!}}}{\prod_{i=1}^{\beta} i^{\frac{k(k+i-2)!}{(i-1)!}}}\right], \quad \beta \geq 2, \end{aligned}$$

where $\theta_k(\beta) = \frac{\beta^k \beta!}{(\beta+k)! \log(\beta+1)}$. To obtain the last expression, of the right hand, expand the product for w then find the pattern of the factors and combine their exponents.

Now, from the result of (15), we find $E(X)$, $E(X^2)$ and $Var(X) = E(X^2) - (E(X))^2$ as special cases. That is, for $k = 1$,

$$\theta_1(\beta) = \frac{\beta}{(\beta+1)\log(\beta+1)}$$

and for $k = 2$,

$$\theta_2(\beta) = \frac{\beta^2}{(\beta+1)(\beta+2)\log(\beta+1)}.$$

Thus, the mean value in (13) and the variance in (14) are, respectively,

$$E(x; \beta) = \theta_1(\beta) \log\left[\frac{(\beta+1)^\beta}{\beta!}\right], \quad \beta \geq 2,$$

$$Var(x; \beta) = \theta_2(\beta) \log\left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}}\right] - (\theta_1(\beta) \log\left[\frac{(\beta+1)^\beta}{\beta!}\right])^2.$$

□

Additional key features of this $\mathcal{H}(x, w; \beta)$ distribution include its covariance and the linear correlation coefficient, which we present in a generalized form in Equations (25) and (26), respectively.

3. Generalization of Theorem 2

Now we present an analogue of Theorem 2 and more general features of the $\mathcal{H}$ -distribution. That is, the Theorem 2, Equation (9), can be generalized over an underlying family of nested $2D$ -lattices $\Omega_\beta(j) = \{w\} \times [0, \beta - j]$ for each $j, w = 0, 1, \dots, \beta - 1$ as follows.

Theorem 3. Consider all the data from Theorem 2 and let $j = 0, 1, \dots, \beta - 1$, then the generalized version of Theorem 2 is given by

$$\mathcal{H}_G(x, w; \beta, j) = \binom{\beta-1}{w} \frac{\beta x^w (\beta-j-x)^{\beta-w-1}}{(\beta-j)^\beta} \frac{\log(\frac{w+2}{w+1})}{\log(\beta+1)}, \quad w = 0, 1, \dots, \beta-1, 0 \leq x \leq \beta-j. \quad (16)$$

For any number system of base $\beta \geq 2$ and each fixed j there is a generalized mixed probability distribution, denoted by $\mathcal{H}_G(\cdot)$, defined in a sample space $\Omega_\beta(j) = \{(x, w) : x \in [0, \beta - j], w = 0, 1, 2, \dots, \beta - 1\}$. In other words, given β , for each $j \in \{0, 1, 2, \dots, \beta - 1\}$ there is a probability distribution $\mathcal{H}_G(x, w; \beta, j)$ which is a homogeneous polynomial of degree $\beta - 1$ with β monomial terms, one term for each value of w . See Tab. 1 of Section 4 and see Fig. 2 for $\beta = 10$. Note that when $j = 0$, Theorem 2 shows up with all its features discussed above. Also, when $j = \beta - 1$ this generalization $\mathcal{H}_G(\cdot)$ becomes a standard Bernstein polynomial of degree $\beta - 1$ defined on the interval $[0, 1]$ or Bézier curves with β control points given by the marginal discrete probability distribution $\mathcal{H}_G(W; \beta, j)$ of Equation (19) and the curves are given by the marginal density probability distribution $\mathcal{H}_G(X, \beta, j)$ of Equation (18).

We have found, from this generalization, sets of families of probability distributions, one family for each fixed base β , with β members in the family. It is well-known that there are infinitely many number systems. Hence, we have infinitely many families of probability distributions available. The news is good!. For example, for the family called the decimal system $\beta = 10$, there are 10 probability distributions, one for each j . Thus, with $j = 0, 1, 2, \dots, 9$, each probability distribution has 10 basis polynomials of degree 9 each, one monomial term for each w with $w = 0, 1, 2, \dots, 9$. So, this example gives us 10^2 basis polynomials. See Tab. 1. We elaborate an example in the next section. All these probability distributions of number systems are sitting over probability spaces $(\Omega_\beta, \mathcal{F}_\beta, \mathbb{P}_\beta)$ of descending sequence of nested $2D$ -lattices (or filtrations)

$$\{w\} \times [0, \beta] \supset \{w\} \times [0, \beta - 1] \supset \dots \supset \{w\} \times [0, \beta - j] \supset \dots \supset \{w\} \times [0, 1], \quad w = 0, 1, \dots, \beta - 1$$

for each β and each $j = 0, 1, 2, \dots, \beta - 1$, respectively. These sequences are dependent on the parameter β of the number system. We point out that these families of sets of descending chains are homotopic topological spaces that hold some interesting properties of topological invariance. We briefly explain these properties in the next section.

Proof. The Theorem 3, Equation (16), must satisfy the universal properties of a probability distribution. That is,

$$(i) \mathcal{H}_G(x, w; \beta, j) \geq 0, \text{ for all } (x, w),$$

$$(ii) \sum_{w=0}^{\beta-1} \int_0^{\beta-j} \binom{\beta-1}{w} \frac{\beta x^w (\beta - j - x)^{\beta-w-1}}{(\beta - j)^\beta} \frac{\log(1 + \frac{1}{w+1})}{\log(\beta + 1)} dx = 1. \quad (17)$$

(i) is obvious. To show (ii), one may proceed as in Theorem 2. We first solve the integral $\int x^w (\beta - j - x)^{\beta-w-1} dx$ which results

$$\frac{(\beta - j)^{\beta-w-1} x^{w+1}}{w+1} \text{hypergeom}([w+1, -\beta+w+1], [w+2], \frac{x}{\beta-j}).$$

Then we evaluate in the limits of the integral $x = \beta - j$ and $x = 0$, then simplifying it results $\frac{(\beta-j)^\beta \Gamma(\beta-w) \Gamma(w+1)}{\Gamma(\beta+1)}$.

Now, replace this result into (17), then run the sum and simplify to get 1. $\square$

Same to Proposition 3, for any base $\beta \geq 2$, the marginal distributions for X and W , respectively, are given by

$$\mathcal{H}_G(x; \beta, j) = \sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \frac{\beta x^w (\beta - j - x)^{\beta-w-1}}{(\beta - j)^\beta \log(\beta + 1)} \log(1 + \frac{1}{w+1}), \quad 0 \leq x \leq \beta - j, \quad j = 0, \dots, \beta - 1, \quad (18)$$

$$\mathcal{H}_G(w; \beta, j) = \frac{1}{\log(\beta + 1)} \log(1 + \frac{1}{w+1}), \quad w = 0, 1, 2, \dots, \beta - 1. \quad (19)$$

Equations (5) and (11) already explain (19). To clarify (18), let's write it as follows

$$\mathcal{H}_G(x; \beta, j) = \sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \frac{\beta x^w (\beta - j - x)^{\beta-w-1}}{(\beta - j)^\beta} \psi_\beta(w), \quad 0 \leq x \leq \beta - j, \quad (20)$$

where $\psi_\beta(w)$ is discussed in Proposition 3. With the expression (20) we have found a family of generalization of Bézier curves of degree $\beta - 1$ with β control points given by $\psi_\beta(w)$ for each w satisfying the condition $\sum_{w=0}^{\beta-1} \psi_\beta(w) = 1$. But, in this case, these expressions are families of probability distributions. Furthermore, the equations

$$\mathcal{B}(x, w; \beta, j) = \binom{\beta-1}{w} \frac{\beta x^w (\beta - j - x)^{\beta-w-1}}{(\beta - j)^\beta}, \quad j, w = 0, 1, 2, \dots, \beta - 1 \quad (21)$$

are, for each fixed j in $\{0, 1, 2, 3, \dots, \beta - 1\}$, the well-known Bernstein basis polynomials in a general form that we have extended over an arbitrary real lattice $\{w\} \times [0, \beta - j]$ which includes the interval $\{w\} \times [0, 1]$ as a special case. *Claim:* All theories developed about Bernstein polynomials, Bézier curves, splines, binomial distribution, and relations to Gamma, Beta, Hypergeometric, etc. and their applications are also applicable for this new distribution of probabilities $\mathcal{H}_G(X, W; \beta, j)$. Explicitly, the distribution in (20) seen from the Bernstein basis polynomial point of view of degree $\beta - 1$ forms a *partition of unity*, as we tested above. Subsequently, we present propositions that delineate the fundamental statistical characteristics of the generalized distribution $\mathcal{H}_G(X, W; \beta, j)$ as defined in (16).

Proposition 5. Let $\mathcal{H}_G(X, W; \beta, j)$ be a generalized p.d. of the random vector (X, W) defined in (16) with marginal distributions given in (18) and (19), respectively. The mean and variance of the random variable X from (18) are given by

$$E(X; \beta, j) = \frac{\beta - j}{(\beta + 1) \log(\beta + 1)} \log\left[\frac{(\beta + 1)^\beta}{\beta!}\right], \quad j = 0, 1, 2, \dots, \beta - 1, \quad (22)$$

$$\text{Var}(X; \beta, j) = \frac{(\beta - j)^2}{(\beta + 1)(\beta + 2) \log(\beta + 1)} \log\left[\frac{(\beta + 1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}}\right] - \left(\frac{\beta - j}{(\beta + 1) \log(\beta + 1)} \log\left[\frac{(\beta + 1)^\beta}{\beta!}\right]\right)^2. \quad (23)$$

Proof. At this point, it is worthwhile to notice that the random variable W is immune to this generalization, however the random variable X is infected, because of the changes in its domain for each value of j . It means that all properties of

$\mathcal{H}(w; \beta)$, inclusive the $E(W^k)$ for $k = 1, 2, 3, \dots$ and variance, remains as in Proposition 1. To prove (22) and (23) we just need to find the k th moment of the marginal p.d. $\mathcal{H}(x; \beta, j)$ given in (18), and proceed as in the case of Proposition 4.

$$E(X^k) = \int_0^{\beta-j} X^k \mathcal{H}_G(x; \beta, j) dx = \int_0^{\beta-j} \sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \frac{\beta x^{k+w} (\beta-j-x)^{\beta-w-1}}{(\beta-j)^\beta \log(\beta+1)} \log\left(\frac{w+2}{w+1}\right) dx.$$

Assuming convergence, first, we solve the integral to get

$$E(X^k) = \frac{(\beta-j)^k \beta!}{(\beta+k)! \log(\beta+1)} \sum_{w=0}^{\beta-1} \frac{(k+w)!}{w!} \log\left(\frac{w+2}{w+1}\right) = \theta_k(\beta, j) \log\left[\frac{(\beta+1)^{\frac{(\beta+k-1)!}{(\beta-1)!}}}{\prod_{i=1}^{\beta} i^{(i+k-2)! \frac{k}{(i-1)!}}}\right], \quad j = 0, 1, \dots, \beta-1, \quad (24)$$

where $\theta_k(\beta, j) = \frac{(\beta-j)^k \beta!}{(\beta+k)! \log(\beta+1)}$ and $k = 1, 2, 3, \dots$

Thus, from this result in (24), we get $E(X; \beta, j)$ for $k = 1$, and $E(X^2)$ for $k = 2$ and $Var(X, \beta, j)$, respectively. $\square$

We clearly observe that Equations (13) and (14) of Proposition 4 emerge from Equations (22) and (23) when j is 0.

Proposition 6. Let $\mathcal{H}_G(x, w; \beta, j)$ be a joint mixed p.d. defined in (16) with $\beta \geq 2$. Then, for $j = 0, 1, \dots, \beta-1$, the covariance and correlation coefficient² between X and W are given, respectively, by

$$Cov(X, W; \beta, j) = \frac{(\beta-j) \log\left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{(2i-2)}}\right]}{(\beta+1) \log(\beta+1)} - \frac{(\beta-j) \log\left[\frac{(\beta+1)^\beta}{\beta!}\right] \log\left[\frac{(\beta+1)^{(\beta-1)}}{\beta!}\right]}{(\beta+1) (\log(\beta+1))^2}, \quad (25)$$

$$Corr(X, W; \beta, j) = \frac{Cov(X, W; \beta, j)}{\sqrt{\left[\frac{(\beta-j)^2 \log\left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}}\right]}{(\beta+1)(\beta+2) \log(\beta+1)} - \frac{(\beta-j)^2 (\log\left[\frac{(\beta+1)^\beta}{\beta!}\right])^2}{(\beta+1)^2 (\log(\beta+1))^2}\right] \left[\frac{\log\left[\frac{(\beta+1)^{(\beta-1)^2}}{\prod_{i=1}^{\beta} i^{(2i-3)}}\right]}{\log(\beta+1)} - \frac{(\log\left[\frac{(\beta+1)^{(\beta-1)}}{\beta!}\right])^2}{(\log(\beta+1))^2}\right]}}. \quad (26)$$

Proof. The calculations for $E(W)$, $Var(W)$ in Equations (6) and (7), as well as $E(X)$, $Var(X)$ in Equations (22) and (23), and $E(XW)$, combined with substitutions in the standard formulas for covariance and correlation coefficients, make up the proof. See the complete proof in the Appendix 4: proofs. $\square$

We have validated the findings with Maple 2024. Note that the effect of replacing Equation (25) in the numerator of Equation (26) produces the cancellation of the factors $(\beta-j)$ in Equation (26). It means that $Corr(X, W; \beta, j)$ depends only on parameter β . That is to say, the relationship between all members of a family β is the same. See details in Appendix 4, Equation 39. In other words, for each given β the β -distributions, for $j = 0, 1, 2, \dots, \beta-1$, have the same correlation coefficient. As it could be, in practice, all members of a family should satisfy the same properties. We illustrate the decimal system in the Section 4.

Finally, but not least, we present the marginal characteristic functions of $\mathcal{H}_G(x, w; \beta, j)$ as an alternative to their cumulative probability distribution functions.

Proposition 7. Let $\mathcal{H}_G(x, w; \beta, j)$ be a generalized p.d. defined in (16) with marginals distributions given in (18) and (19). Then the characteristic function $\varphi: \mathbb{R} \mapsto \mathbb{C}$ are, respectively, given by

$$\varphi_X(t) = E(e^{itX}) = \sum_{w=0}^{\beta-1} {}_1F1[w+1, \beta+1, i(\beta-j)t] \mathcal{H}_G(w; \beta), \quad t \in \mathbb{R}, \quad (27)$$

$$\varphi_W(t) = E(e^{itW}) = \frac{1}{\log(\beta+1)} \log\left[\frac{(\beta+1)^{e^{(\beta-1)it}}}{\prod_{r=2}^{\beta} r^{e^{(r-1)it} - e^{(r-2)it}}}\right]. \quad (28)$$

where i is the complex imaginary unit and $t \in \mathbb{R}$ is the argument of the characteristic function. $\mathcal{H}_G(w, \beta)$ is given in (19), (11) and (5). ${}_1F1[w+1, \beta+1, i(\beta-j)t]$ is the Kummer confluent hypergeometric function, see (Wolfram.Soft, 2025). All other notations are already presented above.

Proof. The proof of the characteristic function given in Equation (27) follows by solving the definite integral $\int_0^{\beta-j} e^{itx} \beta x^w \times (\beta-j-x)^{\beta-w-1} dx$. On this occasion, (Mathematica.Soft, 2024) is our ally. We have

$$\left(\frac{1}{\beta-j}\right)^{-w} (\beta-j)^{\beta-w} \beta \Gamma(\beta-w) \Gamma(w+1) \text{Hypergeometric1F1Regularized}[w+1, \beta+1, i(\beta-j)t],$$

$$\text{if } Re(\beta) > Re(w) > -1.$$

²We compute the covariance $Cov(X, W)$ and correlation coefficient $Corr(X, W)$ using K. Pearson's product moment because both random variables are quantitative. However, due to the mixed nature of the variables, a point multi-serial correlation coefficient could be considered, but it is not the focus of this study. Interested readers can see (Wiley.StatsRef, 2014; Tate, 1954).

Next, consider the relation $Hypergeometric1F1Regularized[w+1, \beta+1, (\beta-j)it] = \frac{{}_1F1[w+1, \beta+1, (\beta-j)it]}{\Gamma(\beta+1)}$. (Wolfram.Soft, 2025) and the result of the integral to replace in

$$\sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \beta \int_0^{\beta-j} \frac{e^{itx} x^w (\beta-j-x)^{\beta-w-1}}{(\beta-j)^\beta} dx \mathcal{H}_G(w; \beta),$$

we have

$$\sum_{w=0}^{\beta-1} \binom{\beta-1}{w} \beta \frac{(\frac{1}{\beta-j})^{-w} (\beta-j)^{\beta-w} \Gamma(\beta-w) \Gamma(w+1)}{(\beta-j)^\beta} \frac{{}_1F1[w+1, \beta+1, (\beta-j)it]}{\Gamma(\beta+1)} \mathcal{H}_G(w; \beta).$$

We now carry out the binomial coefficient formula and express it as discrete gamma function notation

$$\binom{\beta-1}{w} \beta = \frac{(\beta-1)! \beta}{(\beta-w-1)! w!} = \frac{\Gamma(\beta+1)}{\Gamma(\beta-w) \Gamma(w+1)}.$$

Subsequently, substitute this last outcome into the penultimate expression and simplify to complete the proof of Equation (27). Proving (28) is simpler. Follow the proof of Proposition 2. $\square$

Remark 1. To find the mixed joint characteristic function

$$\varphi_{XW}(t, s) = E[\exp(i(tX + sW))],$$

where i is the imaginary unit, is a challenge exercise.

From these characteristic functions, one can derive all traditional concepts of probability distributions. Namely, all the moments of marginals of $\mathcal{H}_G(x, w; \beta, j)$ and joint moments can be generated from $\varphi(\cdot)$. We leave this challenge to the reader.

We may keep computing many more consequences and properties³ to describe the $\mathcal{H}_G(x, w; \beta, j)$ mixed joint probability distribution, but these are enough for our purposes in this paper and the following section. We have purposefully left some algebraic operations for the reader to perform in order to enhance their comprehension of the $\mathcal{H}$ distributions and to help them identify additional properties and applications across a variety of scientific disciplines. This is an important point to emphasize.

4. Example of decimal system $\beta = 10$

In this section, our purpose is to clarify the properties and usefulness of $\mathcal{H}_G(X, W; 10, j)$. We summarize the results in Tab. 1. From (16) for $\beta = 10$, summing over w , we have

$$\mathcal{H}_G(X; 10, j) = \sum_{w=0}^9 \frac{\binom{9}{w} 10x^w (10-j-x)^{9-w}}{(10-j)^{10}} \psi_{10}(w), \quad j = 0, 1, 2, \dots, 9. \quad 0 \leq x \leq 10-j.$$

Is is clear that there are 10 probability distributions; one homogeneous polynomial of degree 9 for each j in $\{0, 1, 2, \dots, 9\}$. See Tab. 1. To get the $E(X; 10, j)$ and $Var(X; 10, j)$ of $\mathcal{H}_G(X; 10, j)$ let's write Equations (22) and (23) as follows

$$(i) \quad E(X; \beta, j) = (\beta-j) \Delta_1(\beta) \quad \text{and} \quad (ii) \quad Var(X; \beta, j) = (\beta-j)^2 \Delta_2(\beta) - [(\beta-j) \Delta_1(\beta)]^2 \quad (29)$$

where

$$\Delta_1(\beta) = \frac{1}{(\beta+1) \log(\beta+1)} \log\left[\frac{(\beta+1)^\beta}{\beta!}\right]$$

and

$$\Delta_2(\beta) = \frac{1}{(\beta+2)(\beta+1) \log(\beta+1)} \log\left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}}\right].$$

Thus for $\beta = 10$ we have

$$\Delta_1(10) = \frac{1}{11 \log(11)} \log\left[\frac{11^{10}}{10!}\right] \approx 0.3364518821$$

and

$$\Delta_2(10) = \frac{1}{132 \log(11)} \log\left[\frac{11^{110}}{1^2 2^4 3^6 4^8 \dots 10^{20}}\right] \approx 0.1883053492.$$

Replacing $\Delta_1(10)$ and $\Delta_2(10)$, in (29), we get 10 different expected values, 10 variances, 10 covariances, one for each j and a single correlation coefficient, due to (40). See Tab. 1. Simplifying (i) and (ii) of (29) with (Maple.Soft, 2024) we get simpler formulas

$$E(x; \beta, j) = (\beta-j) \Delta_1(\beta) = 3.364518821 - 0.3364518821j, \quad j = 0, 1, \dots, 9,$$

³We trust that readers will have the opportunity to continue developing more properties and applications of this machine built for the advancement of probability and statistics and related fields.

$$Var(x; \beta, j) = (\beta - j)^2 [\Delta_2(\beta) - \Delta_1^2(\beta)] = 0.07510548034(-10 + j)^2, \quad j = 0, 1, 2, \dots, 9.$$

Now we describe the relationships between X and W . From Equations (25) and (26) we have

$$Cov(X, W; 10, j) = \frac{(10 - j) \log \left[\frac{(11)^{(9)(10)}}{\prod_{i=1}^{10} i^{(2i-2)}} \right]}{(11) \log(11)} - \frac{(10 - j) \log \left[\frac{(11)^{10}}{10!} \right] \log \left[\frac{(11)^{(9)}}{10!} \right]}{(11)(\log(11))^2}, \text{ for } j = 0, \dots, 9.$$

Evaluating at each j in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and rounding we get:

j	0	1	2	3	4	5	6	7	8	9
$Cov(\cdot)$	6.780	6.102	5.424	4.746	4.068	3.390	2.712	2.034	1.356	0.678

Another way to find these results is by simplifying with (Maple.Soft, 2024) the $Cov(\cdot)$ for j and get

$$Cov(X, W; 10, j) = 6.780137493 - 0.6780137493 \times j \quad \text{for } j = 0, 1, 2, \dots, 9,$$

from which one can verify these results.

Next we calculate the correlation coefficient with Equation (26), for $j = 0, 1, \dots, 9$:

$$Corr(X, W; 10, j) = \frac{Cov(X, W; 10, j)}{\sqrt{\left[\frac{(10-j)^2 \log \left[\frac{(11)^{(10)(11)}}{\prod_{i=1}^{10} i^{(2i-1)}} \right]}{(11)(12) \log(11)} - \frac{(10-j)^2 (\log \left[\frac{(11)^{10}}{10!} \right])^2}{(11)^2 (\log(11))^2} \right] \left[\frac{\log \left[\frac{(11)^{(9)(9)}}{\prod_{i=1}^{10} i^{(2i-3)}} \right]}{\log(11)} - \frac{(\log \left[\frac{(11)^{(9)}}{10!} \right])^2}{(\log(11))^2} \right]}}.$$

Simplifying, with Maple⁴, surprisingly we found that $Corr(X, W; 10, j) = 0.9059142578 \text{csgn}(10 - j)$ or

$$Corr(X, W; 10, j) = \frac{1.336129624(6.780137493 - 0.6780137493j)}{\sqrt{(10 - j)^2}} \quad \text{for } j = 0, 1, 2, \dots, 9.$$

We note that all $Cov(X, W; 10, j)$ are positive and decay linearly as j gets closer to its highest value, $j \rightarrow \beta - 1$, and $Corr(X, W; 10, j) = +0.90591425$ are also positive invariant for all j in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ since $\text{csgn}(z)$ is ± 1 and all covariances are positive. We explain this fact above and further in the lines below of Equation (26).

Additionally, with the Proposition 7, Equation (27), we get the characteristic function⁵

$$\varphi_X(t; \beta, j) = \sum_{w=0}^9 1F1[w + 1, \beta + 1, (\beta - j)it] \mathcal{H}_G(W; \beta) \quad \text{for each } j = 0, 1, 2, \dots, 9,$$

where i is the imaginary unit of complex numbers. In particular, if $\beta = 10$, we have

$$\varphi_X(t; 10, j) = \sum_{w=0}^9 1F1[w + 1, 11, (10 - j)it] \mathcal{H}_G(W; 10) \quad \text{for each } j = 0, 1, 2, \dots, 9.$$

Table 1

Ten probability distributions $\mathcal{H}_G(X; 10, j)$ (left column), one for each j , and some properties for $\beta = 10$

j	$\mathcal{H}_G(X; 10, j)$	$E(X; 10)$	$Var(X; 10)$
0	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (10-x)^{9-w}}{10^{10}} \psi_{10}(w), 0 \leq x \leq 10$	$10\Delta_1(10)$	$100\Delta_2(10) - [10\Delta_1(10)]^2$
1	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (9-x)^{9-w}}{9^{10}} \psi_{10}(w), 0 \leq x \leq 9$	$9\Delta_1(10)$	$81\Delta_2(10) - [9\Delta_1(10)]^2$
2	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (8-x)^{9-w}}{8^{10}} \psi_{10}(w), 0 \leq x \leq 8$	$8\Delta_1(10)$	$64\Delta_2(10) - [8\Delta_1(10)]^2$
3	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (7-x)^{9-w}}{7^{10}} \psi_{10}(w), 0 \leq x \leq 7$	$7\Delta_1(10)$	$49\Delta_2(10) - [7\Delta_1(10)]^2$
4	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (6-x)^{9-w}}{6^{10}} \psi_{10}(w), 0 \leq x \leq 6$	$6\Delta_1(10)$	$36\Delta_2(10) - [6\Delta_1(10)]^2$
5	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (5-x)^{9-w}}{5^{10}} \psi_{10}(w), 0 \leq x \leq 5$	$5\Delta_1(10)$	$25\Delta_2(10) - [5\Delta_1(10)]^2$
6	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (4-x)^{9-w}}{4^{10}} \psi_{10}(w), 0 \leq x \leq 4$	$4\Delta_1(10)$	$16\Delta_2(10) - [4\Delta_1(10)]^2$
7	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (3-x)^{9-w}}{3^{10}} \psi_{10}(w), 0 \leq x \leq 3$	$3\Delta_1(10)$	$9\Delta_2(10) - [3\Delta_1(10)]^2$
8	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (2-x)^{9-w}}{2^{10}} \psi_{10}(w), 0 \leq x \leq 2$	$2\Delta_1(10)$	$4\Delta_2(10) - [2\Delta_1(10)]^2$
9	$\sum_{w=0}^9 \binom{9}{w} \frac{10x^w (1-x)^{9-w}}{1^{10}} \psi_{10}(w), 0 \leq x \leq 1$	$1\Delta_1(10)$	$1\Delta_2(10) - [1\Delta_1(10)]^2$

Based on the results of distinct cases ($\beta = 10$ is included), we dare to **Conjecture**: Let $Cov(X, W; \beta, j) = a \pm bj$ for any base $\beta \geq 2$ and let a and b be real numbers, $j = 0, 1, 2, \dots, \beta - 1$, then $Corr(X, W; \beta, j) = r \cdot \text{csgn}(\beta - j)$, where $|r| \leq 1$. The proof or disproof is in progress.

⁴The $\text{csgn}(z)$ function is used to determine in which half-plane ('left' or 'right') the complex-valued expression or number z lies.

⁵We leave the study and analysis of the characteristic function $\varphi_{XW}(t, s; \beta, j)$ of $\mathcal{H}_G(X, W; \beta, j)$ to the interested reader, which we will develop in another project.

We also illustrate some properties of the marginal distribution $\mathcal{H}_G(W; \beta, j)$ for $\beta = 10$. By Proposition 1 and Equations (6) and (7) we have

$$E(W; 10) = \frac{1}{\log(11)} \log\left[\frac{11^9}{10!}\right] \approx 2.70097070, \quad (30)$$

$$\text{Var}(W; 10) = \frac{1}{\log(11)} \log\left[\frac{11^{81}}{2^1 3^3 4^5 5^7 \dots 10^{17}}\right] - \left(\frac{1}{\log(11)} \log\left[\frac{11^{81}}{10!}\right]\right)^2 \approx 0.08729569, \quad (31)$$

and using Equation (28) from Proposition 7, we get

$$M_W(t) = \frac{1}{\log(11)} \log\left[\frac{11^{e^{9it}}}{2^{e^{it}-1} 3^{e^{2it}-e^{it}} \dots 10^{e^{9it}-e^{8it}}}\right]. \quad (32)$$

We observe that the behavior and properties of the marginal distribution $\mathcal{H}_G(W; \beta, j)$ are independent of j and are solely dependent on the parameter β . Consequently, they are distinct for each number system. In Tab. 1, each row constitutes a probability distribution over its specific sample space, pile nested lattices, $\{w\} \times [10 - j]$, for $w = 0, 1, 2, \dots, 9$ and each j in $\{0, 1, 2, \dots, 9\}$. Each p.d. is a homogeneous polynomial of degree 9, with 10 monomial terms⁶ and they satisfy the universal properties. Each term has an initial original contribution to the probability distribution. For instance, for $j = 0$, $0 \leq x \leq 10$ expanding the probability distribution $\mathcal{H}_G(x, w; 10, 0)$ given in the Tab. 1, then integrating each term, we have

$$\begin{aligned} \mathcal{H}_G(x, w; 10, 0) &= \overbrace{\frac{\log(2)}{10^9 \log(11)} (10 - x)^9}^{w=0} + \overbrace{\frac{9 \log(3/2)}{10^9 \log(11)} x (10 - x)^8}^{w=1} + \dots + \overbrace{\frac{\log(11/10)}{10^9 \log(11)} x^9}^{w=9}, \\ \int_0^{10} \mathcal{H}_G(x, w; 10, 0) dx &= \frac{\ln(2)}{\ln(11)} + \frac{\ln(3/2)}{\ln(11)} + \frac{\ln(4/3)}{\ln(11)} + \dots + \frac{\ln(11/10)}{\ln(11)} = \\ &= 0.2890648263 + 0.1690920837 + 0.1199727426 + \dots + 0.03974743221 = 1 \end{aligned}$$

or $\int_0^{10} \mathcal{H}_G(x, w; 10, 0) dx = 29 \% + 17 \% + 12 \% + \dots + 4 \% = 100 \%$.

See Fig. 2, which is an illustration for $j = 0, 3, 5, 9$. One can verify that $E(x, 10, j)$, $\text{Var}(x; 10, j)$, $\text{Cov}(x; 10, j)$ decay linearly, as $j \rightarrow \beta$, from the left, and β large. Each figure in Fig. 2 shows 10 smooth curves, one for each term of a homogeneous polynomial of degree 9. Each curve represents the path of a polynomial term obtained by replacing each value of $w \in \{0, 1, 2, \dots, 9\}$ in Equation (16). See also Tab. 1.

The distribution $\mathcal{H}_G(x, w; \beta, j)$ or more easily $\mathcal{H}_G(w; \beta)$ gives us precisely the contribution of each value of the random variable W and surprisingly, those are the Benford distribution probabilities, but counting all digits (0 is included) of a decimal system. It means, in finance, each value of W has a role to allocate a proportion of resources or investments. This function is non-negative and satisfies the universal property in its domain $[0, 10]$ for $j = 0$. For others values of $j = 1, 2, \dots, 9$, the same probabilities are redistributed. That is, for example, for $j = 9$ the distribution $\mathcal{H}_G(x, w; 10, 9)$ is on $[0, 1] \times \{w\}$ and integrating⁷ we have

$$\begin{aligned} \mathcal{H}_G(x, w; 10, 9) &= 10 \overbrace{\frac{\log(2)}{\log(11)} (1 - x)^9}^{w=0} + \overbrace{\frac{90 \log(3/2)}{\log(11)} x (1 - x)^8}^{w=1} + \dots + \overbrace{\frac{10 \log(11/10)}{\log(11)} x^9}^{w=9}, \\ \int_0^1 \mathcal{H}_G(x, w; 10, 9) dx &= 0.2890648263 + 0.1690920837 + \dots + 0.03974743221 = 1 \end{aligned}$$

or $\int_0^1 \mathcal{H}_G(x, w; 10, 9) dx = 29 \% + 17 \% + 12 \% + \dots + 4 \% = 100 \%$. We can arrange these integration results in the rows and columns of a 10×10 matrix whose entries are transition probabilities of a Markov chain with $\beta = 10$ states. The family of distributions $\mathcal{H}_G(x, w; 10, j)$ have 100 monomial terms, each of degree 9, and 10 homogeneous polynomials⁸ or probability distributions in their own probability space $(\Omega_{10}, \mathcal{F}_{10}, \mathbb{P}_{10})$ given by the nested, closed, and bounded intervals $[0, 10] \supset [0, 9] \supset \dots \supset [0, 1]$. These intervals are topological spaces with a rich set of properties, see (Munkres, 2000). Because the mixed variables in our study, namely, X is continuous and W is discrete, the underlying probability spaces are filtrations of a series of nested lattices. For instance, if $j = 0$ there is a rectangular lattice, with the right-hand rule orientation, $[0, 10] \times \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ with vertices at $(0, 0)$, $(10, 0)$, $(10, 9)$ and $(0, 9)$ and if $j = 9$ there is a rectangular lattice $[0, 1] \times \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ with vertices at $(0, 0)$, $(1, 0)$, $(1, 9)$ and $(0, 9)$. In Fig. 2, we provide a 3D geometric representation of $\mathcal{H}_G(x, w; 10, j)$ for $j = 0, 3, 5, 9$ whose lattices have right side continuous frontier intervals $0 \leq x \leq 10$, $0 \leq x \leq 7$, $0 \leq x \leq 5$ and $0 \leq x \leq 1$, respectively.

⁶The curve described by each monomial is the path followed by an object with an initial and final state. For $\beta = 10$, there are ten simple paths, which simultaneously change over time according to the change in each value of j and W in the $\mathcal{H}_G(X, W; 10, j)$ distribution. See Fig. 2. We might think that these changes describe random motions over time. But polynomials are smooth functions and follow the rules of standard calculus,

⁷Also, the same allocations are obtained evaluating the marginal $\mathcal{H}_G(w; \beta = 10)$ for each $w = 0, 1, \dots, 9$.

⁸In this research, homogeneous polynomials are smooth functions in a single variable that have nonzero β terms, each of degree $\beta - 1$, and obey to all calculus laws.

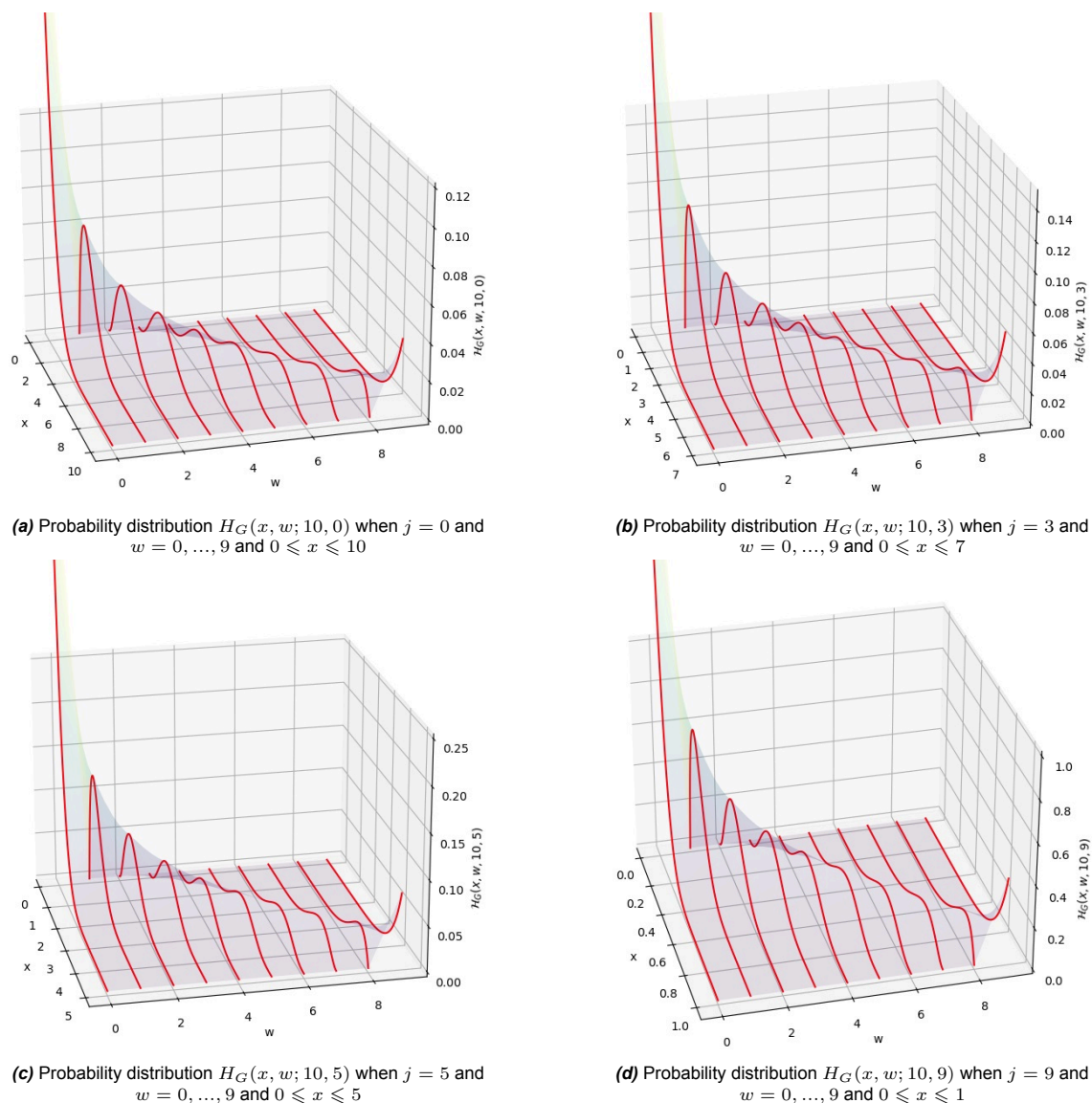


Fig. 2. The subfigures, (a)-(d), illustrates the shapes of four probability distributions for $j = 0, 3, 5, 9$, for a fixed $\beta = 10$

All these closed and bounded intervals are topological spaces in the Euclidean space, and satisfy several topological properties of invariance, such as compactness, connectedness, path-connectedness, locally connectedness, simple connectedness, contractibility, homotopy equivalence, and are simplicial complexes of dimension 1. Any interval $[0, \beta - j]$ for $j = 0, 1, 2, \dots, \beta - 1$ can be deformed continuously into $[0, 1]$. Simply, the variable $s = \frac{x}{\beta - j}$ maps x in $[0, \beta - j]$ to s in $[0, 1]$. Nicely, each space $[0, \beta - j]$ is contractible and behaves as a single point. Therefore, all have the same homotopy type, all are correlated, for any positive integers β and j . For our purpose, let $[0, \beta - j]$ be a contractible topological space with standard topology. Then all maps $\mathcal{H}_G : [0, \beta - j] \mapsto [0, \frac{\beta - j}{2}]$ are homotopic if and only if $[0, \frac{\beta - j}{2}]$ is contractible. In particular, β and j are both even or both odd. With this support, we may construct a family of homotopies as follows.

Let $[0, \beta - j]$ and $[0, \frac{\beta - j}{2}]$ be topological spaces, and let $\mathcal{H}_{G1}(x, w; \beta, j)$ and $\mathcal{H}_{G2}(x, w; \beta, j)$ be polynomials maps $\mathcal{H}_G : [0, \beta - j] \mapsto [0, \frac{\beta - j}{2}]$ of degree $\beta - 1$. Then a polynomial $\mathcal{H}_{G1}$ is homotopic to polynomial $\mathcal{H}_{G2}$ if there exists a continuous map $\mathbb{H} : [0, \beta - j] \times I \mapsto [0, \frac{\beta - j}{2}]$ such that $\mathbb{H}(x, 0) = \mathcal{H}_{G1}(x)$ and $\mathbb{H}(x, 1) = \mathcal{H}_{G2}(x)$ for all x in $[0, \beta - j]$. The map $\mathbb{H}$ is called homotopy from $\mathcal{H}_{G1}(x)$ to $\mathcal{H}_{G2}(x)$. For $(x, t) \in [0, \beta - j] \times I$, we may interpret $t \in I = [0, 1]$ as a parameter measuring time. Then $\mathbb{H}(x, t)$ is 1-parameter family of maps $[0, \beta - j] \mapsto [0, \frac{\beta - j}{2}]$. As time increases from 0 to 1 the map $\mathcal{H}_{G1}(x)$ is deformed continuously into the map $\mathcal{H}_{G2}(x)$.

In particular, for $\beta = 10, j = 0, 2, 4, 6, 8$ if we choose $j = 0$, then $\mathbb{H}(x, t) : [0, 10] \times I \mapsto [0, 5]$ and from Tab. 1 it corresponds to $\mathcal{H}(x, w; 10, 0) = \sum_{w=0}^9 \binom{9}{w} \frac{10x^w(10-x)^{9-w}}{10^{10}} \psi_{10}(w)$ and $\mathcal{H}(x, w; 10, 5) = \sum_{w=0}^9 \binom{9}{w} \frac{10x^w(5-x)^{9-w}}{5^{10}} \psi_{10}(w)$.

Then, a linear homotopy can be given by

$$\mathbb{H}(x, t) = (1 - t) \cdot \mathcal{H}_{G1}(x, w; 10, 0) + t \cdot \mathcal{H}_{G2}(x, w; 10, 5). \quad (33)$$

For every $x \in [0, 10]$, the line segment joining $\mathcal{H}_{G1}(x, w; 10, 0)$ lies completely inside $[0, 5]$. Essentially we move values

of the probability distribution $\mathcal{H}_{G1}(x, w; 10, 0)$ into probability distribution $\mathcal{H}_{G2}(x, w; 10, 5)$ along a straight line. In general, it is well known that homotopy is an equivalence relation. With regret, we leave this topological discussion of the sample space of these distributions, but those interested could read any textbook of algebraic topology. For example, (Munkres, 2000, Chap. 9). This specific example is merely one among hundreds or possibly thousands of other number systems that can be investigated. We hope to keep working on this topic and release its results shortly.

Discussion and conclusions

The outcomes of this research provide a new class of two-dimensional mixed probability distributions and their characteristics. The development of these distributions focused on both the general theoretical foundation and their practical implications in different scientific fields, particularly in statistics.

To achieve our findings: we first consider the F. Benford distribution in (1) and extend its domain from $\{1, 2, \dots, \beta - 1\}$ to $\{0, 1, 2, \dots, \beta - 1\}$ as shown in (5) and then embedding these generalized Benford distributions into an extended form of Bernstein polynomials. It is mathematically demonstrated that these models satisfy all the universal properties that every probability distribution must fulfill.

A key feature of these distributions is that all moments are functions only of their parameter $\beta \geq 2$, which is the base of a number system. Since there are infinitely many number systems, we are faced with an infinite family of probability distributions, all well-defined and supported mathematically. Here, we only present the finite family's case. The cornerstone of the development of these families of distributions lies in the Theorems 2 and (3), whose proofs and the implicated truth propositions guarantee their theoretical rigor. Our study is rigorously documented and is easy to understand for an advanced undergraduate audience in probability theory.

This "machine", known as the $\mathcal{H}$ distribution, is advantageous for the production of statistical congruence data for arbitrary congruence modules under standard arithmetic operations of addition and/or multiplication. The result presented in the context of the decimal system scenario with $\beta = 10$ indicates that we have developed a method for generating doubly stochastic Markov chains. Once the statistical data is made and put into the Markov chain's transition matrix, it can be used to study how a stochastic system changes over time and make predictions, among other things.

We refrain from employing historical data and acknowledge that our theoretical probability distributions require more tangible real-world applications, as they are novel and in the exploratory phase. Based on the results, we modestly believe that these distributions have a significant application impact across several scientific domains.

The main contribution of this study is to give an analytical tool that is easy to use and understand, capable of solving many problems in different fields, especially in statistics. Based on the mix of variables, the type of applications that could be focused on are amphoteric phenomena. That is, experiments or phenomena that behave continuously and discretely at the same time or alternately. This is similar to what occurs in price valuation in finance, with continuous models with discrete jumps. During the writing of the research, the following contributions were observed and noted: first. We have extended the Benford distribution to more general domains by including the digit zero in its definition for any number system with base $\beta \geq 2$. Second. The marginal distribution $\mathcal{H}_G(x; \beta, j)$ in (20) is related to the family of Bézier curves of degree $\beta - 1$ with β control points given by $\psi_\beta(w)$. Third. A new family of probability distributions $\mathcal{H}_G(X, W; \beta, j)$ is introduced into the statistical literature. Fourth. We introduce a methodology using these models to simulate statistical congruence data, specifically focusing on the operations of addition and multiplication within congruence classes across various number systems. Fifth. We provide a mathematical allocation function, which can be represented by either the $\mathcal{H}_G(W, \beta)$ or $\mathcal{H}_G(X, W; \beta, j)$ distribution, to optimize resource distribution in statistical models.

Also, we have demonstrated that all moments (expected value, variance, covariance, correlation coefficient, etc.) of the $\mathcal{H}$ distributions depend only of the parameter β . Therefore, they are easy to calculate. Future research should focus on refining this family of models with real-world data and finding new applications.

Finally, there are other applications of the $\mathcal{H}$ distributions that were planned for inclusion in this project, but there is a limitation of space in the document. Therefore, we are aware that we are leaving several parts of our study open, as they were not included in our objectives, and we believe this will be a source of inspiration for readers who decide to investigate further.

APPENDIX: Proofs

Proof of Proposition 6. We recall the well-known expressions of covariance and the cross-correlation coefficient. Recall that for $\beta \geq 2, j, w = 0, 1, \dots, \beta - 1$:

$$\sigma_{XW} = Cov(X, W; \beta, j) = E(XW) - E(X)E(W), \quad (34)$$

$$\rho_{XW} = Corr(X, W; \beta, j) = \frac{Cov(X, W; \beta, j)}{\sqrt{[Var(X; \beta, j)][Var(W; \beta, j)]}}. \quad (35)$$

We use (Maple.Soft, 2024) to solve some summations and integrations, as needed. Results of $E(X) = E(x; \beta, j)$ and $E(W) = E(w; \beta, j)$ are given in Equation (22) and Equation (6), respectively. Next, we compute $E(XW)$:

$$E(XW) = E(XW, \beta, j) = \frac{(b-j) \log \left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{2i-2}} \right]}{(\beta+1) \log(\beta+1)}. \quad (36)$$

Thus, replacing Equations (6), (22) and (36) into (34) we have the covariance given in Equation (25), which is

$$Cov(X, W; \beta, j) = \frac{(\beta - j) \log \left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{(2i-2)}} \right]}{(\beta+1) \log(\beta+1)} - \frac{(\beta - j) \log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right] \log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right]}{(\beta+1) (\log(\beta+1))^2}, j = 0, 1, \dots, \beta - 1. \quad (37)$$

On the other hand, we have computed the variances $Var(X; \beta, j)$ in Equation (23) and $Var(W; \beta, j)$ in Equation (7).

Thus, we replace Equations (7), (23) and (37) into (35) as follows

$$\begin{aligned} Corr(X, W; \beta, j) &= \\ &= \frac{\frac{(\beta-j) \log \left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{(2i-2)}} \right]}{(\beta+1) \log(\beta+1)} - \frac{(\beta-j) \log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right] \log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right]}{(\beta+1) (\log(\beta+1))^2}}{\sqrt{\left[\frac{(\beta-j)^2 \log \left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}} \right]}{(\beta+1)(\beta+2) \log(\beta+1)} - \frac{(\beta-j)^2 (\log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right])^2}{(\beta+1)^2 (\log(\beta+1))^2} \right] \left[\frac{\log \left[\frac{(\beta+1)^{(\beta-1)^2}}{\prod_{i=1}^{\beta} i^{(2i-3)}} \right]}{\log(\beta+1)} - \frac{(\log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right])^2}{(\log(\beta+1))^2} \right]}}, j = 0, \dots, \beta - 1. \end{aligned} \quad (38)$$

Simplifying the common factors, $(\beta - j)$, of the numerator and denominator, we have

$$\begin{aligned} Corr(X, W; \beta, j) &= \\ &= \frac{\frac{\log \left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{(2i-2)}} \right]}{(\beta+1) \log(\beta+1)} - \frac{\log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right] \log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right]}{(\beta+1) (\log(\beta+1))^2}}{\sqrt{\left[\frac{\log \left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}} \right]}{(\beta+1)(\beta+2) \log(\beta+1)} - \frac{(\log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right])^2}{(\beta+1)^2 (\log(\beta+1))^2} \right] \left[\frac{\log \left[\frac{(\beta+1)^{(\beta-1)^2}}{\prod_{i=1}^{\beta} i^{(2i-3)}} \right]}{\log(\beta+1)} - \frac{(\log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right])^2}{(\log(\beta+1))^2} \right]}}, j = 0, \dots, \beta - 1. \end{aligned} \quad (39)$$

It is clear that the correlation coefficient does not depend on j ; it only depends on the parameter $\beta \geq 2$. That is,

$$Corr(X, W, \beta) = \frac{\frac{\log \left[\frac{(\beta+1)^{(\beta-1)\beta}}{\prod_{i=1}^{\beta} i^{(2i-2)}} \right]}{(\beta+1) \log(\beta+1)} - \frac{\log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right] \log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right]}{(\beta+1) (\log(\beta+1))^2}}{\sqrt{\left[\frac{\log \left[\frac{(\beta+1)^{\beta(\beta+1)}}{\prod_{i=1}^{\beta} i^{2i}} \right]}{(\beta+1)(\beta+2) \log(\beta+1)} - \frac{(\log \left[\frac{(\beta+1)^{\beta}}{\beta!} \right])^2}{(\beta+1)^2 (\log(\beta+1))^2} \right] \left[\frac{\log \left[\frac{(\beta+1)^{(\beta-1)^2}}{\prod_{i=1}^{\beta} i^{(2i-3)}} \right]}{\log(\beta+1)} - \frac{(\log \left[\frac{(\beta+1)^{(\beta-1)}}{\beta!} \right])^2}{(\log(\beta+1))^2} \right]}}. \quad (40)$$

□

One can easily evaluate these formulas using Maple or any other software for any value of $\beta \geq 2$, as shown in the previous sections above.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Barabesi, L., & Pratelli, L. (2020). On the generalized Benford law. *Statistics Probability Letters*, 160, 108702. <https://doi.org/10.1016/j.spl.2020.108702>
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572. <http://www.jstor.org/stable/984802>
- Debnath, L., & Basu, K. (2015). A short history of probability theory and its applications. *International Journal of Mathematical Education in Science and Technology*, 46(1), 13–39. <https://doi.org/10.1080/0020739X.2014.936975>
- Franklin, J. (2015). *The science of conjecture: evidence and probability before Pascal*. Baltimore, USA: Johns Hopkins University Press. <https://books.google.com.pe/books?id=1zDECQAQBAJ&lpg=PP1&hl=es&pg=PP1#v=onepage&q&f=false>
- Fu, D., Shi, Y. Q., & Su, W. (2007). A generalized Benford's law for JPEG coefficients and its applications in image forensics. In Delp III & Wong (Eds.), *Security, steganography, and watermarking of multimedia contents IX* (Vol. 6505, p. 65051L). SPIE. <https://doi.org/10.1117/12.704723>
- Maple.Soft. (2024). *Online Help (Calculus, sec. 5.4, page 133)*. <https://www.maplesoft.com/support/help/Maple/view.aspx?path=csn>.
- Mathematica.Soft. (2024). *Wolfram Language and System*. <https://reference.wolfram.com/language/ref/Integrate.html>.
- Munkres, J. (2000). *Topology* (2nd ed.). Upper Saddle River, USA: Prentice Hall.
- Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1), 39–40. <https://doi.org/10.2307/2369148>
- Pietronero, L., Tosatti, E., Tosatti, V., & Vespignani, A. (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications*, 293(1), 297–304. [https://doi.org/10.1016/S0378-4371\(00\)00633-6](https://doi.org/10.1016/S0378-4371(00)00633-6)
- Singer, I., & Thorpe, J. (1967). *Lecture notes on elementary topology and geometry*. New York, NY: Springer. <https://doi.org/10.1007/978-1-4615-7347-0>
- Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, 25(3), 603–607. <http://www.jstor.org/stable/2236844>
- Wiley.StatsRef. (2014). Point multiseriial (pms) correlation. In Balakrishnan, Colton, Everitt, Piegorsch, Ruggeri, & Teugels (Eds.), *Wiley statsref: Statistics reference online*. John Wiley Sons, Ltd. <https://doi.org/10.1002/9781118445112.stat00688>
- Wolfram.Soft. (2025). *Mathematica. Wolfram language & system documentation center*. <https://reference.wolfram.com/language/ref/Hypergeometric1F1.html>.

Отримано редакцією журналу / Received: 22.02.26
Прорецензовано / Revised: 14.04.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Ксав'єр М. УАРКА, д-р філософії, проф.
ORCID ID: 0000-0002-1564-2343
e-mail: jhuarcao@usmp.pe
Університет Сан-Мартін-де-Поррес, Ліма, Перу

ЙМОВІРНІСНІ РОЗПОДІЛИ ТИПУ Н

Ймовірнісні розподіли є важливими моделями в усіх наукових галузях, однак вони мають вирішальне значення саме в статистиці. Ми пропонуємо новий клас змішаних сумісних розподілів ймовірностей, ймовірнісні простори яких є послідовностями вкладених $2D$ -решіток. Методологія побудови цих моделей передбачає вкладення дискретного розширеного розподілу Бенфорда з параметром β у неперервний модифікований поліном Бернштейна. β є параметром цього сімейства розподілів і визначає основу довільної системи числення. Метою роботи було описати математичну структуру та статистичні властивості взаємодій цих ймовірнісних розподілів. На основі теоретичних розробок та їхніх результатів ми робимо висновок, що зазначені моделі є цінними для численних застосувань у багатьох наукових дисциплінах. Для підтвердження гіпотези наведено конкретний приклад десяткової системи числення.

Ключові слова: розширений β -розподіл Бенфорда, поліноми Бернштейна, система чисел, універсальні властивості ймовірнісного розподілу, гомотопічний семплічний комплекс.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

ДИФЕРЕНЦІАЛЬНІ РІВНЯННЯ, МАТЕМАТИЧНА ФІЗИКА ТА МЕХАНІКА

UDC 517.953

DOI: <https://doi.org/10.17721/1812-5409.2026/1.12>

Iryna ALEXANDROVICH¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-1950-8651
e-mail: ialexandrovich@knu.ua

Larisa ADZHUBEY¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-8487-0884
e-mail: adzhubey@knu.ua

Andrii RYZHENKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-0758-9212
e-mail: andrai@ukr.net

Viktor LYASHKO², PhD Student
ORCID ID: 0000-0003-2647-1188
e-mail: lyashko91@gmail.com

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ROBIN'S PROBLEM FOR A FOURTH-ORDER HYPERBOLIC EQUATION

The Robin problem, also known as a mixed boundary condition, is often applied to partial differential equations. It defines a relationship between the value of a function and its derivative at the boundary of a domain, enabling the modeling of physical processes that affect both the properties of the surface itself and the flow of matter or energy. Fourth-order hyperbolic equations describe complex processes involving wave propagation and their interactions, enabling the modeling of dispersion effects often present in real-world physical processes. Such equations are used in hydrodynamics to model wave propagation at great depths, where both the velocity and the influence of dispersion on wave behavior are taken into account; in elasticity theory to analyze deformations and wave propagation in elastic materials, where dispersion effects and complex boundary conditions are present; and in seismology.

The aim of this work is to solve the Robin problem for a fourth-order hyperbolic equation with conditions imposing a relationship between certain derivatives of the function at the boundary. For this problem, the obtained solution to the Cauchy problem for a second-order hyperbolic equation is used. The solution to the Cauchy problem is obtained using differential operators that transform arbitrary functions into a regular solution of an m -th order hyperbolic equation. By reducing the problem to a solved Cauchy problem for a second-order hyperbolic equation and matching the boundary conditions of these problems, a system of equations arises whose solution exists and is unique as a solution to the Volterra integral equation and as a solution to the inhomogeneous Bessel equation. The solution to the Robin problem is a combination of the obtained solutions to the corresponding Cauchy problems for second-order equations.

Keywords: differential operator, regular solutions, iterated equations of hyperbolic type, Robin's problem.

Classification according to AMS 2020: 35Q74, 35R09.

Introduction

Hyperbolic equations are used to describe wave propagation processes, such as sound, heat, and light waves, as well as in electromagnetism and elasticity theory. They are used to model with high accuracy dynamic phenomena characterized by wave propagation and interaction, as well as the vibration of bridges, structures, and mechanical systems. Mathematical models are formulated as boundary value problems for partial differential equations in bounded and unbounded domains, for which a general theory of function-theoretical methods is considered in (Gilbert, 1969).

In the Riquier and Robin problems, which consider mixed boundary conditions for partial differential equations, solutions are determined taking into account the values of the function and its derivative at the boundary of the domain. Problems for higher-order elliptic equations are considered in (Gomez-Polanco, Guevara, & Molina, 2013; Karachik, 2003). In the Robin problem (Rozkosz, 2025), the existence of a positive renormalized solution was proven and a stochastic representation for a uniformly elliptic divergence operator with a measure on the right-hand side of the equation and an absorbing term on the boundary containing collapsing terms was proposed. In the paper (Guarino Lo Bianco, La Manna, & Velichkov, 2021), a two-phase free limit problem was studied for the first time, in which the solution satisfies the Robin limit condition. The case in which the solution is continuous on the free boundary was considered, and the existence and regularity of minimizers for the associated variational problem were proven.

The use of fourth-order hyperbolic equations allows one to model dispersion effects, which are often present in real physical processes. Wave propagation in dispersive media (liquids, gases, and certain types of solids) requires the formulation of fourth-order hyperbolic equations that take this dependence into account.

This paper is devoted to deepening the results of studying solutions to differential equations containing operators of the

form $D_m = \frac{\partial^m}{\partial x_1 \partial x_2 \dots \partial x_m}$ and their iterations, taking into account the Robin limit condition. Methods for solving such equations

are based on the creation of differential and integral operators that determine the solution of hyperbolic and elliptic equations and systems.

The differential operators constructed in (Lyashko et al., 2024) that transform arbitrary functions of the domain $D \subset \mathbb{R}^2$ into a regular solution to the Robin problem equation

$$Y_k^m W = \left(\frac{\partial^2}{\partial x \partial y} + \frac{k}{2(x+y)} \left(\frac{\partial}{\partial y} - \frac{\partial}{\partial x} \right) \right)^m W = 0, m \in N, k/2 \in Z, \quad (1)$$

are applicable to solving the Robin problem.

1. Main results

Consider the differential equation

$$W_{xy} + \frac{(n-m)\phi'(x)}{\phi(x)+\psi(y)} W_y - \frac{n(m+1)\psi'(y)}{(\phi(x)+\psi(y))^2} W = 0, \quad (2)$$

where $\phi(x), \psi(y)$ are functions satisfying the condition $(\phi(x)+\psi(y))\phi'(x)\psi'(y) \neq 0$, $n, m \in N \cup \{0\}$, and $W(x, y)$ are defined by the equality (Lyashko et al., 2024)

$$W(x, y) = \sum_{k=0}^n (-1)^{n-k} \frac{n!(m+1)_{n-k}}{k!(n-k)!} \frac{(\phi'(x))^{n-k}}{(\phi(x)+\psi(y))^{n-k}} g^{(k)}(x) + \sum_{k=0}^m (-1)^{m-k} \frac{m!(n+1)_{m-k}}{k!(m-k)!} \frac{(\psi'(y))^{m-k}}{(\phi(x)+\psi(y))^{m-k}} f^{(k)}(y), \quad (3)$$

is a solution to equation (2), and the functions $g(x), f(y)$ are arbitrary, differentiable a sufficient number of times in the domain D .

Let us formulate the *Cauchy problem*. In the domain $D = [a, b] \times [a, b]$, find a solution of the equation

$$W_{xy} - \frac{2}{(x+y)^2} W = 0, \quad (4)$$

satisfying the conditions

$$W(x, y)|_{y=x} = \alpha(x), \left(\frac{\partial W}{\partial x} - \frac{\partial W}{\partial y} \right)|_{y=x} = \beta(x), \quad (5)$$

where $\alpha(x) \in C^2([a, b]), \beta(x) \in C^2([a, b])$ are given functions, $x+y \neq 0$ in the domain D .

According to formula (3), satisfying conditions (5), the solution to the Cauchy problem will have the form

$$W(x, y) = (\alpha' - \beta) \frac{x(y-x)}{x+y} + \alpha(x) + \frac{1}{x+y} \int_x^y (y^2 - (y-x)t - t^2) \gamma(t) dt, \quad (6)$$

where $\gamma(t) = \frac{1}{2t}(\alpha'(t) - \beta(t)) + \frac{1}{2}(\alpha'(t) - \beta'(t)) - \frac{1}{2t^2}\alpha(t)$.

We move on to equation (1) for $2k \in Z$. Replacing $W = (x+y)^{n+1}V$ into equation (1), takes the form

$$Y_n^m V = \left(\frac{\partial^2}{\partial x \partial y} + \frac{n(n+1)}{(x+y)^2} \right)^m V = 0, m \in N, n \in N. \quad (7)$$

Let D be the domain in the upper half-plane. Using mathematical induction, it was established (Lyashko et al., 2024) that the solution to equation (7) is

$$V(x, y) = \sum_{i=0}^{m-1} (x+y)^i V_i(x, y), \quad (8)$$

where $V_i(x, y)$, $i = \overline{0, m-1}$, are $2(i+1)$ continuously differentiated solutions of the equation $Y_n V = 0$.

Theorem 1. For any solution of the equation $Y_n^m V = 0$, $m \geq 2$, in the domain D there are functions $g_i(x), f_i(y)$ such that

$$V(x, y) = \sum_{i=0}^{m-1} (x+y)^i \times \left(\sum_{k=0}^n (-1)^{n-k} \frac{(2n-k)!}{k!(n-k)!} \frac{1}{(x+y)^{n-k}} \frac{\partial^k}{\partial x^k} g_i(x) + \sum_{k=0}^n (-1)^{n-k} \frac{(2n-k)!}{k!(n-k)!} \frac{1}{(x+y)^{n-k}} \frac{\partial^k}{\partial y^k} f_i(y) \right) \quad (9)$$

and vice versa, if $g_i(x), f_i(y)$ - are arbitrary in D functions, differentiated the required number of times, then (9) is a solution of equation (7) in D .

Let us apply formula (9) to solve the Robin problem for the equation $Y^2 V = 0$.

Robin problem. Let D be a square $[a, b] \times [a, b]$. Find in the domain D a solution of the equation

$$Y^2 V = \left(\frac{\partial^2}{\partial x \partial y} - \frac{2}{(x+y)^2} \right)^2 V = 0, \quad (10)$$

справджуючи умови

$$\lim_{y \rightarrow x} \left(V + \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) V \right) = a_1(x), \quad (11)$$

$$\lim_{y \rightarrow x} \left(\left(\frac{\partial^2}{\partial x \partial y} - \frac{2}{(x+y)^2} \right) V + \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) \left(\frac{\partial^2}{\partial x \partial y} - \frac{2}{(x+y)^2} \right) V \right) = a_2(x).$$

Here $a_1(x)$, $a_2(x)$ are given functions of sufficient smoothness.

In accordance with statement (8), we seek the solution of problem (10), (11) in the form

$$V(x, y) = V_0(x, y) + (x + y)V_1(x, y).$$

Satisfying boundary conditions (11), we obtain

$$\lim_{y \rightarrow x} \left(V_0 + (x + y)V_1 + \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) (V_0 + (x + y)V_1) \right) = a_1(x),$$

$$\lim_{y \rightarrow x} \left(\left(\frac{\partial^2}{\partial x \partial y} - \frac{2}{(x + y)^2} \right) (V_0 + (x + y)V_1) + \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) \left(\frac{\partial^2}{\partial x \partial y} - \frac{2}{(x + y)^2} \right) (V_0 + (x + y)V_1) \right) = a_2(x).$$

Since V_0, V_1 satisfy to equation

$$\frac{\partial^2 V}{\partial x \partial y} - \frac{2}{(x + y)^2} V = 0,$$

then to find functions $V_0(x, y)$, $V_1(x, y)$ we have the following Cauchy problems

$$I. \frac{\partial^2 V_0}{\partial x \partial y} - \frac{2}{(x + y)^2} V_0 = 0, \quad V_0|_{y=x} = \alpha_1(x), \quad \frac{\partial V_0}{\partial x} - \frac{\partial V_0}{\partial y}|_{y=x} = \beta_1(x), \quad (12)$$

$$II. \frac{\partial^2 V_1}{\partial x \partial y} - \frac{2}{(x + y)^2} V_1 = 0, \quad V_1|_{y=x} = \alpha_2(x), \quad \frac{\partial V_1}{\partial x} - \frac{\partial V_1}{\partial y}|_{y=x} = \beta_2(x). \quad (13)$$

The functions $\alpha_1(x)$, $\beta_1(x)$, $\alpha_2(x)$, $\beta_2(x)$ are determined from the system of equations:

$$\begin{cases} \alpha_1(x) + 2x\alpha_2(x) + \beta_1(x) + 2x\beta_2(x) = a_1(x) \\ 2\alpha_2'(x) - \beta_2(x) + \frac{1}{2x^2}\beta_1(x) + \beta_2'(x) = a_2(x) \\ \alpha_1''(x) - \frac{1}{2x^2}\alpha_1(x) = \beta_1'(x) \\ \alpha_2''(x) - \frac{1}{2x^2}\alpha_2(x) = \beta_2'(x) \end{cases},$$

transforming which, we obtain

$$\begin{cases} 2x^2\alpha_2^{(4)}(x) + (4x + 2x^2)\alpha_2^{(3)}(x) + 4x\alpha_2''(x) - (1 + 1/x)\alpha_2'(x) - \frac{3}{2x^2}\alpha_2(x) = 4x\alpha_2'(x) + 2x^2\alpha_2''(x) \\ \beta_2'(x) = \alpha_2''(x) - \frac{1}{2x^2}\alpha_2(x) \\ \alpha_1''(x) + \alpha_1'(x) - \frac{1}{2x^2}\alpha_1(x) = a_1'(x) - 2x(\alpha_2'(x) + \beta_2'(x)) - 2(\alpha_2(x) + \beta_2(x)) \\ \beta_1'(x) = \alpha_1''(x) - \frac{1}{2x^2}\alpha_1(x). \end{cases}$$

Putting $\alpha_2^{(4)}(x) = \phi(x)$, for $\phi(x)$ we will have an integral Volterra equation of the 2nd kind

$$2x^2\phi(x) + \int_{x_0}^x K(x, t)\phi(t)dt = F(x),$$

where $K(x, t) = 4x + 2x^2 + 4x(x - t) - \frac{x+1}{2x}(x - t)^2 - \frac{1}{4x^2}(x - t)^3$, $F(x) = 4x\alpha_2'(x) + 2x^2\alpha_2''(x)$.

Then the functions $\alpha_2(x)$, $\beta_2(x)$ are expressed by $\phi(x)$:

$$\alpha_2(x) = \frac{1}{6} \int_{x_0}^x \phi(t)(x - t)^3 dt, \quad \beta_2(x) = \int_{x_0}^x \phi(t) \left(\frac{11}{24}x^2 + \frac{1}{4}(x - t) - tx + \frac{1}{4}t^2 \ln \frac{t}{x} + t^2 - \frac{t^3}{12x} \right) dt.$$

The function $\alpha_1(x)$ is a solution of the equation

$$\alpha_1''(x) + \alpha_1'(x) - \frac{1}{2x^2}\alpha_1(x) = G(x),$$

where $G(x) = \alpha_1'(x) - \frac{3}{4}x^3\phi(x) - \int_{x_0}^x \phi(t) \left((x - t)^2 \left(x + \frac{x - t}{3} \right) + \frac{33}{12}x^2 + \frac{3}{2}t^2 + x - \frac{1}{2}t - 4tx + \frac{1}{2}t^2 \ln \frac{t}{x} \right) dt$,

which is a non-homogeneous Bessel equation, and has the form

$$\alpha_1(x) = \sqrt{x} e^{-x/2} J_{\sqrt{3}/2} \left(\frac{ix}{2} \right) \int_{x_0}^x \frac{dx_1}{\exp(x_1 e^{-x_1/2} J_{\sqrt{3}/2} \left(\frac{ix_1}{2} \right))} \int_{x_0}^{x_1} \exp(\sqrt{x_2} e^{-x_2/2} J_{\sqrt{3}/2} \left(\frac{ix_2}{2} \right)) G(x_2) dx_2.$$

Finally,

$$\beta_1(x) = a_1(x) - \alpha_1(x) - 2x \int_{x_0}^x \phi(t) \left(\frac{1}{6}(x-t)^3 + \frac{11}{24}x^2 + \frac{1}{4}(x-t) - tx + \frac{1}{4}t^2 \ln \frac{t}{x} + t^2 - \frac{t^3}{12x} \right) dt.$$

The solution of each Cauchy problem (12), (13), according to the solved problem (4), (5), will be

$$V_0(x, y) = (\alpha'_1(x) - \beta_1(x)) \frac{x(y-x)}{x+y} + \alpha_1(x) + \frac{1}{x+y} \int_x^y (y^2 - (y-x)t - t^2) \gamma_1(t) dt, \quad (14)$$

$$\text{where } \gamma_1(t) = \frac{1}{2t}(\alpha'_1(t) - \beta_1(t)) + \frac{1}{2}(\alpha''_1(t) - \beta'_1(t)) - \frac{1}{2t^2} \alpha_1(t),$$

$$V_1(x, y) = (\alpha'_2(x) - \beta_2(x)) \frac{x(y-x)}{x+y} + \alpha_2(x) + \frac{1}{x+y} \int_x^y (y^2 - (y-x)t - t^2) \gamma_2(t) dt, \quad (15)$$

$$\text{where } \gamma_2(t) = \frac{1}{2t}(\alpha'_2(t) - \beta_2(t)) + \frac{1}{2}(\alpha''_2(t) - \beta'_2(t)) - \frac{1}{2t^2} \alpha_2(t).$$

The solution of the Robin problem (10), (11) is $V(x, y) = V_0(x, y) + (x+y)V_1(x, y)$, where $V_0(x, y)$ and $V_1(x, y)$ are defined in (14) and (15).

Discussion and conclusions

The Robin problem for a fourth-order hyperbolic equation is considered and solved by reducing it to the Cauchy problem for a second-order hyperbolic equation. The solution is obtained using differential operators that transform arbitrary functions into an established, inductively regular solution to the m-th order hyperbolic equation. A theorem for solving the m-th order hyperbolic equation is formulated, which is used to find a solution to the fourth-order equation. The method proposed in this paper allows one to solve problems with complex boundary conditions.

Authors' contribution: Iryna Aleksandrovych – conceptualization; methodology; Larysa Adzhubey – relevance analysis; discussion of the theoretical foundations of the study; Andriy Ryzhenko – analysis of sources, discussion of conclusions; Viktor Lyashko – literature review.

Funding sources. Funding is provided by the authors' own funds and partially provided by Taras Shevchenko National University of Kyiv.

References

- Gilbert, R. P. (1969). *Function theoretic methods in partial differential equation*. Academic Press.
- Gomez-Polanco, A., Guevara, J. M., & Molina, B. (2013). A mimetic iterative scheme for solving biharmonic equations. *Mathematical and Computational Modeling*, 57 (9–10), 2132–2139.
- Karachik, V. (2003). Normalized system of functions with respect to the Laplace and its applications. *Journal of Mathematical Analysis and Applications*, 287, 577–592. [https://doi.org/10.1016/S0022-247X\(03\)00583-3](https://doi.org/10.1016/S0022-247X(03)00583-3)
- Lyashko, S. I., Sydorov, M. V.-S., Lyashko, N. I., & Alexandrovich, I. M. (2024). Differential operators defining solutions to iterated hyperbolic-type equations. *Cybernetics and Systems Analysis*, 60(5), 753–758. <https://doi.org/10.1007/s10559-0-24-00712-4>
- Rozkosz, A. (2025). Robin problem with measure data and singular nonlinearities on the boundary. *arXiv preprint arXiv:2507.07821*. <https://doi.org/10.48550/arXiv.2507.07821>
- Guarino Lo Bianco, S., La Manna, D. A., & Velichkov, B. (2021). A two-phase problem with Robin conditions on the free boundary. *Journal de l'École polytechnique–Mathématiques*, 8, 1–25. <https://doi.org/10.48550/arXiv.2003.14139>

Отримано редакцією журналу / Received: 04.10.25
Прорецензовано / Revised: 26.12.25
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Ірина АЛЕКСАНДРОВИЧ¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-1950-8651
e-mail: ialexandrovich@knu.ua

Лариса АДЖУБЕЙ¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-8487-0884
e-mail: adzhubey@knu.ua

Андрій РИЖЕНКО¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-0758-9212
e-mail: andrai@ukr.net

Віктор ЛЯШКО², PhD Student
ORCID ID: 0000-0003-2647-1188
e-mail: lyashko91@gmail.com

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Інститут кібернетики імені В. М. Глушкова НАН України, Київ, Україна

ЗАДАЧА РОБЕНА ДЛЯ ГІПЕРБОЛІЧНОГО РІВНЯННЯ ЧЕТВЕРТОГО ПОРЯДКУ

Задачу Робена, відому як змішану крайову умову, часто застосовують до диференціальних рівнянь у частинних похідних. Вона визначає зв'язок між значеннями функції та її похідною на межі області, що дає змогу моделювати фізичні процеси, на які впливають як властивості самої поверхні, так і потоки речовини чи енергії. Гіперболічні рівняння четвертого порядку описують складні процеси, що охоплюють поширення хвиль та їхню взаємодію, допомагають моделювати дисперсійні ефекти, які часто наявні в реальних фізичних

процесах. Такі рівняння використовують у задачах гідродинаміки для моделювання поширення хвиль на великих глибинах, де враховують як швидкість, так і вплив дисперсії на поведінку хвиль, теорії пружності для аналізу деформацій та поширення хвиль у пружних матеріалах, коли наявні ефекти дисперсії та складні граничні умови, а також у сейсмології.

Мета пропонованої роботи полягає в розв'язанні задачі Робена для гіперболічного рівняння четвертого порядку з умовами, що накладають зв'язок між певними похідними функції на межі. Для поставленої задачі використовуємо отриманий розв'язок задачі Коші для рівняння гіперболічного типу другого порядку. Розв'язок задачі Коші отримано за допомогою диференціальних операторів, що переводять довільні функції в регулярний розв'язок рівняння гіперболічного типу n -го порядку. Під час зведення задачі до розв'язаної задачі Коші для рівняння гіперболічного типу другого порядку й узгодження крайових умов цих задач виникає система рівнянь, розв'язок якої існує і є єдиним як розв'язок інтегрального рівняння Вольтерра, а також як розв'язок неоднорідного рівняння Бесселя. Розв'язком задачі Робена буде комбінація отриманих розв'язків відповідних задач Коші для рівнянь другого порядку.

Ключові слова: диференціальний оператор, регулярні розв'язки, ітеровані рівняння гіперболічного типу, задача Робена.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 533.6.021.6:534.13

DOI: <https://doi.org/10.17721/1812-5409.2026/1.13>

Сергій АЛЕКСЕЄНКО, д-р техн. наук, проф.

ORCID ID: 0000-0003-0320-989X

e-mail: aleksieienko.s.v@nmu.one

Національний технічний університет "Дніпровська політехніка", Дніпро, Україна

Лілія НАКАШИДЗЕ, д-р техн. наук, доц.

ORCID ID: 0000-0003-3990-6718

e-mail: foton_dnu@ukr.net

Дніпровський національний університет імені Олеся Гончара, Дніпро, Україна

Сергій ЗОЛОТАРЕНКО, асп.

ORCID ID: 0009-0008-5291-4206

e-mail: zolotarenko.s.an@nmu.one

Національний технічний університет "Дніпровська політехніка", Дніпро, Україна

ЧИСЕЛЬНЕ МОДЕЛЮВАННЯ АЕРОАКУСТИЧНИХ ПРОЦЕСІВ ПІД ЧАС ОБТІКАННЯ СЕГМЕНТІВ ЛОПАТЕЙ ВІТРЯНИХ ТУРБІН

Представлено чисельну методику для дослідження аероакустичних процесів, що виникають під час обтікання сегментів лопатей вітрових турбін. Методика ґрунтується на розв'язанні стисливих рівнянь Нав'є – Стокса із застосуванням методу великих вихорів (LES) для точного моделювання турбулентних течій, що дає змогу враховувати складну взаємодію гідродинамічних й акустичних явищ. Запропонований підхід полягає в застосуванні прямого акустичного методу (DAS) для детального аналізу ближнього поля та аналогії Фоукс Вільямса – Гокінгса (FW-H) для ефективного прогнозування шуму в дальньому полі.

З метою верифікації методики чисельні експерименти проведено для аеродинамічного профілю NASA0012 в перехідних режимах обтікання. Відтворено ключові особливості течії, такі як відрив примежового шару, формування відривних зон, перехід примежового шару в турбулентний стан із повторним приєднанням до обтічної поверхні, утворення вихрових структур та їхня взаємодія із задньою крайкою як основне джерело аеродинамічного шуму. Наведено порівняння отриманого розподілу акустичних характеристик у ближньому полі з відомими розрахунковими даними.

Виконано дослідження шумових характеристик профілю NASA0012 у ближньому полі з використанням методу DAS, встановлено зв'язок між аеродинамічними особливостями потоку та механізмом генерації акустичних флуктуацій. Дослідження шумових характеристик у дальньому полі також проведено із використанням аналогії FW-H. Спектральний аналіз, виконаний на основі алгоритму швидкого перетворення Фур'є, продемонстрував узгодження отриманих частот тонального шуму, одержаних методами DAS і FW-H, а також кореляцію із відомими експериментальними даними.

Розроблена методика є ефективним інструментом для вивчення механізмів генерації шуму та може бути застосована у створенні мал шумних конструкцій лопатей вітряних турбін.

Ключові слова: аероакустика, чисельне моделювання, вітрові турбіни, сегменти лопатей, аеродинамічний шум, метод прямого акустичного моделювання, аналогія Фоукс Вільямса – Гокінгса.

Класифікація відповідно до AMS 2020: 76Q05.

Вступ

Аеродинамічний шум, що генерується лопатями вітрових турбін, є однією з ключових проблем сучасної вітроенергетики, яка обмежує більш широке впровадження вітрових технологій, особливо в умовах щільної забудови або в безпосередній близькості до житлових зон. Проблема шуму найбільш критична для малих і середніх вітроустановок, які дедалі частіше застосовують як локальні джерела електроенергії. Такий шум здатен негативно впливати на здоров'я людей, тварин, а також викликати зниження рівня соціального прийняття вітроенергетичних проєктів (Teff-Seker et al., 2022; Zhu, Shen, & Sørensen, 2018).

Складність дослідження аеродинамічного шуму зумовлена комплексними турбулентними течіями навколо лопатей і пов'язаними акустичними явищами, що залежать від конструкції лопаті турбіни, швидкості вітру й умов навколишнього середовища. Традиційні методи вимірювання шуму, такі як польові випробування, є дорогими й обмеженими за точністю через вплив фонового шуму та метеорологічних умов. У цьому контексті чисельні методи моделювання стають ефективними інструментами для прогнозування рівнів аеродинамічного шуму та розробки технологій його зниження (Zhang, Y., Zhang, T., & Li, 2018).

Основна проблема моделювання аеродинамічного шуму пов'язана з необхідністю врахування як його основного джерела, так і динаміки турбулентних течій, що характеризується складною та стохастичною природою. Відомі підходи до моделювання шумових характеристик, засновані на спрощених моделях турбулентності, незважаючи на більш просту практичну реалізацію, мають обмежену точність, що ускладнює їхнє використання під час розроблення нових, тихіших аеродинамічних профілів. Однак сучасні досягнення обчислювальної техніки і їхня дедалі більша обчислювальна потужність відкривають можливості для точного моделювання турбулентних течій, та, відповідно, значно кращого прогнозування шумових характеристик профілів. Проте необхідно зазначити, що для забезпечення надійності чисельних моделей, які розробляються, важливим стає проведення їхньої верифікації шляхом порівняння з експериментальними даними. Такий підхід дає змогу калібрувати численні результати, підвищуючи точність прогнозування шумових характеристик. Тобто чисельне моделювання уможливлює проведення детального аналізу механізмів генерації шуму, тоді як експериментальні дані служать для підтвердження достовірності моделей.

Завдання обчислювальної аероакустики можуть розв'язуватися з використанням таких основних підходів: прямого акустичного моделювання (Direct Acoustic Simulation – DAS) та гібридних методів (акустична аналогія, акустичний / в'язкісний поділ) (Jones et al., 2008; Huang, Yang, & Leung, 2021).

© Алексеевко Сергей, Накашидзе Лилия, Золотаренко Сергей, 2026

Метод DAS дає змогу одночасно отримувати як гідродинамічні, так й акустичні величини шляхом прямого розв'язання рівнянь Нав'є – Стокса для стисливих середовищ. Цей метод є природним способом обчислення генерації звуку, оскільки базові рівняння повністю описують фізику явища (Jones et al., 2008). Також перевагами зазначеного підходу є відсутність обмежень, зумовлених прийнятими припущеннями, можливість уточнення потокоакустичних зв'язків і можливість відстеження процесів генерації та поширення звуку у ближніх і середніх полях. Варто зауважити, що хоча DAS вимагає значних обчислювальних ресурсів, особливо на розв'язання інженерних завдань, з розвитком високопродуктивних обчислювальних технологій методи прямого моделювання стають дедалі доступнішими. Проте бажано використовувати вирішувачі з високою надійністю, сумісністю та можливістю паралельних обчислень (Jones, Sandberg, & Sandham, 2008).

Гібридні методи полягають у використанні двох різних чисельних вирішувачів: нестационарної обчислювальної гідродинаміки як генератора джерела звуку й акустичного вирішувача як методу перенесення. Водночас акустичний вирішувач може бути заснований на розв'язанні системи диференціальних рівнянь у часткових похідних, таких як лінійні рівняння Ейлера або рівняння акустичних збурень (Ewert, & Schröder, 2003), або як інтегральний метод, наприклад, з використанням аналогії FW-H на основі робіт (Lighthill, 1952) та (Ffowcs Williams, & Hawkings, 1969). З обчислювальної позиції інтегральні методи менш затратні і дають змогу отримати явний розв'язок для акустичного тиску, який необхідно чисельно оцінити з початкових умов, одержаних шляхом розв'язання рівнянь Нав'є – Стокса.

У разі використання методу DAS ключовим питанням є моделювання турбулентності. Натеper найточнішим методом моделювання поля турбулентної течії є метод великих вихорів (LES) (Wagner et al., 1996), проте цей підхід обчислювально витратний. Однак значне збільшення обчислювальної потужності, що відбулося останнім часом, та використання технологій паралельних обчислень дають змогу використовувати LES під час моделювання течій, пов'язаних із шумом задньої крайки.

LES ґрунтується на просторовій фільтрації, а не на усередненні за часом, як у випадку з RANS-моделюванням. Просторова фільтрація забезпечує пряме розділення масштабів турбулентності, що перевищують розмір сітки, без будь-яких припущень моделювання. Для масштабів турбулентності, менших за розмір сітки, застосовують спеціальну модель турбулентності, відому як модель підсіткового масштабу (SGS).

Крім згаданих основних підходів до розв'язання задач обчислювальної аероакустики, варто виокремити метод прямого чисельного моделювання (Direct Numerical Simulation – DNS), який є найточнішим методом, оскільки розв'язує повні рівняння Нав'є – Стокса без використання будь-яких моделей турбулентності, даючи змогу повністю розв'язати всі масштаби турбулентності (Sandberg, & Jones, 2010). Цей підхід забезпечує детальний аналіз турбулентних течій та акустичних явищ, але вимагає надзвичайно великих обчислювальних ресурсів, що робить його вкрай обмеженим для практичних застосувань.

Недоліки сучасних методів, зокрема висока обчислювальна складність DNS, обмежена точність FW-H у близькому полі та ресурсоємність DAS, обґрунтовують необхідність розроблення методології, яка б інтегрувала їхні переваги для ефективного моделювання аеродинамічного шуму.

Об'єктом дослідження у пропонованій роботі є аеродинамічне й акустичне поле навколо сегмента лопаті з профілем NACA0012. Мета дослідження – розроблення чисельного інструменту для моделювання аеродинамічного шуму сегментів лопатей вітрових турбін, призначеного для дослідження й оптимізації їхніх акустичних характеристик.

Розділ 1 роботи присвячено опису методик дослідження, розд. 2 і 3 містять ключові теоретичні положення LES, DAS та FW-H, описану постановку задачі та методику розв'язання. Розділи 4 і 5 присвячені результатам чисельних експериментів і їхній верифікації. У розд. 6 наведено обговорення, висновки та перспективи подальших досліджень.

1. Методика дослідження

Запропонована методика являє собою комплексний чисельний підхід до моделювання аероакустичних процесів під час обтікання сегментів лопатей вітрових турбін. Вона інтегрує метод LES для детального опису турбулентних течій, DAS для аналізу ближнього акустичного поля та аналогію FW-H для прогнозування шуму в дальньому полі. Основою слугують стисливі рівняння Нав'є – Стокса, що розв'язуються з використанням LES для фільтрації великих масштабів турбулентності та моделі Смагоринського для підсіткових напружень. Залучення DAS, на відміну від обмеження лише FW-H, допомагає безпосередньо встановлювати зв'язок між гідродинамічними особливостями потоку (такими як відрив примежового шару, формування вихрових структур і турбулентні флуктуації) та механізмами генерації шуму в близькому полі, де гідродинаміка й акустика тісно переплітаються. Проте не всі гідродинамічні характеристики впливають на шум, ключову роль відіграють лише нестационарні компоненти, такі як періодичні вихори чи пульсації, які FW-H ефективно екстраполює в дальнє поле як інтегральний метод, мінімізуючи обчислювальні витрати.

Реалізація методики передбачає чисельний алгоритм розв'язання рівнянь Нав'є – Стокса з дискретизацією третього порядку для точнішого відтворення турбулентних структур і високої роздільної здатності сітки в зонах турбулентності. Для верифікації підходу застосовуємо відомі тестові випадки обтікання аеродинамічних профілів у перехідних режимах, що дає змогу перевірити відтворення ключових особливостей течії, таких як утворення вихрових структур і їхня взаємодія з поверхнею. Обробка результатів передбачає спектральний аналіз (з використанням FFT) для порівняння частотних характеристик між DAS (повний сигнал із сумішшю гідродинамічних й акустичних флуктуацій) і FW-H (ізолюваний акустичний внесок), що допомагає ідентифікувати компоненти гідродинаміки, відповідальні за генерацію шуму, та долати складність розділення акустичної складової у близькому полі. Такий комплексний аналіз у близькому та дальньому полях забезпечує повноту, коли DAS фокусується на локальних механізмах генерації, а FW-H – на поширенні та рівнях шуму.

Загалом методика може слугувати інструментом для вивчення механізмів шумоутворення та розробки технологій зниження шуму, таких як оптимізація геометрії лопатей чи пасивний або активний контроль потоку. Серед переваг можна зазначити баланс між точністю й обчислювальною ефективністю, а також можливість глибокого аналізу взаємозв'язків між гідродинамікою та акустикою. Водночас необхідно зауважити, що, як і в інших чисельних підходах до аероакустики, будь-які особливості розрахункової схеми можуть призводити до розбіжностей у результатах досліджень, тому існує проблема забезпечення відповідності отриманих чисельних результатів реальним експериментальним даним. З огляду

на це ключовим пріоритетом методики є забезпечення її надійності, що гарантує отримання реалістичних результатів для ефективного практичного використання в розробленні технологій шумозниження.

2. Математична постановка задачі

У межах запропонованої методики математична постановка задачі базується на розв'язанні рівнянь Нав'є – Стокса, що враховують стисливість потоку та дають змогу детально описувати джерела шуму. Ці рівняння охоплюють рівняння неперервності, імпульсу й енергії і мають такий вигляд:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (1)$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) = -\nabla p + \nabla \cdot \tau + \rho \mathbf{f}, \quad (2)$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot (\mathbf{u}(\rho E + p)) = \nabla \cdot (\mathbf{u} \cdot \tau) + \rho \mathbf{f} \cdot \mathbf{u} + \nabla \cdot (k \nabla T), \quad (3)$$

де ρ – густина, $\mathbf{u}$ – вектор швидкості, p – тиск, τ – тензор в'язких напружень, E – повна питома енергія, $\mathbf{f}$ – масові (об'ємні) сили, k – коефіцієнт теплопровідності, T – температура.

Система рівнянь (1)–(3) замикається рівнянням стану, що пов'язує тиск, густину та температуру.

У LES великі масштаби турбулентності розв'язують прямим чисельним методом, тоді як дрібні масштаби моделюють за допомогою підсіткових моделей. Фільтровані рівняння Нав'є – Стокса неперервності та імпульсу для LES мають вигляд:

$$\frac{\partial \bar{\rho}}{\partial t} + \nabla \cdot (\bar{\rho} \bar{\mathbf{u}}) = 0 \quad (4)$$

$$\frac{\partial (\bar{\rho} \bar{\mathbf{u}})}{\partial t} + \nabla \cdot (\bar{\rho} \bar{\mathbf{u}} \otimes \bar{\mathbf{u}}) = -\nabla \bar{p} + \nabla \cdot \bar{\tau} - \nabla \cdot \tau^{SGS}, \quad (5)$$

де риска над символом позначає просторове фільтрування відповідної величини; $\bar{\mathbf{u}}$ – швидкість, відфільтрована за Фавром; τ^{SGS} – підсітковий тензор напружень. Для моделювання τ^{SGS} запропоновано використовувати модель Смагоринського:

$$\tau_{ij}^{SGS} = -2\nu_t \bar{S}_{ij}, \quad \nu_t = (C_S \Delta)^2 |\bar{S}|, \quad |\bar{S}| = \sqrt{2\bar{S}_{ij} \bar{S}_{ij}}, \quad (6)$$

де ν_t – турбулентна в'язкість, C_S – константа Смагоринського (приймаємо 0,1), Δ – ширина фільтра, $\bar{S}_{ij}$ – тензор деформації.

Наведені рівняння описують динаміку стисливого потоку, зокрема турбулентність й акустичні хвилі, що є критично важливим для моделювання аеродинамічного шуму (Wagner et al., 1996).

У разі застосування аналогії FW-H, що використовує інтегральний підхід для обчислення акустичного тиску на основі даних про течію, рівняння для визначення акустичного збурення тиску в точці спостереження має вигляд:

$$p' = \frac{\partial^2}{\partial x_i \partial x_j} [T_{ij} H(f)] + \frac{\partial}{\partial x_i} [L_i \delta(f)] + \frac{\partial}{\partial t} [\dot{M} \delta(f)], \quad (7)$$

де p' – акустичне збурення тиску, T_{ij} – тензор Лайтхілла, $H(f)$ і $\delta(f)$ – функції Хевісайда і дельта-функція Дірака, L_i – векторна щільність силових джерел на поверхні, а $\dot{M}$ – щільність джерела масового потоку.

Формула (7) справедлива за умови, якщо потік за межами інтеграційної поверхні вважати незбуреним (тобто гідродинамічні збурення є малими, а їхній внесок у генерацію шуму є неістотним), а сама поверхня може бути рухомою або деформованою.

Метод DAS передбачає пряме розв'язання повних стисливих рівнянь Нав'є – Стокса без додаткових припущень, що забезпечує високу точність у ближньому полі, де гідродинамічні й акустичні ефекти тісно пов'язані (Belamadi et al., 2022).

Для використання зазначених методів необхідно обирати точки спостереження в просторі, де буде обчислюватися акустичний тиск. Вибір точок спостереження залежить від методу і цілей аналізу.

3. Чисельний метод

Чисельний метод охоплює дискретизацію рівнянь Нав'є – Стокса для стисливих середовищ, моделювання турбулентних течій за допомогою методу LES, обчислення акустичного поля з використанням аналогії FW-H для дальнього поля та DAS для ближнього поля, а також налаштування обчислювальних сіток, граничних умов і постобробки результатів.

У запропонованій методиці застосовано неявний чисельний алгоритм розв'язання рівнянь Нав'є – Стокса зі зв'язуванням тиску та швидкості. Просторову дискретизацію виконано третьопорядковою схемою MUSCL для забезпечення вищої точності відтворення турбулентних структур і, відповідно, акустичних полів. Для часової дискретизації використано неявну схему другого порядку, що гарантує стабільність обчислень. Часовий крок Δt було обрано так, щоб число Курранта – Фрідрікса – Леві (CFL) було меншим 1, з урахуванням умови $\Delta t \leq \min\left(\frac{\Delta x}{c \cdot CFL}, \frac{1}{10 f_{max}}\right)$, де Δx – мінімальний просторовий крок, c – швидкість звуку, а f_{max} – верхня межа частотного спектру, що визначає мінімальну точність дискретизації в часі.

При цьому під час моделювання використовується сітка з високою роздільною здатністю в зонах зі значною турбулентністю, таких як задня крайка і слід за сегментом лопаті, де розмір комірки був меншим за 1/10 масштабу турбулентних вихорів. Під час розрахунків для досягнення статистично стійкого стану течії було забезпечено достатню кількість часових кроків, прийняту як 20 періодів обтікання сегменту лопаті.

У разі застосування FW-H до нерухомої твердотілої поверхні, рівняння (7) спрощується до так званої акустичної інтегральної формули:

$$4\pi p'(\mathbf{X}, t) = \frac{\partial}{\partial t} \int_S \frac{\rho_0 U_n}{r} dS + \int_S \frac{\partial}{\partial n} \left(\frac{p - p_0}{r} \right) dS, \quad (8)$$

де $\mathbf{X}$ – точка спостереження, r – відстань між джерелом і приймачем, U_n – нормальна складова швидкості на поверхні S , p_0 – тиск спокою. Нормаль n орієнтована зовні від поверхні обтікання.

Із використанням DAS виконано пряме розв'язання повних стисливих рівнянь Нав'є – Стокса та здійснено аналіз флуктуацій тиску в обраних точках спостереження в просторі.

Для примежового шару забезпечено виконання вимоги $y^+ \leq 0.2$ для першої комірки біля поверхні, що гарантує правильне розв'язання в'язкого підшару. Тут y^+ – це безрозмірна відстань від стінки до центру першої сіткової комірки, що враховує в'язкі характеристики потоку.

4. Обробка результатів розрахунків

Ключовим етапом реалізації методики чисельного моделювання аероакустичних процесів у роботі є оцінка шумових характеристик сегментів лопатей, що досліджуються, з подальшим аналізом та локалізацією джерел шуму. Запропонований підхід передбачає аналіз акустичного тиску, отриманого за допомогою DAS та аналогії FW-H, з подальшим обчисленням спектрів шуму, рівнів звукового тиску (SPL), візуалізацією джерел шуму та порівнянням результатів. Хоча обидва методи мають спільні елементи постобробки, такі як використання FFT й обчислення SPL, процеси отримання даних різняться: у DAS акустичний тиск є частиною розв'язку повних стисливих рівнянь Нав'є – Стокса, тоді як у FW-H його визначають через інтегральні співвідношення на основі даних про течію, отриманих із розрахунків із використанням LES (Wagner, Bareiß, & Guidati, 1996). У методиці передбачається порівняння результатів DAS і FW-H за піковими частотами спектрів шуму, а також верифікація через порівняння з експериментальними даними у ближньому полі.

Точки спостереження для DAS були розташовані у ближньому полі поблизу задньої крайки сегмента лопаті, для FW-H – у дальньому полі.

Спектри шуму використано для аналізу частотних характеристик аеродинамічного шуму, що дає змогу ідентифікувати домінуючі частоти та їхні джерела. Для обох використаних методів спектри отримують шляхом застосування FFT до часових рядів акустичного тиску.

У методі DAS акустичне поле є складовою частиною загального розв'язку рівнянь Нав'є – Стокса для стисливого середовища (Huang, Yang, & Leung, 2021). Це дає змогу безпосередньо отримувати часові ряди акустичного тиску $p'(t)$ у вибраних точках спостереження.

Для аналізу частотних характеристик шуму до цих сигналів застосовуємо швидко перетворення Фур'є, що допомагає обчислити комплексну спектральну амплітуду:

$$P(f) = \int_{-\infty}^{\infty} p'(t) e^{-i2\pi f t} dt, \quad (9)$$

де $P(f)$ – комплексна спектральна компонента акустичного тиску на частоті f , $p'(t)$ – часовий ряд акустичного тиску в точці спостереження. Цей процес дає змогу визначити частотний розподіл шуму, наприклад, тональні компоненти, пов'язані з вихреутворенням, або широкосмуговий шум від турбулентності (Belamadi et al., 2022).

У методі FW-H акустичний тиск визначають окремо на основі отриманих даних про течію. При цьому основне рівняння FW-H (7) для твердої поверхні сегментів лопатей спрощується до вигляду (8).

Рівень звукового тиску (SPL) є загальноприйнятим показником рівня шуму, що допомагає порівнювати чисельні результати з експериментальними даними або нормативними вимогами.

У DAS, оскільки акустичний тиск $p'(t)$ є частиною розв'язку, SPL обчислюється безпосередньо із часових рядів тиску в обраних точках спостереження:

$$SPL = 20 \log_{10} \left(\frac{p_{rms}}{p_0} \right), \quad p_{rms} = \sqrt{\frac{1}{T} \int_0^T p'^2(t) dt}, \quad (10)$$

де p_{rms} – середньоквадратичне значення акустичного тиску за період спостереження T , $p_0 = 20 \mu Pa$ – опорний тиск, що відповідає пороговому значенню чутності для людини.

Така метрика надає можливість прямо порівнювати результати моделювання з експериментальними даними або нормативними вимогами щодо шуму.

У методі FW-H після обчислення акустичного тиску $p'(t)$ за допомогою інтегральних формул у точках спостереження SPL розраховують за тією ж формулою.

5. Верифікація чисельної методики

З метою верифікації запропонованого підходу було виконано серію обчислювальних експериментів. Як тестовий було обрано відомий розрахунковий випадок обтікання симетричного профілю NACA0012 при числі Рейнольдса $Re = 50\,000$ і числі Маха $M = 0.4$. Розрахунки проведено для трьох значень кута атаки: 3° , 5° та 7° , що допомогло охопити діапазон режимів обтікання профілю з утворенням відривного пухиря на поверхні всмоктування.

Обчислювальна область містила 1 440 000 елементів ортогональної O-сітки, сформованої навколо аеродинамічного профілю з адіабатичною стінкою без проковзування. На відстані 40 довжин хорди від профілю застосовували умови вільного потоку із поступовим розрідженням сітки до межі для ефективного усунення штучних відбиттів.

Отримані результати були зіставлені з відомими чисельними даними (Arif et al., 2020; 2022), що дало можливість підтвердити коректність реалізації математичної й обчислювальної моделей та їхню здатність відтворювати складні вихрові структури течії, взаємодія яких із обтічною поверхнею профілю є основним джерелом акустичного випромінювання.

Унаслідок моделювання за запропонованою методикою було відтворено складну динаміку течії поблизу поверхні профілю, зокрема формування відривного пухиря, його пульсації та подальше утворення регулярних вихрових структур у задньому шлейфі (рис. 1).

Під дією негативного градієнта тиску ламінарний примежовий шар відривається та під впливом збурень, що нарастають, переходить у турбулентний режим. При цьому в примежовому шарі посилюється перемішування та передача імпульсу в напрямку, перпендикулярному до поверхні. Унаслідок втрати енергії примежовий шар знову приєднується до обтічної поверхні, утворюючи так званий "відривний пухир". У зоні повторного приєднання через нестійкість, що розвивається під час обтікання поверхні, зароджуються вихори. Далі, взаємодіючи із задньою крайкою профілю, ці вихори генерують тональний шум.

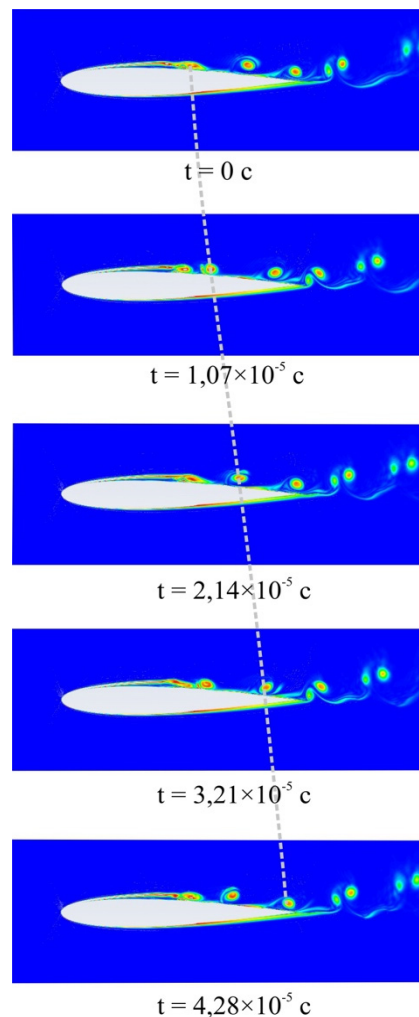


Рис. 1. Еволюція розподілу завихреності в часі під час обтікання профілю NASA0012 під кутом атаки $\alpha = 5^\circ$

На рис. 1 проілюстровано переміщення таких вихрових структур протягом одного циклу, що відповідає, у цьому випадку, виникненню трьох флуктуацій тиску у ближньому полі. Картина течії та динаміка переміщення вихорів є подібними до результатів, опублікованих у (Arif, 2020).

Крім того, негативний градієнт тиску з боку нагнітання поблизу задньої крайки призводить до викривлення профілів швидкості у примежовому шарі, які також переходять у зону відривного потоку. Нестійкість, що поширюється вниз за потоком, істотно посилюється в цій області й може охоплювати задню крайку, беручи участь у формуванні зворотної кінцевої вихору. Як зазначалося, ці вихори взаємодіють із задньою крайкою, створюючи коливальне поле навколо профілю на тій самій частоті, що й спостережувана нестійкість, і, відповідно, спостережуваний дискретний тон.

Окрім того, на основі розподілу моментальних полів завихреності було проведено ідентифікацію точок відриву та повторного приєднання примежового шару, із формуванням "відривного пухиря", а також виконано порівняння з відомими даними (рис. 2). Спостерігається хороше узгодження даних.

Отже, запропонована методика моделювання аеродинамічного шуму з використанням прямого розв'язання стисливих рівнянь Нав'є – Стокса (DAS) із застосуванням моделі LES, що також лежить в основі методу із використанням інтегрального підходу FW-H, дає змогу відтворити характерну структуру обтікання профілю та фізичні механізми генерації аеродинамічного шуму.

6. Результати розрахунків

Наступні дослідження процесів розповсюдження аеродинамічного шуму були проведені також на прикладі профілю NASA0012, що обтікається під кутом атаки 5° із характеристиками потоку $M = 0,4$ і $Re = 5 \times 10^4$. На рис. 3 наведено миттєвий розподіл поверхневого коефіцієнта тиску та відносних коливань тиску в момент взаємодії вихрової структури із задньою крайкою профілю. На рис. 3 б показано область поблизу профілю, що охоплює ближнє та перехідне до дальнього поле. Спостерігаємо значні флуктуації тиску на задній частині верхньої поверхні профілю, що зумовлені послідовним сходженням за потоком вихрових структур після точки повторного приєднання відокремленого примежового шару (рис. 3 а). Окрім того, на нижній частині профілю також бачимо коливання коефіцієнта поверхневого тиску (але зі значно меншою амплітудою) із частотою, що збігається із частотою сходження вихрових структур із задньої крайки профілю. Унаслідок аналізу поля тиску в ближньому полі робимо висновок про те, що основним джерелом акустичних коливань є задня крайка профілю (рис. 3 б), оскільки центр концентричних колоподібних хвиль розміщується на цій крайці, а коливання тиску вниз і вгору від крайки – у протифазі.

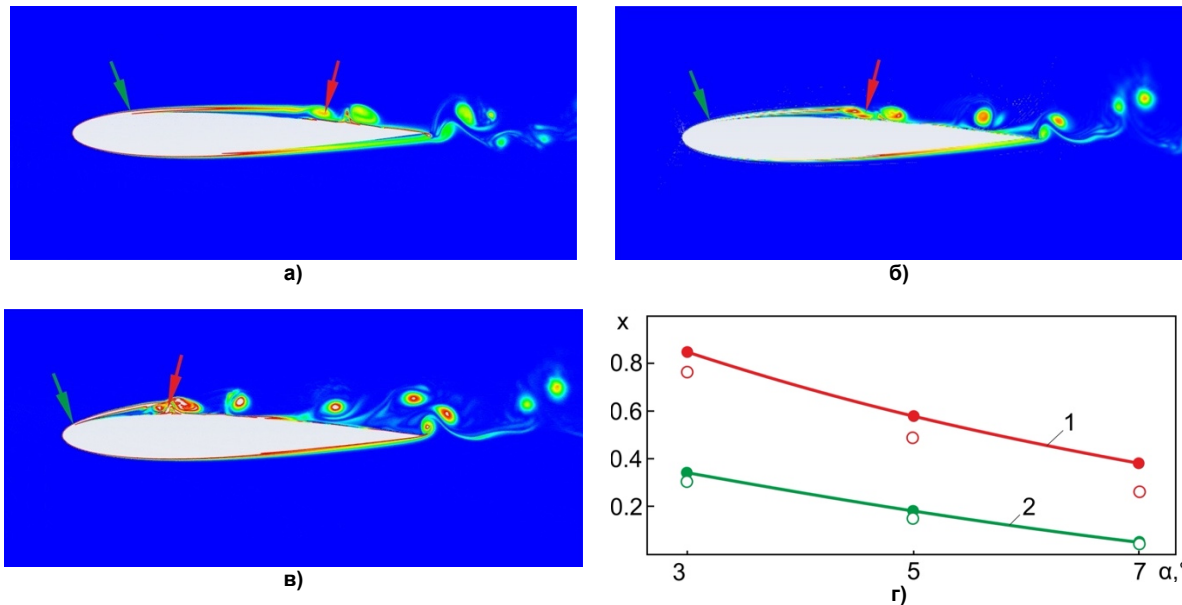


Рис. 2. Положення точок відриву та повторного приєднання примежового шару за перехідного режиму обтікання профілю NASA0012 з утворенням "відривного пухиря", для кута атаки $\alpha = 3^\circ$ (а), 5° (б) та 7° (в), (г) – крива 1 відповідає положенню точок повторного приєднання, крива 2 – положенню точок відриву примежового шару, • – (Arif et al., 2020), ° – результати цієї роботи

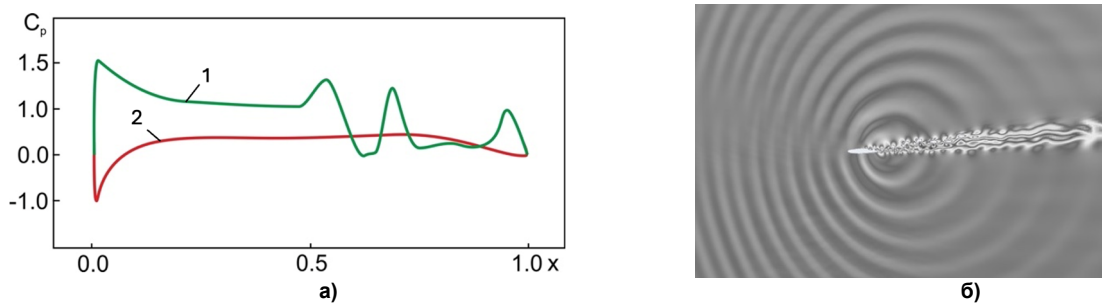


Рис. 3. Миттєвий розподіл коефіцієнта тиску по обтічній поверхні: (а) – крива 1 відповідає верхній частині профілю, крива 2 – нижній частині профілю; миттєві відносні коливання тиску в області поблизу профілю – візуалізація задньої крайки профілю як джерела аеродинамічного шуму (б)

На рис. 4 наведено кругову діаграму розподілу відносного тиску p/p_∞ (де p_∞ – тиск потоку, що набігає) та середньоквадратичне значення акустичного тиску за період спостереження p'_{rms} по колу на відстані трьох довжин хорди від задньої крайки профілю. Також виконано порівняння із відомими розрахунковими даними (Arif et al., 2020), спостерігається розходження p/p_∞ в межах 0,5 % та якісне і кількісне узгодження p'_{rms} із відхиленням не більше 25 %.

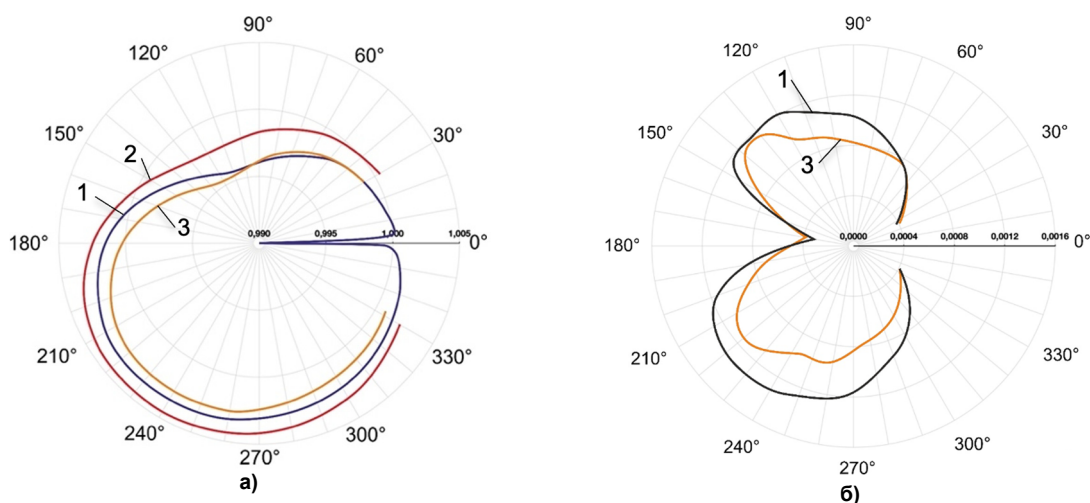


Рис. 4. Просторовий розподіл акустичних характеристик: азимутальний графік усередненого за часом p/p_∞ (а) та азимутальний графік p'_{rms} (б) для $r = 3$, $\alpha = 5^\circ$; 1, 2 – результати роботи (Arif et al., 2020), 3 – результати цієї роботи

Крім цього, чисельні дослідження були проведені для профілю NACA0012 із довжиною хорди $c = 0,1$ м із швидкістю набігаючого потоку $V_\infty = 24$ м/с під кутом атаки $\alpha = 0^\circ$, що відповідало числу Рейнольдса $Re \sim 160\,000$. Точка вимірювання коливань тиску розташовувалася на відстані $0,1$ м над задньою крайкою профілю. Спектральний аналіз результатів, отриманих унаслідок прямого моделювання, було проведено за допомогою швидкого перетворення Фур'є із частотною роздільною здатністю 12 Гц. Отримані результати зіставили з відомими експериментальними даними (Yakhina et al., 2020). Результати порівняння отриманих спектрів шуму (рис. 5) показали хорошу кореляцію частот піків: основний тон – 1300 Гц (експеримент) проти 1240 Гц (чисельне моделювання); вторинні тони – 1410 Гц, 1520 Гц, 1200 Гц (експеримент) проти 1410 Гц, 1130 Гц, 1130 Гц (чисельне моделювання), але водночас відсутність вторинного піку із частотою 844 Гц (чисельне моделювання) в експерименті. Проте рівень звукового тиску у виконаних двовимірних розрахунках виявився вищим на ~ 15 дБ, про що також є згадка, наприклад, в (Nguyen et al., 2021).

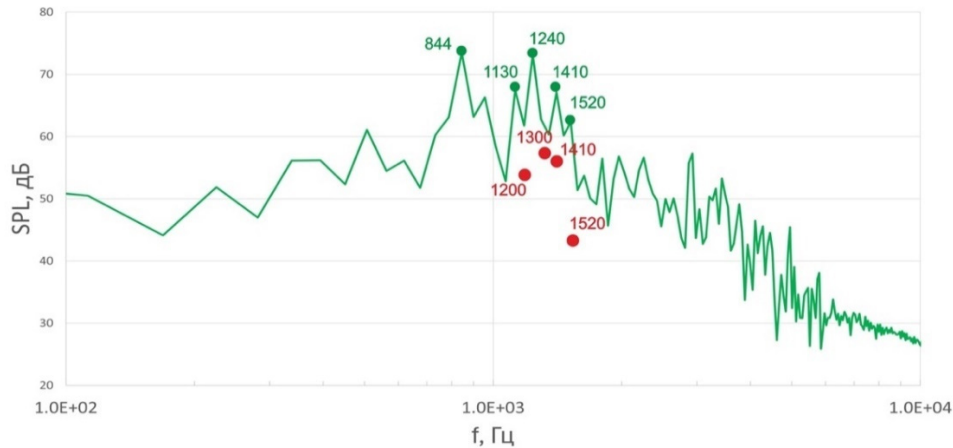


Рис. 5. Спектри шуму для моделі сегменту лопаті з профілем NACA0012 зі швидкістю набігаючого потоку 24 м/с і кутом атаки 0° , порівняння розрахункових і відомих експериментальних даних

З іншого боку, для дослідження відповідності результатів, отриманих методами DAS і FW-H, було проведено дослідження частотних спектрів аеродинамічного шуму під час обтікання профілю NACA0012 із довжиною хорди $c = 0,215$ м зі швидкістю набігаючого потоку $V_\infty = 10$ м/с під кутом атаки $\alpha = 0^\circ$. Точка спостереження у ближньому полі розміщується на відстані $0,2$ м над задньою крайкою профілю, а у дальньому полі – $2,1$ м.

Отримані результати наведено на рис. 6. Спостерігаємо практично повну відповідність частотних піків у шумових спектрах у ближньому і дальньому полях. Отже, у роботі показано, що метод DAS під час дослідження процесів генерації та розповсюдження аеродинамічного шуму в ближньому полі сприяє отриманню абсолютно коректних частотних розподілів, і, відповідно, є ефективним і достатньо точним інструментом для розроблення технологій зниження аеродинамічних шумів. Однак розрахункове значення рівня звукового тиску не може бути напряму зіставлене із фізичними значеннями.

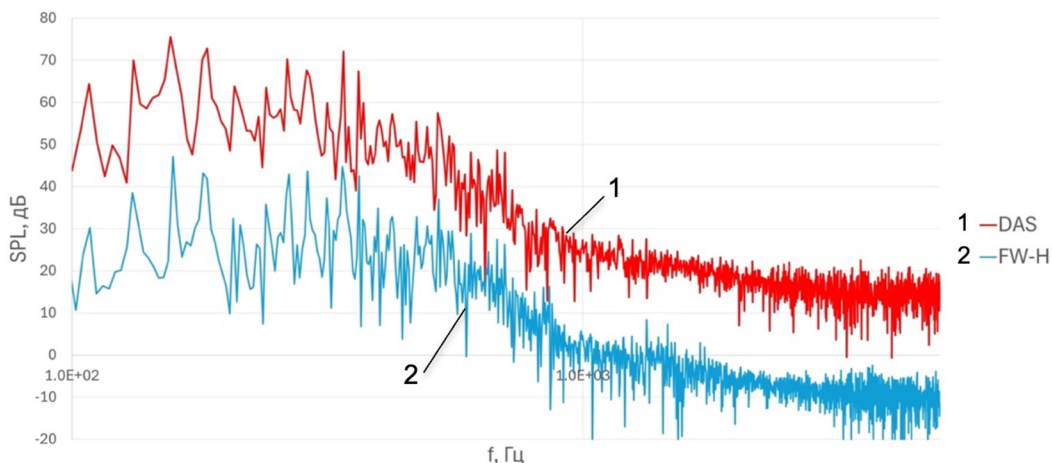


Рис. 6. Спектри шуму для моделі сегменту лопаті з профілем NACA0012 довжиною хорди $0,215$ м зі швидкістю набігаючого потоку 20 м/с і кутом атаки 0° , порівняння розрахункових даних, отриманих методами DAS і FW-H

Використання двовимірної постановки задачі дало змогу в межах проведених досліджень застосувати детальну сітку та врахувати припущення про ламінарний характер течії на значній частині профілю з можливими відривними зонами. Також, як зазначено в (Desquesnes, Terracol, & Sagaut, 2007), двовимірні розрахунки здатні відтворювати фізику, притаманну утворенню вихровим структурам під час обтікання перехідних профілів, та пікові тони аеродинамічного шуму.

Отже, чисельне моделювання за допомогою розробленої методики на прикладі двовимірної постановки сприяло успішному відтворенню основних особливостей течії, охоплюючи відрив ламінарного пограничного шару, з його повторним приєднанням та наступним утворенням вихрових структур. Ці структури, взаємодіючи із задньою крайкою, генерують

пульсації тиску, що є джерелом аеродинамічного шуму. Хороша кореляція частот піків між отриманими чисельними та відомими експериментальними даними підтверджує здатність моделі точно відтворювати механізми генерації шуму.

Розбіжність у рівнях звукового тиску може бути пов'язана з обмеженнями двовимірної моделі, яка не враховує тривимірні ефекти, що впливають на інтенсивність вихрових структур. Хоча тривимірні моделі можуть забезпечити точніші прогнози рівнів шуму, проте двовимірна модель виявилася також ефективною для аналізу частотного складу й ідентифікації джерел шуму, що є важливим для розроблення стратегій зменшення шумності.

Дискусія і висновки

У роботі розроблено та реалізовано чисельну методику моделювання аеродинамічного шуму, що виникає під час обтікання сегментів лопатей вітрових турбін. Методика базується на розв'язанні стисливих рівнянь Нав'є – Стокса із застосуванням методу LES та полягає у застосуванні підходів DAS й інтегральної аналогії FW-H, що дає можливість відтворювати як ближнє, так і дальнє акустичне поле. Проведене дослідження показало, що основним джерелом шуму під час обтікання перехідних профілів є періодичні вихрові структури, що утворюються внаслідок нестійкостей у примежовому шарі та які взаємодіють із задньою крайкою профілю, формуючи тональний компонент акустичного випромінювання.

Чисельні розрахунки для профілю NACA0012 за малих кутів атаки та числі Рейнольдса до 160 000 допомогли відтворити характерну структуру течії, включно із формуванням відривного пухиря, періодичним вихреутворенням і коливаннями підйомної сили. Аналіз флуктуацій тиску у ближньому полі виявив концентрацію акустичної активності поблизу задньої крайки, що повністю відповідає теоретичним уявленням про механізм генерації тонального шуму через акустично-гідродинамічний зворотний зв'язок. Спектральний аналіз отриманих часових сигналів акустичного тиску продемонстрував високу точність прогнозування частот основного та вторинних тонів, що добре збігаються з результатами відомих експериментальних вимірювань (відхилення не перевищує 3–5 %).

Водночас встановлено, що рівень звукового тиску, розрахований у двовимірній постановці задачі, перевищує експериментальний приблизно на 15 дБ, що, вочевидь, пояснюється відсутністю тривимірної дисипації в моделі.

Незважаючи на це, отримані результати є достатньо точними для якісного аналізу механізмів шумоутворення та дають змогу ефективно ідентифікувати джерела шуму під час обтікання сегменту лопаті вітряної турбіни. Запропонована методика є гнучким і достовірним інструментом, який може бути використаний як для фундаментальних досліджень, так і для практичного проектування мал шумних профілів технічних пристроїв. Подальші дослідження передбачають розширення моделі до тривимірної постановки, а також вивчення впливу геометричних модифікацій на акустичну ефективність лопаті.

Внесок авторів: Сергій Алексєнко – концептуалізація, методика, теоретичні основи дослідження; Лілія Накашидзе – методологія, формальний аналіз, редагування тексту; Сергій Золотаренко – аналіз даних, візуалізація.

Джерела фінансування. Роботу виконано за підтримки зовнішнього інструменту допомоги Європейського Союзу для виконання зобов'язань України у Рамковій програмі Європейського Союзу з наукових досліджень та інновацій "Горизонт 2020", договір PH/54-2024 з Міністерством освіти і науки України, № державної реєстрації: 0124U004405.

Список використаних джерел

- Arif, I., Lam, G. C. Y., Leung, R. C. K., & Naseer, M. R. (2022). Distributed surface compliance for airfoil tonal noise reduction at various loading conditions. *Physics of Fluids*, 34, 046113. <https://doi.org/10.1063/5.0087350>
- Arif, I., Lam, G. C. Y., Wu, D., & Leung, R. C. K. (2020). Passive airfoil tonal noise reduction by localized flow-induced vibration of an elastic panel. *Aerospace Science and Technology*, 107, 106319. <https://doi.org/10.1016/j.ast.2020.106319>
- Belamadi, R., Settar, A., Chetehouna, K., & Ilinca, A. (2022). Numerical modeling of horizontal axis wind turbine: aerodynamic performances improvement using an efficient passive flow control system. *Energies*, 15(13), 4872. <https://doi.org/10.3390/en15134872>
- Desquesnes, M., Terracol, M., & Sagaut, P. (2007). Numerical investigation of the tone noise mechanism over laminar airfoils. *Journal of Fluid Mechanics*, 591, 155–182. <https://doi.org/10.1017/S0022112007007896>
- Ewert, R., & Schröder, W. (2003). Acoustic perturbation equations based on flow decomposition via source filtering. *Journal of Computational Physics*, 188(2), 365–398. [https://doi.org/10.1016/S0021-9991\(03\)00168-2](https://doi.org/10.1016/S0021-9991(03)00168-2)
- Ffowcs Williams, J. E., & Hawkings, D. L. (1969). Sound generation by turbulence and surfaces in arbitrary motion. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 264(1151), 321–342. <https://doi.org/10.1098/rsta.1969.0031>
- Huang, D., Yang, Z., & Leung, R. C. K. (2021). Implementation of direct acoustic simulation using ANSYS Fluent. *InterNoise and Noise-Con Congress and Conference Proceedings*, 263(3), 1243–1252. <https://doi.org/10.3397/IN-2021-1787>
- Jones, L. E., Sandberg, R. D., & Sandham, N. D. (2008). Direct numerical simulations of forced and unforced separation bubbles on an airfoil at incidence. *Journal of Fluid Mechanics*, 602, 175–207. <https://doi.org/10.1017/S0022112008000864>
- Lighthill, M. J. (1952). On sound generated aerodynamically. I. General theory. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 211(1107), 564–587. <https://doi.org/10.1098/rspa.1952.0060>
- Nguyen, L., Golubev, V., Mankbadi, R., Yakhina, G., & Roger, M. (2021). Numerical investigation of tonal trailing-edge noise radiated by low Reynolds number airfoils. *Applied Sciences*, 11(5), 2257. <https://doi.org/10.3390/app11052257>
- Sandberg, R. D., & Jones, L. E. (2010). Direct numerical simulations of airfoil self-noise. *Procedia Engineering*, 6, 274–282. <https://doi.org/10.1016/j.proeng.2010.09.029>
- Teff-Seker, Y., Berger-Tal, O., Lehnardt, Y., & Teschner, N. (2022). Noise pollution from wind turbines and its effects on wildlife: A cross-national analysis of current policies and planning regulations. *Renewable and Sustainable Energy Reviews*, 168, 112801. <https://doi.org/10.1016/j.rser.2022.112801>
- Wagner, S., Bareiß, R., & Guidati, G. (1996). *Wind Turbine Noise*. Springer. <https://doi.org/10.1007/978-3-642-88710-9>
- Yakhina, G., Roger, M., Moreau, S., Nguyen, L., & Golubev, V. (2020). Experimental and analytical investigation of the tonal trailing-edge noise radiated by low Reynolds number airfoils. *Acoustics*, 2(2), 310–329. <https://doi.org/10.3390/acoustics2020018>
- Zhang, Y., Zhang, T., & Li, X. (2018). Influences of operating parameters on the aerodynamics and aeroacoustics of a horizontal-axis wind turbine. *Energy*, 160, 597–611. <https://doi.org/10.1016/j.energy.2018.07.048>
- Zhu, W. J., Shen, W. Z., & Sørensen, J. N. (2018). Wind turbine noise generation and propagation modeling at DTU Wind Energy: A review. *Renewable and Sustainable Energy Reviews*, 88, 133–150. <https://doi.org/10.1016/j.rser.2018.02.029>

References

- Arif, I., Lam, G. C. Y., Leung, R. C. K., & Naseer, M. R. (2022). Distributed surface compliance for airfoil tonal noise reduction at various loading conditions. *Physics of Fluids*, 34, 046113. <https://doi.org/10.1063/5.0087350>
- Arif, I., Lam, G. C. Y., Wu, D., & Leung, R. C. K. (2020). Passive airfoil tonal noise reduction by localized flow-induced vibration of an elastic panel. *Aerospace Science and Technology*, 107, 106319. <https://doi.org/10.1016/j.ast.2020.106319>
- Belamadi, R., Settar, A., Chetehouna, K., & Ilinca, A. (2022). Numerical modeling of horizontal axis wind turbine: aerodynamic performances improvement using an efficient passive flow control system. *Energies*, 15(13), 4872. <https://doi.org/10.3390/en15134872>
- Desquesnes, M., Terracol, M., & Sagaut, P. (2007). Numerical investigation of the tone noise mechanism over laminar airfoils. *Journal of Fluid Mechanics*, 591, 155–182. <https://doi.org/10.1017/S0022112007007896>
- Ewert, R., & Schröder, W. (2003). Acoustic perturbation equations based on flow decomposition via source filtering. *Journal of Computational Physics*, 188(2), 365–398. [https://doi.org/10.1016/S0021-9991\(03\)00168-2](https://doi.org/10.1016/S0021-9991(03)00168-2)
- Ffowcs Williams, J. E., & Hawkings, D. L. (1969). Sound generation by turbulence and surfaces in arbitrary motion. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 264(1151), 321–342. <https://doi.org/10.1098/rsta.1969.0031>
- Huang, D., Yang, Z., & Leung, R. C. K. (2021). Implementation of direct acoustic simulation using ANSYS Fluent. *InterNoise and Noise-Con Congress and Conference Proceedings*, 263(3), 1243–1252. <https://doi.org/10.3397/IN-2021-1787>
- Jones, L. E., Sandberg, R. D., & Sandham, N. D. (2008). Direct numerical simulations of forced and unforced separation bubbles on an airfoil at incidence. *Journal of Fluid Mechanics*, 602, 175–207. <https://doi.org/10.1017/S0022112008000864>
- Lighthill, M. J. (1952). On sound generated aerodynamically. I. General theory. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 211(1107), 564–587. <https://doi.org/10.1098/rspa.1952.0060>
- Nguyen, L., Golubev, V., Mankbadi, R., Yakhina, G., & Roger, M. (2021). Numerical investigation of tonal trailing-edge noise radiated by low reynolds number airfoils. *Applied Sciences*, 11(5), 2257. <https://doi.org/10.3390/app11052257>
- Sandberg, R. D., & Jones, L. E. (2010). Direct numerical simulations of airfoil self-noise. *Procedia Engineering*, 6, 274–282. <https://doi.org/10.1016/j.proeng.2010.09.029>
- Teff-Seker, Y., Berger-Tal, O., Lehnardt, Y., & Teschner, N. (2022). Noise pollution from wind turbines and its effects on wildlife: A cross-national analysis of current policies and planning regulations. *Renewable and Sustainable Energy Reviews*, 168, 112801. <https://doi.org/10.1016/j.rser.2022.112801>
- Wagner, S., Bareiß, R., & Guidati, G. (1996). *Wind Turbine Noise*. Springer. <https://doi.org/10.1007/978-3-642-88710-9>
- Yakhina, G., Roger, M., Moreau, S., Nguyen, L., & Golubev, V. (2020). Experimental and analytical investigation of the tonal trailing-edge noise radiated by low Reynolds number airfoils. *Acoustics*, 2(2), 310–329. <https://doi.org/10.3390/acoustics2020018>
- Zhang, Y., Zhang, T., & Li, X. (2018). Influences of operating parameters on the aerodynamics and aeroacoustics of a horizontal-axis wind turbine. *Energy*, 160, 597–611. <https://doi.org/10.1016/j.energy.2018.07.048>
- Zhu, W. J., Shen, W. Z., & Sørensen, J. N. (2018). Wind turbine noise generation and propagation modeling at DTU Wind Energy: A review. *Renewable and Sustainable Energy Reviews*, 88, 133–150. <https://doi.org/10.1016/j.rser.2018.02.029>

Отримано редакцією журналу / Received: 09.06.25

Прорецензовано / Revised: 18.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Serhii ALEKSIENKO, DSc (Engin.), Prof.
ORCID ID: 0000-0003-0320-989X
e-mail: aleksienko.s.v@nmu.one
Dnipro University of Technology, Dnipro, Ukraine

Liliya NAKASHYDZE, DSc (Engin.), Assoc. Prof.
ORCID ID: 0000-0003-3990-6718
e-mail: foton_dnu@ukr.net
Oles Honchar Dnipro National University, Dnipro, Ukraine

Serhii ZOLOTARENKO, PhD Student
ORCID ID: 0009-0008-5291-4206
e-mail: zolotarenko.s.an@nmu.one
Dnipro University of Technology, Dnipro, Ukraine

NUMERICAL MODELING OF AEROACOUSTIC PROCESSES DURING THE FLOW AROUND WIND TURBINE BLADE SEGMENTS

The paper presents a numerical methodology for studying aeroacoustic processes that occur during flow around wind turbine blade segments. The methodology is based on solving the compressible Navier – Stokes equations using the large eddy simulation (LES) method for accurate modeling of turbulent flows, which allows taking into account the complex interaction of hydrodynamic and acoustic phenomena. The proposed approach consists in applying the direct acoustic method (DAS) for detailed analysis of the near field and the Fawkes Williams – Hawkings (FW-H) analogy for effective prediction of noise in the far field.

To verify the methodology, numerical experiments were conducted for the NACA0012 airfoil in transient flow regimes. Key flow features are reproduced, such as boundary layer separation, formation of separation zones, transition of the boundary layer to a turbulent state with reattachment to the streamlined surface, formation of vortex structures and their interaction with the trailing edge as the main source of aerodynamic noise. A comparison of the obtained distribution of acoustic characteristics in the near field with known calculation data is presented.

The noise characteristics of the NACA0012 profile in the near field were studied using the DAS method, establishing a connection between the aerodynamic features of the flow and the mechanism of acoustic fluctuations generation. The noise characteristics in the far field were also studied using the FW-H analogy. Spectral analysis, performed based on the fast Fourier transform algorithm, demonstrated the agreement of the obtained tonal noise frequencies obtained by the DAS and FW-H methods, as well as correlation with known experimental data.

The developed methodology is an effective tool for studying noise generation mechanisms and can be applied in creating low-noise designs of wind turbine blades.

Key words: aeroacoustics, numerical modeling, wind turbines, blade segments, aerodynamic noise, Direct Acoustic Simulation, Ffowcs Williams – Hawkings analogy.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 539.3

DOI: <https://doi.org/10.17721/1812-5409.2026/1.14>

Олександр ГРИГОРЕНКО¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-4109-2672
e-mail: ayagrigorenko1991@gmail.com

Володимир ТРАЧ², д-р техн. наук, проф.
ORCID ID: 0000-0001-9500-2743
e-mail: v.m.trach@nuwm.edu.ua

Андрій ПОДВОРНИЙ², д-р техн. наук, доц.
ORCID ID: 0000-0001-8518-4395
e-mail: a.v.podvorny@nuwm.edu.ua

Наталія ЖУКОВА¹, канд. фіз.-мат. наук, ст. наук. співроб.
ORCID ID: 0000-0002-2699-3771
e-mail: zukowa_n@ukr.net

¹Інститут механіки ім. С. П. Тимошенка НАН України, Київ, Україна

²Національний університет водного господарства та природокористування, Рівне, Україна

НАПРУЖЕНО-ДЕФОРМОВАНІЙ СТАН АНІЗОТРОПНИХ ТОВСТОСТІННИХ ЦИЛІНДРИЧНИХ ОБОЛОНОК ПІД ЗМІННИМ НАВАНТАЖЕННЯМ

Товстостінні циліндричні оболонкові конструкції є важливими елементами різноманітних сучасних інженерних споруд, таких як трубопроводи, резервуари, реакторні установки й аерокосмічні корпуси. Експлуатація цих конструкцій в умовах складного силового або температурного навантаження потребує достовірного аналізу їхнього напружено-деформованого стану, особливо у разі дії розподіленого навантаження, що може змінюватися за коловим напрямком. Подібні впливи можуть бути зумовлені нерівномірним тиском робочого середовища, як-от несиметричний турбулентний потік у трубопроводах або асиметричне аеродинамічне навантаження на корпуси літальних апаратів. Подібні чинники можуть суттєво впливати на міцність, стійкість і довговічність конструкцій.

У пропонованій роботі розглянуто просторовий підхід до розв'язку задачі дослідження напружено-деформованого стану для анізотропної товстостінної циліндричної оболонки з боропластика, на яку діє розподілений тиск, періодично змінний у коловому напрямку. Анізотропія пружних властивостей матеріалу конструкції обумовлена незбігом його власних осей ортотропії із криволінійною системою координат циліндричної оболонки.

Для побудови тривимірної системи диференціальних рівнянь застосовано модифікацію варіаційного принципу Ху – Васілізу. Розв'язок отриманої системи рівнянь здійснюється з використанням аналітичного методу Бубнова – Галеркіна. Згідно з ним усі компоненти напружено-деформованого стану системи рівнянь розкладаються у подвійні ряди Фур'є так, щоб за координатою уздовж твірної оболонки вони задовольняли граничним умовам на торцях, а у коловому напрямку була врахована їхня періодичність. Результуюча система неоднорідних диференціальних рівнянь розв'язується чисельним методом дискретної ортогоналізації. Представлений підхід дає змогу отримати достовірні результати для конструкцій, що функціонують в умовах складного навантаження, і може бути використаний для підвищення їхньої надійності та безпеки.

Ключові слова: товстостінна анізотропна оболонка, просторова постановка, напружено-деформований стан, змінне розподілене навантаження.

Класифікація відповідно до AMS 2020: 74G10, 74G15.

Вступ

Товстостінні циліндричні оболонки є ключовими елементами багатьох інженерних конструкцій, таких як трубопроводи, резервуари, корпуси космічних апаратів та інших споруд, що експлуатуються в умовах силового навантаження або температурного впливу. Дослідження їхнього напружено-деформованого стану (НДС) набуває особливої актуальності у врахуванні змінного в коловому напрямку навантаження, що часто виникає в реальних умовах експлуатації. Такі навантаження, спричинені, наприклад, нерівномірним розподілом тиску, можуть суттєво впливати на міцність, стійкість і довговічність конструкції.

Аналіз НДС товстостінних циліндричних оболонок у просторовій постановці (Григоренко, Василенко, & Панкратова, 1991) дає змогу найповніше врахувати характер деформування та взаємодії напружень, що не завжди можливо забезпечити в межах класичних чи уточнених теорій. Використання просторового підходу є необхідним для забезпечення функціонування конструкцій, особливо в умовах, коли їхнє руйнування може призвести до значних економічних або екологічних наслідків.

Аналіз останніх досліджень і публікацій вказує на те, що наукові роботи із цієї тематики практично відсутні. У дослідженні (Григоренко, Василенко, & Панкратова, 1991) представлено підхід до розв'язку подібної задачі, але лише для окремих видів граничних умов на торцях циліндричних оболонок.

У представленій роботі автори пропонують підхід до розв'язання задачі про НДС анізотропної циліндричної оболонки (рис. 1) у просторовій постановці, що перебуває в полі дії розподіленого бокового тиску, змінного в коловому напрямку. Матеріалом, з якого виготовлена оболонка, є волокнистий композит – боропластик. Анізотропія пружних властивостей пов'язана з незбігом його власних осей ортотропії з криволінійною системою координат циліндричної оболонки.

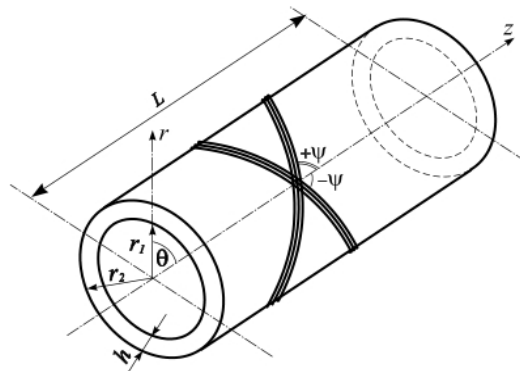


Рис. 1. Нетонка анізотропна циліндрична оболонка

1. Основні означення та результати

Відповідно до варіаційного принципу Ху – Васідзу (Washizu, 1982) з умови стаціонарності функціоналу Π_1 , що визначається з інтегралу:

$$\Pi_1 = \iiint_V \left\{ W(e_{ij}) - \sigma_{ij} \left[e_{ij} - \frac{1}{2} (u_{i,j} + u_{j,i}) \right] \right\} dV + \iint_{S_1} \Psi(u_i) dS_1 - \iint_{S_2} p_i (u_i - \bar{u}_i) dS_2, \quad (1)$$

можливо здобути необхідні для розв'язку поставленої задачі рівняння рівноваги, співвідношення пружності та граничні умови. У виразі (1), без додаткових умов, варіюються переміщення u_i , деформації e_{ij} , напруження σ_{ij} , напруження p_i на поверхні S_2 , що викликані переміщеннями $\bar{u}_i$. Також у цьому функціоналі $W(e_{ij})$ – потенціальна енергія деформації, $\Psi(u_i)$ – потенціал поверхневих навантажень, крапка з комою перед параметрами i, j позначає коваріантну похідну за координатою з відповідним індексом $i, j = 1, 2, 3$, де 1 відповідає координаті z , 2 – θ , 3 – r ; V – об'єм матеріалу оболонки.

Згідно з підходом, представленим у (Semenyuk, Trach, & Podvornyi, 2023), використовуючи умову стаціонарності виразу для варіації функціонала (1) та із застосуванням геометричних співвідношень (Novozhilov, 1961), отримано тривимірну систему із шести диференціальних рівнянь рівноваги для циліндричних анізотропних оболонок у частинних похідних виду (Semenyuk, Trach, & Podvornyi, 2023).

2. Доведення

З метою зменшення розмірності здобутої системи рівнянь скористаємося процедурою аналітичного методу Бубнова – Галеркіна (Semenyuk, Trach, & Podvornyi, 2023), відповідно до якого розкладемо функції напружень і переміщень у подвійні тригонометричні ряди Фур'є з урахуванням їхньої періодичності за коловим напрямом, а також уздовж твірної циліндричної оболонки так, щоб вони задовольняли умовам на торцях оболонки.

Після відповідних математичних перетворень отримаємо одновимірну систему диференціальних рівнянь дванадцятого порядку в нормальній формі Коші:

$$\frac{d\bar{y}}{dr} = T(r)\bar{y} + f, \quad T(r) = t_{n,l}(r); \quad (n, l = 1 \div 12), \quad (2)$$

де $\bar{y} = \{y_{1,p}; y_{2,p}; y_{3,p}; y_{4,p}; y_{5,p}; y_{6,p}; y_{1,m}'; y_{2,m}'; y_{3,m}'; y_{4,m}'; y_{5,m}'; y_{6,m}'\}$ – розв'язуюча вектор-функція, а f – навантаження, $T(r)$ – квадратна матриця із змінними коефіцієнтами, що залежить від аргументу r , $t_{n,l}(r)$ – ненульові елементи матриці $T(r)$, вони ж є коефіцієнтами за невідомих систем (2).

Систему рівнянь НДС (2) за умовами на торцях оболонки

$$\sigma_{zz} = u_r = u_\theta = 0 \quad (3)$$

розв'язуємо з використанням чисельного методу дискретної ортогоналізації (Григоренко, Василенко, & Панкратова, 1991).

3. Приклади

Для представлення можливостей запропонованого підходу розглянемо циліндричну товстостінну анізотропну оболонку, виготовлену з волокнистого композиту – боропластику (рис. 1). Оболонкова конструкція перебуває під дією розподіленого зовнішнього тиску $q = -q_0 \sin(\pi z / L) \cos k\theta$, за $q_0 = 1,0$ МПа та параметра хвилеутворення $k = 4$ за коловим напрямком θ . Осі ортотропії матеріалу боропластику можуть бути повернуті на кут $0^\circ \leq \psi \leq 90^\circ$ відносно твірної оболонки (рис. 1). Геометричні параметри циліндра: радіус внутрішньої поверхні $r_1 = 0,54$ м; зовнішньої – $r_2 = 0,66$ м; довжина уздовж твірної $L = 1,2$ м. Механічні характеристики однонапрявленого волокнистого композиту боропластику: $E_{zz} = 280 E_0$; $E_{\theta\theta} = E_{rr} = 31 E_0$; $G_{z\theta} = G_{r\theta} = 10,5 E_0$; $G_{rz} = 21,2 E_0$; $\nu_{\theta z} = 0,25$; $\nu_{z\theta} = 0,0277$; $E_0 = 1000$ МПа.

На рис. 2 представлено залежність величин нормальних напружень $\sigma_{\theta\theta}$ від кута θ на рівні зовнішньої (завантаженої) поверхні оболонки. Крива 1 відповідає куту $\psi = 0^\circ$, крива 2 для $\psi = 30^\circ$, 3 – $\psi = 60^\circ$, 4 – $\psi = 90^\circ$.

З аналізу результатів, наведених на рис. 2, бачимо, що зміна величин нормальних напружень $\sigma_{\theta\theta}$ за коловим напрямком на зовнішній навантаженій поверхні оболонки відбувається циклічно відповідно до параметру хвилеутворення k

та залежить від величини кута повороту ψ головних напрямів пружності волокнистого матеріалу конструкції. Показано, що за кута $\psi = 0^\circ$ величини $\sigma_{\theta\theta}$ набувають мінімальних значень, зростаючи за збільшення кута ψ . Розбіжність між мінімальними величинами $\sigma_{\theta\theta}$ за $\psi = 0^\circ$ та максимальними за $\psi = 90^\circ$ досягає 43 %.

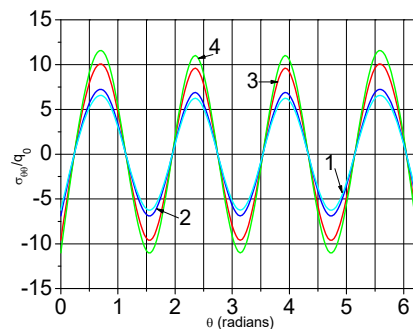


Рис. 2. Графік залежності нормальних напружень $\sigma_{\theta\theta}$ від кута повороту ψ

Дискусія і висновки

У пропонованій роботі шляхом модифікації варіаційного принципу Ху – Васіду отримано тривимірну систему диференціальних рівнянь рівноваги в частинних похідних, що описують напружено-деформований стан анізотропних товстостінних циліндричних оболонок. Для приведення такої системи рівнянь до одновимірної використано процедуру аналітичного методу Бубнова – Галеркіна. Розв'язок одновимірної системи звичайних диференціальних рівнянь напружено-деформованого стану в нормальній формі Коші здійснюється за радіальним напрямком циліндра за використання чисельного методу дискретної ортогоналізації.

Розглянуто зміну одного з параметрів напружено-деформованого стану за коловим напрямком анізотропної товстостінної циліндричної оболонки для різних кутів повороту головних напрямів пружності попередньо ортотропного волокнистого матеріалу конструкції.

Внесок авторів: Олександр Григоренко – концептуалізація, методика, збір і перевірка емпіричних даних, емпіричне дослідження; Володимир Трач – концептуалізація, методика, збір і перевірка емпіричних даних, емпіричне дослідження; Андрій Подворний – програмне забезпечення, емпіричне дослідження; Наталія Жукова – програмне забезпечення, аналіз джерел, підготовка огляду літератури.

Джерела фінансування: Дослідження було проведено власним коштом авторів.

Список використаних джерел

- Григоренко, Я. М., Василенко, А. Т., & Панкратова, Н. Д. (1991). *Задачи теории упругости неоднородных тел*. Наукова думка.
 Novozhilov, V. V. (1961). *Theory of elasticity*. Elsevier.
 Semenyuk, N. P., Trach, V. M., & Podvornyi, A. V. (2023). Stress-strain state of a thick-walled anisotropic cylindrical shell. *International Applied Mechanics*, 59(1), 79–89. <https://doi.org/10.1007/s10778-023-01201-5>
 Washizu, K. (1982). *Variational Methods in Elasticity and Plasticity*. Pergamon Press. <https://doi.org/10.1002/zamm.19840640121>

References

- Grigorenko, Y. M., Vasilenko, A. T., & Pankratova, N. D. (1991). *Problems of the theory of elasticity of inhomogeneous bodies*. Naukova dumka [in Russian].
 Novozhilov, V. V. (1961). *Theory of elasticity*. Elsevier.
 Semenyuk, N. P., Trach, V. M., & Podvornyi, A. V. (2023). Stress-strain state of a thick-walled anisotropic cylindrical shell. *International Applied Mechanics*, 59(1), 79–89. <https://doi.org/10.1007/s10778-023-01201-5>
 Washizu, K. (1982). *Variational Methods in Elasticity and Plasticity*. Pergamon Press. <https://doi.org/10.1002/zamm.19840640121>

Отримано редакцією журналу / Received: 24.12.25

Прорецензовано / Revised: 17.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Alexander GRIGORENKO¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-4109-2672
e-mail: ayagrigorenko1991@gmail.com

Volodymyr TRACH², DSc (Engin.), Prof.
ORCID ID: 0000-0001-9500-2743
e-mail: v.m.trach@nuwm.edu.ua

Andrii PODVORNYI², DSc (Engin.), Assoc. Prof.
ORCID ID: 0000-0001-8518-4395
e-mail: a.v.podvorny@nuwm.edu.ua

Natalia ZHUKOVA¹, PhD (Phys. & Math.), Senior Research
ORCID ID: 0000-0002-2699-3771
e-mail: zukowa_n@ukr.net

¹S. P. Timoshenko Institute of Mechanics, Kyiv, Ukraine

²National University of Water and Environmental Engineering, Rivne, Ukraine

STRESS-STRAIN STATE OF ANISOTROPIC THICK-WALLED CYLINDRICAL SHELLS UNDER VARIABLE LOADING

Thick-walled cylindrical shell structures are essential components of various modern engineering systems, including pipelines, storage tanks, reactor installations, and aerospace bodies. The operation of such structures under complex mechanical or thermal loading requires reliable analysis of their stress-strain state, particularly in the presence of distributed loads varying along the circumferential direction. Such effects may arise from non-uniform working medium pressure, for example, asymmetric turbulent flow in pipelines or asymmetric aerodynamic loads acting on aerospace hulls. These factors can significantly influence the strength, stability, and service life of the structures.

This study presents a three-dimensional approach to the analysis of the stress-strain state of an anisotropic thick-walled cylindrical boroplastic shell subjected to distributed pressure periodically varying in the circumferential direction. The anisotropy of the elastic properties of the material is determined by the misalignment of its orthotropy axes with the curvilinear coordinate system of the cylindrical shell.

To derive the three-dimensional system of differential equations, a modification of the Hu – Washizu variational principle is applied. The resulting system is solved using the analytical Bubnov – Galerkin method. Within this framework, all stress-strain components are expanded into double Fourier series, ensuring that the longitudinal coordinate satisfies boundary conditions at the shell ends, while circumferential periodicity is preserved. The final system of nonhomogeneous differential equations is solved numerically using the discrete orthogonalization method.

The proposed approach provides accurate results for structures operating under complex loading conditions and can be effectively applied to improve their reliability and safety.

Keywords: *thick-walled anisotropic shell, three-dimensional formulation, stress-strain state, variable distributed load.*

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.652+517.518.85

DOI: <https://doi.org/10.17721/1812-5409.2026/1.15>

Volodymyr DENYSIUK, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-3210-7350
e-mail: kvomden@ukr.net
State University "Kyiv Aviation Institute", Kyiv, Ukraine

Tetiana OLESHKO, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-8054-1178
e-mail: tetiana.oleshko@npp.kai.edu.ua
State University "Kyiv Aviation Institute", Kyiv, Ukraine

Liudmyla RYBACHUK, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-7652-587X
e-mail: rybachuk.liudmyla@iit.kpi.ua

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

INCOMPLETE TRIGONOMETRIC SPLINES IN THE PROBLEMS OF CONSTRUCTING APPROXIMATE SOLUTIONS OF THE SECOND ORDER LINEAR DIFFERENTIAL EQUATIONS

This study introduces a numerical analytical method for deriving approximate solutions to the first boundary value problem associated with second-order linear differential equations. The approach is founded upon the utilization of incomplete fundamental trigonometric splines, both even and odd, whose structural properties closely parallel those of classical polynomial splines and have been extensively investigated in prior works. The proposed methodology employs residual minimization through the established collocation technique, thereby reducing the original problem to the resolution of systems of linear algebraic equations with respect to the spline-defining parameters. A distinctive feature of this method lies in the fact that the values of the sought solution themselves serve as the determining parameters of the fundamental trigonometric splines. Boundary conditions are incorporated by prescribing specific values for the corresponding spline parameters. The theoretical exposition is supplemented with illustrative examples.

It should be noted that incomplete trigonometric splines are functions with a certain evenness. In our opinion, in many cases this property is convenient when setting boundary conditions for the desired solution. The given method assumes generalization. Thus, in particular, the minimization of the residual can be carried out by the method of discrete least squares and its variations. This method can also be applied to linear differential equations of arbitrary orders.

Keywords: *fundamental trigonometric splines, incomplete trigonometric splines, even fundamental splines, odd fundamental splines, boundary value problems, linear differential equations.*

AMS 2020 classification: 65D05, 41A05.

Introduction

The primary objective of this study is to develop a methodology for constructing approximate solutions to the first boundary value problem for linear differential equations, utilizing incomplete trigonometric splines of both even and odd types. Linear differential equations frequently arise in scientific and engineering applications. Although a variety of theoretical approaches to their solution have been established, obtaining exact solutions is often challenging and, in many cases, unattainable. In such cases, the problem of finding approximate solutions to such equations is considered; such solutions are found using numerical methods.

Numerical methods for finding approximate solutions to differential equations can be divided into two classes: numerical and numerical-analytical methods.

Numerical methods allow us to find discrete values of the desired approximate solution on the interval of the differential equation. There are many such methods known, among which we mention the Runge – Kutta, Adams, Euler, Krylov, etc. methods and their modifications. Such methods are quite simple to implement on modern computers and, if applicable, lead to satisfactory results.

When applying numerical-analytical methods, an approximate solution is sought in the form of some sufficiently smooth function that depends on several parameters; these parameters are chosen based on certain conditions. Since we will be dealing with methods of this class in the future, let's consider the use of such methods in more detail.

Suppose we have a linear differential operator of the 2nd order of the form

$$L_2 u = u''(x) + p_1(x)u'(x) + p_2(x)u(x), \quad (1)$$

where $x \in [a, b]$, $p_1(x), p_2(x)$ are some functions that are generally continuous on $[a, b]$.

Consider the equation

$$L_2 u = f(x). \quad (2)$$

It is known from the theory of differential equations that the general solution $u(x)$ of this equation contains 2 arbitrary constants. To determine these arbitrary constants, additional 2 conditions are added to (2). We will consider the case when these conditions are imposed on the values of the desired solution at points a and b of the segment $[a, b]$, and have the form

$$u(a) = a_0; u(b) = b_0. \quad (3)$$

Such conditions are called boundary conditions and are said to be the first boundary value problem.

Let $C_{[a,b]}^k$ be the class of functions that have a continuous derivative of order k on $[a, b]$. Note that the classes $C_{[a,b]}^0$ contain continuous functions, and the classes $C_{[a,b]}^{-1}$ contain piecewise continuous functions with a finite number of discontinuities of the first kind.

Let's look for a solution to this equation along with the boundary conditions in the form of a function $\varphi(x, \alpha_1, \alpha_2, \dots, \alpha_N)$ that satisfies the conditions:

$$\text{a) } \varphi(x, \alpha_1, \alpha_2, \dots, \alpha_N) \in C_{[a,b]}^r, \text{ where } r \geq 2;$$

$$\text{b) } \varphi(a, \alpha_1, \alpha_2, \dots, \alpha_N) = a_0; \varphi(b, \alpha_1, \alpha_2, \dots, \alpha_N) = b_0;$$

and $\alpha_1, \alpha_2, \dots, \alpha_N$ are some unspecified parameters ($N = 1, 2, \dots$).

Substituting the function $\varphi(x, \alpha_0, \alpha_1, \dots, \alpha_N)$ into (2), we get a residual

$$L_n \varphi(x, \alpha_0, \alpha_1, \dots, \alpha_N) - f(x) = \varepsilon(x, \alpha_0, \alpha_1, \dots, \alpha_N). \quad (4)$$

By minimizing the residual in one sense or another, it is easy to obtain the conditions for determining the parameters of $\alpha_0, \alpha_1, \dots, \alpha_N$.

There are various methods for minimizing the residual: the collocation method, the least squares method (continuous and discrete variants), the method of moments and its variants, the subdomain method (Fletcher, 2012), etc. In this paper, without losing generality, we will limit ourselves to the collocation method. As is known, this method consists of the following.

A certain system of points $a \leq x_1 < x_2 < \dots < x_N \leq b$ is set on $[a, b]$ and it is required that the following conditions be met

$$\varepsilon(x_i, \alpha_1, \alpha_2, \dots, \alpha_N) = 0 \quad (i = 1, 2, \dots, N). \quad (5)$$

From this system of N equations, it is possible to determine N parameters $\alpha_0, \alpha_1, \dots, \alpha_N$.

When applying numerical-analytical methods for constructing approximate solutions, the method of constructing the function $\varphi(x, \alpha_1, \alpha_2, \dots, \alpha_N)$ plays an important role. In many cases, a generalized polynomial is chosen as this function

$$P_N(x) = \alpha_1 \omega_1(x) + \alpha_2 \omega_2(x) + \dots + \alpha_N \omega_N(x), \quad (6)$$

where $\omega_1(x), \omega_2(x), \dots, \omega_N(x)$ is a system of linearly independent functions.

In the last decades, splines have been widely chosen as such systems of functions. The fundamental basis of spline theory and its application to variational problems was established in the classical works of (Ahlberg, Nilson, & Walsh, 1967; Prenter, 1975; Schumaker, 2010). Significant contributions to the development of spline collocation methods for differential equations were also made by (Fairweather, & Meade, 1989). However, polynomial splines have limitations when approximating functions with oscillating nature. For such problems, the use of trigonometric splines, first introduced by (Schoenberg, 1964), is more natural. The fundamental theory of functions on grids, which is essential for such approximations, is detailed in (Guter, Kudryavtsev, & Levitan, 1966).

Recently, there has been a resurgence of interest in using trigonometric B-splines for solving boundary value problems. (Abbas et al., 2014) and (Nazir et al., 2016) demonstrated the unconditional stability and effectiveness of cubic trigonometric B-splines collocation for diffusion and advection problems. (Heilat, 2023) extended this to systems of second-order boundary value problems, while (Yaseen, Abbas, & Riaz, 2021) applied it to fractional diffusion equations. To improve accuracy, researchers have introduced hybrid and extended B-splines with free parameters, as seen in the work of (Khan, & Misro, 2025), where optimization algorithms are used to tune shape parameters. Advanced applications to nonlinear and singular problems have been explored by (Kukreja, & Shallu, 2025; Roul, 2024), while (Zou, 2025) proposed optimization-based trigonometric interpolation for mixed boundary conditions.

Despite these advances, most B-spline methods use control points as coefficients, which complicates the satisfaction of boundary conditions (3). In this study, we propose a method based on incomplete fundamental trigonometric splines. Classes of such splines were introduced by (Denysiuk, 2018), and their representation forms were analyzed by (Denysiuk, & Rybachuk, 2024). A distinctive feature of our method is that the coefficients α_k in (6) are the values of the solution itself. This allows boundary conditions (3) to be satisfied trivially by setting $\alpha_1 = a_0$ and $\alpha_N = b_0$, significantly simplifying the computational algorithm.

1. Main definitions and results

In (Denysiuk, & Rybachuk, 2024), three forms of representation of trigonometric splines were given: through discrete (finite) Fourier series, through trigonometric B-splines, and through fundamental trigonometric splines.

When developing methods for constructing approximate solutions to boundary value problems for linear differential equations based on trigonometric splines, the form of representation of trigonometric splines and the interpretation of the parameters of trigonometric splines in each case play an important role. Thus, in the case of representation of trigonometric splines through discrete (finite) Fourier series, discrete Fourier coefficients act as defining parameters; both periodic and non-periodic splines can be constructed.

In the case of representing splines through trigonometric B-splines, the defining parameters are the coefficients at these B-splines, which are difficult to interpret in any reasonable way; at the same time, only periodic splines are currently plotted.

Finally, in the case of representing trigonometric splines through fundamental splines, the determining parameters are the values of the desired solution at the nodes of the given collocation grids; both periodic and non-periodic splines can be constructed.

Although all three forms can be used for methods of constructing approximate solutions, in our opinion, the most convenient form is the one using fundamental trigonometric splines. This is because using the values of the desired solution as parameters greatly simplifies the enforcement of the boundary conditions for this solution. Therefore, in the following, we will limit ourselves to considering the case when the desired solution is constructed using fundamental trigonometric splines.

Let $[a, b]$ be a certain segment of the real axis, and let there be a certain grid $\Delta_N = \{x_j\}_{j=1}^N$ on this segment, and $a \leq x_1 < x_2 < \dots < x_N \leq b$. The system of functions $\varphi_k(x)$ ($k = 1, 2, \dots, N$) on the grid Δ_N is called fundamental if the following conditions are met (Guter, Kudryavtsev, & Levitan, 1966).

$$\varphi_k(x_j) = \begin{cases} 1, k = j; \\ 0, k \neq j, k, j = 1, 2, \dots, N. \end{cases} \quad (7)$$

Given a system of fundamental functions, the function $\Phi_N(x)$ interpolating a given function $f(x)$ at the nodes of the grid Δ_N can be represented as

$$\Phi_N(x) = \sum_{k=1}^N f(x_k) \varphi_k(x).$$

Similarly, systems of fundamental polynomial splines are introduced.

Systems of trigonometric fundamental splines are of three types: complete fundamental splines, as well as even and odd fundamental splines (Denysiuk, & Rybachuk, 2024).

Complete fundamental splines are periodic functions and can be used to find periodic solutions of linear differential equations if such solutions exist; recall that the existence of such solutions is discussed. In this paper, we will not consider the problem of finding periodic solutions of differential equations.

Even and odd fundamental splines can be used to find nonperiodic solutions to differential equations. Let us consider this issue in more detail.

Let us define uniform grids $\Delta_2^{(I)} = \{x_j^{(I)}\}_{j=1}^N$ on $[0, \pi]$, where $x_j^{(0)} = \frac{\pi}{N-1}(j-1)$, $x_j^{(1)} = \frac{\pi}{2N}(2j-1)$ and $\Delta_3^{(I)} = \{x_j^{(I)}\}_{j=1}^N$, where $x_j^{(0)} = \frac{\pi}{N+1}j$, $x_j^{(1)} = \frac{\pi}{2N}(2j-1)$, I is the grid indicator ($I = 0, 1$), and $N = 1, 2, 3, \dots$. We will not introduce separate notations for the nodes of these grids, since in each case we will determine which grid we are talking about.

Non-periodic even and odd trigonometric splines are constructed on the grids $\Delta_2^{(I)}$ and $\Delta_3^{(I)}$, respectively (Denysiuk, & Rybachuk, 2024). Even and odd trigonometric splines have different properties; in particular, even trigonometric splines take non-zero values at the points 0 and π , and their derivatives take zero values at these points. Odd splines, on the other hand, take on zero values at the points 0 and π , and their derivatives at these points take on non-zero values. Taking this into account, and wanting to deal with uniform grids, we will limit ourselves to considering even splines with the same stitching and interpolation grids $\Delta_2^{(0)}$ and odd splines with the same stitching and interpolation grids $\Delta_3^{(1)}$.

Here are the expressions for constructing fundamental even splines on the grid $\Delta_2^{(0)}$; here and further, in the notation of even and odd fundamental splines, we will omit the indices of the stitching and interpolation grids, as well as the index of the type of convergence factors, which is equal to $\sigma 1(r, j) = j^{-(1+r)}$ (r, j are natural numbers).

The paired fundamental splines $stc(r, q, k, t)$ of order r ($r = 1, 2, \dots$), and their derivatives of order q ($q = 1, 2, \dots, r-1$) on the grid $\Delta_2^{(0)}$ are calculated by the expressions

$$stc(r, q, k, t) = \frac{2}{N-1} \left[0.5 \cdot I(q) + \sum_{j=1}^{N-2} \frac{c(\sigma 1, r, q, j, t) \cos(x_k^{(0)})}{H(\sigma 1, r, j)} + 0.5 \frac{c(\sigma 1, r, q, N-1, t) \cos(x_k^{(0)})}{H(\sigma 1, r, N-1)} \right], \quad (8)$$

where

$$I(q) = \begin{cases} 1, \text{if } q = 0, \\ 0, \text{otherwise} \end{cases}$$

$$H(\sigma 1, r, k) = \sigma 1(r, k) + \sum_{m=1}^{\infty} [\sigma 1(r, 2m(N-1) + k) + \sigma 1(r, 2m(N-1) - k)],$$

$$c(\sigma 1, r, q, j, t) = \sigma 1(r, j) j^q \cos(jt + q \frac{\pi}{2}) +$$

$$+ \sum_{m=1}^{\infty} \left[\sigma 1(r, 2m(N-1) + j) \left(2m(N-1) + j \right)^q \cos((2m(N-1) + j)t + q \frac{\pi}{2}) + \right. \\ \left. + \sigma 1(r, 2m(N-1) - j) \left(2m(N-1) - j \right)^q \cos((2m(N-1) - j)t + q \frac{\pi}{2}) \right].$$

Graphs of paired trigonometric splines are shown on the Fig. 1 ($r = 1$ and $N = 9$ for clarity). Note that here and in the following graphs, the vertical lines coincide with the nodes of the grid $\Delta_2^{(0)}$.

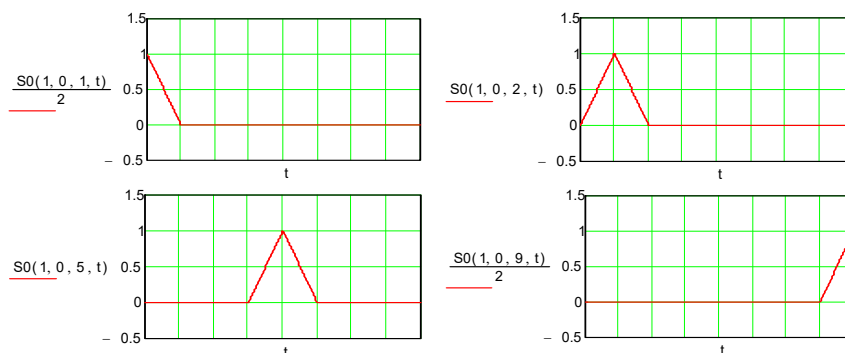


Fig. 1. Paired fundamental splines on the grid $\Delta_2^{(0)}$; $r = 1$, $N = 9$

Similarly, the odd fundamental splines $sts^{(0)}(r, q, k, t)$ of order r ($r = 1, 2, \dots$), and their derivatives of order q ($q = 1, 2, \dots, r-1$) on the grid $\Delta_3^{(0)}$, are calculated by the expression

$$sts^{(0)}(r, q, k, t) = \frac{2}{N+1} \sum_{j=1}^N \frac{s(\sigma 1, r, q, j, t) \sin(jx_k^{(0)})}{H(\sigma 1, r, j)}, \quad (9)$$

where

$$\begin{aligned} H(\sigma 1, r, k) &= \sigma 1(r, k) + \sum_{m=1}^{\infty} [\sigma 1(r, 2m(N+1) + k) + \sigma 1(r, 2m(N+1) - k)], \\ s(\sigma 1, r, q, j, t) &= \sigma 1(r, j) j^q \sin(jt + q \frac{\pi}{2}) + \\ &+ \sum_{m=1}^{\infty} (-1)^m [\sigma 1((r, 2m(N+1) + j)) ((2m(N+1) + j)^q \sin((2m(N+1) + j)t + q \frac{\pi}{2}) - \\ &- \sigma 1((r, 2m(N+1) - j)) ((2m(N+1) - j)^q \sin((2m(N+1) - j)t + q \frac{\pi}{2}))]. \end{aligned}$$

As before, we present the graphs of odd trigonometric splines (Fig. 2), assuming $r = 1$, $N = 9$. Note that the vertical lines in these graphs coincide with the nodes of the grid $\Delta 3_N^{(0)}$.

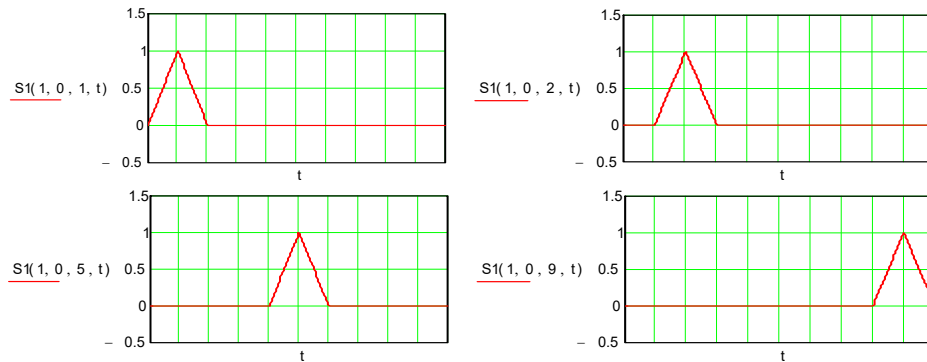


Fig. 2. Odd fundamental splines $sts^{(0)}(1, 0, k, t)$ on the grid $\Delta 3_N^{(0)}$; $r = 1$, $N = 9$

On the grid $\Delta 3_N^{(1)}$, the odd fundamental splines $sts^{(1)}(r, q, k, t)$ of order r ($r = 1, 2, \dots$), and their derivatives of order q ($q = 1, 2, \dots, r-1$), are calculated by the expressions

$$sts^{(1)}(r, q, k, t) = \frac{2}{N} \left[\sum_{j=1}^{N-1} \frac{s(\sigma 1, r, q, j, t) \sin(jx_k^{(1)})}{H(\sigma 1, r, j)} + 0.5 \frac{s(\sigma 1, r, q, N, t) \sin(Nx_k^{(1)})}{H(\sigma 1, r, N)} \right], \quad (10)$$

where

$$\begin{aligned} H(\sigma 1, r, k) &= \sigma 1(r, k) + \sum_{m=1}^{\infty} [\sigma 1(r, 2mN + k) + \sigma 1(r, 2mN - k)], \\ s(\sigma 1, r, q, j, t) &= \sigma 1(r, j) j^q \sin(jt + q \frac{\pi}{2}) + \\ &+ \sum_{m=1}^{\infty} (-1)^m [\sigma 1(r, 2mN + j) (2mN + j)^q \sin((2mN + j)t + q \frac{\pi}{2}) - \\ &- \sigma 1(r, 2mN - j) (2mN - j)^q \sin((2mN - j)t + q \frac{\pi}{2})]. \end{aligned}$$

Graphs of odd trigonometric splines are shown on the Fig. 3, assuming $r = 1$, $N = 9$. Note that the vertical lines in these graphs coincide with the nodes of the grid $\Delta 3_N^{(0)}$.

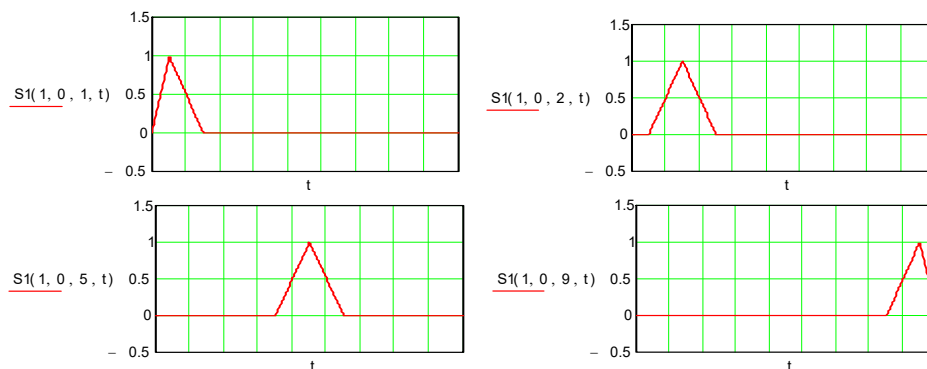


Fig. 3. Odd fundamental splines $sts^{(1)}(1, 0, k, t)$ on the grid $\Delta 3_N^{(0)}$; $r = 1$, $N = 9$

Given the even and odd fundamental splines $stc(r, q, k, t)$, $sts^{(0)}(1, 0, k, t)$ and $sts^{(1)}(1, 0, k, t)$, the splines $Stc(r, q, t)$ and $Sts(r, q, t)$ interpolating a continuous function $f(x)$ at the nodes of the grids $\Delta 2_N^{(0)}$, $\Delta 3_N^{(0)}$ and $\Delta 3_N^{(1)}$ can be represented as

$$Stc(r, q, t) = 0.5f(x_1)stc(r, q, 1, t) + \sum_{k=2}^{N-1} f(x_k)stc(r, q, k, t) + 0.5f(x_N)stc(r, q, N, t), \quad (11)$$

$$Sts^{(0)}(r, q, t) = \sum_{k=1}^N f(x_k)sts^{(0)}(r, q, k, t), \quad (12)$$

$$Sts^{(1)}(r, q, t) = \sum_{k=1}^N f(x_k)sts^{(1)}(r, q, k, t). \quad (13)$$

It is clear that interpolation splines $Stc(r, q, k, t)$, $Sts^{(0)}(r, q, k, t)$, $Sts^{(1)}(r, q, k, t)$ and fundamental splines $stc(r, q, k, t)$, $sts^{(0)}(r, q, k, t)$, $sts^{(1)}(r, q, k, t)$ belong to the class $C_{[0, \pi]}^{r-1}$.

Let us now consider the first boundary value problem for (2) with boundary conditions (3), which, as we have already mentioned, is formulated as follows. Find the function $u(x), u(x) \in C_{[0, \pi]}^r$ ($r \geq 2$), $x \in [0, \pi]$, which satisfies the equation

$$u''(x) + p_1(x)u'(x) + p_2(x)u(x) = f(x) \quad (14)$$

and boundary conditions

$$u(0) = a_0; u(\pi) = b_0. \quad (15)$$

There are three ways to find an approximate solution $u^*(x)$ to this problem.

a) On the grid $\Delta 2_N^{(0)}$ via paired splines

$$u^*(x) = 0.5\alpha_1 stc(r, q, 1, x) + \sum_{k=2}^{N-1} \alpha_k stc(r, q, k, x) + 0.5\alpha_N stc(r, q, N, x), \quad (16)$$

where $r \geq 3$ and the parameters α_k ($k = 2, 2, \dots, N-1$) are determined from the collocation conditions.

It is necessary to take into account the fact that when finding an approximate solution in the form of (16), additional boundary conditions of the form $u'(0) = u'(\pi) = 0$ are also fulfilled. This is due to the fact that when differentiating (16), the functions $\cos(x)$ are transformed into functions $-\sin(x)$, which turn to 0 at the ends of the segment $[0, \pi]$.

b) On the grid $\Delta 3_N^{(0)}$ through odd splines in the form

$$u^{*(0)}(x) = \sum_{k=1}^N \alpha_k sts^{(0)}(r, q, k, x), \quad (17)$$

and on the grid $\Delta 3_N^{(1)}$ in the form of

$$u^{*(1)}(x) = \sum_{k=1}^N \alpha_k sts^{(1)}(r, q, k, x), \quad (18)$$

where $r \geq 3$, and the parameters α_k ($k = 1, 2, \dots, N$) are determined from the collocation conditions. Note that such splines are conveniently used for problems with zero boundary conditions.

In cases where the solution to the equation is sought on the segment $[a, b]$, this segment is mapped to the segment $[0, \pi]$ by linearly replacing the variables.

Consider the case when the solution to problems (14), (15) is sought in the form (17). It is clear that in this case, the nodes $x_1 = 0$ and $x_N = \pi$ of the $\Delta 2_N^{(0)}$ mesh coincide with the ends of the segment $[0, \pi]$, and, therefore, it follows from (15) that $\alpha_1 = a_0$ and $\alpha_N = b_0$. To determine the parameters of $\alpha_2, \alpha_2, \dots, \alpha_{N-1}$, we substitute (17) into (14) and equate the values of the left and right parts at the nodes $x_2, x_2, \dots, x_{N-1}$ of the grid $\Delta 2_N^{(0)}$; as a result, we obtain the following system of linear algebraic equations with respect to the desired parameters $\alpha_2, \alpha_2, \dots, \alpha_{N-1}$:

$$0.5a_0 A(r, 1, x_j) + \sum_{k=2}^{N-1} \alpha_k A(r, k, x_j) + 0.5b_0 A(r, N, x_j) = f(x_j) \quad (j = 2, 3, \dots, N-1), \quad (19)$$

where

$$A(r, k, x_j) = stc(r, 2, k, x_j) + p_1(x_j)stc(r, 1, k, x_j) + p_2(x_j)stc(r, 0, k, x_j) \quad (k = 1, 2, \dots, N; j = 2, 3, \dots, N-1).$$

2. Examples

To illustrate these provisions, we will provide examples.

Example 1. Let us consider a boundary value problem

$$u''(x) + \frac{C}{1+x} u'(x) - \frac{x}{1+x} u(x) = \frac{C-2-x^2(1+x)}{(1+x)^3}, \quad x \in [0, 1];$$

$$u(0) = 0; u(1) = 0.5, \quad (C = 0, 1, 10).$$

The exact solution to this problem is

$$u_0(x) = \frac{x}{1+x}.$$

By composing a system of equations of the type (19) at $r = 3$, and solving it, we obtain approximate solutions $u_1(x)$ of this problem for the values $C = 0, 1, 10$. Graphs of exact solution and approximate solutions for different values of the parameter C are shown on the Fig. 4.

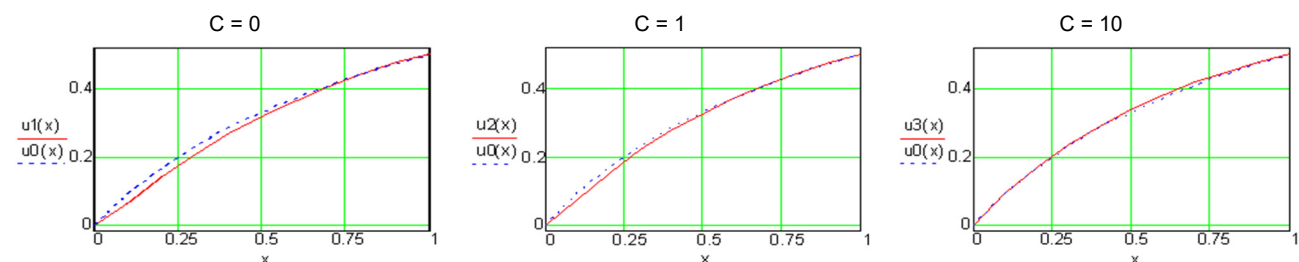


Fig. 4. The exact solution $u_0(x)$ and approximate solutions $u_1(x)$, $u_2(x)$ and $u_3(x)$, corresponding to the values of the parameter $C = 0, 1, 10$

Note that the largest deviations of the exact solution from the approximate solution do not exceed 0.048 when $C = 0$, 0.044 when $C = 1$, and 0.042 when $C = 10$.

Example 2. Let us consider a boundary value problem

$$u''(x) + u(x) = \cos(x) \cos(2x), \quad x \in [0, \pi];$$

$$u(0) = 1; u(\pi) = -1.$$

The exact solution to this problem is

$$u_0(x) = 1.0625 \cos(x) - 0.4 \sin(x) - 0.0625 \cos(3x) + 0.25x \sin(x).$$

By composing systems of equations of the type (19) at $r = 3, 4, 5$, and solving them, we obtain approximate solutions $u_3(x)$, $u_4(x)$, $u_5(x)$ of this problem corresponding to the given values of the parameter r . Graphs of exact solution and approximate solutions for different values of the parameter r are shown on the Fig. 5.

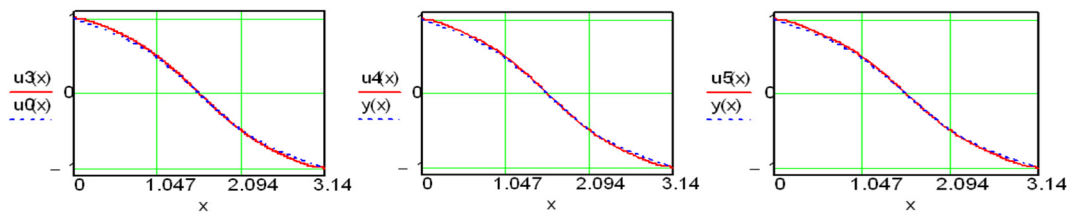


Fig. 5. The exact solution $u_0(x)$ and approximate solutions $u_3(x)$, $u_4(x)$ and $u_5(x)$, corresponding to the values of the parameter $r = 3, 4, 5$

The largest deviations of the exact solution from the approximate solution do not exceed 0.048 at $r = 3$, 0.045 at $r = 4$, and 0.043 at $r = 5$.

Let us now consider the case when the solution to problems (14), (15) is sought in the form (18), with (15) having the form

$$u(0) = 0; u(\pi) = 0. \quad (20)$$

It is clear that the nodes x_1 and x_N of the $\Delta 3_N^{(0)}$ mesh do not coincide with the ends of the segment $[0, \pi]$. Since (18) is equal to zero at points 0 and π , the boundary conditions are automatically satisfied. By drawing up a system of equations of the form

$$\sum_{k=1}^N \alpha_k A(r, k, x_j) = f(x_j), \quad (j = 2, 3, \dots, N-1), \quad (21)$$

where

$$A(r, k, x_j) = sts^{(0)}(r, 2, k, x_j) + p_1(x_j)stc^{(0)}(r, 1, k, x_j) + p_2(x_j)stc^{(0)}(r, 0, k, x_j), \\ (k = 1, 2, \dots, N; j = 1, 2, 3, \dots, N)$$

and finding the value of α_k ($k = 1, 2, \dots, N$), we construct an approximate solution $u^{*(0)}(x)$.

Reasoning similarly, and using the splines $sts^{(1)}(r, q, k, x)$ to construct the values $A(r, k, x_j)$, it is easy to construct the solution $u^{*(1)}(x)$.

As before, to illustrate the above provisions, we will provide an example.

Example 3. Let us consider a boundary value problem

$$u''(x) + u(x) = -x, \quad x \in [0, \pi] \quad (x \in [0, 1]), \\ u(0) = 0; u(1) = 0.$$

The exact solution to this problem is

$$u(x) = \frac{\sin(x)}{\sin(1)} - t.$$

By composing systems of equations of the type (21) at $r = 3, 4, 5$, and solving them, we obtain approximate solutions $u_3^{(0)}(x)$, $u_4^{(0)}(x)$, $u_5^{(0)}(x)$ of this problem corresponding to the given values of the parameter r . Graph of exact $u(x)$ and approximate $u_4^{(0)}(x)$ solutions to this problem is shown on the Fig. 6; we also show the errors of other solutions for different values of the parameter r .

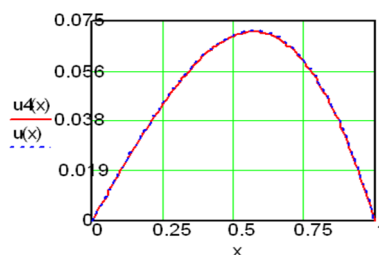


Fig. 6. The exact solution $u(x)$ and the approximate solution $u_4^{(0)}(x)$, corresponding to the values of the parameter $r = 4$

The largest deviations of the exact solution from the approximate solution do not exceed 0.00157 at $r = 3$, 0.00105 at $r = 4$, and 0.00098 at $r = 5$.

Similar calculations were performed for the approximate solutions $u_3^{(1)}(x)$, $u_4^{(1)}(x)$, $u_5^{(1)}(x)$; the largest deviations of the exact solution from the approximate solution do not exceed 0.0005 at $r = 3, 4$, but increase to 0.00063 at $r = 5$.

Discussion and conclusions

The study introduces methodologies for constructing approximate solutions to the first boundary value problem associated with linear differential equations through the application of incomplete trigonometric splines of both even and odd forms.

The utilization of even trigonometric splines is particularly appropriate in cases where the first derivative of the solution is subject to homogeneous boundary conditions, whereas odd trigonometric splines are recommended when the solution itself satisfies zero boundary conditions. To substantiate the theoretical framework, illustrative examples are provided.

Furthermore, it is demonstrated that the computational error exhibits dependence on the parameter r , which encapsulates the differential properties inherent to the employed trigonometric splines.

The proposed approaches underscore the necessity for continued investigation to refine and extend the methods for approximating solutions to the first boundary value problem in the context of linear differential equations.

Authors' contribution: Volodymyr Denysiuk – conceptualization, methodology and theoretical foundations of research, empirical data collection and validation; Tetiana Oleshko – analysis of sources, preparation of literature review, empirical research; Liudmyla Rybachuk – analysis of sources, preparation of literature review, empirical research, writing (review and editing).

Sources of funding. All scientific research described in this work was carried out at the authors' own expense.

References

- Abbas, M., Majid, A. A., Ismail, A. I. M., & Rashid, A. (2014). The application of cubic trigonometric B-spline to the numerical solution of the hyperbolic problems. *Applied Mathematics and Computation*, 239, 74–88. <https://doi.org/10.1016/j.amc.2014.04.031>
- Ahlberg, J. H., Nilson, E. N., & Walsh, J. L. (1967). *The theory of splines and their applications*. Academic Press.
- Denysiuk, V. (2018). Generalized trigonometric functions and their applications. *IOSR Journal of Mathematics*, 14(6), 19–25. <https://doi.org/10.9790/5728-1406011925>
- Denysiuk, V., & Rybachuk, L. (2024). *Forms of representation of interpolation trigonometric splines*. arXiv. <https://doi.org/10.48550/arXiv.2407.19275>
- Fairweather, G., & Meade, D. (1989). A survey of spline collocation methods for the numerical solution of differential equations. In *Mathematics for large scale computing* (Vol. 120, pp. 297–341). Marcel Dekker. <https://doi.org/10.1201/9780429332760-12>
- Fletcher, C. A. J. (2012). *Computational Galerkin methods*. Springer. <https://doi.org/10.1007/978-3-642-85949-6>
- Guter, R. S., Kudryavtsev, L. D., & Levitan, B. M. (1966). *Elements of the theory of functions*. Pergamon Press.
- Heilat, A. S. (2023). Cubic trigonometric B-spline method for solving a linear system of second order boundary value problems. *European Journal of Pure and Applied Mathematics*, 16(4), 2384–2396. <https://doi.org/10.29020/nybg.ejpm.v16i4.4947>
- Khan, S. N. K. A., & Misro, M. Y. (2025). Hybrid B-spline collocation method with particle swarm optimization for solving linear differential problems. *AIMS Mathematics*, 10(3), 5399–5420. <https://doi.org/10.3934/math.2025249>
- Kukreja, V. K., & Shallu. (2025). Highly accurate spline collocation technique for the numerical solution of generalized Burgers-Fisher's problem. *Computational Methods for Differential Equations*, 13(2), 578–591. <https://doi.org/10.22034/CMDE.2024.49824.2071>
- Nazir, T., Abbas, M., Ismail, A. I. M., Majid, A. A., & Rashid, A. (2016). The numerical solution of advection–diffusion problems using new cubic trigonometric B-splines approach. *Applied Mathematical Modelling*, 40(7–8), 4586–4611. <https://doi.org/10.1016/j.apm.2015.11.041>
- Prenter, P. M. (1975). *Splines and variational methods*. Wiley.
- Roul, P. (2024). A high-order B-spline collocation method for solving a class of nonlinear singular boundary value problems. *Journal of Mathematical Chemistry*, 62, 1308–1322. <https://doi.org/10.1007/s10910-024-01590-z>
- Schoenberg, I. J. (1964). On trigonometric spline interpolation. *Journal of Mathematics and Mechanics*, 13(5), 795–825.
- Schumaker, L. L. (2010). *Spline functions: Basic theory* (3rd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511618994>
- Yaseen, M., Abbas, M., & Riaz, M. B. (2021). A collocation method based on cubic trigonometric B-splines for the numerical simulation of the time-fractional diffusion equation. *Advances in Difference Equations*, 2021, 210. <https://doi.org/10.1186/s13662-021-03360-6>
- Zou, X. (2025). *Trigonometric interpolation based optimization for second order non-linear ODE with mixed boundary conditions*. arXiv. <https://doi.org/10.48550/arXiv.2504.19280>

Отримано редакцією журналу / Received: 01.12.25

Прорецензовано / Revised: 17.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Володимир ДЕНИСЮК, д-р фіз.-мат. наук, проф.

ORCID ID: 0000-0002-3210-7350

e-mail: kvomden@ukr.net

Державний університет "Київський авіаційний інститут", Київ, Україна

Тетяна ОЛЕШКО, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-8054-1178

e-mail: tetiana.oleshko@npp.kai.edu.ua

Державний університет "Київський авіаційний інститут", Київ, Україна

Людмила РИБАЧУК, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-7652-587X

e-mail: rybachuk.liudmyla@iil.kpi.ua

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

НЕПОВНІ ТРИГОНОМЕТРИЧНІ СПЛАЙНИ В ЗАДАЧАХ ПОБУДОВИ НАБЛИЖЕНИХ РОЗВ'ЯЗКІВ ЛІНІЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ ДРУГОГО ПОРЯДКУ

Представлено числовий аналітичний метод для отримання наближених розв'язків першої крайової задачі, пов'язаної з лінійними диференціальними рівняннями другого порядку. Підхід базується на використанні неповних фундаментальних тригонометричних сплайнів, як парних, так і непарних, структурні властивості яких дуже схожі на властивості класичних поліноміальних сплайнів та були широко досліджені в попередніх роботах. Запропонована методологія використовує мінімізацію залишків за допомогою встановленого методу колокації, тим самим зводячи вихідну задачу до розв'язання систем лінійних алгебраїчних рівнянь відносно параметрів, що визначають сплайн. Відмінна особливість цього методу полягає в тому, що самі значення шуканого розв'язку є визначальними параметрами фундаментальних тригонометричних сплайнів. Гранічні умови враховано шляхом призначення конкретних значень для відповідних параметрів сплайна. Теоретичний виклад доповнено ілюстративними прикладами.

Варто зазначити, що неповні тригонометричні сплайни є функціями з певною парністю. На нашу думку, у багатьох випадках ця властивість зручна під час задання граничних умов для шуканого розв'язку. Наведений метод передбачає узагальнення. Зокрема, мінімізацію залишку можна здійснити методом дискретних найменших квадратів і його варіаціями. Цей метод також можна застосовувати до лінійних диференціальних рівнянь довільного порядку.

Ключові слова: фундаментальні тригонометричні сплайни, неповні тригонометричні сплайни, парні фундаментальні сплайни, непарні фундаментальні сплайни, крайові задачі, лінійні диференціальні рівняння.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 532.59: 534.29

DOI: <https://doi.org/10.17721/1812-5409.2026/1.16>

Олександр ЖУК, д-р фіз.-мат. наук, ст. наук. співроб.

ORCID: 0009-0004-7253-6127

e-mail: opzhuk1939@gmail.com

Інститут механіки ім. С. П. Тимошенка НАН України, Київ, Україна

Ярослав ЖУК, д-р фіз.-мат. наук, проф., чл.-кор. НАН України

ORCID: 0000-0002-2726-8395

e-mail: yaroslavzhuk@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

Марія КАШТАЛЯН, д-р філософії, проф.

ORCID: 0000-0003-1950-5677

e-mail: m.kashtalyan@abdn.ac.uk

Абердинський університет, Абердин, Шотландія, Велика Британія

ДІЯ СИЛ АКУСТИЧНОГО ВИПРОМІНЕННЯ В ІДЕАЛЬНІЙ РІДИНІ НА В'ЯЗКУ КРАПЛЮ

Досліджено дію акустичної радіаційної сили на сферичну краплю в'язкої рідини, розташовану в ідеальній рідині в полі плоскої звукової хвилі. Акустична радіаційна сила є мірою взаємодії між падаючою акустичною хвилею та краплею в'язкої рідини, що визначається як середнє за часом інтегральне значення звукового тиску на поверхню в'язкої краплі. Розроблено процедуру розв'язання задачі, що реалізується у два етапи. На першому етапі для визначення тиску розроблено підхід, що передбачає застосування потенціалу поля швидкості, отриманого як рішення лінійної задачі дифракції первинної хвилі на перешкоді. Під час формулювання граничних умов тиск поверхневого натягу краплі не враховували, а її радіус вважали незмінним, тобто розглядали випадок малих пульсаційних коливань. Задоволення граничних умов на поверхні в'язкої краплі й умов Зоммерфельда на нескінченності разом із використанням властивості ортогональності поліномів Лежандра дало змогу отримати нескінченну систему алгебраїчних рівнянь відносно невідомих коефіцієнтів у розвиненнях потенціалів у ряди Фур'є. На другому етапі розв'язування задачі за допомогою знайдених потенціалів обчислено тиск і гідродинамічну силу, викликану акустичним полем. Наступне осереднення гідродинамічної сили за часом допомогло обчислити безпосередньо акустичну радіаційну силу. Числові розрахунки проведено для випадку, коли зовнішньою рідиною є вода, а крапля, утворена із нерозчинної у воді в'язкої рідини, – гліцерин. Проведено серію розрахунків для множини значень радіусу в'язкої краплі. Встановлено, що акустична радіаційна сила не є монотонною функцією частоти первинної хвилі. У випадку твердої сферичної частинки напрям дії радіаційної сили, як відомо, збігається з напрямом поширення акустичної хвилі. У випадку в'язкої сферичної краплі напрям дії радіаційної сили може як збігатися з напрямом поширення хвилі, так і бути протилежним. Виявлено, що існують частоти падаючої хвилі, які забезпечують нульову радіаційну силу, що діє на краплю, яка в цьому випадку є нерухомою.

Ключові слова: акустична радіаційна сила, сферична крапля рідини, ідеальна стислива рідина, в'язка рідина, плоска гармонічна хвиля, звукове поле.

Класифікація відповідно до AMS 2020: 76Q05, 74F10.

Вступ

Уперше явище акустичного тиску, що виникає під дією звукової хвилі, було описане у класичній роботі (King, 1934). Подальший розвиток теорії спрямований на дослідження акустичних радіаційних сил (APC), які діють на тверді, рідинні та газові об'єкти, що перебувають у полі звукової хвилі (Zhuk O., Zhuk Y., & Klimchuk, 2023; Cantrell, 2018). Ці сили є результатом перерозподілу імпульсу хвилі внаслідок розсіювання на межах поділу середовищ і зумовлюють сталу компоненту тиску (King, 1934; Doinkinov, 2005). З практичного погляду, ефект акустичного випромінення широко застосовують у сучасних технологіях – зокрема, для ультразвукової левітації, очищення рідин, акустичного транспортування мікрокрапель, маніпуляцій із мікрочастинками тощо. Для опису таких систем необхідно враховувати як властивості зовнішнього середовища, так і внутрішню в'язкість краплі, що впливає на розподіл тиску та швидкості в рідині.

У пропонованій статті досліджено дію акустичного випромінення на сферичну краплю в'язкої рідини, розміщену в ідеальній стисливій рідині під дією плоскої гармонічної звукової хвилі. Метою роботи є встановлення закономірностей впливу параметрів звукового поля та фізичних характеристик рідин на величину і напрям акустичної сили. Використано аналітичний підхід, що базується на розв'язанні лінійної задачі дифракції хвилі методом розділення змінних і подальшому визначенні гідродинамічних та радіаційних сил шляхом усереднення за період хвилі (Zhuk O., Zhuk Y., & Klimchuk, 2023).

1. Постановка задачі

Уведемо з початком у центрі сферичної краплі дві системи координат: прямокутну декартову $Ox_1x_2x_3$ і сферичну $O\theta\varphi$. Динамічний процес у зовнішній ідеальній рідині, визначений на першому етапі під час розв'язання лінійної задачі дифракції, описується потенціалом Φ_i первинної хвилі та потенціалом Φ_s розсіяної хвилі на краплі, що є розв'язками лінійного хвильового рівняння

$$\Delta\Phi - \frac{1}{a_0^2} \frac{\partial^2 \Phi}{\partial t^2} = 0. \quad (1)$$

$$\Phi_i = A \exp[i(kx - \omega t)], \quad \Phi_i = A \sum_{n=0}^{\infty} (2n+1) i^n j_n(kr) P_n(\cos\theta) \exp(-i\omega t), \quad \Phi_s = \sum_{n=0}^{\infty} A_n h_n^{(1)}(kr) P_n(\cos\theta) \exp(-i\omega t),$$

$$p = -\rho_0 \frac{\partial \Phi}{\partial t}, \quad \vec{v} = \text{grad } \Phi, \quad \Phi = \Phi_i + \Phi_s, \quad k = \frac{\omega}{a_0}.$$

© Жук Олександр, Жук Ярослав, Кашталян Марія, 2026

Потенціал Φ_s задовольняє умови Зоммерфельда на нескінченності.

Динамічний процес у в'язкій рідині краплі на першому етапі під час розв'язування лінійної задачі дифракції опишемо за допомогою лінеаризованої теорії в'язкої стисливої рідини, використавши потенціали швидкості потенціальної $\bar{\Phi}_d$ і вихрової $\bar{\Psi}_d$ складових потенціалу швидкості хвильового поля, що є розв'язками рівнянь:

$$\left[\left(1 + \frac{\lambda' + 2\mu'}{\bar{\rho}_0 \bar{a}_0^2} \frac{\partial}{\partial t} \right) \Delta - \frac{1}{\bar{a}_0^2} \frac{\partial^2}{\partial t^2} \right] \bar{\Phi}_d = 0, \quad \left(v' \Delta - \frac{\partial}{\partial t} \right) \bar{\Psi}_d = 0, \quad (2)$$

$$\bar{\Phi}_d = \sum_{n=0}^{\infty} \bar{A}_n j_n(\alpha r) P_n(\cos \theta) \exp(-i\omega t), \quad \bar{\Psi}_d = \sum_{n=0}^{\infty} \bar{B}_n j_n(\beta r) P_n(\cos \theta) \exp(-i\omega t). \quad (3)$$

$$\alpha = \bar{k} + i\bar{k}_1, \quad \bar{k} = i\bar{k}_1, \quad \bar{k}_1 = (\lambda' + 2\mu')\omega / (2\bar{\rho}_0 \bar{a}_0^2), \quad \beta = (i\omega/v')^{1/2}, \quad \bar{v} = \nabla \bar{\Phi}_d + \nabla \times (\nabla \times \bar{e}_r r \bar{\Psi}_d),$$

$$\hat{\sigma} = (-\bar{p} + \lambda' \nabla \cdot \bar{v}) \hat{E} + 2\mu' \hat{e}, \quad 2\hat{e} = \nabla \bar{v} + (\nabla \bar{v})^T, \quad \bar{p} = \bar{\rho}_0 \left(\frac{\lambda' + 2\mu'}{\bar{\rho}_0} \Delta - \frac{\partial}{\partial t} \right) \bar{\Phi}_d, \quad \sigma_{rr} = \bar{e}_r \cdot \hat{\sigma} \cdot \bar{e}_r, \quad \sigma_{r\theta} = \bar{e}_r \cdot \hat{\sigma} \cdot \bar{e}_\theta.$$

Із множин розв'язків для $\Phi_s, \bar{\Phi}_d$ і $\bar{\Psi}_d$ вибір необхідних співвідношень для потенціалів – обчислення невідомих коефіцієнтів в їхніх розвиненнях у ряди Фур'є – проведено шляхом дотримання граничних умов на поверхні сферичної краплі в'язкої рідини:

$$v_r|_{r=R} = \bar{v}_r|_{r=R}, \quad p|_{r=R} = \sigma_{rr}|_{r=0}, \quad \sigma_{r\theta}|_{r=R} = 0. \quad (4)$$

Під час формулювання граничних умов тиск поверхневого натягу краплі не враховували, а її радіус R вважали незмінним – малі пульсаційні коливання.

2. Методика розв'язання задачі та властивості рідин

Використавши ортогональність поліномів Лежандра в рівняннях (4), одержано нескінченну систему алгебраїчних рівнянь відносно невідомих коефіцієнтів $A_n, \bar{A}_n, \bar{B}_n$ в розвиненнях потенціалів у ряди Фур'є

$$\begin{cases} k \frac{dh_n^{(1)}(q)}{dq} A_n + i \frac{\omega}{R} \beta_{1n} \bar{A}_n + i \frac{\omega}{R} \beta_{2n} \bar{B}_n = -A(2n+1) i^n k \frac{dj_n(q)}{dq} \\ i\omega \rho_0 h_n^{(1)}(q) A_n - \frac{2\mu'}{R^2} \alpha_{1n} \bar{A}_n - \frac{2\mu'}{R^2} \alpha_{2n} \bar{B}_n = -A(2n+1) i^{n+1} \omega \rho_0 j_n(q) \\ \varepsilon_{1n} \bar{A}_n + \varepsilon_{2n} \bar{B}_n = 0, \quad q = kR. \end{cases} \quad (5)$$

У системі рівнянь (5) уведено такі позначення:

$$\alpha_{1n} = \left[n^2 - n - \frac{(\beta R)^2}{2} \right] j_n(\alpha R) + 2(\alpha R) j_{n+1}(\alpha R), \quad \alpha_{2n} = n(n+1) [(n-1) j_n(\beta R) - (\beta R) j_{n+1}(\beta R)], \quad \beta_{2n} = n(n+1) j_n(\beta R),$$

$$\beta_{1n} = n j_n(\alpha R) - (\alpha R) j_{n+1}(\alpha R), \quad \varepsilon_{1n} = (n-1) j_n(\alpha R) - (\alpha R) j_{n+1}(\alpha R), \quad \varepsilon_{2n} = \left[n^2 - 1 - \frac{(\beta R)^2}{2} \right] j_n(\beta R) - (\beta R) j_n(\beta R).$$

Система рівнянь (5) розв'язана для випадку малої в'язкості. Уведено параметр $\varepsilon = (v'/(\omega R^2))^{1/2} \ll 1$ і з точністю до ε^2 одержано: $\alpha R \approx \bar{k} R$, $\beta R \approx (1+i)/(\varepsilon\sqrt{2}) = \varepsilon^{-1} \exp(i\pi/4)$.

Під час обчислення сферичних функцій за великих значень аргументів використано їхні асимптотичні представлення. Визначенням потенціалів акустичного поля завершується розв'язування задачі дифракції первинної хвилі на в'язкій краплі.

На другому етапі розв'язування задачі обчислено гідродинамічну силу, викликану акустичним полем. Симетрія акустичного поля відносно осі Ox_3 обумовлює напрям її дії на краплю вздовж цієї осі. Вона дорівнює

$$F_{x3} = -R^2 \int_0^{\pi} \int_0^{2\pi} p \sin \theta \cos \theta d\theta d\varphi. \quad (6)$$

Відповідно до прийнятого методу дослідження, тиск P у формулі (5) для гідродинамічної сили обчислено, використавши потенціал $\Phi = \Phi_i + \Phi_s$ акустичного поля, одержаний на першому етапі розв'язування задачі. До того ж відповідно до застосованого методу дослідження тиск P у формулі (6) для гідродинамічної сили визначено з точністю до величин другого порядку

$$p = -\rho_0 \frac{\partial \Phi}{\partial t} - \frac{1}{2} \rho_0 (grad \Phi)^2 + \frac{1}{2} \frac{\rho_0}{a_0^2} \left(\frac{\partial \Phi}{\partial t} \right)^2. \quad (7)$$

Структура формули (7) передбачає під час обчислення тиску P використання дійсної складової потенціалу Φ . Тоді у формулі (7) для тиску необхідно покласти

$$\frac{\partial \Phi}{\partial t} = \frac{d\Phi}{dt} - V_{x3} \cos \theta \frac{\partial \Phi}{\partial r} + V_{x3} \frac{\sin \theta}{r} \frac{\partial \Phi}{\partial \theta}. \quad (8)$$

Для визначення швидкості коливального руху краплі вздовж осі Ox_3 застосовано рівняння

$$m \frac{dV_{x3}}{dt} = - \iint_S p \cos \theta dS, \quad p = -\bar{p}_0 \frac{\partial \Phi}{\partial t}, \quad m = \frac{4}{3} \pi R^3 \bar{\rho}_0. \quad (9)$$

Інтеграл у правій частині рівняння (9) береться по всій поверхні сферичної краплі й визначає силу, що діє на неї з боку оточуючої ідеальної рідини.

Вважаючи радіус сферичної краплі малим і не наводячи громіздкі викладки під час обчислення гідродинамічної сили (7) і в її осередненні в часі, одержимо:

$$\begin{aligned} \langle F_{x3} \rangle &= \langle F_{x3}^{(1)} \rangle + \langle N_{x3}^{(1)} \rangle + \langle N_{x3}^{(2)} \rangle + \langle F_{x3}^{(3)} \rangle. \\ \langle F_{x3}^{(1)} \rangle &= 2\pi\rho_0\eta \left[\frac{1}{3} q \left(K_1(q) \frac{dK_0(q)}{dq} + L_1(q) \frac{dL_0(q)}{dq} \right) + \frac{2}{15} q \left(K_1(q) \frac{dK_2(q)}{dq} + L_1(q) \frac{dL_2(q)}{dq} \right) + \frac{2}{5} (K_1(q)K_2(q) + L_1(q)L_2(q)) \right], \\ \langle N_{x3}^{(1)} \rangle &= 2\pi\rho_0q^2 \sum_{n=0}^{\infty} \frac{n+1}{(2n+1)(2n+3)} \left(\frac{dK_n(q)}{dq} \frac{dK_{n+1}(q)}{dq} + \frac{dL_n(q)}{dq} \frac{dL_{n+1}(q)}{dq} \right), \\ \langle N_{x3}^{(2)} \rangle &= 4\pi\rho_0 \sum_{m=0}^{\infty} \frac{n(n+1)(n+2)}{(2n+1)(2n+3)} (K_n(q)K_{n+1}(q) + L_n(q)L_{n+1}(q)), \\ \langle F_{x3}^{(3)} \rangle &= -2\pi\rho_0q^2 \sum_{n=0}^{\infty} \frac{n+1}{(2n+1)(2n+3)} (K_n(q)K_{n+1}(q) + L_n(q)L_{n+1}(q)). \end{aligned} \quad (10)$$

У формулах для складових APC позначено:

$$\begin{aligned} K_n(q) &= (2n+1)A_j(q) \cos\left(n\frac{\pi}{2}\right) + A_{n1}j_n(q) - A_{n2}y_n(q), \quad L_n(q) = (2n+1)A_j(q) \sin\left(n\frac{\pi}{2}\right) + A_{n1}y_n(q) + A_{n2}j_n(q), \\ A_n &= \text{Re } A_n + i \text{Im } A_n = A_{n1} + iA_{n2}. \end{aligned}$$

Аргумент q в обчисленні наведених вище співвідношень дорівнює kR .

3. Результати розрахунків й аналіз

Для числових досліджень дії APC на сферичну краплю в'язкої рідини в ідеальній рідині використано формулу (10). Зовнішньою рідиною є вода (густина 1000 кг/м^3 , швидкість звуку 1500 м/с), а, крапля утворена із нерозчинної у воді в'язкої рідини, – гліцерин (густина 1264 кг/м^3 , швидкість звуку 1920 м/с , коефіцієнт динамічної в'язкості $1,412 \text{ Па}\cdot\text{с}$). Амплітуда A первинної хвилі Φ_i становить $0,9 \cdot 10^{-4} \text{ м}^2/\text{с}$. Результати розрахунків зображено на рис. 1, 2. Цифри, якими позначено криві, вказують радіус краплі в міліметрах.

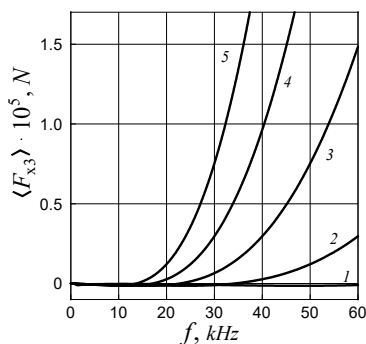


Рис. 1. Загальна залежність APC від нуля

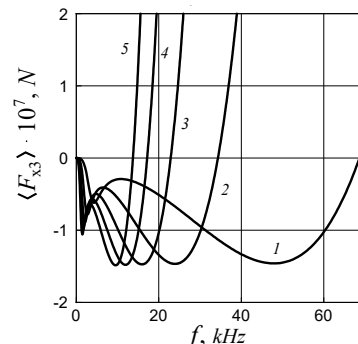


Рис. 2. Особливість поведінки APC в околі частоти первинної хвилі

Бачимо, що радіаційна сила не є монотонною функцією частоти первинної хвилі. На відміну від випадку твердої сферичної частинки, напрям дії радіаційної сили може як збігатися з напрямом поширення акустичної хвилі, так і бути протилежним.

Дискусія і висновки

Встановлено, що акустична радіаційна сила не є монотонною функцією частоти первинної хвилі. У випадку твердої сферичної частинки напрям дії APC, як відомо, збігається з напрямом поширення акустичної хвилі. У випадку в'язкої сферичної краплі напрям дії радіаційної сили може як збігатися з напрямом поширення хвилі, так і бути протилежним. Виявлено, що існують частоти падаючої хвилі, які забезпечують нульову APC, що діє на краплю, яка в цьому випадку є нерухомою.

Внесок авторів: Олександр Жук – концептуалізація, методика; Ярослав Жук – збір і перевірка емпіричних даних, емпіричне дослідження, програмне забезпечення; Марія Кашталян – аналіз джерел, підготовка огляду літератури або теоретичних основ дослідження.

Джерела фінансування. Робота фінансувалася частково власним коштом і коштом КНУ імені Тараса Шевченка.

Список використаних джерел

- Cantrell, J. H. (2018). *Acoustic radiation pressure in fluids*. NASA/TM–2018-219806.
 Doinikov, A. A. (2005). *Bubble and Particle Dynamics in Acoustic Fields: Modern Trends and Applications*. Research Signpost.
 King, L. V. (1934). On the acoustic radiation pressure on spheres. *Proceedings of Royal Society of London A*, 147, 212–240. <https://doi.org/10.1098/rspa.1934.0215>
 Zhuk, O. P., Zhuk, Y., & Klimchuk, T. (2023). On the Acoustic Radiation Force Affecting Two Liquid Drops Located in the Wave Field. *Axioms*, 12(10), 940. <https://doi.org/10.3390/axioms12100940>

References

- Cantrell, J. H. (2018). *Acoustic radiation pressure in fluids*. NASA/TM–2018-219806.
 Doinikov, A. A. (2005). *Bubble and Particle Dynamics in Acoustic Fields: Modern Trends and Applications*. Research Signpost.
 King, L. V. (1934). On the acoustic radiation pressure on spheres. *Proceedings of Royal Society of London A*, 147, 212–240. <https://doi.org/10.1098/rspa.1934.0215>
 Zhuk, O. P., Zhuk, Y., & Klimchuk, T. (2023). On the Acoustic Radiation Force Affecting Two Liquid Drops Located in the Wave Field. *Axioms*, 12(10), 940. <https://doi.org/10.3390/axioms12100940>

Отримано редакцією журналу / Received: 30.10.25
 Прорецензовано / Revised: 18.03.26
 Схвалено до друку / Accepted: 16.04.26
 Опубліковано / Published: 05.06.26

Oleksandr ZHUK, DSc (Phys. & Math.), Senior Researcher
 ORCID: 0009-0004-7253-6127
 e-mail: opzhuk1939@gmail.com
 S. P. Timoshenko Institute of Mechanics, NAS of Ukraine, Kyiv, Ukraine

Yaroslav ZHUK, DSc (Phys. & Math.), Prof., Corresponding Member of the National Academy of Sciences of Ukraine
 ORCID: 0000-0002-2726-8395
 e-mail: yaroslavzhuk@knu.ua
 Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Maria KASHTALYAN, PhD, Prof.
 ORCID: 0000-0003-1950-5677
 e-mail: m.kashtalyan@abdn.ac.uk
 University of Aberdeen, Aberdeen, Scotland, UK

THE EFFECT OF ACOUSTIC RADIATION IN AN IDEAL FLUID ON THE VISCOSITY OF A DROP

The effect of acoustic radiation force on a spherical drop of viscous liquid, located in turn in an ideal liquid in a plane sound wave field, is investigated. Acoustic radiation force is a measure of the interaction between an incident acoustic wave and a drop of viscous liquid, defined as the time-averaged integral value of the sound pressure over the surface of the viscous drop. A procedure for solving the problem has been developed, which is implemented in two steps. At the first step, an approach was developed to determine the pressure, which involves the use of the velocity field potential obtained as a solution to the linear problem of incident wave scattering at an obstacle. When formulating the boundary conditions, the surface tension pressure of the drop was not taken into account, and its radius was considered constant, i.e., the case of small pulsation oscillations was considered. Satisfying the boundary conditions on the surface of a viscous drop and Sommerfeld's conditions at infinity, together with the use of the orthogonality property of Legendre polynomials, made it possible to obtain an infinite system of algebraic equations with respect to unknown coefficients in the Fourier series expansions of potentials. At the second step of solving the problem, the pressure and hydrodynamic force caused by the acoustic field were calculated using the found potentials. The subsequent averaging of the hydrodynamic force over time allowed us to calculate the acoustic radiation force directly. Numerical calculations were performed for the case where the external fluid is water and the drop is formed from a viscous fluid insoluble in water - glycerin. A series of calculations was performed for a set of values of the radius of the viscous drop. It was found that the acoustic radiation force is not a monotonic function of the incident wave frequency. In the case of a solid spherical particle, the direction of the radiation force, as is known, coincides with the direction of propagation of the acoustic wave. In the case of a viscous spherical drop, the direction of the radiation force can be either coincident with the direction of wave propagation or opposite to it. It has been found that there are frequencies of the incident wave that provide zero radiation force acting on the drop, which in this case is stationary.

Key words: *acoustic radiation force, spherical liquid drop, ideal compressible fluid, viscous fluid, plane harmonic wave, sound field.*

Один з авторів (Я. Жук) є членом редколегії видання, тому він не брав участі у рецензуванні та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

One of the authors (Ya. Zhuk) is a member of the journal editorial board; he did not take part in the peer-review process or in the decision to publish this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 517.9

DOI: <https://doi.org/10.17721/1812-5409.2026/1.17>

Oleksiy KAPUSTYAN¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-9373-6812
e-mail: kapustyan@knu.ua

Anna SUKRETN¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-2985-1250
e-mail: sukretna.a.v@knu.ua

Dmytro KAPUSTIAN¹, Student
ORCID ID: 0009-0008-2992-4545
e-mail: dimonkapus@gmail.com

Myrzagali OSPANOV², PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0009-0009-3616-8511
e-mail: ospanov_mn@enu.kz

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
²L. N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan

Dedicated to the memory of academician Mykola Perestyuk

STABILITY OF INFECTION-FREE ALMOST PERIODIC SOLUTION IN IMPULSIVE SEIR-MODEL

Mathematical models of the infection spread dynamics have remained a relevant area of research for the past few decades. Moreover, these models have also proven to be suitable for modeling DDoS attacks in computer networks. The qualitative behaviour of solutions in such models is usually arranged as follows: there is some threshold value at which the global asymptotic stability of the solution with a zero component, which is responsible for the number of infected individuals, changes to the stability of the non-zero equilibrium position. From a meaningful point of view, this means that the self-exhaustion and disappearance of the infection changes to a constant circulation of the infection in the population.

However, the presence of an impulsive perturbation can make the infection-free solution asymptotically stable regardless of the threshold value, and such impulsive perturbations allow a natural interpretation within the framework of the epidemiological model as the process of vaccination of a certain part of individuals.

Such problems have already been considered for the SIR- and SEIR-models in the case of a simple structure of impulsive points, due to which the impulsive problem was actually reduced to a problem without impulsive perturbation. However, a more complex structure of the set and magnitudes of impulsive effects requires the application of the theory of impulsive differential equations, which is done in this work.

Namely, the article considers the SEIR-system of differential equations with impulsive perturbations that have an almost periodic character. The existence and properties of an infection-free impulsive almost periodic solution are investigated, and its asymptotic stability is proven.

Key words: SEIR-model of differential equations, impulsive perturbation, infection-free solution, periodic solution, almost periodic solution.

AMS 2020 classification: 34C25, 34C27, 34D10, 34D20.

Introduction

The classical SIR-model is a three-dimensional system of differential equations that describes the infection spread dynamics in a population consisting of potentially Susceptible, Infectious and Recovered individuals. Various modifications of this system, as well as a qualitative analysis, are contained in the review (Hethcote, 1989). SEIR-model is a four-dimensional system that includes Exposed individuals of infection as part of the population (Liu, Hethcote, & Levin, 1987). This model can also be adapted to describe the dynamics of certain types of DDoS attacks in computer networks (Balarezo et al., 2022). A qualitative analysis of this type of model has been investigated in works (Li, & Muldowney, 1995; Li, & Muldowney, 1996). As in SIR-models, it is shown that if the contact number $\sigma \leq 1$, then there are physically justified invariant regions where the only globally asymptotically stable equilibrium position is $I = 0$ (infection-free). If $\sigma > 1$, then another equilibrium position with $I > 0$ becomes asymptotically stable.

It turns out that the presence of an impulsive perturbation can radically change this qualitative picture. The theory of differential equations systems with an impulsive perturbation was created in the works of prominent mathematicians A. M. Samoilenko and M. O. Perestyuk (Samoilenko, & Perestyuk, 1995). In models of the infection spread, the impulsive perturbation has a natural interpretation as a vaccination procedure of a certain part of potentially susceptible individuals. For the first time, such a problem was considered in (Shulgin, Stone, & Agur, 1998). In the simplest situation, when moments of an impulsive perturbation are equidistant, the asymptotic stability of an infection-free impulsive periodic solution was proven. The same impulsive problem for the SEIR-model was solved in (D'Onofrio, 2002). It should be noted that the simple structure of the impulsive moments ($t_n = nT$) actually allows to reduce such a problem to the study of a periodic problem without impulses, that makes it problematic to use the approach (Shulgin, Stone, & Agur, 1998), (D'Onofrio, 2002) to more complex impulsive problems (Feketa, Klinshov, & Lücken, 2021).

In this paper, we investigate the impulsive problem when at the moments of time $t_n = nT + \alpha_n$, $\{\alpha_n\}$ – almost periodic sequence) trajectories of SEIR-model undergo impulsive perturbation which can be interpreted as a vaccination process. Using the general impulsive systems theory (Samoilenko, & Perestyuk, 1995), we prove the existence of an almost periodic impulsive infection-free solution and investigate its dependence on the parameter T . This allows us to establish the main result of the paper that for sufficiently small T this solution is asymptotically stable.

© Kapustyan Oleksiy, Sukretna Anna, Kapustian Dmytro, Ospanov Myrzagali, 2026

1. Formulation of the problem

We consider a classical SEIR-model respect to the unknown functions $S(t)$ (Susceptible is the fraction of the population that can become infected), $E(t)$ (Exposed is the fraction of the population that is carried the infection), $I(t)$ (Infectious is the fraction of the population that is infected) and $R(t)$ (the fraction of the population that Recovered and has immune), for which equality

$$S(t) + E(t) + I(t) + R(t) = 1, \quad t \geq 0$$

and an ordinary differential equations system

$$\begin{cases} \dot{S} = -\lambda SI + \nu - \nu S + \delta R, \\ \dot{E} = \lambda SI - (\varepsilon + \nu)E, \\ \dot{I} = \varepsilon E - (\gamma + \nu)I, \\ \dot{R} = \gamma I - (\delta + \nu)R, \end{cases}$$

holds. Here $\lambda, \nu, \delta, \varepsilon, \gamma$ are positive parameters that determine the indicators of the spread of infection, namely:

$\lambda > 0$ is the natural frequency of contacts,

$\nu > 0$ is the proportion of deaths and newborns,

$\delta > 0$ is the proportion of recovered people who lose immunity,

$\varepsilon > 0$ is the proportion of infection carriers who get sick,

$\gamma > 0$ is the proportion of sick people who recover.

Also important for analysing the problem is the contact number

$$\sigma = \frac{\lambda \varepsilon}{(\varepsilon + \nu)(\gamma + \nu)}.$$

Substituting the expression $R = 1 - S - E - I$ into the system, we have the main object of study

$$\begin{cases} \dot{S} = -\lambda SI + \nu - \nu S + \delta(1 - S - E - I), \\ \dot{E} = \lambda SI - (\varepsilon + \nu)E, \\ \dot{I} = \varepsilon E - (\gamma + \nu)I. \end{cases} \quad (1)$$

Let's

$$D = \{(S, E, I) \in \mathbb{R}_+^3 \mid 0 \leq S + E + I \leq 1\}$$

then the following result is known.

Theorem 1 (Li, & Muldowney, 1996). *For any initial data $(S(0), E(0), I(0)) \in D$ there is only one corresponding solution of (1) on the interval $[0, +\infty)$ and for all $t \geq 0$ the inclusion $(S(t), E(t), I(t)) \in D$ takes place.*

If $\sigma \leq 1$, then where is the only one equilibrium point $P_0^ = \{S_0^* = 1, E_0^* = 0, I_0^* = 0\}$ (infection free) in D , which is globally asymptotically stable in D .*

If $\sigma > 1$, then P_0^ is an unstable equilibrium point and another equilibrium point $P_+^* = \{S_1^* = \frac{1}{\sigma}, E_1^* \in (0, 1), I_1^* \in (0, 1)\}$ appears, which is globally asymptotically stable in $int D$.*

Modelling the vaccination process as a sharp decrease in the susceptible to infection population proportion we arrive at an impulsive problem

$$\Delta S|_{t=t_n} := S(t_n + 0) - S(t_n) = -b_n S(t_n), \quad (2)$$

where $\{b_n\} \subset (0, 1)$, $\{t_n\}$ are the given values and the moments of the impulsive perturbation.

By solution (1), (2) according to (Samoilenko, & Perestyuk, 1995) we will understand a left-continuous function, which is satisfies the system (1), when $t \neq t_n$, and at points $\{t_n\}$ satisfies the equality (2).

In papers (Shulgin, Stone, & Agur, 1998), (D'Onofrio, 2002) in the case $b_n = b$, $t_n = nT$ it is shown that when $\sigma > 1$ in problem (1), (2) there is a periodic impulsive solution $\bar{P}(t) = (\bar{S}(t), 0, 0)$, which under condition

$$\frac{1}{T} \int_0^T \bar{S}(t) dt < \frac{1}{\sigma}$$

is asymptotically stable.

In this paper we generalize this result to the case of almost periodic (Samoilenko, & Perestyuk, 1995) sequences $\{b_n\}$, $\{t_n = nT + \alpha_n\}$. We prove the existence of an almost periodic impulsive infection-free solution and show that for sufficiently small $T > 0$ this solution is asymptotically stable.

2. Main results

Let us first consider a periodic problem, namely, let there exist $p \in \mathbb{N}$ and $T > 0$ such that

$$\begin{aligned} 0 < t_1 < t_2 < \dots < t_p < T, \quad t_{n+p} = t_n + T, \quad n \geq 1, \\ b_{n+p} = b_n, \quad n \geq 1. \end{aligned} \quad (3)$$

Then the following statement is true.

Lemma 1. *Under conditions (3), problem (1), (2) has a periodic infection-free impulsive solution, and also*

$$\sup_{t \geq 0} \bar{S}(t) \rightarrow 0, \quad T \rightarrow 0. \quad (4)$$

Proof of Lemma 1. Let's find $\bar{S}(t)$ as a T -periodic solution to the problem

$$\begin{cases} \dot{S} = (\nu + \delta)(1 - S), \\ \Delta S|_{t=t_n} = -b_n S(t_n). \end{cases} \quad (5)$$

On the interval of continuity $[\tau, t]$ we have

$$S(t) = (S(\tau) - 1)e^{-(v+\delta)(t-\tau)} + 1,$$

then

$$S(t_1 + 0) = [(S(0) - 1)e^{-(v+\delta)t_1} + 1](1 - b_1) = S(0)e^{-(v+\delta)t_1}(1 - b_1) + \alpha_1, \quad 0 < \alpha_1 < 1 - e^{-(v+\delta)t_1},$$

and so on

$$S(t_p + 0) = S(0)e^{-(v+\delta)t_p} \prod_{i=1}^p (1 - b_i) + \alpha_p, \quad 0 < \alpha_p < 1 - e^{-(v+\delta)t_p}.$$

Therefore,

$$S(T) = S(0)e^{-(v+\delta)T} \prod_{i=1}^p (1 - b_i) + (\alpha_p - 1)e^{-(v+\delta)(T-t_p)} + 1$$

and the initial data for this periodic solution is determined as follows

$$S(0) = \frac{(\alpha_p - 1)e^{-(v+\delta)(T-t_p)} + 1}{1 - e^{-(v+\delta)T} \prod_{i=1}^p (1 - b_i)}. \quad (6)$$

At the same time, it follows from (6) that

$$S(0) = \frac{1 - e^{-(v+\delta)T}}{1 - e^{-(v+\delta)T} \prod_{i=1}^p (1 - b_i)},$$

which means that $S(T) \rightarrow 0, T \rightarrow 0$. Then from (5) we deduce that for any $\varepsilon > 0$ there exists $T^* > 0$ such that for all $T \in (0, T^*)$ and for all $t \geq 0$ we have $S(t) < \varepsilon$, hence (4) holds. Lemma 1 is proven. $\square$

Remark 1. It is easy to see that the periodic solution $\bar{S}(t)$ is asymptotically stable in problem (5), that is, for an arbitrary initial data $S(0) \in (0, 1)$ and the corresponding solution $S(t)$ of problem (5) holds

$$S(t) - \bar{S}(t) \rightarrow 0, \quad t \rightarrow \infty. \quad (7)$$

Now we consider an almost periodic impulsive problem. We will use the following definitions (Samoilenko & Perestyuk, 1995).

Definition 1. A sequence $\{\tau_i\}_{i \in \mathbb{Z}}$ is almost periodic if

$$\forall \varepsilon > 0 \quad \exists N = N(\varepsilon) \quad \forall k \in \mathbb{Z} \quad \exists p \in [k, k + N] \cap \mathbb{Z}: \quad \forall i \in \mathbb{Z} \quad |\tau_{i+p} - \tau_i| < \varepsilon. \quad (8)$$

The number p is called ε -almost period of the sequence $\{\tau_i\}_{i \in \mathbb{Z}}$.

In the sense of Definition 1, it is stated that for any $\varepsilon > 0$ there exists a relatively dense set of ε -almost periods.

Remark 2. If the sequence $\{t_i\}_{i \in \mathbb{Z}}$ is such that equality $t_{i+p} = t_i + T$ holds for any $i \in \mathbb{Z}$, then the sequence $\{\tau_i = t_{i+1} - t_i\}_{i \in \mathbb{Z}}$ is almost periodic.

In the following, we consider the situation when the sequence $\{t_i\}_{i \in \mathbb{Z}}$ satisfies the condition

$$\dots < t_i < t_{i+1} < \dots, \quad i \in \mathbb{Z}; \quad t_i \rightarrow \pm\infty, \quad i \rightarrow \pm\infty.$$

Let $N(t, t + s)$ denotes the number of the sequence $\{t_i\}_{i \in \mathbb{Z}}$ members belonging to the interval $(t, t + s]$.

It is known (Samoilenko, & Perestyuk, 1995) that if a sequence $\{t_{i+1} - t_i\}_{i \in \mathbb{Z}}$ is almost periodic, then uniformly over $t \in \mathbb{R}$ there exists a limit

$$\lim_{s \rightarrow \infty} \frac{N(t, t + s)}{s} = \theta. \quad (9)$$

In particular, if $t_{i+p} = t_i + T$ then $\theta = \frac{p}{T}$.

Definition 2. For a given sequence $\{t_i\}_{i \in \mathbb{Z}}$, we put $\tau_i^j = t_{i+j} - t_i, i, j \in \mathbb{Z}$. We will say that the sequence $\{\tau_i^j\}$ is uniformly almost periodic if for an arbitrary $\varepsilon > 0$ there is a relatively dense set of ε -almost periods common to all j , that is

$$\forall \varepsilon > 0 \quad \exists N \quad \forall k \in \mathbb{Z} \quad \exists p \in [k, k + N] \cap \mathbb{Z}: \quad \forall j \in \mathbb{Z} \quad \forall i \in \mathbb{Z} \quad |\tau_{i+p}^j - \tau_i^j| < \varepsilon.$$

Lemma 2 (Samoilenko, & Perestyuk, 1995). Let the sequence $\{t_i\}_{i \in \mathbb{Z}}$ be such that for some $T > 0$ the equalities

$$t_i = iT + \alpha_i, \quad i \in \mathbb{Z},$$

hold, where the sequence $\{\alpha_i\}_{i \in \mathbb{Z}}$ is almost periodic and $\sup_{i \in \mathbb{Z}} |\alpha_i| = \alpha < \frac{T}{2}$. Then the sequence $\{\tau_i^j = t_{i+j} - t_i\}$ is uniformly almost periodic.

Definition 3. A piecewise continuous function $\varphi: \mathbb{R} \rightarrow \mathbb{R}^d$ with discontinuities of the first kind at points $\{t_i\}_{i \in \mathbb{Z}}$ is called almost periodic if the sequence $\{\tau_i^j = t_{i+j} - t_i\}$ is uniformly almost periodic and the following properties are satisfied:

- 1) $\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall i \in \mathbb{Z} \quad \forall t', t'' \in (t_i, t_{i+1}): |t' - t''| < \delta \Rightarrow \|\varphi(t') - \varphi(t'')\| < \varepsilon;$
- 2) for arbitrary $\varepsilon > 0$ there exists set Γ for which there exists such $l > 0$ that for all $a \in \mathbb{R}: [a, a + l] \cap \Gamma \neq \emptyset$ and for each $\tau \in \Gamma, t \in \mathbb{R}$ inequality $|t - t_i| > \varepsilon$ leads to $\|\varphi(t + \tau) - \varphi(t)\| < \varepsilon$.

Now consider the almost periodic impulsive problem (2), where

$$t_i = iT + \alpha_i, \quad i \in \mathbb{Z},$$

$$\text{the sequence } \{\alpha_i\}_{i \in \mathbb{Z}} \text{ is almost periodic and } \sup_{i \in \mathbb{Z}} |\alpha_i| = \alpha < \frac{T}{2}, \quad (10)$$

$$\text{the sequence } \{b_i\}_{i \in \mathbb{Z}} \text{ is almost periodic, } \{b_i\}_{i \in \mathbb{Z}} \in (0, 1) \text{ and } \inf_{i \in \mathbb{Z}} |b_i| \geq b > 0.$$

Lemma 3. Under conditions (10), problem (1), (2) has an almost periodic infection-free solution $\bar{P}(t) = \{\bar{S}(t), 0, 0\}$ in the sense of Definition 3 and

$$\sup_{t \geq 0} \bar{S}(t) \rightarrow 0, \quad T \rightarrow 0.$$

Remark 3. It also follows from (Samoilenko, & Perestyuk, 1995, Theorem 24.1) that the almost periodic solution $\bar{S}(t)$ is asymptotically stable for problem (5)9, i.e., the limit equality (7) holds.

Remark 4. For the correctness of the definitions, we consider the impulsive perturbation on the entire real axis, but we are interested in the dynamics on the half-axis $[0, +\infty)$.

Proof of Lemma 3. Since the inequality

$$-(\nu + \delta) + \theta \ln \left(\sup_{i \in \mathbb{Z}} (1 - b_i) \right) \leq -(\nu + \delta) < 0$$

holds, where θ is determined from (8), then from (Samoilenko, & Perestyuk, 1995, Theorem 24.1) problem (5) has a unique almost periodic solution

$$\bar{S}(t) = (\nu + \delta) \int_{-\infty}^t X(t, s) ds, \quad (11)$$

where for $t_k \leq t < t_{k+1}$, $t_{m-1} \leq s < t_m$, $s < t$ we have

$$X(t, s) = e^{-(\nu + \delta)(t - t_k)} \prod_{s \leq t_i < t} (1 - b_i) e^{-(\nu + \delta)(t_i - t_{i-1})} (1 - b_m) e^{-(\nu + \delta)(t_m - s)}.$$

From formula (11) we deduce that $\bar{S}(t) \in (0, 1)$ for an arbitrary $t \in \mathbb{R}$.

For a more precise estimate of $\bar{S}(t)$, we note that from the existence of a uniform with respect to $t \in \mathbb{R}$ limits in (8) we have

$$\theta = \lim_{k \rightarrow \infty} \frac{N \left(0, kT + \frac{T}{2} \right)}{kT + \frac{T}{2}} = \lim_{k \rightarrow \infty} \frac{k + \min\{0, \text{sign } \alpha_0\}}{kT + \frac{T}{2}} = \frac{1}{T}.$$

Then for an arbitrary $\varepsilon > 0$ there exists such $K = K(\varepsilon) \geq 1$ that for all $t \geq s$ we have

$$X(t, s) \leq K e^{(\varepsilon - \nu - \delta + \frac{1}{T} \ln(1 - b))(t - s)},$$

and therefore (Samoilenko, & Perestyuk, 1995, Theorem 24.1) for $\varepsilon = \frac{\nu + \delta}{2}$ and for $K \geq 1$, which does not depend on T for sufficiently small T number $\beta = -\frac{\nu + \delta}{2} + \frac{1}{T} \ln(1 - b) < 0$ is negative and so for $t \geq 0$ we have

$$\bar{S}(t) \leq K(\nu + \delta) \cdot \left(\frac{1}{-\beta} \right) = \frac{K(\nu + \delta)}{\frac{\nu + \delta}{2} + \frac{1}{T} |\ln(1 - b)|},$$

therefore $\sup_{t \geq 0} \bar{S}(t) \rightarrow 0$, $T \rightarrow 0$. Lemma 3 is proven. $\square$

The main result of the paper is the following theorem.

Theorem 2. Let in problem (1), (2) impulsive perturbations satisfy conditions (3) or (10) and $\bar{P}(t) = \{\bar{S}(t), 0, 0\}$ is a corresponding infection-free impulsive almost periodic solution. Then there is such $T^* > 0$ that for $T < T^*$ the solution $\bar{P}(t)$ is globally (in $\text{Int } D$) asymptotically stable.

Proof of Theorem 2. With fixed $E(t)$ and $I(t)$, consider the impulsive problem for first equation of system (1).

$$\begin{cases} \dot{S} = -\lambda SI + (\nu + \delta)(1 - S) - \delta(E + I), \\ \Delta S|_{t=t_n} = -b_n S(t_n), \quad S|_{t=0} = S_0 \in (0, 1). \end{cases} \quad (12)$$

Since there are $E(t) \geq 0$, $I(t) \geq 0$, then for an arbitrary $t \geq 0$ we have

$$S(t) \leq X(t), \quad (13)$$

where $X(t)$ is the solution of (5) with the initial condition $X(0) = S_0$ and therefore, according to the remarks 1, 3, we have

$$X(t) - \bar{S}(t) \rightarrow 0, \quad t \rightarrow \infty. \quad (14)$$

Now, with a fixed $S(t)$, consider second and third equations of system (1)

$$\begin{cases} \dot{E} = \lambda SI - (\varepsilon + \nu)E, \\ \dot{I} = \varepsilon E - (\gamma + \nu)I, \\ E(0) = E_0 \in (0, 1), \quad I(0) = I_0 \in (0, 1). \end{cases} \quad (15)$$

Since system (15) is cooperative, that is, there are inequalities

$$\frac{\partial \dot{E}}{\partial I} = \lambda S(t) > 0, \quad \frac{\partial \dot{I}}{\partial E} = \varepsilon > 0,$$

hold, then according to Kamke's theorem, we have

$$0 \leq E(t) \leq Y(t), \quad 0 \leq I(t) \leq Z(t),$$

where $\{Y, Z\}$ are the solutions of the system

$$\begin{cases} \dot{Y} = \lambda XZ - (\varepsilon + \nu)Y, \\ \dot{Z} = \varepsilon Y - (\gamma + \nu)Z, \\ Y(0) = E_0, \quad Z(0) = I_0, \end{cases} \quad (16)$$

and $X(t)$ is satisfied by the limiting relation (14). So, for sufficiently large t , we have that

$$\frac{d}{dt} (Y(t) + Z(t)) = -\nu Y(t) - (\gamma + \nu - \lambda X(t))Z(t) \leq -\nu (Y(t) + Z(t)), \quad (17)$$

if T^* is chosen so that the inequality

$$\sup_{t \geq 0} \bar{S}(t) < \frac{\nu}{\lambda} \quad (18)$$

holds for $T < T^*$.

Thus, for arbitrary $t \geq 0$ we have

$$E(t) + I(t) \leq K(E_0 + I_0)e^{-\nu t}, \quad (19)$$

where $K > 0$ is some constant.

Let's go to the new variables x, y, z in (1), (2), where

$$\begin{aligned} S(t) &= \bar{S}(t) + x(t), \\ E(t) &= y(t), \\ I(t) &= z(t). \end{aligned}$$

Then the stability of $\bar{P}(t)$ is equal to the stability of the zero solution of the system

$$\begin{cases} \dot{x} = -\lambda \bar{S}(t)z - \lambda xz - \nu x + \delta(-x - y - z), \\ \dot{y} = \lambda \bar{S}(t)z + \lambda xz - (\varepsilon + \nu)y, \\ \dot{z} = \varepsilon y - (\gamma + \nu)z, \\ \Delta x|_{t=t_n} = -b_n x(t_n). \end{cases} \quad (20)$$

Under condition (18), for each $t \geq 0$ we have

$$y(t) + z(t) \leq (y_0 + z_0)e^{-\nu t}. \quad (21)$$

Further, on an arbitrary interval of continuity (τ, t) , from the first equation in (20) we obtain

$$\frac{1}{2} \frac{d}{dt} x^2(t) \leq -(\nu + \delta)x^2(t) - \delta x(t)y(t) - (\lambda \bar{S}(t) + \delta)x(t)z(t).$$

Choosing $T^* > 0$ so that for $T < T^*$ condition

$$\sup_{t \geq 0} \bar{S}(t) < \frac{\nu}{\lambda} \quad (22)$$

is fulfilled, we obtain

$$\frac{1}{2} \frac{d}{dt} x^2(t) \leq -(\nu + \delta)x^2(t) + (\nu + \delta)|x(t)|(y(t) + z(t)) \leq -\frac{\nu + \delta}{2}x^2(t) + \frac{\nu + \delta}{2}(y(t) + z(t))^2.$$

Thus, on an arbitrary interval of continuity (τ, t) , using (21), we derive

$$\begin{aligned} \frac{d}{dt} x^2(t) &\leq -(\nu + \delta)x^2(t) + (\nu + \delta)(y(t) + z(t))^2, \\ x^2(t) &\leq x^2(\tau)e^{-(\nu + \delta)(t - \tau)} + (\nu + \delta) \int_{\tau}^t (y_0 + z_0)^2 e^{-2\nu s} e^{-(\nu + \delta)(t - s)} ds \leq \\ &\leq x^2(\tau)e^{-(\nu + \delta)(t - \tau)} + \frac{\nu + \delta}{\delta} (y_0 + z_0)^2 e^{-\nu t}. \end{aligned} \quad (23)$$

Then

$$\begin{aligned} x^2(t_1) &\leq x_0^2 e^{-(\nu + \delta)t_1} + \frac{\nu + \delta}{\delta} (y_0 + z_0)^2 e^{-\nu t_1}, \\ x^2(t_2) &\leq x^2(t_1)(1 - b_1)^2 e^{-(\nu + \delta)(t_2 - t_1)} + \frac{\nu + \delta}{\delta} (y_0 + z_0)^2 e^{-\nu t_2} \leq \\ &\leq x_0^2 (1 - b_1)^2 e^{-(\nu + \delta)t_2} + \frac{\nu + \delta}{\delta} (y_0 + z_0)^2 ((1 - b_1)^2 + 1) e^{-\nu t_2}, \end{aligned}$$

continuing further at the k -th step, we obtain

$$x^2(t_k) \leq x_0^2 (1 - b)^{2k} e^{-(\nu + \delta)t_k} + \frac{\nu + \delta}{\delta(1 - b^2)} (y_0 + z_0)^2 e^{-\nu t_k}, \quad k \geq 1. \quad (24)$$

From estimates (24) we deduce that $x^2(t) \rightarrow 0, t \rightarrow \infty$, and this, together with (21), gives the desired asymptotic stability.

Theorem 2 is proven. $\square$

Discussion and conclusions

The article considers the SEIR system of differential equations with impulsive perturbations of an almost periodic nature. The existence and properties of an almost periodic impulsive solution without infections are investigated, and its asymptotic stability is also proven. From a practical point of view, this means that by choosing a certain realistic vaccination schedule (which takes into account possible deviations from a strictly periodic nature) and supporting vaccination of a certain proportion of the population, it is possible to achieve the stability of the infection-free state of the system, which will mean the attenuation of the epidemic process. The obtained results can also be applied to high-dimensional systems of differential equations, which are used for describing DDoS attacks in computer networks. The study generalizes previously known results by considering a more mathematically complex model that allows for a more realistic practical interpretation. It should be noted that the linear impulse perturbation condition (2) is due to empirical data on the cessation of the circulation of infections in the population upon reaching a sufficiently high level (80–90 %, and sometimes more than 95%) of vaccinated people in the population (see, for example, <https://ourworldindata.org/vaccination>). As a rule, achieving such levels requires mass vaccination, so we really neglect the throughput of the medical system, considering it proportional to the population size. In our further studies, we plan to consider the case of a nonlinear dependence of the magnitude of the impulse disturbance on S , which is a more realistic case from the point of view of the practical vaccination process, which depends on the capabilities of the system.

Authors' contribution: Olexiy Kapustyan – conceptualization; methodology, writing and editing; Dmytro Kapustyan – analysis of sources, preparation of literature review and technical support; Anna Sukretna – validation of results, writing, editing and technical support; Myrzagali Ospanov – preparation of theoretical foundations of research.

Sources of funding. This work is supported by the National Research Foundation of Ukraine (NRFU), Project 2023.03/0074 "Infinite-Dimensional Evolutionary Equations with Multivalued and Stochastic Dynamics".

References

- Hethcote, H. W. (1989). Three Basic Epidemiological Models. In S. A. Levin, T. G. Hallam, L. J. Gross (Eds.), *Applied Mathematical Ecology. Biomathematics*, vol 18. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-61317-3_5
- Liu, W. H., Hethcote, H. W., & Levin, S. A. (1987). Dynamical Behavior of Epidemiological Models with Nonlinear Incidence Rates. *Journal of Mathematical Biology*, 25, 359–380. <https://doi.org/10.1007/bf00277162>
- Balarezo, J. F., Wang, S., Chavez, K. G., Al-Hourani, A. & Kandeepan, S. (2022). A Survey on DOS/DDOS Attacks Mathematical Modelling for Traditional, SDN and Virtual Networks. *Engineering Science and Technology, an International Journal*, 31, Article ID: 101065. <https://doi.org/10.1016/j.jestch.2021.09.011>
- Li, M. Y., & Muldowney J. S. (1995). Global stability for the SEIR model in epidemiology. *Math. Biosci.*, 125, 155–168. [https://doi.org/10.1016/0025-5564\(95\)92756-5](https://doi.org/10.1016/0025-5564(95)92756-5)
- Li, M. Y., & Muldowney, J. S. (1996). A Geometric Approach to Global-Stability Problems. *SIAM Journal on Mathematical Analysis*, 27, 1070–1083. <http://dx.doi.org/10.1137/S0036141094266449>
- Samoilenko, A. M., & Perestyuk, N. A. (1995). *Impulsive Differential Equations*. World Scientific, Singapore. <https://doi.org/10.1142/2892>
- Shulgin, B., Stone L., & Agur, Z. (1998). Pulse Vaccination Strategy in the SIR Epidemic Model. *Bulletin of Mathematical Biology*, 60(6), 1123–1148. [https://doi.org/10.1016/S0895-7177\(02\)00177-2](https://doi.org/10.1016/S0895-7177(02)00177-2)
- D'Onofrio, A. (2002). Stability properties of pulse vaccination strategy in SEIR epidemic model. *Mathematical Biosciences*, 179(1), 57–72. [https://doi.org/10.1016/S0025-5564\(02\)00095-0](https://doi.org/10.1016/S0025-5564(02)00095-0)
- Feketa, P., Klinshov, V., & Lücken, L. (2021). A survey on the modelling of hybrid behaviors: How to account for impulsive jumps properly. *Communications in Nonlinear Science and Numerical Simulation*, 103, 105955. <https://doi.org/10.1016/j.cnsns.2021.105955>

Отримано редакцією журналу / Received: 29.12.25

Прорецензовано / Revised: 09.04.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Олексій КАПУСТЯН¹, д-р фіз.-мат. наук, проф.

ORCID ID: 0000-0002-9373-6812

e-mail: kapustyan@knu.ua

Анна СУКРЕТНА¹, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-2985-1250

e-mail: sukretna.a.v@knu.ua

Дмитро КАПУСТЯН¹, студ.

ORCID ID: 0009-0008-2992-4545

e-mail: dimonkapus@gmail.com

Мирзгалі ОСПАНОВ², канд. фіз.-мат. наук, доц.

ORCID ID: 0009-0009-3616-8511

e-mail: ospanov_mn@enu.kz

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Євразійський національний університет імені Л. Н. Гумільова, Астана, Республіка Казахстан

ПРО СТІЙКІСТЬ МАЙЖЕ ПЕРІОДИЧНОГО ВІЛЬНОГО ВІД ІНФЕКЦІЇ РОЗВ'ЯЗКУ В ІМПУЛЬСНІЙ SEIR-МОДЕЛІ

Математичні моделі динаміки розповсюдження інфекції є актуальним напрямком досліджень упродовж кількох останніх десятиліть. Ба більше, ці моделі показали себе придатними також для моделювання DDoS-атак у комп'ютерних мережах. Якісна поведінка розв'язків у таких моделях зазвичай улаштована так: існує деяке порогове значення, у разі досягнення якого глобальна асимптотична стійкість розв'язку з нульовою компонентою, що відповідає за кількість інфікованих осіб, змінюється на стійкість ненульового положення рівноваги. Зі змістовної позиції це означає, що самостійне вичерпання та зникнення інфекції змінюється на постійну циркуляцію інфекції в популяції.

Однак наявність імпульсного збурення може зробити вільний від інфекції розв'язок асимптотично стійким незалежно від порогового значення, причому такі імпульсні збурення допускають природну інтерпретацію у межах епідеміологічної моделі як процес вакцинації певної частини індивідумів.

Такі задачі вчені вже розглядали для SIR- і SEIR-моделей у випадку простої структури імпульсних точок, завдяки чому імпульсна задача фактично зводилася до задачі без імпульсного збурення. Проте складніша структура множини та величин імпульсних впливів вимагає застосування теорії імпульсних диференціальних рівнянь, що і зроблено у цій роботі. Зокрема, у статті розглянуто SEIR-систему диференціальних рівнянь з імпульсними збуреннями, що мають майже періодичний характер. Досліджено існування та властивості вільного від інфекції імпульсного майже періодичного розв'язку та доведено його асимптотичну стійкість.

Ключові слова: SEIR-модель диференціальних рівнянь, імпульсне збурення, вільний від інфекції розв'язок, періодичний розв'язок, майже періодичний розв'язок.

Один з авторів (О. Капустян) є членом редколегії видання, тому він не брав участі у рецензуванні та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

One of the authors (O. Kapustyan) is a member of the journal editorial board; he did not take part in the peer-review process or in the decision to publish this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 517.9

DOI: <https://doi.org/10.17721/1812-5409.2026/1.18>

Andrii LOVEIKIN, PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0002-7988-8350

email: andrii.loveikin@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Andrii KRENEVYCH, PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0002-0720-786X

e-mail: krenevych@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

SINGULARITY OF THE POTENTIAL FIELD AT THE VERTEX OF A SYMMETRIC TIP WITH FOUR IDENTICAL PERPENDICULAR WINGS OF WEDGE-SHAPED FORM

The study addresses a spatial problem of the potential theory outside a symmetric tip with four identical perpendicular wings of wedge-shaped form. The main objective is to determine the singularity of the solution at the tip's vertex.

By utilizing the symmetric shape of the tip, the problem is reduced to finding a harmonic function in a quarter-space that satisfies mixed boundary conditions on its sides.

The homogeneous solutions of the problem are expressed in spherical coordinates as a Mehler – Fock-type integral involving Legendre functions. Using the method of integral transforms, the problem is reduced to a Wiener – Hopf-type functional equation, which is transformed into an infinite system of linear algebraic equations by the Wiener – Hopf method. The solutions of this system allow for the derivation of analytical representations for the homogeneous solutions of the original boundary-value problem. The constructed homogeneous solutions enable the determination of the singularity of the potential field at the tip's vertex.

As a result of solving the given problem, a relationship between the singularity exponent and the deviation angle of the tip's wing edges has been established.

The constructed homogeneous solutions enable the determination of the nature of the potential field singularity at the tip's vertex and establish a quantitative relationship between the singularity exponent and the deviation angle of the tip's wing edges. The obtained results provide an assessment of the qualitative behavior of the potential field at the tip's vertex, as well as determine the quantitative characteristics describing this behavior. It makes it possible to find correct solutions of boundary-value problems in complex domains with corner points with the same geometry, especially when numerical methods are used.

Key words: *potential theory, mixed boundary-value problem, solution singularity, integral transformation method, Wiener – Hopf technique.*

AMS 2020 classification: 31B20, 31B25.

Introduction

Studying physical fields, defining their qualitative and quantitative characteristics, and describing technical and technological processes often involve the analysis of boundary-value problems of mathematical physics in domains with complex shapes. When a field or process is studied in a domain or body with irregular inclusions, or when its boundary contains edges, vertices, or other singular points, the field or process has some distributional singularities in the regions surrounding them, which consequently affect its qualitative and quantitative characteristics. Therefore, identifying the peculiarities of field behavior is an important task from both qualitative and quantitative perspectives with respect to the field or process under study. Furthermore, the identification of these peculiarities makes it possible to improve existing numerical algorithms for solving problems of mathematical physics and to develop new, more advanced ones.

One of the approaches to constructing solutions for problems in domains with internal inclusions and boundary irregularities is based on the use of integral transforms and Wiener–Hopf methods. This approach is quite effective for solving problems in domains of complex geometry. It became especially popular in two-dimensional problems involving the propagation of waves of different types. In particular, the work (Nethercote, Kisil, & Assier, 2022) considers the problem of acoustic wave diffraction on an array of point scatterers. The works (Daniele, & Lombardi, 2021) and (Kuryliak, & Lysechko, 2022) are devoted to the construction of solutions to problems involving the propagation of electromagnetic waves in two-dimensional wedge-shaped domains with boundary conditions of various types. It is also worth paying attention to the works (Nagasaka, Kishiguchi, & Kobayashi, 2022) and (He, & Kobayashi, 2023), in which the authors, using the Wiener – Hopf method, solve two-dimensional problems of electromagnetic wave diffraction on a semi-axis with impedance-type boundary conditions and in a waveguide with a piecewise homogeneous filling. Besides wave problems, the Wiener – Hopf method also makes it possible to solve two-dimensional problems in potential theory and elasticity theory for domains with corner points on the boundary and for domains containing inclusions of irregular geometry (Daniele, & Lombardi, 2022; Loveikin, 2022; 2023). In problems with complex geometry, constructing the exact solutions requires solving the vector Wiener – Hopf equations. The author of the works (Idemen, 1979) presents a method for obtaining the exact solution of such vector equations in some cases. In the works (Kurilyak, 1998; Kurilyak, 1999), vector Wiener – Hopf equations make it possible to construct the solutions of diffraction problems for wedge domains with different boundary conditions.

The present work addresses a spatial problem concerning the singular behavior of a potential field at the vertex of an inclusion shaped as a compound tip composed of four identical sectoral plates (wings), which share a common edge and are arranged in perpendicular planes (see Fig. 1). Locally, at the tip's vertex, the problem exhibits a purely three-dimensional character. Although the nature of the singularity along the wings can be determined using results from two-dimensional potential theory, it is not feasible to do so at the vertex of the tip. Thus, the three-dimensional problem must be solved. A similar conclusion was drawn by Maurizi M. and Berto F. (2020), who investigated spatial singularities in problems of elasticity theory for domains with corner points.

© Loveikin Andrii, Krenevykh Andrii, 2026

1. Problem formulation

We consider an inclusion shaped as a tip (or spike), with its wings lying in the Oxz i Oyz planes (see Fig. 1). Their common edge lies along the positive semi-axis Oz_+ , while other edges deviate from vertical axis by an angle α ($0 < \alpha \leq \frac{\pi}{2}$).

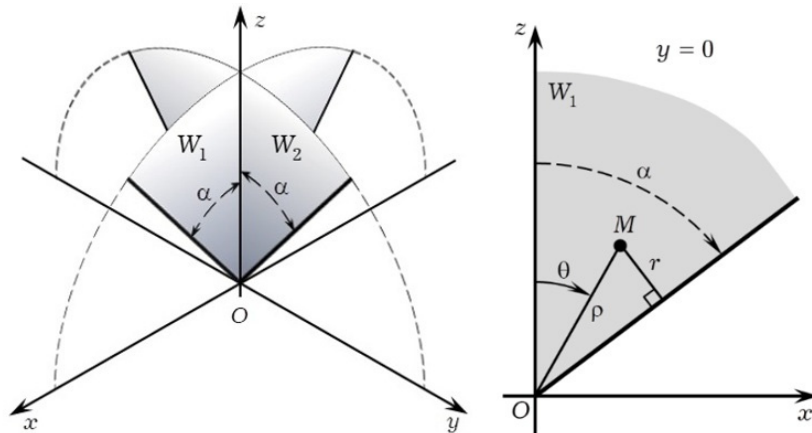


Fig. 1. Form of the tip and wing geometry

To identify the singular behavior of the potential field – electrostatic, for instance – we seek homogeneous solutions of the problem, defined as nontrivial solutions of the corresponding homogeneous boundary-value problem.

$$\Delta u = 0, (x, y, z) \in \mathbb{R}^3 \setminus W, \quad u|_W = 0, \quad (1)$$

where W is the union of all four wings. From a geometrical point of view, the problem is symmetric with respect to the x and y coordinates. Hence, we seek its solution in the coordinate quarter where $x > 0, y > 0$. The main singularity of the behavior of the potential field is determined by the solutions that are even with respect to both the x and y coordinates. For such solutions, the boundary-value problem takes the form:

$$\Delta u = 0, x > 0, y > 0, \quad (2)$$

$$u|_{y=0} = 0, (x, z) \in \overline{W}_1, \quad u|_{x=0} = 0, (y, z) \in \overline{W}_2, \quad (3)$$

$$\frac{\partial u}{\partial y} \Big|_{y=0} = 0, (x, z) \notin \overline{W}_1, \quad \frac{\partial u}{\partial x} \Big|_{x=0} = 0, (y, z) \notin \overline{W}_2, \quad (4)$$

where $W_{1,2}$ are two-dimensional open domains representing the wings lying in the half-planes Ox_+z and Oy_+z respectively (see Fig. 1). Conditions (4) express the evenness of the homogeneous solutions with respect to the y and x coordinates.

Taking into account the geometry of the inclusion, it is more convenient to solve problem (2)–(4) in the spherical coordinate system (ρ, θ, φ) , in which it takes the following form:

$$\Delta u = 0, \rho > 0, 0 < \theta < \pi, 0 < \varphi < \frac{\pi}{2}, \quad (5)$$

$$u|_{\varphi=0} = 0, \rho > 0, 0 < \theta \leq \alpha, \quad \frac{\partial u}{\partial \varphi} \Big|_{\varphi=0} = 0, \rho > 0, \alpha < \theta < \pi. \quad (6)$$

$$u|_{\varphi=\frac{\pi}{2}} = 0, \rho > 0, 0 < \theta \leq \alpha, \quad \frac{\partial u}{\partial \varphi} \Big|_{\varphi=\frac{\pi}{2}} = 0, \rho > 0, \alpha < \theta < \pi. \quad (7)$$

Boundary conditions (6) and (7) are symmetric with respect to the bisectoral half-plane $\varphi = \frac{\pi}{4}$, which divides the domain of the problem (5)–(7) into two symmetric subdomains. As a result, the homogeneous solutions split into two series – even and odd with respect to the half-plane. The main feature of the singular behavior of the potential field at the vertex of the tip is determined by finding even solutions that must satisfy the following condition:

$$\frac{\partial u}{\partial \varphi} \Big|_{\varphi=\frac{\pi}{4}} = 0. \quad (8)$$

The present article is devoted to the construction of such solutions.

2. Construction of homogeneous solutions.

We seek nontrivial solutions of problem (5)–(7), which satisfy condition (8), in the form of a Mehler – Fock-type integral with respect to coordinate θ (Parfenenko, & Ulitko, 1988).

$$u = \rho^{s-\frac{1}{2}} \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \mu A_s(\mu) \cos\left(\mu\left(\frac{\pi}{4} - \varphi\right)\right) G_s(\mu) P_{s-\frac{1}{2}}^{-\mu}(\cos \theta) d\mu, \rho > 0, 0 < \theta < \pi, 0 \leq \varphi \leq \frac{\pi}{2}, \quad (9)$$

where $A_s(\mu)$ is an unknown coefficient, $G_s(\mu) = \Gamma\left(\frac{1}{2} - s + \mu\right) \Gamma\left(\frac{1}{2} + s + \mu\right)$, $\Gamma(\xi)$ – Euler's gamma function, $P_v^{-\mu}(t)$ is the Legendre function of the first kind (Bateman, & Erdelyi, 1953). The parameter s is assumed to be a purely imaginary number.

The equation (9) can be viewed as Mellin transformation of the desired harmonic function up to a factor $\rho^{s-\frac{1}{2}}$.

For the kernel of the Mehler – Fock-type integral it's possible to obtain the following asymptotic (Bateman, & Erdelyi, 1953)

$$\mu G_s(\mu) P_{s-\frac{1}{2}}^{-\mu}(\cos \theta) \Big|_{\mu \rightarrow \pm i\infty} \sim \Gamma(1 + \mu) \left(\tan \frac{\theta}{2}\right)^\mu, \mu \rightarrow \pm i\infty, s - \text{fixed number};$$

$$\left| \Gamma(1 + \mu) \left(\tan \frac{\theta}{2} \right)^\mu \right|_{\mu=i\tau} = |\Gamma(1 + i\tau)| \sim \sqrt{\tau} e^{-\frac{\pi}{2}\tau}, \quad \tau \rightarrow \pm\infty.$$

With this choice of the function u , condition (8) is automatically satisfied.

Therefore, to determine the unknown coefficient $A_s(\mu)$, we substitute expression (9) into the mixed boundary conditions (6), which are extended over the entire interval $0 < \theta < \pi$ by introducing the unknown functions as follows:

$$\begin{aligned} \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \mu A_s(\mu) \cos \frac{\pi\mu}{4} G_s(\mu) P_{s-\frac{1}{2}}^{-\mu}(\cos \theta) d\mu &= \begin{cases} 0, & 0 < \theta \leq \alpha, \\ v(\theta), & \alpha < \theta < \pi, \end{cases} \\ \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \mu^2 A_s(\mu) \sin \frac{\pi\mu}{4} G_s(\mu) P_{s-\frac{1}{2}}^{-\mu}(\cos \theta) d\mu &= \begin{cases} \sigma(\theta) \sin \theta, & 0 < \theta < \alpha, \\ 0, & \alpha < \theta < \pi. \end{cases} \end{aligned} \quad (10)$$

The unknown function v defines the values of the sought harmonic function u on the extension of the wings, while the unknown function σ represents the normal derivative of u on their surfaces.

We apply direct Mehler – Fock-type transform formulas to equalities (10) (Parfenenko, & Ulitko, 1988).

Using the formula for the Legendre function with a negative argument and expressing these functions in terms of the hypergeometric function (Bateman, & Erdelyi, 1953), we obtain:

$$\begin{aligned} \mu A_s(\mu) \cos \frac{\pi\mu}{4} &= \frac{a^{-\mu} X_s(\mu)}{\Gamma(1 + \mu)}, \\ \mu A_s(\mu) \sin \frac{\pi\mu}{4} &= \frac{1}{\sin(\pi\mu)} \left(-\cos(\pi s) \frac{a^\mu Y_s(\mu)}{\Gamma(1 + \mu)} + \frac{\pi}{G_s(\mu)} \frac{a^{-\mu} Y_s(-\mu)}{\Gamma(1 - \mu)} \right), \quad |Re \mu| < \frac{1}{2}, \end{aligned} \quad (11)$$

where $X_s(\mu), Y_s(\mu)$ are unknown functions of a complex argument, defined in terms of v and σ as follows:

$$\begin{aligned} X_s(\mu) &= \mu \int_0^{\pi-\alpha} \frac{v(\pi-t)}{\sin t} \left(a \cdot \tan \frac{t}{2} \right)^\mu F\left(\frac{1}{2} - s, \frac{1}{2} + s, 1 + \mu, \sin^2 \frac{t}{2}\right) dt, \\ Y_s(\mu) &= \int_0^\alpha \sigma(t) \left(\frac{1}{a} \tan \frac{t}{2} \right)^\mu F\left(\frac{1}{2} - s, \frac{1}{2} + s, 1 + \mu, \sin^2 \frac{t}{2}\right) dt, \end{aligned}$$

$F(\cdot, \cdot, \cdot)$ denotes the hypergeometric function, $a = \tan \frac{\alpha}{2}$ ($0 < a \leq 1$).

Taking into account the root singularity of the harmonic function u along the wing edge and applying Watson's lemma (Olver, 1974), we obtain the following for the unknown functions of a complex argument: $X_s(\mu)$ and $Y_s(\mu)$ are analytic in the half-plane $Re \mu > -\frac{1}{2}$, as functions of the argument μ for each fixed value s , and behave like $Const \cdot \mu^{-\frac{1}{2}}$ as μ tends to infinity.

Excluding the unknown $A_s(\mu)$ from the equalities (11), we obtain the Wiener–Hopf-type equations (Noble, 1958) with respect to the analytic functions $X_s(\mu), Y_s(\mu)$

$$\frac{G_s(\mu)}{\Gamma(1 + \mu)} \left(X_s(\mu) + \frac{\cos(\pi s) a^{2\mu} Y_s(\mu)}{4 \cos \frac{\pi\mu}{2} \sin^2 \frac{\pi\mu}{4}} \right) = \frac{\pi}{\sin(\pi\mu)} \frac{\cos \frac{\pi\mu}{4}}{\sin \frac{\pi\mu}{4}} \frac{Y_s(-\mu)}{\Gamma(1 - \mu)}, \quad |Re \mu| < \frac{1}{2}. \quad (12)$$

Factorization of the equation coefficient (12) (Bateman, & Erdelyi, 1953)

$$\frac{\pi}{\sin(\pi\mu)} \frac{\cos \frac{\pi\mu}{4}}{\sin \frac{\pi\mu}{4}} = -4\Gamma(\mu)\Gamma(-\mu) \frac{\Gamma\left(1 + \frac{\mu}{4}\right)\Gamma\left(1 - \frac{\mu}{4}\right)}{\Gamma\left(\frac{1}{2} + \frac{\mu}{4}\right)\Gamma\left(\frac{1}{2} - \frac{\mu}{4}\right)}.$$

Applying factorization and separately collecting the functions that are analytic for $Re \mu > -\frac{1}{2}$ and $Re \mu < \frac{1}{2}$, we rewrite equation (12) in the following form:

$$\Phi_s(\mu) + \frac{\cos(\pi s) a^{2\mu}}{4 \cos \frac{\pi\mu}{2} \sin^2 \frac{\pi\mu}{4}} \Psi_s(\mu) = -K_s(-\mu) \Psi_s(-\mu), \quad |Re \mu| < \frac{1}{2}, \quad (13)$$

where, for simplicity of reasoning, we have introduced the following notation for the unknown functions:

$$\begin{pmatrix} \Phi_s(\mu) \\ \Psi_s(\mu) \end{pmatrix} = \frac{\Gamma\left(\frac{1}{2} + \frac{\mu}{4}\right)}{\Gamma\left(1 + \frac{\mu}{4}\right)} \frac{\mu G_s(\mu)}{\Gamma^2(1 + \mu)} \begin{pmatrix} X_s(\mu) \\ Y_s(\mu) \end{pmatrix} \quad (14)$$

and the coefficient:

$$K_s(\mu) = \frac{4\Gamma^2\left(1 + \frac{\mu}{4}\right)\Gamma^2(\mu)}{\Gamma^2\left(\frac{1}{2} + \frac{\mu}{4}\right)G_s(\mu)}.$$

The functions $\Phi_s(\mu), \Psi_s(\mu)$ are analytic in the half-plane $Re \mu > -\frac{1}{2}$, as functions of the variable μ for each fixed value of s . As $\mu \rightarrow \infty$, they behave like $Const \cdot \mu^{-\frac{1}{2}}$. At the point $\mu = 0$, the functions $\Phi_s(\mu)$ and $\Psi_s(\mu)$ have zeros of at least first order. The coefficient $K_s(\mu)$ has a pole of second order at the point $\mu = 0$, and $K_s(\mu) \rightarrow 1$ as $\mu \rightarrow \infty$.

After factorization, the second term in the left-hand side of equation (13) has a simple pole at $\mu = 0$, which lies in the strip $|Re \mu| < \frac{1}{2}$. This pole will be compensated by the same pole of the function $K_s(-\mu)\Psi_s(-\mu)$ in the right-hand side.

The right-hand side of equation (13) is analytic in the half-plane $Re \mu < \frac{1}{2}$, except for a simple pole at the point $\mu = 0$. The first term on the left-hand side of the equation is analytic for $Re \mu > -\frac{1}{2}$. The second term has simple poles at $\mu = 1 + 2k$, $k = 0, 1, \dots$, second-order poles at $\mu = 4j$, $j = 1, 2, \dots$, and simple pole at $\mu = 0$.

By eliminating these singularities, we can rewrite equation (13) in the following form:

$$\begin{aligned} \Phi_s(\mu) + \frac{\cos(\pi s) a^{2\mu}}{4 \cos \frac{\pi \mu}{2} \sin^2 \frac{\pi \mu}{4}} \Psi_s(\mu) + \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{\mu - 1 - 2k} x_k - \\ - \frac{4}{\pi^2} \cos(\pi s) \left[\frac{C_0}{\mu} + \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{(\mu - 4j)^2} + \frac{y_j \ln a^2 + z_j}{\mu - 4j} \right) \right] = -K_s(-\mu) \Psi_s(-\mu) + \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{\mu - 1 - 2k} x_k - \\ - \frac{4}{\pi^2} \cos(\pi s) \left[\frac{C_0}{\mu} + \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{(\mu - 4j)^2} + \frac{y_j \ln a^2 + z_j}{\mu - 4j} \right) \right], |Re \mu| < \frac{1}{2}, \end{aligned} \quad (15)$$

where

$$C_0 = \Psi'_s(0), \quad x_k = \Psi_s(1 + 2k), k = 0, 1, \dots, \quad y_j = \Psi_s(4j), \quad z_j = \Psi'_s(4j), j = 1, 2, \dots, \quad (16)$$

and $x_k \sim \frac{\text{Const}}{1+2k}$, $k \rightarrow +\infty$, $y_j \sim \frac{\text{Const}}{j}$, $z_j \sim \frac{\text{Const}}{j^2}$, $j \rightarrow +\infty$.

In the present work, we consider the case where the deviation angle of the tip blades α is bounded by $0 < \alpha \leq \frac{\pi}{2}$. Under these restrictions, the parameter $a = \tan \frac{\alpha}{2}$ belongs to the interval $(0, 1]$, which ensures the convergence of the series in equation (15).

In equality (15), the additional terms were chosen so that the left-hand side is analytic in the half-plane $Re \mu > -\frac{1}{2}$, and the right-hand side is analytic in $Re \mu < \frac{1}{2}$, respectively. As $\mu \rightarrow \infty$, both sides of equation (15) tend to zero as $\text{Const} \cdot \mu^{-1}$. Under these conditions, according to the Wiener–Hopf method (Noble, 1958) both the left-hand and right-hand sides of equation (15) reduce into the zero function. As a result, to determine the unknown analytic functions $\Phi_s(\mu)$, $\Psi_s(\mu)$ we write the following:

$$\begin{aligned} \Phi_s(\mu) + \frac{\cos(\pi s) a^{2\mu}}{4 \cos \frac{\pi \mu}{2} \sin^2 \frac{\pi \mu}{4}} \Psi_s(\mu) + \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{\mu - 1 - 2k} x_k - \\ - \frac{4}{\pi^2} \cos(\pi s) \left[\frac{C_0}{\mu} + \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{(\mu - 4j)^2} + \frac{y_j \ln a^2 + z_j}{\mu - 4j} \right) \right] = 0, \quad Re \mu > -\frac{1}{2}; \end{aligned} \quad (17)$$

$$K_s(-\mu) \Psi_s(-\mu) + \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{\mu + 1 + 2k} x_k + \frac{4}{\pi^2} \cos(\pi s) \left[\frac{C_0}{\mu} + \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{(\mu + 4j)^2} - \frac{y_j \ln a^2 + z_j}{\mu + 4j} \right) \right] = 0, \quad Re \mu > -\frac{1}{2}. \quad (18)$$

For the complete determination of the analytic functions $\Phi_s(\mu)$ and $\Psi_s(\mu)$ we must find the unknown quantities C_0, x_k , $k = 0, 1, \dots$, y_j, z_j , $j = 1, 2, \dots$, which appear in expressions (17) and (18). To accomplish this, taking into account expression (16), we consecutively substitute $\mu = 1 + 2m$, $m = 0, 1, \dots$ into equality (18). We obtain equations that determine x_k , for $k = 0, 1, \dots$,

$$\begin{aligned} K_s(1 + 2m) x_m + \frac{\cos(\pi s)}{2\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{1 + m + k} x_k - \\ - \frac{4}{\pi^2} \cos(\pi s) \left[\frac{C_0}{1 + 2m} - \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{(1 + 2m + 4j)^2} - \frac{y_j \ln a^2 + z_j}{1 + 2m + 4j} \right) \right] = 0, \quad m = 0, 1, 2, \dots \end{aligned} \quad (19)$$

After that, we consecutively substitute $\mu = 4n$, $n = 1, 2, \dots$, into (18). As a result, we obtain equations for determining y_j , $j = 1, 2, \dots$,

$$K_s(4n) y_n + \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{4n + 1 + 2k} x_k - \frac{\cos(\pi s)}{\pi^2} \left[\frac{C_0}{n} - \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{4(n + j)^2} - \frac{y_j \ln a^2 + z_j}{n + j} \right) \right] = 0, \quad n = 1, 2, \dots \quad (20)$$

Finally, we differentiate expression (18) and consecutively substitute $\mu = 4n$, $n = 1, 2, \dots$, into the resulting equality. Hence, we obtain equations that define the unknown values z_j for $j = 1, 2, \dots$,

$$\begin{aligned} K_s(4n) \left(z_n + \left\{ \frac{1}{2} \left(\psi(n) - \psi\left(\frac{1}{2} + n\right) \right) + 2\psi(1 + 4n) - \psi\left(\frac{1}{2} - s + 4n\right) - \psi\left(\frac{1}{2} - s + 4n\right) \right\} y_n \right) - \\ - \frac{\cos(\pi s)}{\pi} \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{(4n + 1 + 2k)^2} x_k + \frac{\cos(\pi s)}{4\pi^2} \left[\frac{C_0}{n^2} - \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{2(n + j)^3} - \frac{y_j \ln a^2 + z_j}{(n + j)^2} \right) \right] = 0, \quad n = 1, 2, \dots, \end{aligned} \quad (21)$$

Here, $\psi(t)$ denotes the logarithmic derivative of the gamma function (Bateman, & Erdelyi, 1953).

Equations (19)–(21) do not close the infinite system with respect to all unknown values, as they don't include an equation for determining coefficient C_0 . To obtain the final missing equation, we express the function $\Psi_s(\mu)$ from equality (18). The resulting expression is then substituted into equation (17), and we take the limit as $\mu \rightarrow 0$. Thus, we obtain the last equation of the system.

$$\left[\cos(\pi s) \left(\ln a^2 + \frac{1}{2} \psi\left(\frac{1}{2}\right) + 2\psi\left(\frac{1}{2} + s\right) - \frac{5}{2} \psi(1) \right) - \pi \sin(\pi s) \right] C_0 - \quad (22)$$

$$-\frac{\pi}{2} \cos(\pi s) \sum_{k=0}^{+\infty} \frac{(-1)^k a^{2(1+2k)}}{1+2k} x_k - \frac{\cos(\pi s)}{4} \sum_{j=1}^{+\infty} a^{8j} \left(\frac{y_j}{2j^2} - 2 \frac{y_j \ln a^2 + z_j}{j} \right) = 0.$$

As a result of the solution, the Wiener – Hopf-type equation (17) has been reduced to an infinite system of linear algebraic equations (19) – (22). In the case $0 < a = \tan \frac{\alpha}{2} < 1$ ($0 < \alpha < \frac{\pi}{2}$), the resulting system is quasi-regular with meromorphic coefficients for all complex values of s . In the case $a = \tan \frac{\alpha}{2} = 1$ ($\alpha = \frac{\pi}{2}$), the reduction to the quasi-regular system (19)–(22) can be achieved by the change of variables: $x_k = \frac{\tilde{x}_k}{\sqrt{1+2k}}$, $k = 0, 1, \dots$, $y_j = \frac{\tilde{y}_j}{\sqrt{j}}$, $z_j = \frac{\tilde{z}_j}{\sqrt{j}}$, $j = 1, 2, \dots$. The regularity of the infinite system (19)–(22) allows us to analytically continue it with respect to the parameter s from the imaginary axis to the entire complex plane. For the analysis of the solutions of this system and their numerical approximation (if needed), the method of reduction to the finite system can be applied.

4. Explicit form of the homogeneous solution and analysis of the field singularity at the tip vertex.

For the final construction of the homogeneous solutions, we write their representation in terms of the solutions of the infinite system of linear algebraic equations (19)–(22). For this purpose, we determine the coefficient $A_s(\mu)$ from the first equality in (11) and substitute it into expression (9). Taking into account equality (14), we obtain:

$$u = \frac{\rho^{s-\frac{1}{2}}}{2\pi i} \int_{-i\infty}^{+i\infty} a^{-\mu} \Phi_s(\mu) \Gamma(\mu) \frac{\Gamma\left(\frac{1}{2} + \frac{\mu}{4}\right) \cos\left(\mu\left(\frac{\pi}{4} - \varphi\right)\right)}{\Gamma\left(1 + \frac{\mu}{4}\right) \cos\frac{\pi\mu}{4}} P_{s-\frac{1}{2}}^{-\mu}(\cos \theta) d\mu, \quad \rho > 0, 0 < \theta < \pi, 0 \leq \varphi \leq \frac{\pi}{2}. \quad (23)$$

In the case $0 < \theta \leq \alpha$, the integral (23) can be evaluated using the Residue theorem. If we close the integration contour in the half-plane $\operatorname{Re} \mu > 0$ and compute sum of residues at all simple poles $\mu = 2 + 4l$, $l = 0, 1, \dots$, of the integrand in this half-plane, then we obtain:

$$u = \frac{4}{\pi} \rho^{s-\frac{1}{2}} \sum_{l=0}^{+\infty} \Phi_s(2+4l) \frac{\Gamma(1+l) \Gamma(2+4l)}{\Gamma\left(\frac{3}{2}+l\right) a^{2+4l}} P_{s-\frac{1}{2}}^{-2-4l}(\cos \theta) \sin((2+4l)\varphi), \quad \rho > 0, 0 < \theta \leq \alpha, 0 \leq \varphi \leq \frac{\pi}{2}. \quad (24)$$

From equalities (17), (18) it follows that the values $\Phi_s(2+4l)$, $l = 0, 1, \dots$, are expressed linearly in terms of the solutions of the infinite system (19)–(22): $C_0, x_k, k = 0, 1, \dots, y_j, z_j, j = 1, 2, \dots$. Thus, the explicit form of the sought homogeneous solutions of the boundary-value problem (5) – (8) is completely defined by the solutions of the system (19)–(22).

A similar representation can also be obtained for the homogeneous solutions in the interval $\alpha < \theta < \pi$.

Up to this point, we have assumed that the parameter s is purely imaginary. Since both the infinite system (19)–(22) and its solutions can be analytically continued from the imaginary axis to the entire complex plane, representation (24) remains valid for the sought homogeneous solutions for all arbitrary complex values of s .

The system (19)–(22) is a linear homogeneous system. Therefore, it has the obvious trivial solution for all complex values of s . However, to determine the homogeneous solutions of the boundary-value problem (5)–(8), this system must have a nontrivial solution, which is only possible for certain values of the parameter s . These specific values of s determine the final form of the homogeneous solutions of problem (5)–(8). Since the infinite system of linear algebraic equations (19)–(22) is quasi-regular for all complex values of s , it can be replaced by a reduced (finite) system for the approximate determination of the possible values of s . The reduced system will have nontrivial solutions if and only if its determinant is equal to zero. The numerical analysis shows that the parameter s can only take real values symmetric with respect to zero.

The singularity of the potential field at the tip vertex at the point O (see Fig. 1) is determined by positive values of s . If we consider only the homogeneous solution corresponding to the smallest positive value $s = s_1$, then, taking into account expression (24), the derivative of the potential on blade W_1 (see Fig. 1) along the normal $\vec{n}$ can be written as follows:

$$\left. \frac{\partial u}{\partial \vec{n}} \right|_{W_1} = \frac{1}{\rho \sin \theta} \left. \frac{\partial u}{\partial \varphi} \right|_{\varphi=0} = \frac{4}{\pi} \rho^{\gamma_1} \sum_{l=0}^{\infty} \frac{\Phi_{s_1}(2+4l)}{\sin \theta} \frac{\Gamma(1+l) \Gamma(3+4l)}{\Gamma\left(\frac{3}{2}+l\right) a^{2+4l}} P_{s_1-\frac{1}{2}}^{-2-4l}(\cos \theta), \quad \rho > 0, 0 < \theta < \alpha, \quad (25)$$

where $\gamma_1 = s_1 - \frac{3}{2}$.

It follows from (17) and (18) that the values $\Phi_{s_1}(2+4l)$, $l = 0, 1, 2, \dots$, satisfy the following asymptotic expressions:

$$\Phi_{s_1}(2+4l) \sim \frac{F_1}{2+4l}, \quad l \rightarrow \infty, \quad F_1 = \text{Const.}$$

Using it together with the properties of Gamma function and Legendre function $P_v^{-\mu}(\cos \theta)$ (Bateman, & Erdelyi, 1953), expression (25) could be rewritten as follows:

$$\left. \frac{\partial u}{\partial \vec{n}} \right|_{W_1} = \rho^{\gamma_1} \left(\frac{Q_1 \sin \theta}{\sqrt{\sin(\alpha - \theta)}} + w_1(\theta) \right), \quad \rho > 0, 0 < \theta < \alpha, \quad (26)$$

where $Q_1 = \text{Const}$, and $w_1(\theta)$ denotes the sum of a uniformly convergent series on $0 \leq \theta \leq \alpha$ and, therefore, is a continuous function.

The following Tab. 1 represents approximate values of s_1 and the singularity exponent depending on the angle deviation α of the blade edges.

Analyzing equality (26) and numerical results presented in the table, we can conclude that as the tip vertex (point O , see Fig. 1) is approached while $\rho \rightarrow 0$, and $\theta \in (0, \alpha)$ is a fixed value, the derivative of the potential along the normal grows like

$$\left. \frac{\partial u}{\partial \vec{n}} \right|_{W_1} \sim K_0(\theta) \rho^{\gamma_1}, \quad \rho \rightarrow 0+.$$

The obtained result shows that the potential field has a power-law singularity at the tip vertex, with the exponent γ_1 depending on the deviation angle α of the wing edges. For sufficiently sharp deviation angles of the edges $0 < \alpha < \alpha_*$

$\approx 0.36894 \pi$ the singularity exponent satisfies the estimate $\gamma_1 < -\frac{1}{2}$. This means that the power-law singularity at the tip vertex is stronger than the classical power-law singularity that occurs along the wing edges. In the case $\alpha = \alpha_*$, the classical square-root singularity occurs at the tip vertex, while for $\alpha_* < \alpha \leq \frac{\pi}{2}$ the singularity at the vertex becomes weaker than that along the wing edges.

Table 1

Approximate values of s_1 – the smallest positive zero of the determinant of the reduced system (19)–(22) and the singularity exponent γ_1 of the solution's normal derivative

α/π	s_1	γ_1	α/π	s_1	γ_1	α/π	s_1	γ_1
0.01	0.61075	– 0.88926	0.2	0.81605	– 0.68395	0.36894	1	– 0.5
0.05	0.671	– 0.82901	0.25	0.8651	– 0.63493	0.4	1.0416	– 0.45836
0.1	0.7224	– 0.77761	0.3	0.91795	– 0.58209	0.45	1.1168	– 0.38332
0.15	0.76925	– 0.73079	0.35	0.97615	– 0.52385	0.5	1.2103	– 0.29568

Based on equality (26), we can analyze the behavior of the square-root singularity intensity coefficient that arise along the wing edges. If we fix the value of ρ and let θ approach $\alpha -$, then from (26) we obtain

$$\left. \frac{\partial u}{\partial n} \right|_{w_1} \sim \frac{K_1(\rho)}{\sqrt{r}}, \theta \rightarrow \alpha - ,$$

where $r = \rho \sin(\alpha - \theta)$ is the distance to the edge (see Fig.1), and the following equality holds for the intensity coefficient of the singularity:

$$K_1(\rho) = \rho^{\gamma_1 + \frac{1}{2}} Q_1 \sin \alpha, \quad \rho > 0.$$

From the obtained equality for the coefficient K_1 and numerical results presented in the table, we conclude that for sharp angles $0 < \alpha < \alpha_* \approx 0.36894 \pi$, the intensity coefficient of the singularity on the edge tends to infinity as the tip vertex O is approached along the edge. This is due to the fact that the singularity at the tip vertex is stronger than those along the edges. In the case $\alpha = \alpha_*$, when the classical singularity occurs at point O , the coefficient K_1 remains bounded while approaching this point. On the other hand, if $\alpha_* < \alpha \leq \frac{\pi}{2}$, the intensity coefficient tends to zero as the tip vertex O is approached, since for such values of the angle α , the singularity at the vertex is weaker than that along the wing edges.

Discussion and conclusions

In the present work, an exact solution to the three-dimensional problem of potential theory was obtained for determining the singularity of the potential field at the tip vertex formed by four mutually perpendicular wedge-shaped blades sharing a common edge.

To construct the homogeneous solutions of the problem that make it possible to identify the singularity, spherical coordinates are used together with Mehler–Fock-type integral transforms involving Legendre functions.

The problem was reduced to a Wiener–Hopf-type functional equation, which, using the Wiener–Hopf method, was further reduced to a quasi-completely regular infinite system of linear algebraic equations. The analysis of the obtained infinite system allowed us to determinate the explicit form of the homogeneous solutions of the problem.

The constructed homogeneous solutions make it possible to determine and analyze the singularity of the potential field at the tip vertex. It has been established that the behavior of the potential field near the tip vertex has a complex three-dimensional character, and the singularity of this behavior can only be determined by solving the spatial problem considered in the present work.

Authors' contribution: Andrii, Loveikin – conceptualization; methodology; analysis of results; Andrii, Krenevych – software; numerical analysis, preparation of literature review.

Sources of funding. Funding is partially provided by Taras Shevchenko National University of Kyiv.

References

- Bateman, H., & Erdelyi, A. (1953). *Higher Transcendental Functions*, Vol. 1. McGraw-Hill.
- Daniele, V.G., & Lombardi, G. (2021). The generalized Wiener–Hopf equations for wave motion in angular regions: electromagnetic application. *Proceedings of the Royal Society A*, 477, 20210040. <https://doi.org/10.1098/rspa.2021.0040>
- Daniele, V.G., & Lombardi, G. (2022). The generalized Wiener–Hopf equations for the elastic wave motion in angular regions. *Proceedings of the Royal Society A*, 478, 20210624. <https://doi.org/10.1098/rspa.2021.0624>
- He, K., & Kobayashi, K. (2023). Diffraction by a Semi-Infinite Parallel-Plate Waveguide with Five-Layer Material Loading: Rigorous Wiener–Hopf Analysis. *Progress In Electromagnetics Research B*, 98, 125–145. <http://dx.doi.org/10.2528/PIERB22112401>
- Idemen, M. (1979). A New Method to Obtain Exact Solutions of Vector Wiener–Hopf Equations. *Zeitschrift für Angewandte Mathematik und Mechanik*, 59(11), 656–658. <https://doi.org/10.1002/zamm.19790591115>
- Kurilyak, D. B. (1998). Mixed boundary-valued diffraction problem for the wedge with Dirichlet and Neumann boundary conditions at the faces. *Reports of the National Academy of the Sciences of Ukraine*, 2, 107–112. [in Ukrainian]. [Куриляк, Д.Б. (1998) Змішана крайова задача дифракції для клина з умовами Діріхле і Неймана на гранях. *Доповіди НАН України*, 2, 107–112.]
- Kurilyak, D. B. (1999). The vector-valued Wiener–Hopf equation for a mixed boundary-value problem of diffraction at a wedge intersected by a Half-Plane. *Journal of Mathematical Sciences*, 96(1), 2873–2877. <https://doi.org/10.1007/BF02168998>
- Kurilyak, D. B., & Lysechko, V. O. (2022). Wave diffraction from the truncated hollow wedge: analytical regularization and Wiener–Hopf analysis. *Zeitschrift für angewandte Mathematik und Physik*, 73(208). <https://doi.org/10.1007/s00033-022-01841-6>
- Loveikin, A.V. (2022). Equilibrium of an Elastic Half Plane with Rigidly Fixed Boundary Weakened by an Oblique Cut. *Journal of Mathematical Sciences*, 261, 176–193. <https://doi.org/10.1007/s10958-022-05746-7>
- Loveikin, A.V. (2023). Plane Potential Field Outside a Symmetric T-Shaped Profile. *Journal of Mathematical Sciences*, 272, 93–111. <https://doi.org/10.1007/s10958-023-06402-4>

- Maurizi, M., & Berto, F. (2020). 3D effects on Fracture Mechanics: corner point singularities. *Procedia Structural Integrity*, 26, 336–347. <https://doi.org/10.1016/j.prostr.2020.06.043>
- Nagasaka, T., Kishiguchi, K., & Kobayashi, K. (2022). Diffraction by a Fractional Semi-Infinite Plate: Comparison with Impedance Boundary Conditions. *URSI Radio Science Letters*, 4, <https://doi.org/10.46620/22-0043>
- Nethercote, M. A., Kisil, A. V. & Assier, R. C. (2022). Diffraction of acoustic waves by a wedge of point scatterers. *SIAM Journal on Applied Mathematics*, 82(3), 872–898. <https://doi.org/10.1137/21M1438608>
- Noble, B. (1958). *Methods Based on the Wiener–Hopf Technique for the Solution of Partial Differential Equations*. Pergamon.
- Olver, F. (1974). *Asymptotics and Special Functions*. Academic Press.
- Parfenenko, D. M., & Ulitko, A. F. (1988). About new integral transformation and its application to contact problems of theory of elasticity. *Reports of the Academy of Sciences of Ukraine. Serie A*, 11, 51–56 [in Ukrainian]. [Парфененко, Д. М., & Улітко, А. Ф. (1988). Про одне нове інтегральне перетворення та його застосування в контактних задачах теорії пружності. *Доповіді АН УРСР. Серія А*, 11, 51–56].

Отримано редакцією журналу / Received: 04.06.25

Прорецензовано / Revised: 14.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Андрій ЛОВЕЙКІН, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-7988-8350

e-mail: andrii.loveikin@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

Андрій КРЕНЕВИЧ, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-0720-786X

e-mail: krenevych@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

ОСОБЛИВІСТЬ ПОТЕНЦІАЛЬНОГО ПОЛЯ У ВЕРШИНІ СИМЕТРИЧНОГО НАКОНЕЧНИКА ІЗ ЧОТИРМА ОДНАКОВИМИ ПЕРПЕНДИКУЛЯРНИМИ КЛИНОПОДІБНИМИ КРИЛАМИ

Розв'язано просторову задачу теорії потенціалу зовні включення у формі симетричного наконечника із чотирма однаковими перпендикулярними крилами клиноподібної форми. Основна мета – визначення особливості поведінки розв'язку у вершині наконечника.

Використовуючи симетричну форму наконечника, задачу зведено до відшукування гармонічної функції у чверті простору, що має задовольняти мішані крайові умови на його сторонах.

Однорідні нетривіальні розв'язки задачі подано у сферичних координатах інтегралом типу Мелера – Фока по функціях Лежандра. Методом інтегральних перетворень задачу зведено до функціонального рівняння типу Вінера – Гопфа, що методом Вінера – Гопфа зводиться до однорідної нескінченної системи лінійних алгебраїчних рівнянь. Розв'язки отриманої системи дають змогу записати аналітичні зображення для однорідних розв'язків вихідної крайової задачі.

Побудовані однорідні розв'язки допомагають встановити характер особливості потенціального поля у вершині наконечника. Показники особливості визначаються з умов існування нетривіальних розв'язків однорідної нескінченної системи лінійних алгебраїчних рівнянь.

Унаслідок розв'язання поставленої задачі встановлено залежність між показником особливості та кутом відхилення ребер крил наконечника. Отримані результати дають змогу оцінити якісну картину поведінки потенціального поля у вершині наконечника, а також визначити кількісні характеристики, що описують цю поведінку.

Ключові слова: теорія потенціалу, мішана крайова задача, особливість розв'язку, метод інтегральних перетворень, метод Вінера – Гопфа.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 536.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.19>

Генадій МІШУРІС¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0003-2565-1961
e-mail: ggim@aber.ac.uk

Юрій ПРОЦЕРОВ², канд. фіз.-мат. наук, доц.
ORCID ID: 0009-0009-2659-2802
e-mail: protserov@onu.edu.ua

Наталія ВАЙСФЕЛЬД³, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-8082-2503
e-mail: natalya.vaysfeld@kcl.ac.uk

Зінаїда ЖУРАВЛЬОВА², канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-3271-8864
e-mail: z.zhuravlova@onu.edu.ua

¹Університет Абериствіта, Абериствіт, Сполучене Королівство
²Одеський національний університет імені І. І. Мечникова, Одеса, Україна
³Кінгс-коледж Лондон, Лондон, Сполучене Королівство

ЗАДАЧА ТЕПЛОПРОВІДНОСТІ ДЛЯ СКЛАДЕНОГО ЦИЛІНДРУ З НЕІДЕАЛЬНИМИ УМОВАМИ КОНТАКТУ МІЖ ЙОГО ШАРАМИ

Розглянуто задачу теплопровідності для багат шарового циліндра з урахуванням впливу міжфазних контрастів, що істотно змінюють механізми теплоперенесення й ускладнюють визначення ефективної теплопровідності. Метою дослідження є отримання неідеальних умов контакту між шарами для різних порядків коефіцієнтів теплопровідності за наявності прошарку. На основі асимптотичного підходу виведено три типи умов спряження, що відповідають характерним режимам взаємодії шарів. Запропонований метод забезпечує можливість коректного опису ключових теплових і механічних ефектів за допомогою спрощених співвідношень, придатних для подальших інженерних застосувань.

Ключові слова: складений циліндр, теплопровідність, неідеальні умови спряження.

Класифікація відповідно до AMS 2020: 80A11, 74A15.

Вступ

Наявність прошарків та інтерфейсів у багат шарових матеріалах суттєво змінює механізми теплоперенесення й ускладнює опис ефективної теплопровідності. На відміну від однорідних тіл, де теплопровідність визначається лише внутрішніми властивостями матеріалу, у шарованих структурах важливу роль відіграють термічний опір межі, мікроструктура інтерфейсів, а також можливі дефекти та неоднорідності на контакт. У сучасних композитах, тонкоплівкових системах і наноструктурах інтерфейсний тепловий опір часто стає домінуючим фактором, що визначає теплову поведінку всієї системи.

Дослідження у цьому напрямку виконувалися багатьма авторами, переважно в межах спрощених постановок, представлених у класичних монографіях (Коваленко, 1965; Nowacki, 1975). Попри здобуті результати, питання коректного формулювання умов спряження за наявності контрастних властивостей у шарованих структурах залишалося відкритим. Наявні роботи переважно ґрунтувалися на підходах класичної теорії пружності, зокрема на результатах, отриманих у працях (Bigoni et al, 1998; Mishuris et al, 2006; Dumont et al, 2022) та подальших дослідженнях, присвячених інтерфейсним ефектам.

Мета запропонованої роботи полягає у врахуванні неідеальних умов контакту, що виникають унаслідок температурних недосконалостей, та у виведенні й подальшому аналізі практично значущих умов спряження між шарами. У межах задачі теплопровідності для складеного циліндра отримано три типи умов спряження, що відповідають різним порядкам співвідношень коефіцієнтів теплопровідності шарів і властивостей міжфазного прошарку між ними.

1. Основні результати

Розглянемо в циліндричній системі координат циліндр, що складається із трьох шарів:

$$\Omega_1: 0 \leq r < R, -H_1 < z < -h/2; \Omega_0: 0 \leq r < R, -h/2 < z < h/2; \Omega_2: 0 \leq r < R, h/2 < z < H_2,$$

де $h \ll H_1, H_2$ товщина міжфазного включення.

Через u_j, k_j позначимо температуру та коефіцієнти теплопровідності в областях $\Omega_j, j = 0, 1, 2$. Функції u_j мають задовольняти рівняння теплопровідності

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u_j}{\partial r} \right) + \frac{\partial^2 u_j}{\partial z^2} = 0, (r, z) \in \Omega_j. \quad (1)$$

Граничні умови, не обмежуючи загальності міркувань, вважаємо такими: циліндрична поверхня $r = R$ циліндра та його нижня основа $z = -H_1$ є теплоізовольовані, а на верхній основі $z = H_2$ задано температуру

$$\frac{\partial u_j}{\partial r} \Big|_{r=R} = 0, j = 0, 1, 2; \frac{\partial u_1}{\partial z} \Big|_{z=-H_1} = 0; u_2 \Big|_{z=H_2} = f(r). \quad (2)$$

Умовами спряження шарів циліндру будуть рівність температур і теплових потоків на їхніх спільних границях

$$u_1 \Big|_{z=-h/2} = u_0 \Big|_{z=-h/2}, k_1 \frac{\partial u_1}{\partial z} \Big|_{z=-h/2} = k_0 \frac{\partial u_0}{\partial z} \Big|_{z=-h/2} \text{ та } u_2 \Big|_{z=h/2} = u_0 \Big|_{z=h/2}, k_2 \frac{\partial u_2}{\partial z} \Big|_{z=h/2} = k_0 \frac{\partial u_0}{\partial z} \Big|_{z=h/2}. \quad (3)$$

В області Ω_0 введемо нову змінну $\zeta = \frac{z}{\varepsilon}$, де ε – мала величина, $\zeta \in (-h_0/2; h_0/2)$, $h_0 = \frac{h}{\varepsilon}$, та розглянемо рівняння відносно $u_0(r, \zeta)$

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u_0}{\partial r} \right) + \frac{1}{\varepsilon^2} \frac{\partial^2 u_0}{\partial \zeta^2} = 0, 0 \leq r < R, -h_0/2 < \zeta < h_0/2.$$

© Мішуріс Генадій, Процеров Юрій, Вайсфельд Наталія, Журавльова Зінаїда, 2026

Ураховуючи аналіз порядків його членів відносно нової змінної, отримуємо, що $\frac{\partial^2 u_0}{\partial \zeta^2} = 0$, звідки $\frac{\partial u_0}{\partial \zeta} = A(r)$ та $u_0(r, \zeta) = A(r)\zeta + B(r)$. Умови спряження (3) перепишемо так:

$$u_0|_{\zeta=-h_0/2} = u_1|_{z=-h/2}; u_0|_{\zeta=h_0/2} = u_2|_{z=h/2} \text{ та } \frac{k_0}{\varepsilon} \frac{\partial u_0}{\partial \zeta}|_{\zeta=-h_0/2} = k_1 \frac{\partial u_1}{\partial z}|_{z=-h/2}; \frac{k_0}{\varepsilon} \frac{\partial u_0}{\partial \zeta}|_{\zeta=h_0/2} = k_2 \frac{\partial u_2}{\partial z}|_{z=h/2}. \quad (4)$$

Розглянемо три випадки умов спряження залежно від порядків співвідношення коефіцієнтів теплопровідності k_0 і k_1, k_2 шарів циліндру.

I випадок. Коефіцієнт теплопровідності k_0 має порядок, що дорівнює порядкам коефіцієнтів k_1, k_2 . Зобразимо функцію $u_0(r, \zeta)$ як асимптотичний ряд за ступенями ε : $u_0 = v_0 + \varepsilon v_1 + \varepsilon^2 v_2 + \dots$.

Розглянемо спочатку другу умову спряження із (4)

$$\frac{k_0}{\varepsilon} \left(\frac{\partial v_0}{\partial \zeta} + \varepsilon \frac{\partial v_1}{\partial \zeta} + \dots \right) |_{\zeta=h_0/2} = k_2 \frac{\partial v_2}{\partial z} |_{z=h/2}; \frac{k_0}{\varepsilon} \left(\frac{\partial v_0}{\partial \zeta} + \varepsilon \frac{\partial v_1}{\partial \zeta} + \dots \right) |_{\zeta=-h_0/2} = k_1 \frac{\partial v_1}{\partial z} |_{z=-h/2}.$$

У правих частинах містяться величини порядку $O(1)$, отже $\frac{\partial v_0}{\partial \zeta} |_{\zeta=\pm h_0/2} = 0$. Звідси $A(r) = 0$ і $u_0 = (r, \zeta) = B(r)$. З урахуванням цього маємо, що стрибок температури $\langle u_0 \rangle = u_0|_{\zeta=-h_0/2} - u_0|_{\zeta=h_0/2} = 0$ з першої з умов (4) $u_1|_{z=-h/2} = u_2|_{z=h/2}$. Це надає першу з умов ідеального теплового контакту – рівність температур.

Рівняння теплопровідності (1) для цього випадку записано у формі

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v_0}{\partial r} \right) + \frac{1}{\varepsilon} \frac{\partial^2 v_1}{\partial \zeta^2} = 0.$$

Ураховуючи порядки величин доданків, маємо $\frac{\partial^2 v_1}{\partial \zeta^2} = 0$, звідки $\frac{\partial v_1}{\partial \zeta} = C(r)$ і

$$\frac{\partial u_0}{\partial \zeta} |_{\zeta=\pm h_0/2} = \frac{\partial v_0}{\partial \zeta} |_{\zeta=\pm h_0/2} + \frac{\partial v_1}{\partial \zeta} |_{\zeta=\pm h_0/2} = C(r).$$

З огляду на це стрибок теплового потоку

$$\left\langle \frac{k_0}{\varepsilon} \frac{\partial u_0}{\partial \zeta} \right\rangle = \frac{k_0}{\varepsilon} \frac{\partial u_0}{\partial \zeta} |_{\zeta=-h_0/2} - \frac{k_0}{\varepsilon} \frac{\partial u_0}{\partial \zeta} |_{\zeta=h_0/2} = 0.$$

Із другої з умов (4) отримано, що $k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}$. Це є другою умовою ідеального теплового контакту – рівність теплових потоків.

Отже, випадок I зводиться до умов класичного ідеального теплового контакту

$$u_1|_{z=-h/2} = u_2|_{z=h/2}; k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}. \quad (5)$$

II випадок. Нехай тепер коефіцієнт теплопровідності k_0 має порядок набагато менший ніж порядки коефіцієнтів k_1, k_2 , тобто $k_0 = \varepsilon k$, де k має порядок, такий як і коефіцієнти k_1, k_2 .

З умови (4) для теплових потоків знаходимо

$$\varepsilon k \frac{1}{\varepsilon} \left(\frac{\partial v_0}{\partial \zeta} + \varepsilon \frac{\partial v_1}{\partial \zeta} + \dots \right) |_{\zeta=-h_0/2} = k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2}; \varepsilon k \frac{1}{\varepsilon} \left(\frac{\partial v_0}{\partial \zeta} + \varepsilon \frac{\partial v_1}{\partial \zeta} + \dots \right) |_{\zeta=h_0/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}.$$

І з урахуванням того, що $\frac{\partial v_0}{\partial \zeta} = A(r)$, отримуємо $k \frac{\partial v_0}{\partial \zeta} |_{\zeta=-h_0/2} = kA(r) = k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2}$, звідки

$$A(r) = \frac{k_1}{k} \frac{\partial u_1}{\partial z} |_{z=-h/2} \text{ або } A(r) = \frac{k_2}{k} \frac{\partial u_2}{\partial z} |_{z=h/2}.$$

Отже, стрибок температури набуває вигляду

$$\langle u_0 \rangle = u_0|_{\zeta=h_0/2} - u_0|_{\zeta=-h_0/2} = A(r) \frac{h_0}{2} + B(r) + A(r) \frac{h_0}{2} - B(r) = A(r)h_0 = \frac{k_1}{k} h_0 \frac{\partial u_1}{\partial z} |_{z=-h/2},$$

де величина $\frac{k_1}{k} h_0 = \frac{\varepsilon k_1}{k_0} \frac{h}{\varepsilon} = \frac{k_1 h}{k_0}$ має порядок $O(1)$. Ураховуючи тепер першу з умов (4), отримаємо першу умову спряження

$$u_2|_{z=h/2} - u_1|_{z=-h/2} = \frac{k_1 h}{k_0} \frac{\partial u_1}{\partial z} |_{z=-h/2} \text{ або } u_2|_{z=h/2} - u_1|_{z=-h/2} = \frac{k_2 h}{k_0} \frac{\partial u_2}{\partial z} |_{z=h/2}.$$

Для одержання другої умови спряження розглянемо стрибок теплового потоку

$$\langle k_0 \frac{\partial u_0}{\partial \zeta} \rangle = \langle k \frac{\partial v_0}{\partial \zeta} \rangle = k \frac{\partial v_0}{\partial \zeta} |_{\zeta=-h_0/2} - k \frac{\partial v_0}{\partial \zeta} |_{\zeta=h_0/2} = kA(r) - kA(r) = 0.$$

Тоді з другої умови (4) отримуємо, що $k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}$. Зауважимо, що ця умова також впливає з попереднього спряження. Отже, другий випадок зводиться до таких умов спряження:

$$u_2|_{z=h/2} - u_1|_{z=-h/2} = \frac{k_1 h}{k_0} \frac{\partial u_1}{\partial z} |_{z=-h/2}; k_1 \frac{\partial u_1}{\partial z} |_{z=-h/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}. \quad (6)$$

III випадок. Коефіцієнт теплопровідності k_0 має порядок набагато більший ніж порядки коефіцієнтів k_1, k_2 , тобто $k_0 = \frac{k}{\varepsilon}$, де k має порядок такий, як і коефіцієнти k_1, k_2 . Тоді умову (4) відносно теплових потоків записано у формі

$$\frac{k}{\varepsilon} \frac{1}{\varepsilon} \left(\frac{\partial v_0}{\partial \zeta} + \varepsilon \frac{\partial v_1}{\partial \zeta} + \varepsilon^2 \frac{\partial v_2}{\partial \zeta} + \dots \right) |_{\zeta=h_0/2} = k_2 \frac{\partial u_2}{\partial z} |_{z=h/2}.$$

Звідси, ураховуючи порядки величин доданків, отримуємо, що $\frac{\partial v_0}{\partial \zeta} |_{\zeta=\pm h_0/2} = \frac{\partial v_1}{\partial \zeta} |_{\zeta=\pm h_0/2} = 0$. Тоді $A(r) = \frac{\partial v_0}{\partial \zeta} = 0$ та $v_0(r, \zeta) = B(r)$. Із цього випливає, що стрибок температури

$$\langle v \rangle = \langle v_0 \rangle = v_0|_{\zeta=h_0/2} - v_0|_{\zeta=-h_0/2} = B(r) - B(r) = 0.$$

Ураховуючи першу з умов (4), отримуємо першу з умов спряження $u_1|_{z=-h/2} = u_2|_{z=h/2}$. Рівняння (1) відносно функції $v_2(r, \zeta)$ у цьому випадку набуває вигляду

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v_0}{\partial r} \right) + \frac{\partial^2 v_2}{\partial \zeta^2} = 0.$$

Після інтегрування за змінною ζ отримуємо $\frac{\partial v_2}{\partial \zeta} = -\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v_0}{\partial r} \right) \zeta + C(r)$. Стрибок теплового потоку знайдено:

$$\langle k \frac{\partial v_2}{\partial \zeta} \rangle = k \frac{\partial v_2}{\partial \zeta} |_{\zeta=h_0/2} - k \frac{\partial v_2}{\partial \zeta} |_{\zeta=-h_0/2} = -\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v_0}{\partial r} \right) k h_0.$$

Використаємо, що $k h_0 = \varepsilon k_0 \frac{h}{\varepsilon} = k_0 h$ має порядок $O(1)$ та у зазначеному випадку правильними є співвідношення

$$v_0 = u_1|_{z=-h/2} = u_2|_{z=h/2} \text{ та } k_0 \frac{\partial v_2}{\partial z}|_{z=h_0/2} = k_2 \frac{\partial u_2}{\partial z}|_{z=h/2}; k_0 \frac{\partial v_2}{\partial z}|_{z=-h_0/2} = k_1 \frac{\partial u_1}{\partial z}|_{z=-h/2}.$$

Друга умова спряження набуває вигляду

$$k_2 \frac{\partial u_2}{\partial z}|_{z=h/2} - k_1 \frac{\partial u_1}{\partial z}|_{z=-h/2} = -k_0 h \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u_1}{\partial r} \right) |_{z=-h/2}.$$

Отже, третій випадок зводиться до таких умов спряження:

$$u_1|_{z=-h/2} = u_2|_{z=h/2}; k_2 \frac{\partial u_2}{\partial z}|_{z=h/2} - k_1 \frac{\partial u_1}{\partial z}|_{z=-h/2} = -k_0 h \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u_1}{\partial r} \right) |_{z=-h/2}. \quad (7)$$

Для зведення розглянутих крайових задач до одновимірних задач застосуємо скінчене інтегральне перетворення Ганкеля за змінною r

$$u_{jn}(z) = \int_0^R u_j(r, z) J_0(\lambda_n r) r dr, j = 1, 2; n = 0, 1, 2, \dots$$

із формулою обернення

$$u_j(r, z) = \frac{2}{R^2} \sum_{n=0}^{\infty} u_{jn}(z) \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)},$$

де $\lambda_0 = 0$, а $\lambda_n, n = 1, 2, \dots$ – це додатні корені рівняння $J_1(\lambda R) = 0$, де $J_m(x), m = 0, 1$ – циліндричні функції першого роду.

Застосуємо це інтегральне перетворення до рівнянь (1) та крайових умов (2) за змінною z , при цьому крайові умови за змінною r будуть задовольнятися автоматично. Зважаючи на це, отримуємо такі крайові задачі:

$$u_{1n}''(z) - \lambda_n^2 u_{1n}(z) = 0, -H_1 < z < -h/2, u_{1n}'(-H_1) = 0, \quad (8)$$

$$u_{2n}''(z) - \lambda_n^2 u_{2n}(z) = 0, h/2 < z < H_2, u_{2n}(H_2) = f_n, f_n = \int_0^R f(r) J_0(\lambda_n r) r dr. \quad (9)$$

Додаткові крайові умови одержуємо пізніше, після застосування інтегрального перетворення Ганкеля до умов спряження (5)–(7).

Розв'язком крайової задачі (8) є функції

$$u_{10}(z) = B_0 \text{ та } u_{1n}(z) = B_n \frac{\cosh \lambda_n(z+H_1)}{\cosh \lambda_n H_1}, \quad (10)$$

де B_0, B_n – невідомі сталі, а розв'язком крайової задачі (9) – функції

$$u_{20}(z) = f_0 - C_0(H_2 - z) \text{ та } u_{2n}(z) = f_n \frac{\cosh \lambda_n z}{\cosh \lambda_n H_2} - C_n \frac{\sinh \lambda_n(H_2 - z)}{\cosh \lambda_n H_2}, \quad (11)$$

де C_0, C_n – невідомі сталі.

Розглянемо спочатку випадок I, тобто ідеальний тепловий контакт. Застосуємо інтегральне перетворення Ганкеля до умов спряження (5):

$$u_{1n}(-h/2) = u_{2n}(h/2); k_1 u_{1n}'(-h/2) = k_2 u_{2n}'(h/2), n = 0, 1, 2, \dots$$

Підставимо вирази (10) і (11) та знайдемо

$$B_0 = f_0, C_0 = 0, B_n = \frac{k_2 f_n}{\Delta_n} \cosh \lambda_n H_1, C_n = \frac{f_n}{\Delta_n} \left[k_1 \cosh \lambda_n \frac{h}{2} \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_2 \sinh \lambda_n \frac{h}{2} \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \right].$$

Тут $\Delta_n = k_1 \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \sinh \lambda_n \left(H_2 - \frac{h}{2} \right) + k_2 \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \cosh \lambda_n \left(H_2 - \frac{h}{2} \right)$.

За формулою обернення інтегрального перетворення Ганкеля отримаємо:

$$u_1(r, z) = \frac{2}{R^2} f_0 + \frac{2k_2}{R^2} \sum_{n=1}^{\infty} f_n \frac{\cosh \lambda_n(z+H_1)}{\Delta_n} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)},$$

$$u_2(r, z) = \frac{2}{R^2} f_0 + \frac{2}{R^2} \sum_{n=1}^{\infty} \frac{f_n}{\cosh \lambda_n H_2} \left\{ \cosh \lambda_n z - \frac{\sinh \lambda_n(H_2 - z)}{\Delta_n} \left[k_1 \cosh \lambda_n \frac{h}{2} \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_2 \sinh \lambda_n \frac{h}{2} \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \right] \right\} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)}.$$

Розглянемо тепер випадок II. Умови спряження (6) у просторі трансформант Ганкеля мають такий вигляд:

$$u_{2n}(h/2) - u_{1n}(-h/2) = \frac{k_1}{k_0} h u_{1n}'(-h/2); k_1 u_{1n}'(-h/2) = k_2 u_{2n}'(h/2).$$

Підставивши вирази (10) та (11), знайдемо

$$B_0 = f_0, C_0 = 0, B_n = \frac{k_2 f_n}{\Delta_n^*} \cosh \lambda_n H_1,$$

$$C_n = \frac{f_n}{\Delta_n^*} \left\{ k_1 \cosh \lambda_n \frac{h}{2} \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_2 \sinh \lambda_n \frac{h}{2} \left[\cosh \lambda_n \left(H_1 - \frac{h}{2} \right) + \frac{k_1}{k_0} h \lambda_n \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \right] \right\}.$$

Тут $\Delta_n^* = k_1 \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \sinh \lambda_n \left(H_2 - \frac{h}{2} \right) + k_2 \cosh \lambda_n \left(H_2 - \frac{h}{2} \right) \left[\cosh \lambda_n \left(H_1 - \frac{h}{2} \right) + \frac{k_1}{k_0} h \lambda_n \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \right]$.

За формулою обернення інтегрального перетворення Ганкеля отримаємо

$$u_1(r, z) = \frac{2}{R^2} f_0 + \frac{2k_2}{R^2} \sum_{n=1}^{\infty} f_n \frac{\cosh \lambda_n(z+H_1)}{\Delta_n^*} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)},$$

$$u_2(r, z) = \frac{2}{R^2} f_0 + \frac{2}{R^2} \sum_{n=1}^{\infty} \frac{f_n}{\cosh \lambda_n H_2} \left\{ \cosh \lambda_n z - \frac{\sinh \lambda_n(H_2 - z)}{\Delta_n^*} \left[k_1 \cosh \lambda_n \frac{h}{2} \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_2 \sinh \lambda_n \frac{h}{2} \left[\cosh \lambda_n \left(H_1 - \frac{h}{2} \right) + \frac{k_1}{k_0} h \lambda_n \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \right] \right] \right\} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)}.$$

Розглянемо тепер випадок III. Застосовувавши інтегральне перетворення Ганкеля до умов спряження (7), отримуємо

$$u_{1n}(-h/2) = u_{2n}(h/2); k_2 u_{2n}'(h/2) - k_1 u_{1n}'(-h/2) = k_0 h \lambda_n^2 u_{1n}(-h/2).$$

Підставимо до них вирази (10) і (11) та знайдемо

$$B_0 = f_0, C_0 = 0, B_n = \frac{k_2 f_n}{\Delta_n^{**}} \cosh \lambda_n H_1,$$

$$C_n = -\frac{f_n}{\Delta_n^{**}} \left\{ k_2 \sinh \lambda_n \frac{h}{2} \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_1 \cosh \lambda_n \frac{h}{2} \left[\sinh \lambda_n \left(H_1 - \frac{h}{2} \right) + h \lambda_n \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \right] \right\}.$$

Тут $\Delta_n^{**} = k_2 \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \cosh \lambda_n \left(H_2 - \frac{h}{2} \right) + k_1 \sinh \lambda_n \left(H_2 - \frac{h}{2} \right) \left[\sinh \lambda_n \left(H_1 - \frac{h}{2} \right) + h \lambda_n \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) \right]$.

За формулою обернення інтегрального перетворення Ганкеля маємо:

$$u_1(r, z) = \frac{2}{R^2} f_0 + \frac{2k_2}{R^2} \sum_{n=1}^{\infty} f_n \frac{\cosh \lambda_n(z+H_1)}{\Delta_n^{**}} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)},$$

$$u_2(r, z) = \frac{2}{R^2} f_0 + \frac{2}{R^2} \sum_{n=1}^{\infty} \frac{f_n}{\cosh \lambda_n H_2} \left\{ \cosh \lambda_n z + \frac{\sinh \lambda_n(H_2 - z)}{\Delta_n^{**}} \left\{ k_2 \sinh \lambda_n \frac{h}{2} \cosh \lambda_n \left(H_1 - \frac{h}{2} \right) - k_1 \cosh \lambda_n \frac{h}{2} \left[\sinh \lambda_n \left(H_1 - \frac{h}{2} \right) + h \lambda_n \sinh \lambda_n \left(H_1 - \frac{h}{2} \right) \right] \right\} \right\} \frac{J_0(\lambda_n r)}{J_0^2(\lambda_n R)}.$$

Дискусія і висновки

У проведеному дослідженні враховано неідеальні умови контакту, зумовлені температурними недоскональностями на межах шарів скінченного циліндра. Отримано три типи умов спряження, що відповідають різним порядкам співвідношень коефіцієнтів теплопровідності шарів і характеристик міжфазного прошарку. Запропонований підхід ґрунтується на асимптотичному методі, що дає змогу адекватно відобразити ключові механічні ефекти й водночас використовувати спрощені співвідношення для подальших інженерних розрахунків.

Внесок авторів: Геннадій Мішуріс – концептуалізація, методика, Наталя Вайсфельд, Юрій Процеров, Зінаїда Журавльова – теоретичні дослідження та числові розрахунки.

Подяки та джерела фінансування. Дослідження підтримано європейським проектом, що фінансується Рамковою програмою з досліджень та інновацій Горизонт 2020 (2014–2020) (H2020-MSCA-RISE-2020) Грантова угода № 101008140 EffectFact "Ефективні методи факторизації для матричних функцій: Розвиток теорії, чисельних методів і ефективних застосувань", Наталя Вайсфельд є вдячною гранту Вольфсона Королівського товариства R3/233003.

Список використаних джерел

- Коваленко, А. Д. (1965). *Введение в термоупругость*. Наукова думка.
- Bigoni, D., Serkov, S. K., Valentini, M., & Movchan, A. B. (1998). Asymptotic models of dilute and densely packed composites with imperfectly bonded inclusions. *International Journal of Solids and Structures*, 35, 3239–3258.
- Dumont, S., Lebon, F., Rizzoni, R., & Serpilli, M. (2022). Derivation of Imperfect Interface Laws for Multi-Physic Composites by a Multiscale Approach: Theoretical and Numerical Studies. In: *The 8th European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS Congress, ECCOMAS*, Oslo, Norway.
- Mishuris, G. S., Movchan, N. V., & Movchan, A. B. (2006). Steady-state motion of a mode-III crack on imperfect interfaces. *The Quarterly Journal of Mechanics and Applied Mathematics*, 59(4), 487–516. <https://doi.org/10.1093/qjmam/hbl013>
- Nowacki, W. (1975). *Dynamic Problems of Thermoelasticity*. Springer Dordrecht.

References

- Bigoni, D., Serkov, S. K., Valentini, M., & Movchan, A. B. (1998). Asymptotic models of dilute and densely packed composites with imperfectly bonded inclusions. *International Journal of Solids and Structures*, 35, 3239–3258.
- Dumont, S., Lebon, F., Rizzoni, R., & Serpilli, M. (2022). Derivation of Imperfect Interface Laws for Multi-Physic Composites by a Multiscale Approach: Theoretical and Numerical Studies. In: *The 8th European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS Congress, ECCOMAS*, Oslo, Norway.
- Kovalenko, A. D. (1965). *Introduction in thermoelasticity*. Naukova dumka [in Russian].
- Mishuris, G. S., Movchan, N. V., & Movchan, A. B. (2006). Steady-state motion of a mode-III crack on imperfect interfaces. *The Quarterly Journal of Mechanics and Applied Mathematics*, 59(4), 487–516. <https://doi.org/10.1093/qjmam/hbl013>
- Nowacki, W. (1975). *Dynamic Problems of Thermoelasticity*. Springer Dordrecht.

Отримано редакцією журналу / Received: 02.12.25

Прорецензовано / Revised: 08.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Gennady MISHURIS¹, DSc (Phys. & Math.), Prof.

ORCID ID: 0000-0003-2565-1961

e-mail: ggmm@aber.ac.uk

Yurii PROTSEROV², PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0009-0009-2659-2802

e-mail: protserov@onu.edu.ua

Natalya VAYSFELD³, DSc (Phys. & Math.), Prof.

ORCID ID: 0000-0001-8082-2503

e-mail: natalya.vaysfeld@kcl.ac.uk

Zinaida ZHURAVLOVA², PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0002-3271-8864

e-mail: z.zhuravlova@onu.edu.ua

¹Aberystwyth University, Ceredigion, United Kingdom

²Odesa I. I. Mechnikov National University, Odesa, Ukraine

³King's College London, London, United Kingdom

HEAT CONDUCTION PROBLEM FOR A COMPOSITE CYLINDER WITH IMPERFECT CONTACT CONDITIONS BETWEEN ITS LAYERS

The paper considers the thermal conductivity problem for a multilayer cylinder, taking into account the influence of interface contrasts, which significantly change the mechanisms of heat transfer and complicate the determination of effective thermal conductivity. The aim of the study is to obtain non-ideal contact conditions between layers for different orders of thermal conductivity coefficients in the presence of an interlayer. Based on an asymptotic approach, three types of coupling conditions are derived that correspond to the characteristic modes of interaction between the layers. The proposed method provides the possibility of correctly describing key thermal and mechanical effects using simplified relationships suitable for further engineering applications.

К е у о р д с : composite cylinder, heat conduction, imperfect contact conditions.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 517.9

DOI: <https://doi.org/10.17721/1812-5409.2026/1.20>

Viktoriia NAZARCHUK, Master's Student
ORCID ID: 0009-0007-2252-4584

e-mail: nazarchukviktoriia@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

CONTROLLABILITY OF PSEUDOPARABOLIC INTEGRO-DIFFERENTIAL SYSTEMS WITH MEMORY

The paper studies linear pseudoparabolic integro-differential equations with a Volterra-type integral term, which provide an effective mathematical framework for describing dynamic processes with memory and nonlocal interactions. Such mathematical models are characteristic of various areas of applied research, where the evolution of a system depends not only on current conditions but also on its previous state. Typical examples include the behavior of viscoelastic materials (such as polymers, biopolymers, and high-temperature metals), the processes of groundwater filtration and contaminant transport in porous media, as well as in problems related to the transport of heat and nutrients in tissues.

The aim of the study is to generalize known theoretical results to a broader class of pseudoparabolic equations by introducing additional terms into both the differential and integral parts of the operators. To achieve this goal, the methodology of a priori estimates in negative norms, developed by S. I. Lyashko and his colleagues, is employed. The proposed approach enables the investigation of well-posedness, generalized solvability, and issues of controllability and optimal control for systems with distributed parameters.

The main result of the study is the establishment of a priori estimates in negative norms that account for the new terms introduced into the pseudoparabolic integro-differential equation. A key distinction of this work from previous studies is that the obtained estimates are proven without imposing non-negativity assumptions on the lower-order terms of the differential operator. This substantially broadens the applicability of the theory, allowing the extension of theorems on the well-posedness of initial-boundary value problems, as well as results on controllability and the existence of optimal control, to a wider class of pseudoparabolic-type systems.

Key words: pseudoparabolic integro-differential equation, Volterra operator, memory effect, a priori estimates, well-posedness, generalized solvability, controllability, optimal control.

AMS 2020 classification: 35R09, 35K70, 93C20, 93C23, 93B05, 45K05, 35A01.

Introduction

Partial differential equations appear in the analysis of many physical phenomena. They serve as a fundamental tool for mathematical modeling, providing a framework for describing dynamical systems that evolve over time. While classical PDEs typically focus on processes governed by local interactions, there exists a broad class of equations that extend this perspective by incorporating terms accounting for nonlocal and memory effects.

Temporal nonlocality is widely observed in physics, biology, chemistry, ecology, materials engineering, and mechanism design. For example, in many physical and biological systems, complex models account for the interaction of chemical reactions, momentum transport, diffusion processes, and mechanical stresses (Vromans, van de Ven, & Muntean, 2019), and such phenomena are frequently modeled by PDEs with delay (see, e.g., (Guo, Feng, & Li, 2025; Zhang, & Li, 2025; Pokojovy, & Anikushyn, 2019, 2017; Liu et al., 2025; Wu, & Yin, 2025; Ezzinbi, & Staili, 2025; Salhi, 2025; Demchenko, Anikushyn, & Pokojovy, 2019)). Another important class of temporal nonlocal models arises from fractional time derivatives, which encode history-dependent effects through power-law memory kernels. Such equations naturally emerge in the modeling of anomalous diffusion, viscoelasticity, and heterogeneous transport processes. Several derivations have been proposed using stochastic time-change, continuous-time random walks, and homogenization techniques (see, e.g., (Metzler, & Klafter, 2000; Mainardi, 1996; Mainardi et al., 2006; Caputo, & Mainardi, 1971; Kubica, Ryszewska, & Yamamoto, 2020; Gal, & Warma, 2020; Jin, & Zhou, 2024)).

Beyond delay and fractional models, a broad and versatile framework for temporal nonlocality is provided by PDEs with memory. These systems incorporate hereditary effects, where prior states influence present dynamics through convolution terms involving suitable memory kernels. Such behavior is commonly observed in processes involving porous media, including groundwater flow, pollutant transport in soil, and the movement of fluids and gases through fractured materials (e.g., (Lyashko, 2002; Cueto-Felgueroso, & Juanes, 2009; Di, Shang, & Zheng, 2016; Jhangeer et al., 2024)). Memory effects also arise in viscoelastic materials, which are extensively used in construction and industrial manufacturing. For example, in carbon sequestration applications, the prior history of gas injection and deformation of rock formations can significantly influence storage efficiency. Moreover, biological systems often exhibit memory-driven transport of heat and nutrients in tissues, which plays a crucial role in medical procedures such as hyperthermia-based cancer treatments (Ragab et al., 2021).

These memory effects are especially prominent in the mechanical response of certain materials (Anikushyn, Lyashko, & Samosonok, 2024). In the modeling of viscoelastic substances, for instance, one typically arrives at equations of motion formulated as linear hyperbolic differential equations modified by a dissipative Volterra-type integral term (Conti, Gatti, & Pata, 2008; Cannarsa, & Sforza, 2011; Duvaut, & Lions, 1976; Hranishak, & Anikushyn, 2024; Anikushyn, 2024). Materials that demonstrate such viscoelastic characteristics include amorphous and semicrystalline polymers, biopolymers, metals exposed to high temperatures, and bituminous substances (Lakes, 1998).

Numerous applied mathematical models governed by linear integro-differential equations have been rigorously explored in classical monographs (Renardy, Hrusa, & Nohel, 1987; Fabrizio, & Morro, 1992), where the authors offer formal justifications for these formulations and analyze their mathematical behavior.

Over the past years, various formulations for linear and nonlinear *pseudoparabolic* integro-differential equations and their systems have attracted significant attention across multiple research communities. The scope of this progress is vast and extends beyond what can be concisely summarized here. We provide just a few examples.

Di *et al.* (Di, Shang, & Yu, 2020) considered a nonlinear integro-differential equation of pseudoparabolic type given by

$$u_t - \Delta u_t - \Delta u + \int_0^t g(t - \tau) \Delta u(\tau) d\tau = f(u)$$

where the Volterra-type integral term accounts for the memory behavior of the system. Under homogeneous Dirichlet boundary conditions, the authors demonstrated the existence of finite-time blow-up solutions for initial data with non-negative and positive energy levels.

In a related study (Liu, Kim, & Fang, 2025), Liu *et al.* investigated the well-posedness and blow-up behavior of a semilinear pseudoparabolic equation with a nonlinear convolution memory term. The integral operator considered in this setting commonly arises in models from filtration theory, population dynamics, and semiconductor physics. The authors established threshold conditions that distinguish between global existence and blow-up scenarios for initial states with low energy.

Maqbul & Raheem (Maqbul, & Raheem, 2020) analyzed a one-dimensional semilinear pseudoparabolic equation, where the temporal nonlocal operator was applied directly to the function u , which differs from the previously examined cases where Δu served as the argument of the integral term. The problem was studied under homogeneous integral conditions, and the Rothe time-discretization method was applied to establish the existence and uniqueness of strong solutions.

Further research has focused on developing numerical methods for solving initial-boundary value problems of this class. For instance, in (Wang, 2024), efficient two-grid finite element algorithms were proposed for two-dimensional nonlinear pseudo-parabolic integro-differential problems. Optimal error estimates were obtained for fully discrete schemes in the H_1 and L_p norms. The analysis relied on a combined temporal-spatial error splitting strategy.

Amar *et al.* (Amar *et al.*, 2019) studied a degenerate pseudo-parabolic problem with memory

$$\begin{aligned} -\operatorname{div} \left(C \nabla u_t + A \nabla u + \int_0^t B(x, t - \tau) \nabla u d\tau \right) &= f, \\ u &= 0, x \in \partial\Omega, \\ x|_{t=0} &= u_0(x) \text{ on } \Omega. \end{aligned}$$

Here A, B, C are symmetric matrices, A is positive definite and C is only positive semi-definite. The authors obtained the existence and uniqueness result.

In most physical and biological systems, diffusion is positive as it represents the spontaneous spreading of particles, heat, or other quantities from regions of high concentration to regions of low concentration, driven by the natural tendency toward entropy maximization. From a mathematical perspective, a positive diffusion coefficient ensures that the corresponding equation is parabolic and dissipative, implying that perturbations decay smoothly over time and that solutions remain stable.

However, negative (or effective negative) diffusion may arise in certain contexts, such as active media, reaction-diffusion systems, phase separation processes, or linearizations around unstable states. In such cases, feedback mechanisms, energy input from external sources, or coupling between system components lead to local aggregation rather than dispersion. Stability can often be recovered through additional effects, including nonlinear responses, higher-order corrections, or memory terms. For instance, effective negative diffusivity appears in the continuum limits of discrete models describing social aggregation (Popescu, & Dietrich, 2004; Miller *et al.*, 2024, 2025).

In the pseudo-parabolic context, typical physical models leading to such equations (e.g., second-order fluids, consolidation of clays, double-porosity diffusion, viscoelastic diffusion with small relaxation time, Sobolev-type corrections) usually yield a positive definite coefficient matrix $B = \{b_{ij}\}_{i,j=1}^n$ (defined later in (5)), corresponding to positive diffusivity. This is the standard and physically natural case (Cannon, & Lin, 1988).

A negative sign of the matrix B in the linear constant-coefficient setting corresponds to an energy input or anti-diffusion. Such behavior is usually physically unstable or non-physical for simple diffusion or filtration models. Negative diffusivity was considered, for example, in (Bertsch, Smarrazzo, & Tesei, 2013, 2016), where the authors analyzed the pseudo-parabolic regularization

$$u_t = \Delta \varphi(u) + \varepsilon \Delta [\psi(u)]_t,$$

associated with the forward-backward parabolic equation

$$u_t = \Delta \varphi(u).$$

The key feature of that model is a Perona-Malik-type nonlinearity $\varphi(u)$ whose derivative $\varphi'(u)$ changes sign, for some $\alpha > 0$:

$$\varphi'(u) > 0 \quad \text{for } u \in [0, \alpha), \quad \varphi'(u) < 0 \quad \text{for } u \in (\alpha, \infty).$$

The derivative $\varphi'(u)$ thus acts as an effective diffusion coefficient multiplying Δu in

$$\Delta \varphi(u) = \nabla \cdot (\varphi'(u) \nabla u) = \varphi''(u) |\nabla u|^2 + \varphi'(u) \Delta u.$$

Hence, regions where $\varphi'(u) < 0$ correspond to effective negative diffusivity, since the coefficient multiplying Δu becomes negative. Whenever $\varphi'(u) < 0$, the equation $u_t = \Delta\varphi(u)$ becomes backward parabolic and thus ill-posed in the classical sense. The pseudo-parabolic regularization term $\varepsilon \Delta[\psi(u)]_t$ compensates for this instability and restores well-posedness by introducing higher-order temporal smoothing.

Analyzing such complex systems, particularly those that may be ill-posed or lack classical stability, requires a mathematical framework that allows one to establish well-posedness and controllability. One suitable approach is the methodology of a priori inequalities in negative norms, developed by S.I. Lyashko and his colleagues, the principles of which are presented in the monograph (Lyashko, 2002). The application of this theory has proven effective for issues of well-posedness of initial-boundary value problems, optimal control, controllability of systems with distributed parameters, etc.

This methodology has been comprehensively developed and applied in a number of subsequent works, particularly (Lyashko, & Semenov, 2001; Nomirowskyi, & Anikushyn, 2011; Nomirowskyi, 2004; Lyashko, & Nomirowskii, 2003; Lyashko, Nomirowskii, & Sergienko, 2001; Lyashko, & Nomirowskyi, 2003; Nomirowskyi, 2006; Nomirowskii, 2007; Nomirowskyi, & Vostrikov, 2016; Tymchyshyn, & Nomirowskii, 2021). Later, it has also proven effective for application to Dirichlet problems for various classes of integro-differential equations with Volterra-type integral terms. In particular, hyperbolic integro-differential equations were studied in (Hranishak et al., 2025; Anikushyn, 2024; Anikushyn, Lyashko, & Samosonok, 2024; Hranishak, & Anikushyn, 2024), parabolic equations in (Hulianytskyi, & Anikushyn, 2014; Anikushyn, 2014, 2025), and pseudoparabolic type equations were presented in (Hulianytskyi, & Semenov, 2012; Anikushyn, & Andaral, 2025; Anikushyn, & Nazarchuk, 2024).

In the present paper, we consider a system described by a linear pseudo-parabolic integro-differential equation with a Volterra-type integral operator

$$a(x)u_t - \sum_{i,j=1}^n (a_{ij}(x)u_{tx_j})_{x_i} + b(x)u + \sum_{i=1}^n b_i(x)u_{x_i} - \sum_{i,j=1}^n (b_{ij}(x)u_{x_j})_{x_i} + \\ + \int_0^t \left(\sum_{i,j=1}^n (K_{ij}(x, t, \tau)u_{x_j})_{x_i} + \sum_{i=1}^n K_i(x, t, \tau)u_{x_i} + K(x, t, \tau)u \right) d\tau = f(x, t)$$

with Dirichlet-type conditions

$$u|_{x \in \partial\Omega} = 0, \quad u|_{t=0} = 0.$$

A specific case of this initial-boundary value problem ($K_i = K = b_i = 0$) was studied in (Anikushyn, & Nazarchuk, 2024), where a priori estimates in negative norms were established. Using the approach developed in (Lyashko, 2002), the authors proved the well-posedness of the problem and formulated a theorem on the existence of an optimal control for the corresponding system.

In (Hulianytskyi, & Semenov, 2012), a priori estimates in negative norms were also obtained for a structurally related pseudoparabolic integro-differential equation containing an additional Volterra-type term. These estimates served as the basis for proving existence and uniqueness theorems for generalized solutions of the associated boundary value problems. The study additionally addressed the problem of impulse-point controllability for systems governed by such equations.

In the present work, we extend the analysis of the pseudoparabolic integro-differential equation considered in (Anikushyn, & Nazarchuk, 2024) by introducing additional terms with coefficients b_i in the differential part of the operator, as well as kernel functions K_{ij} , K_i , and K in the integral operator. We derive a priori estimates in negative norms for this extended equation, with full consideration of the new terms. The assumptions on the coefficients are retained from (Anikushyn, & Nazarchuk, 2024): the self-adjointness of the differential part, the non-negative definiteness of the matrix $\{a_{ij}\}_{i,j=1}^n$, and the non-negativity of the function a .

The derived inequalities allow us to generalize the results of (Anikushyn, & Nazarchuk, 2024) and (Hulianytskyi, & Semenov, 2012) to a wider class of equations. This generalization covers well-posedness, generalized solvability, the optimal control problem, and controllability, while crucially extending the framework from (Anikushyn, & Andaral, 2025) by removing the assumption of non-negative definiteness for the matrix $\{b_{ij}\}_{i,j=1}^n$ and non-negativity for the function b .

1. Main notations and spaces

Let us consider the cylindrical domain $Q = \Omega \times (0, T)$, where Ω is a bounded domain in $\mathbb{R}^n$ with a smooth boundary $\partial\Omega$, and the linear integro-differential equation with a Volterra-type integral operator

$$\mathcal{L}u \equiv L_D u + L_I u = f(x, t). \quad (1)$$

Here, $u(x, t)$ is the unknown function describing the state of the system in the domain Q , and the operators L_D and L_I are given by

$$L_D u \equiv (\mathcal{A}u)_t + \mathcal{B}u, \quad (2)$$

$$L_I u \equiv \int_0^t \left(\sum_{i,j=1}^n (K_{ij}(x, t, \tau)u_{x_j})_{x_i} + \sum_{i=1}^n K_i(x, t, \tau)u_{x_i} + K(x, t, \tau)u \right) d\tau, \quad (3)$$

where

$$\mathcal{A}u \equiv - \sum_{i,j=1}^n (a_{ij}(x)u_{x_j})_{x_i} + a(x)u, \quad (4)$$

$$\mathcal{B}u \equiv - \sum_{i,j=1}^n (b_{ij}(x) u_{x_j})_{x_i} + \sum_{i=1}^n b_i(x) u_{x_i} + b(x)u. \quad (5)$$

We assume that the coefficients $\{a_{ij}\}_{i,j=1}^n, \{b_{ij}\}_{i,j=1}^n \subset C^1(\overline{\Omega})$, $a, b, b_i \in C(\overline{\Omega})$ satisfy, for all $x \in \Omega$,

$$a_{ij}(x) = a_{ji}(x), b_{ij}(x) = b_{ji}(x), a(x) \geq 0, \quad (6)$$

and there exists a constant $\alpha > 0$ such that, for all $x \in \Omega$ and $\xi_i \in \mathbb{R}$, $i = \overline{1, n}$,

$$\sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j \geq \alpha \sum_{i=1}^n \xi_i^2. \quad (7)$$

We also assume that the kernels $K_{ij}(x, t, \tau)$, $K_i(x, t, \tau)$ are continuously differentiable, and $K(x, t, \tau)$ is continuous. In particular, there exists a constant $M > 0$ such that $|K_{ij}(x, t, \tau)| < M$, $|K_i(x, t, \tau)| < M$, and $|K(x, t, \tau)| < M$ for all $x \in \Omega$ and $t, \tau \in [0, T]$. For simplicity, we additionally assume that all functions $|a_{ij}|, |b_{ij}|, |b_i|, a, |b|$ in $\overline{\Omega}$ are bounded by the same constant M .

The domain of the operator $\mathcal{L}$ is taken to be the space of functions infinitely differentiable in $\overline{Q}$ and satisfying homogeneous initial and Dirichlet boundary conditions

$$u|_{t=0} = 0, \quad (8)$$

$$u|_{x \in \partial\Omega} = 0. \quad (9)$$

We denote this set of functions by C_{BR}^∞ .

We define the space W_{BR} as the completion of C_{BR}^∞ with respect to the norm

$$\|u\|_W = \left(\int_Q u_t^2 + \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2}. \quad (10)$$

The adjoint operator (in $L_2(Q)$) is defined as $\mathcal{L}^*v \equiv L_D^*v + L_I^*v$, where

$$L_D^*v \equiv -(\mathcal{A}v)_t + \mathcal{B}^*v,$$

$$L_I^*v \equiv \int_t^T \left(\sum_{i,j=1}^n (K_{ij}(x, t, \tau) v_{x_i})_{x_j} - \sum_{i=1}^n (K_i(x, \tau, t) v)_{x_i} + K(x, \tau, t) v \right) d\tau.$$

The domain of $\mathcal{L}^*$ is taken to be the space of functions infinitely differentiable in $\overline{Q}$ and satisfying

$$v|_{t=T} = 0, \quad (11)$$

$$v|_{x \in \partial\Omega} = 0. \quad (12)$$

We denote this set by C_{BR+}^∞ , and by W_{BR+} we denote its completion with respect to the norm from Equation (10).

Furthermore, we denote by H_{BR} and H_{BR+} the completions of C_{BR}^∞ and C_{BR+}^∞ , respectively, with respect to the norm

$$\|u\|_H = \left(\int_Q u^2 + \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2}, \quad (13)$$

and by $W_{BR}^-, W_{BR+}^-, H_{BR}^-, H_{BR+}^-$ the corresponding negative spaces constructed with respect to the associated positive spaces relative to $L_2(Q)$.

Remark 1. It can be shown that the completions of C_{BR}^∞ and C_{BR+}^∞ with respect to the norm (13) coincide. Hence, $H_{BR} = H_{BR+}$ and $H_{BR}^- = H_{BR+}^-$.

2. Generalized solvability and controllability

Consider a pseudoparabolic-type system in which the state function u satisfies the equation

$$\mathcal{L}u = f + \mathcal{B}h, \quad f \in W_{BR+}^-, \quad (14)$$

where the operator $\mathcal{L}$ is defined according to (1)–(5), and $\mathcal{B}$ describes the action of the control h from the admissible set $\mathcal{U}$ of the control space $\mathcal{H}$ on the system. The function $u(x, t)$ is subject to the initial-boundary conditions (8), (9).

In the next section, we establish a priori estimates that play a key role in justifying the well-posedness and generalized solvability of the initial-boundary value problem, as well as in formulating the controllability problem for the obtained solutions. We first present these estimates in the form of the following theorem.

Theorem 1. *There exist constants $C_1 > 0$ and $C_2 > 0$ such that, for all $u \in C_{BR}^\infty$ and $v \in C_{BR+}^\infty$, the following inequalities hold*

$$\begin{aligned} C_1 \|u\|_{H_{BR}} &\leq \|\mathcal{L}u\|_{W_{BR+}^-} \leq C_2 \|u\|_{W_{BR}}, \\ C_1 \|v\|_{H_{BR+}} &\leq \|\mathcal{L}^*v\|_{W_{BR}^-} \leq C_2 \|v\|_{W_{BR+}}. \end{aligned}$$

Below, we present an approach to defining a generalized solution to problem (14), following the ideas of (Lyashko, 2002; Anikushyn, & Andaral, 2025). Other possible approaches to defining a generalized solution can also be found in the cited literature.

Definition 1. A function $u \in H_{BR}$ is called a generalized solution of problem (14) with the right-hand side $F = f + \mathcal{B}h \in W_{BR+}^-$ if there exists a sequence $u_i \in C_{BR}^\infty$, $i = 1, 2, \dots$, such that

$$\|u - u_i\|_{H_{BR}} \xrightarrow{i \rightarrow \infty} 0, \quad \|\mathcal{L}u_i - F\|_{W_{BR+}^-} \xrightarrow{i \rightarrow \infty} 0.$$

Applying Theorem 3, and Lemma 3 of (Lyashko, 2002, pp. 8,10), along with inequalities from Theorem 1, the following statement holds.

Theorem 2. For any function $F \in W_{BR+}^-$, there exists a unique solution $u \in H_{BR}$ of problem (14) in the sense of Definition 1. Moreover, there exists a constant C independent of F such that

$$\|u\|_{H_{BR}} \leq C \|F\|_{W_{BR+}^-}.$$

Following (Lyashko, & Semenov, 2001), we focus on asymptotic controllability.

Definition 2. System (14) is said to be asymptotically controllable in H_{BR} by the set of controls $\mathcal{U}$ if, for any $u^* \in H_{BR}$, there exists a sequence $h_i \in \mathcal{U}$, $i = 1, 2, \dots$, such that

$$\|u(h_i) - u^*\|_{H_{BR}} \xrightarrow{i \rightarrow \infty} 0,$$

where $u(h_i)$ is the generalized solution of equation (14) with the right-hand side $F = f + \mathcal{B}h_i \in W_{BR+}^-$.

Based on the a priori estimates from Theorem 1, the results of Theorem 2, and the approach of (Lyashko, 2002), we obtain the following result.

Theorem 3. If the set $\mathcal{B}(\mathcal{U})$ is dense in W_{BR+}^- , then system (14) is asymptotically controllable in the space H_{BR} by the set of admissible controls $\mathcal{U}$.

We now consider an application of Theorem 3 to the problem of controllability for a system with impulsive action. Suppose that the operator $\mathcal{B}$ is given by

$$\mathcal{B}h = \sum_{i=1}^s \delta(t - t_i) \otimes g_i(x), \quad t_i \in [0, T], \quad g_i(x) \in L_2(\Omega), \quad (15)$$

where the control is $h = \{(t_i, g_i(x))\}_{i=1}^s$, a finite sequence of elements from the set $\mathcal{H} = ([0, T] \times L_2(\Omega))^s$.

It is known (Lyashko, & Semenov, 2001), that the set of functions

$$\left\{ \sum_{i=1}^s \delta(t - t_i) \otimes g_i(x) \mid h \in \mathcal{H} \right\}$$

is dense in W_{BR+}^- . Therefore, we can formulate the following theorem.

Theorem 4. Let the state of the system be described by the solution of problem (14) with the operator (15) and the control space $\mathcal{H}$, and suppose that $\mathcal{B}(\mathcal{U})$ is dense in W_{BR+}^- . Then system (14) is asymptotically (impulsively) controllable in H_{BR} by the set $\mathcal{U}$.

It is worth noting that similar results on density, and hence controllability, can also be obtained for other types of generalized actions. In particular, let us consider the case of impulsive-pointwise control, as studied in (Lyashko, & Semenov, 2001; Lyashko, Semenov, & Sergienko, 2002).

Suppose that the domain Ω is cylindrical with respect to the variable x_1 , i.e., $\Omega = (0, 1) \times \Omega'$, $\Omega' \in \mathbb{R}^{n-1}$. And let the operator $\mathcal{B}$ is given by the following equation

$$\mathcal{B}h = \sum_{i=1}^s \delta(t - t_i) \otimes \delta(x_1 - x_{1,i}) \otimes g_i(x_2, \dots, x_n), \quad t_i \in [0, T], \quad x_{1,i} \in [0, 1], \quad g_i \in L_2(\Omega'), \quad (16)$$

where the control is $h = \{(t_i, x_{1,i}, g_i(x_2, \dots, x_n))\}_{i=1}^s$, a finite sequence of elements from the set $\mathcal{H} = ([0, T] \times [0, 1] \times L_2(\Omega'))^s$.

The set of functions

$$\left\{ \sum_{i=1}^s \delta(t - t_i) \otimes \delta(x_1 - x_{1,i}) \otimes g_i(x_2, \dots, x_n) \mid h \in \mathcal{H} \right\}$$

is dense in W_{BR+}^- (Hulianytskyi, & Semenov, 2012; Lyashko, 2002), which ensures the applicability of the asymptotic controllability theorems to this type of generalized action as well.

3. A priori inequalities in negative norms

We now proceed to prove the a priori inequalities from Theorem 1 for the operators $\mathcal{L}$ and $\mathcal{L}^*$. We shall assume that C may denote different constants where its exact value is immaterial.

Lemma 1. *There exists a constant $C > 0$ such that, for any function $u \in C_{BR}^\infty$, the following inequality holds:*

$$C\|u\|_{W_{BR}} \geq \|\mathcal{L}u\|_{W_{BR}^-}.$$

Proof. For any function $v \in C_{BR+}^\infty$, consider the scalar product $(L_D u, v)_{L_2(Q)}$. Applying the Cauchy-Schwarz and Poincaré inequalities to estimate the term containing b_i , we obtain

$$\begin{aligned} \left| \int_Q v \sum_{i=1}^n b_i u_{x_i} dQ \right| &\leq \int_Q \left| v \sum_{i=1}^n b_i u_{x_i} \right| dQ \leq M \left(\int_Q n v^2 dQ \int_Q \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2} \leq \\ &\leq C_p^2 M \sqrt{n} \left(\int_Q v_i^2 dQ \right)^{1/2} \left(\int_Q \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2} \leq C_p^2 M \sqrt{n} \|v\|_{W_{BR+}} \|u\|_{W_{BR}}. \end{aligned}$$

The remaining terms of the differential part of the problem operator are estimated as in (Lyashko, 2002).

Thus, for some constant C , we have

$$|(L_D u, v)_{L_2(Q)}| \leq C \|v\|_{W_{BR+}} \|u\|_{W_{BR}}. \quad (17)$$

Next, for any $v \in C_{BR+}^\infty$, we estimate $(L_I u, v)_{L_2(Q)}$:

$$\begin{aligned} |(L_I u, v)_{L_2(Q)}| &\leq \left| \int_Q \int_0^t \sum_{i,j=1}^n (K_{ij}(x, t, \tau) u_{x_i})_{x_j} d\tau v(x, t) dQ \right| + \\ &+ \left| \int_Q \int_0^t \sum_{i=1}^n K_i(x, t, \tau) u_{x_i} d\tau v(x, t) dQ \right| + \left| \int_Q \int_0^t K(x, t, \tau) u(x, \tau) d\tau v(x, t) dQ \right|. \end{aligned}$$

By repeatedly applying the Cauchy-Schwarz and Poincaré inequalities, and taking into account conditions (8), (9), we obtain the following estimate for the first term:

$$\begin{aligned} \left| \int_Q \int_0^t \sum_{i,j=1}^n (K_{ij}(x, t, \tau) u_{x_i})_{x_j} d\tau v(x, t) dQ \right| &= \left| \int_Q \int_0^t \sum_{i,j=1}^n K_{ij}(x, t, \tau) u_{x_i}(x, \tau) d\tau v_{x_j}(x, t) dQ \right| \leq \\ &\leq \sum_{i,j=1}^n \int_Q \int_0^t |K_{ij}(x, t, \tau) u_{x_i}(x, \tau) v_{x_j}(x, t)| d\tau dQ \leq \\ &\leq M \sum_{i,j=1}^n \int_Q |v_{x_j}(x, t)| \int_0^t |u_{x_i}(x, \tau)| d\tau dQ \leq MT \sum_{i,j=1}^n \left(\int_Q v_{x_j}^2 dQ \int_Q u_{x_i}^2 dQ \right)^{1/2} \leq \\ &\leq n M T C_p^2 \left(\int_Q \sum_{j=1}^n v_{x_j}^2 dQ \right)^{1/2} \left(\int_Q \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2} \leq C \|v\|_{W_{BR+}} \|u\|_{W_{BR}}. \end{aligned}$$

For the second term, we have

$$\begin{aligned} \left| \int_Q \int_0^t \sum_{i=1}^n K_i(x, t, \tau) u_{x_i} d\tau v(x, t) dQ \right| &\leq \sum_{i=1}^n \int_Q \int_0^t |K_i(x, t, \tau) u_{x_i}(x, \tau) v(x, t)| d\tau dQ \leq \\ &\leq MT \left(\int_Q n v^2 dQ \right)^{1/2} \left(\sum_{i=1}^n \int_Q u_{x_i}^2 dQ \right)^{1/2} \leq \\ &\leq M T C_p^2 \sqrt{n} \left(\int_Q v_i^2 dQ \right)^{1/2} \left(\int_Q \sum_{i=1}^n u_{x_i}^2 dQ \right)^{1/2} \leq C \|v\|_{W_{BR+}} \|u\|_{W_{BR}}. \end{aligned}$$

For the third term, we obtain

$$\begin{aligned} \left| \int_Q \int_0^t K(x, t, \tau) u(x, \tau) d\tau v(x, t) dQ \right| &\leq \int_Q \int_0^t |K(x, t, \tau) u(x, \tau) v(x, t)| d\tau dQ \leq \\ &\leq MT \left(\int_Q v^2 dQ \right)^{1/2} \left(\int_Q u^2 dQ \right)^{1/2} \leq \end{aligned}$$

$$\leqslant MTC_p^2 \left(\int_Q v_i^2 dQ \right)^{1/2} \left(\int_Q u_i^2 dQ \right)^{1/2} \leqslant C \|v\|_{W_{BR+}} \|u\|_{W_{BR}}.$$

Hence, for some constant C , we have

$$|(L_I u, v)_{L_2(Q)}| \leqslant C \|v\|_{W_{BR+}} \|u\|_{W_{BR}}. \quad (18)$$

From (17) and (18), we arrive at the overall estimate

$$|(\mathcal{L}u, v)_{L_2(Q)}| = |(L_D u, v)_{L_2(Q)} + (L_I u, v)_{L_2(Q)}| \leqslant |(L_D u, v)_{L_2(Q)}| + |(L_I u, v)_{L_2(Q)}| \leqslant C \|v\|_{W_{BR+}} \|u\|_{W_{BR}},$$

for some constant C .

Since the set C_{BR+}^∞ is dense in the space W_{BR+} , dividing the last inequality by $\|v\|_{W_{BR+}}$ and taking the supremum over v yields the desired inequality. $\square$

The following lemma is proved in a similar way.

Lemma 2. *There exists a constant $C > 0$ such that, for any function $v \in C_{BR+}^\infty$, the following inequality holds:*

$$C \|v\|_{W_{BR+}} \geqslant \|\mathcal{L}^* v\|_{W_{BR}^-}.$$

Remark 2. The inequalities established in Lemmas 1 and 2 allow the operators $\mathcal{L}$ and $\mathcal{L}^*$ to be extended by continuity from their domains of definition to the spaces W_{BR} and W_{BR+} , respectively. These inequalities remain valid for all $u \in W_{BR}$ and $v \in W_{BR+}$, while preserving the previous notation for the extended operators.

To prove the following inequalities, we shall use the following auxiliary results, established in (Anikushyn, & Andaral, 2025) and (Hulianytskyi, & Anikushyn, 2014).

Lemma 3. *Let $f \in C([0, T])$ and suppose that, for any $t, \tau \in [0, T]$, the inequality $|K(t, \tau)| \leqslant M$ holds. Then, for any constant $c > 0$, the following estimate is valid:*

$$\left| \int_0^T \int_0^t K(t, \tau) e^{c\tau} f(\tau) d\tau f(t) dt \right| \leqslant M \sqrt{\frac{T}{c}} \int_0^T e^{ct} f^2(t) dt.$$

Lemma 4. *Let $f \in C^1([0, T])$, $f(T) = 0$, and suppose that, for any $t, \tau \in [0, T]$, the inequality $|K(t, \tau)| \leqslant M$ holds. Then, for any constant $c > 0$, the following estimate is valid:*

$$\left| \int_0^T \int_0^t K(t, \tau) e^{c\tau} f'(\tau) d\tau f(t) dt \right| \leqslant \frac{MT}{c} \int_0^T e^{ct} (f'(t))^2 dt.$$

Lemma 5. *There exists a constant $C > 0$ such that, for any function $u \in W_{BR}$, the following inequality holds:*

$$\|\mathcal{L}u\|_{W_{BR+}^-} \geqslant C \|u\|_{H_{BR}}.$$

Proof. Consider any function $u \in C_{BR}^\infty$. Let

$$v(x, t) = -e^{ct} \int_T^t e^{-2c\tau} u(x, \tau) d\tau. \quad (19)$$

As in (Anikushyn, & Nazarchuk, 2024), we have $v \in W_{BR+}$. Differentiating with respect to t , we see that

$$u(x, t) = u^{(1)}(x, t) + u^{(2)}(x, t),$$

where

$$u^{(1)}(x, t) = -e^{ct} v_t(x, t), \quad (20)$$

$$u^{(2)}(x, t) = ce^{ct} v(x, t). \quad (21)$$

The positive constant c will be specified later. We now show that, for some positive number C , the inequality holds:

$$(\mathcal{L}u, v)_{L_2(Q)} \geqslant C \|v\|_{W_{BR+}}^2.$$

We first consider the scalar product $(L_D u, v)_{L_2(Q)}$. By applying integration by parts, the boundary condition (8), and noting that the function v defined by (19) satisfies $v|_{t=T} = 0$, we obtain

$$\begin{aligned} \int_Q au_t v dQ &= \int_Q (auv)_t dQ - \int_Q av_t u dQ = \int_\Omega \int_0^T (auv)_t dt d\Omega - \int_Q auv_t dQ = \\ &= \int_\Omega ((auv)|_{t=T} - (auv)|_{t=0}) d\Omega - \int_Q auv_t dQ = - \int_Q auv_t dQ = \end{aligned}$$

$$= - \int_Q au^{(1)}v_t dQ - \int_Q au^{(2)}v_t dQ. \quad (22)$$

Taking into account (19) and condition (8), we have

$$\begin{aligned} \sum_{i,j=1}^n \int_Q ((a_{ij}u_{x_j})_{x_i}v)_t dQ &= \sum_{i,j=1}^n \int_\Omega \int_0^T ((a_{ij}u_{x_j})_{x_i}v)_t dt d\Omega = \\ &= \sum_{i,j=1}^n \int_\Omega ((a_{ij}u_{x_j}|_{t=T})_{x_i}v|_{t=T} - (a_{ij}u_{x_j}|_{t=0})_{x_i}v|_{t=0}) d\Omega = 0. \end{aligned}$$

Now, using the Gauss formula and the last equality, we obtain

$$\begin{aligned} - \int_Q \sum_{i,j=1}^n (a_{ij}u_{x_j})_{x_i}v dQ &= - \int_Q \sum_{i,j=1}^n ((a_{ij}u_{x_j})_{x_i}v)_t dQ + \int_Q \sum_{i,j=1}^n (a_{ij}u_{x_j})_{x_i}v_t dQ = \\ &= - \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}v_{x_i}t dQ = - \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}^{(1)}v_{x_i}t dQ - \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}^{(2)}v_{x_i}t dQ. \end{aligned} \quad (23)$$

Similarly, applying the Gauss formula in the standard way to the terms obtained from the operator $\mathcal{B}$, we arrive at

$$\begin{aligned} (L_D u, v)_{L_2(Q)} &= - \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}^{(1)}v_{x_i}t dQ - \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}^{(2)}v_{x_i}t dQ - \\ &- \int_Q au^{(1)}v_t dQ - \int_Q au^{(2)}v_t dQ + \int_Q \sum_{i,j=1}^n b_{ij}u_{x_j}^{(1)}v_{x_i} dQ + \int_Q \sum_{i,j=1}^n b_{ij}u_{x_j}^{(2)}v_{x_i} dQ + \\ &+ \int_Q \sum_{i=1}^n b_i u_{x_i}^{(1)}v dQ + \int_Q \sum_{i=1}^n b_i u_{x_i}^{(2)}v dQ + \int_Q bu^{(1)}v dQ + \int_Q bu^{(2)}v dQ. \end{aligned}$$

In (Anikushyn, & Nazarchuk, 2024) it was shown that, taking into account condition (6) and the Poincaré and Cauchy inequalities, the terms involving $u^{(1)}$ can be estimated as

$$- \int_Q au^{(1)}v_t dQ = \int_Q ae^{ct}v_t^2 dQ \geq 0, \quad (24)$$

$$- \int_Q \sum_{i,j=1}^n a_{ij}u_{x_j}^{(1)}v_{x_i}t dQ = \int_Q \sum_{i,j=1}^n a_{ij}e^{ct}v_{x_j}t v_{x_i}t dQ \geq \alpha \int_Q \sum_{i=1}^n e^{ct}v_{x_i}^2 t dQ, \quad (25)$$

$$\int_Q bu^{(1)}v dQ = \frac{1}{2} \int_\Omega bv^2|_{t=0} d\Omega + \frac{c}{2} \int_Q be^{ct}v^2 dQ \geq -\frac{MC_p}{2} \sum_{i=1}^n \int_\Omega v_{x_i}^2|_{t=0} d\Omega - \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct}v_{x_i}^2 dQ, \quad (26)$$

$$\begin{aligned} \int_Q \sum_{i,j=1}^n b_{ij}u_{x_j}^{(1)}v_{x_i} dQ &= \frac{1}{2} \int_\Omega \sum_{i,j=1}^n b_{ij}v_{x_j}|_{t=0}v_{x_i}|_{t=0} d\Omega + \frac{c}{2} \int_Q \sum_{i,j=1}^n b_{ij}e^{ct}v_{x_j}v_{x_i} dQ \geq \\ &\geq -\frac{nM}{2} \sum_{i=1}^n \int_\Omega v_{x_i}^2|_{t=0} d\Omega - \frac{nMc}{2} \sum_{i=1}^n \int_Q e^{ct}v_{x_i}^2 dQ. \end{aligned} \quad (27)$$

For the term with b_i we apply integration by parts, obtaining

$$\begin{aligned} \int_Q \sum_{i=1}^n b_i u_{x_i}^{(1)}v dQ &= - \int_Q \sum_{i=1}^n b_i e^{ct}v_{x_i}t v dQ = - \int_Q \sum_{i=1}^n (b_i e^{ct}v_{x_i}v)_t dQ + \int_Q \sum_{i=1}^n (b_i e^{ct}v)_t v_{x_i} dQ = \\ &= \int_\Omega \sum_{i=1}^n b_i v_{x_i}|_{t=0}v|_{t=0} d\Omega + c \int_Q \sum_{i=1}^n b_i e^{ct}v_{x_i}v dQ + \int_Q \sum_{i=1}^n b_i e^{ct}v_{x_i}v_t dQ. \end{aligned} \quad (28)$$

Recalling that the constant M bounds all coefficients $|b|$, $\{|b_i|\}_{i=1}^n$, $\{|b_{ij}|\}_{i,j=1}^n$ in the domain Ω , and using the Poincaré and Cauchy inequalities, we estimate the terms in (28) as follows:

$$\begin{aligned} \int_\Omega \sum_{i=1}^n b_i v_{x_i}|_{t=0}v|_{t=0} d\Omega &\geq -M \int_\Omega \sum_{i=1}^n |v_{x_i}|_{t=0}v|_{t=0}| d\Omega \geq -\frac{M}{2} \sum_{i=1}^n \int_\Omega (v_{x_i}^2|_{t=0} + v^2|_{t=0}) d\Omega \geq \\ &\geq -\frac{M}{2} \sum_{i=1}^n \int_\Omega v_{x_i}^2|_{t=0} d\Omega - \frac{MC_p}{2} \sum_{i=1}^n \int_\Omega v_{x_i}^2|_{t=0} d\Omega, \end{aligned} \quad (29)$$

$$\begin{aligned} c \int_Q \sum_{i=1}^n b_i e^{ct} v_{x_i} v dQ &\geq -Mc \int_Q \sum_{i=1}^n e^{ct} |v_{x_i} v| dQ \geq -\frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} (v_{x_i}^2 + v^2) dQ \geq \\ &\geq -\frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ - \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ, \end{aligned} \quad (30)$$

$$\begin{aligned} \int_Q \sum_{i=1}^n b_i e^{ct} v_{x_i} v_t dQ &\geq -M \int_Q \sum_{i=1}^n e^{ct} |v_{x_i} v_t| dQ \geq -\frac{M}{2} \sum_{i=1}^n \int_Q e^{ct} \left(kv_{x_i}^2 + \frac{v_t^2}{k} \right) dQ \geq \\ &\geq -\frac{Mk}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ - \frac{MC_p}{2k} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ. \end{aligned} \quad (31)$$

Here $k > 0$ is an arbitrary number to be chosen later.

Overall, taking into account (24)–(31), for the terms with $u^{(1)}$ we obtain

$$\begin{aligned} - \int_Q \sum_{i,j=1}^n a_{ij} u_{x_j}^{(1)} v_{x_i} dQ - \int_Q a u^{(1)} v_t dQ - \int_Q \sum_{i,j=1}^n (b_{ij} u_{x_j}^{(1)})_{x_i} v dQ + \int_Q \sum_{i=1}^n b_i u_{x_i}^{(1)} v dQ + \int_Q b u^{(1)} v dQ &\geq \\ &\geq \alpha \int_Q \sum_{i=1}^n e^{ct} v_{x_i}^2 dQ - \left(\frac{MC_p}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} d\Omega + \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ \right) - \\ &- \left(\frac{nM}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} d\Omega + \frac{nMc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ \right) - \left(\frac{M}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} d\Omega + \frac{MC_p}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} d\Omega \right) - \\ &- \left(\frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ + \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ \right) - \left(\frac{Mk}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ + \frac{MC_p}{2k} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ \right). \end{aligned} \quad (32)$$

We now turn to the terms with $u^{(2)}$. Integrating by parts with respect to t and applying the boundary condition (11), for the term in (22) we have

$$\begin{aligned} - \int_Q a v_t u^{(2)} dQ &= -c \int_Q a v_t e^{ct} v dQ = -\frac{c}{2} \int_Q (a e^{ct} v^2)_t dQ + \frac{c^2}{2} \int_Q a e^{ct} v^2 dQ = \\ &= -\frac{c}{2} \int_Q \int_0^T (a e^{ct} v^2)_t dt d\Omega + \frac{c^2}{2} \int_Q a e^{ct} v^2 dQ = \\ &= -\frac{c}{2} \int_Q (a e^{cT} v^2|_{t=T} - a v^2|_{t=0}) d\Omega + \frac{c^2}{2} \int_Q a e^{ct} v^2 dQ = \\ &= \frac{c}{2} \int_Q a v^2|_{t=0} d\Omega + \frac{c^2}{2} \int_Q a e^{ct} v^2 dQ. \end{aligned} \quad (33)$$

Since $a(x) \geq 0$, this expression is non-negative.

Similarly, for the corresponding term in (23), we apply integration by parts with respect to t . This yields

$$\begin{aligned} -c \int_Q \sum_{i,j=1}^n a_{ij} e^{ct} v_{x_j} v_{x_i} dQ &= \frac{c}{2} \int_Q \sum_{i,j=1}^n (a_{ij} e^{ct} v_{x_j} v_{x_i})_t dQ + \frac{c^2}{2} \int_Q \sum_{i,j=1}^n a_{ij} e^{ct} v_{x_j} v_{x_i} dQ = \\ &= \frac{c}{2} \int_Q \sum_{i,j=1}^n a_{ij} v_{x_j} |_{t=0} v_{x_i} |_{t=0} d\Omega + \frac{c^2}{2} \int_Q \sum_{i,j=1}^n a_{ij} e^{ct} v_{x_j} v_{x_i} dQ. \end{aligned}$$

Applying the uniform ellipticity condition (7), we obtain the estimate

$$- \int_Q \sum_{i,j=1}^n a_{ij} u_{x_j}^{(2)} v_{x_i} dQ \geq \frac{\alpha c}{2} \int_Q \sum_{i=1}^n v_{x_i}^2|_{t=0} d\Omega + \frac{\alpha c^2}{2} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ. \quad (34)$$

The remaining terms involving $u^{(2)}$ are estimated directly using the Cauchy and Poincaré inequalities:

$$\int_Q b u^{(2)} v dQ = c \int_Q b e^{ct} v^2 dQ \geq -cMC_p \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ, \quad (35)$$

$$- \int_Q \sum_{i,j=1}^n (b_{ij} u_{x_j}^{(2)})_{x_i} v dQ = \int_Q \sum_{i,j=1}^n b_{ij} u_{x_j}^{(2)} v_{x_i} dQ = c \int_Q \sum_{i,j=1}^n b_{ij} e^{ct} v_{x_j} v_{x_i} dQ \geq -nMc \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 dQ. \quad (36)$$

The term with b_i is estimated similarly to (30):

$$\int_Q \sum_{i=1}^n b_i u_{x_i}^{(2)} v \, dQ = c \int_Q \sum_{i=1}^n b_i e^{ct} v_{x_i} v \, dQ \geq -\frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ. \quad (37)$$

In view of (33)–(37), for the terms containing $u^{(2)}$, we obtain

$$\begin{aligned} & - \int_Q \sum_{i,j=1}^n a_{ij} u_{x_j}^{(2)} v_{x_i} \, dQ - \int_Q a u^{(2)} v_t \, dQ - \int_Q \sum_{i,j=1}^n (b_{ij} u_{x_j}^{(2)})_{x_i} v \, dQ + \int_Q \sum_{i=1}^n b_i u_{x_i}^{(2)} v \, dQ + \int_Q b u^{(2)} v \, dQ \geq \\ & \geq \frac{\alpha c}{2} \int_Q \sum_{i=1}^n v_{x_i}^2|_{t=0} \, d\Omega + \frac{\alpha c^2}{2} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ - cMC_p \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \\ & - nMc \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \left(\frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ + \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ \right). \end{aligned} \quad (38)$$

Thus, adding Equations (32) and (38), we obtain

$$\begin{aligned} (L_D u, v)_{L_2(Q)} & \geq -\frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - cMC_p \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{nMc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \\ & - nMc \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \\ & - \frac{cMC_p}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \frac{Mc}{2} \sum_{i=1}^n \int_Q e^{ct} v_{x_i}^2 \, dQ - \\ & - \frac{MC_p}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} \, d\Omega - \frac{nM}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} \, d\Omega - \frac{M}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} \, d\Omega - \\ & - \frac{MC_p}{2} \sum_{i=1}^n \int_Q v_{x_i}^2|_{t=0} \, d\Omega + \frac{\alpha c}{2} \int_Q \sum_{i=1}^n v_{x_i}^2|_{t=0} \, d\Omega + \frac{\alpha c^2}{2} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ - \\ & - \frac{MC_p}{2k} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ + \alpha \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ = \\ & = \frac{1}{2} (\alpha c^2 - Mc(5C_p + 3n + 2) - Mk) \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ + \\ & + \frac{1}{2} (\alpha c - M(2C_p + n + 1)) \int_Q \sum_{i=1}^n v_{x_i}^2|_{t=0} \, d\Omega + \left(\alpha - \frac{MC_p}{2k} \right) \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ. \end{aligned}$$

Choosing $k = \frac{MC_p}{\alpha}$ and a sufficiently large c such that $\alpha c \geq M(2C_p + n + 1)$, we have

$$(L_D u, v)_{L_2(Q)} \geq \frac{\alpha}{2} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ + \frac{1}{2} \left(\alpha c^2 - Mc(5C_p + 3n + 2) - \frac{M^2 C_p}{\alpha} \right) \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ. \quad (39)$$

Next, consider the scalar product $(L_I u, v)_{L_2(Q)}$. Applying the Gauss formula to the terms containing $K_{ij}(x, t, \tau)$ and $K_i(x, t, \tau)$, and taking into account the initial-boundary conditions, we have

$$\begin{aligned} (L_I u, v)_{L_2(Q)} & = - \sum_{i,j=1}^n \int_Q \int_0^t K_{ij}(x, t, \tau) u_{x_i}^{(1)}(x, \tau) \, d\tau v_{x_i}(x, t) \, dQ - \sum_{i,j=1}^n \int_Q \int_0^t K_{ij}(x, t, \tau) u_{x_i}^{(2)}(x, \tau) \, d\tau v_{x_i}(x, t) \, dQ + \\ & + \sum_{i=1}^n \int_Q \int_0^t K_i(x, t, \tau) u_{x_i}^{(1)}(x, \tau) \, d\tau v(x, t) \, dQ + \sum_{i=1}^n \int_Q \int_0^t K_i(x, t, \tau) u_{x_i}^{(2)}(x, \tau) \, d\tau v(x, t) \, dQ + \\ & + \int_Q \int_0^t K(x, t, \tau) u^{(1)}(x, \tau) \, d\tau v(x, t) \, dQ + \int_Q \int_0^t K(x, t, \tau) u^{(2)}(x, \tau) \, d\tau v(x, t) \, dQ. \end{aligned}$$

Using Lemmas 3 and 4, together with the Cauchy-Schwarz and Poincaré inequalities for each $x \in \Omega$, we estimate the terms of the integral part of the operator.

For the terms with $u^{(1)}$, we obtain

$$\left| \sum_{i,j=1}^n \int_Q \int_0^t K_{ij}(x, t, \tau) e^{c\tau} v_{x_i \tau}(x, \tau) \, d\tau v_{x_j}(x, t) \, dQ \right| \leq \frac{nMT}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 \, dQ, \quad (40)$$

$$\left| \sum_{i=1}^n \int_Q \int_0^t K_i(x, t, \tau) e^{c\tau} v_{x_i\tau}(x, \tau) d\tau v(x, t) dQ \right| \leq \frac{MTC_p}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ, \quad (41)$$

$$\left| \sum_{i=1}^n \int_Q \int_0^t K(x, t, \tau) e^{c\tau} v_\tau(x, \tau) d\tau v(x, t) dQ \right| \leq \frac{MT}{c} \int_Q e^{ct} v_t^2 dQ \leq \frac{MTC_p}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ. \quad (42)$$

For the terms with $u^{(2)}$, we have

$$\left| \sum_{i,j=1}^n \int_Q \int_0^t K_{ij}(x, t, \tau) e^{c\tau} cv_{x_i}(x, \tau) d\tau v_{x_j}(x, t) dQ \right| \leq nM\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ, \quad (43)$$

$$\left| \sum_{i=1}^n \int_Q \int_0^t K_i(x, t, \tau) e^{c\tau} cv_{x_i}(x, \tau) d\tau v(x, t) dQ \right| \leq MC_p\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ, \quad (44)$$

$$\left| \sum_{i=1}^n \int_Q \int_0^t K(x, t, \tau) e^{c\tau} cv(x, \tau) d\tau v(x, t) dQ \right| \leq cM\sqrt{\frac{T}{c}} \int_Q e^{ct} v^2 dQ \leq MC_p\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ. \quad (45)$$

Combining (39) and (40)–(45), we obtain

$$\begin{aligned} (\mathcal{L}u, v)_{L_2(Q)} &= (L_D u, v)_{L_2(Q)} + (L_I u, v)_{L_2(Q)} \geq (L_D u, v)_{L_2(Q)} - |(L_I u, v)_{L_2(Q)}| \geq \\ &\geq \frac{\alpha}{2} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ + \frac{1}{2} \left(\alpha c^2 - Mc(5C_p + 3n + 2) - \frac{M^2 C_p}{\alpha} \right) \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ - \\ &- \frac{nMT}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ - \frac{MTC_p}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ - \frac{MTC_p}{c} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ - \\ &- nM\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ - MC_p\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ - MC_p\sqrt{T}c \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ = \\ &= \left(\frac{\alpha}{2} - \frac{MT}{c} (2C_p + n) \right) \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ + \\ &+ \frac{1}{2} \left(\alpha c^2 - Mc(5C_p + 3n + 2) - 2M\sqrt{T}c(2C_p + n) - \frac{M^2 C_p}{\alpha} \right) \int_Q e^{ct} \sum_{i=1}^n v_{x_i}^2 dQ. \end{aligned}$$

Choosing c sufficiently large, we have

$$(\mathcal{L}u, v)_{L_2(Q)} \geq \frac{\alpha}{4} \int_Q e^{ct} \sum_{i=1}^n v_{x_i t}^2 dQ \geq C \|v\|_{W_{BR+}}^2. \quad (46)$$

Note that u and v satisfy the inequality

$$\|u\|_{H_{BR}}^2 \leq 2e^{2cT} (1 + c^2 C_p) \|v\|_{W_{BR+}}^2.$$

Therefore, in view of (46) and, in a standard way, applying the Schwarz inequality, we obtain

$$C_1 \|u\|_{H_{BR}} \|v\|_{W_{BR+}} \leq C \|v\|_{W_{BR+}}^2 \leq (\mathcal{L}u, v)_{L_2(Q)} \leq \|\mathcal{L}u\|_{W_{BR+}^-} \|v\|_{W_{BR+}},$$

for some constant C_1 . This completes the proof for all $u \in C_{BR}^\infty$. The statement of the lemma for all other functions $u \in W_{BR}$ is obtained by passing to the limit, as described in (Anikushyn, & Andaral, 2025). $\square$

Lemma 6. *There exists a constant $C > 0$ such that for any function $v \in W_{BR+}$ the following inequality holds:*

$$\|\mathcal{L}^* v\|_{W_{BR}^-} \geq C \|v\|_{H_{BR+}}.$$

Proof. The proof is analogous to that of Lemma 5, considering the auxiliary function

$$u(x, t) = e^{-ct} \int_0^t e^{2c\tau} v(x, \tau) d\tau$$

and using analogues of Lemmas 3 and 4. $\square$

Discussion and conclusions

We extend the analysis of pseudoparabolic integro-differential equations with Volterra-type terms to a wider class that includes additional terms with coefficients b_i in the differential operator $\mathcal{B}$ and kernels K_{ij} , K_i , and K in the integral part. A key contribution of this work is the establishment of a priori estimates in negative norms under the structural assumptions

of (Anikushyn, & Nazarchuk, 2024), namely the self-adjointness of the differential part, the non-negative definiteness of $\{a_{ij}\}_{i,j=1}^n$, and the non-negativity of a . Crucially, we achieve this generalization without requiring the non-negative definiteness of the differential operator $\mathcal{B}$. This significantly extends the framework presented in the recent work (Anikushyn, & Andaral, 2025), where such non-negativity was assumed, thus broadening the applicability of these results. These estimates enable us to extend the theorems on the well-posedness of the initial-boundary value problem and the controllability results from (Anikushyn, & Nazarchuk, 2024) and (Hulianytskyi, & Semenov, 2012).

Although this issue is not addressed in the present paper, the same a priori estimates make it possible to extend the result concerning the existence of an optimal control to our setting, in a manner analogous to that presented in (Anikushyn, & Nazarchuk, 2024).

Acknowledgments. The author would like to thank the anonymous referee(s) for their review of the manuscript.

Sources of funding. The study was partially financially supported by Taras Shevchenko National University of Kyiv.

References

- Amar, M., Andreucci, D., Gianni, R., & Timofte, C. (2019). A degenerate pseudo-parabolic equation with memory. *Communications in Applied and Industrial Mathematics*, 10(1), 71–77. <https://doi.org/10.2478/caim-2019-0013>
- Anikushyn, A. (2014). Generalized optimal control for systems described by linear integro-differential equations with nonnegative definite integral operators. *Journal of Automation and Information Science*, 46(6), 58–67. <https://doi.org/10.1615/JAutomatInfScien.v46.i6.60>
- Anikushyn, A. (2024). Pulse and point optimal control of linear hyperbolic integro-differential systems with mixed boundary conditions. *Journal of Optimization, Differential Equations and Their Applications*, 32(2), 137–155. <https://dx.doi.org/10.15421/142412>
- Anikushyn, A. (2025). Mass transfer processes with memory effect: Optimal control and modelling. In A. Luntovskyy, M. Klymash, I. Melnyk, M. Beshley, & D. Gütter (Eds.), *Networks and Sustainability. TCSET 2024. Lecture Notes in Electrical Engineering*, vol 1473. (pp. 521–545). Springer, Cham. https://doi.org/10.1007/978-3-032-02272-1_23
- Anikushyn, A., & Andaral, A. (2025). Generalized optimal control of pseudo-parabolic integro-differential systems. *Journal of Mathematical Sciences (United States)*, 291(2), 901–918. <https://doi.org/10.1007/s10958-025-07867-1>
- Anikushyn, A., Lyashko, V., & Samosonok, O. (2024). Optimal control of the systems governed by linear hyperbolic integro-differential equations. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics*, 78(1), 111–118. <https://doi.org/10.17721/1812-5409.2024/1.22>
- Anikushyn, A., & Nazarchuk, V. (2024). A priori inequalities and generalized optimal control of pseudo-parabolic integro-differential systems with non-coercive diffusion term. *Nonlinear Oscillations*, 27(4), 454–468 [in Ukrainian]. [Анікушин, А., & Назарчук, В. (2024). Априорні нерівності та узагальнене оптимальне керування псевдопараболічними інтегро-диференціальними системами з некоерцитивним дифузійним членом. *Нелінійні коливання*, 27(4), 454–468]. <https://doi.org/10.3842/nosc.v27i4.1486>
- Bertsch, M., Smarrazzo, F., & Tesei, A. (2013). Pseudoparabolic regularization of forward-backward parabolic equations: A logarithmic nonlinearity. *Analysis & PDE*, 6(7), 1719–1754. <https://doi.org/10.2140/apde.2013.6.1719>
- Bertsch, M., Smarrazzo, F., & Tesei, A. (2016). Pseudo-parabolic regularization of forward-backward parabolic equations: Power-type nonlinearities. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2016(712), 51–80. <https://doi.org/10.1515/crelle-2013-0123>
- Cannarsa, P., & Sforza, D. (2011). Integro-differential equations of hyperbolic type with positive definite kernels. *Journal of Differential Equations*, 250(12), 4289–4335. <https://doi.org/10.1016/j.jde.2011.03.005>
- Cannon, J. R., & Lin, Y. (1988). Classical and weak solutions for one-dimensional pseudo-parabolic equations with typical boundary data. *Annali di Matematica Pura ed Applicata*, 152(1), 375–385. <https://doi.org/10.1007/BF01766158>
- Caputo, M., & Mainardi, F. (1971). Linear models of dissipation in anelastic solids. *La Rivista del Nuovo Cimento*, 1, 161–198. <https://doi.org/10.1007/BF02820620>
- Conti, M., Gatti, S., & Pata, V. (2008). Uniform decay properties of linear volterra integro-differential equations. *Mathematical Models and Methods in Applied Sciences*, 18(1), 21–45. <https://doi.org/10.1142/S0218202508002590>
- Cueto-Felgueroso, L., & Juanes, R. (2009). A phase field model of unsaturated flow. *Water Resources Research*, 45(10), W10409. <https://doi.org/10.1029/2009WR007945>
- Demchenko, H., Anikushyn, A., & Pokojovy, M. (2019). On a kelvin-voigt viscoelastic wave equation with strong delay. *SIAM Journal on Mathematical Analysis*, 51(6), 4382–4412. <https://doi.org/10.1137/18M1219308>
- Di, H., Shang, Y., & Yu, J. (2020). Blow-up analysis of a nonlinear pseudo-parabolic equation with memory term. *AIMS Mathematics*, 5(4), 3408–3422. <https://doi.org/10.3934/math.2020220>
- Di, H., Shang, Y., & Zheng, X. (2016). Global well-posedness for a fourth order pseudo-parabolic equation with memory and source terms. *Discrete and Continuous Dynamical Systems - B*, 21(3), 781–801. <https://doi.org/10.3934/dcdsb.2016.21.781>
- Duvaut, G., & Lions, J.-L. (1976). *Inequalities in mechanics and physics*. Springer. <https://dx.doi.org/10.1007/978-3-642-66165-5>
- Ezzinbi, K., & Staili, Y. (2025). Viability for a class of partial functional differential equations with non-dense domain. *Journal of Mathematical Analysis and Applications*, 548(1), 129350. <https://doi.org/10.1016/j.jmaa.2025.129350>
- Fabrizio, M., & Morro, A. (1992). *Mathematical problems in linear viscoelasticity*. Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611970807>
- Gal, C., & Warma, M. (2020). *Fractional-in-time semilinear parabolic equations and applications*. Springer Cham. <https://doi.org/10.1007/978-3-030-45043-4>
- Guo, W., Feng, Y., & Li, W. (2025). Stabilization of coupled nonlinear systems with lévy noise and time-varying delays. *Communications in Nonlinear Science and Numerical Simulation*, 149, 108878. <https://doi.org/10.1016/j.cnsns.2025.108878>
- Hranishak, K., & Anikushyn, A. (2024). Point control of linear hyperbolic integro-differential systems. *Bulletin of Taras Shevchenko National University of Kyiv. Physical and Mathematical Sciences*, 79(2), 20–28. <https://doi.org/10.17721/1812-5409.2024/2.4>
- Hranishak, K., Anikushyn, A., Lyashko, V., & Samosonok, O. (2025). Optimal control of hyperbolic integro-differential systems. *Cybernetics and Systems Analysis*, 61(1), 66–76. <https://doi.org/10.1007/s10559-025-00744-4>
- Hulianytskyi, A., & Anikushyn, A. (2014). Generalized solvability of parabolic integro-differential equations. *Differential Equations*, 50(1), 98–109. <https://doi.org/10.1134/S0012266114010133>
- Hulianytskyi, A., & Semenov, V. (2012). Integro-differential pseudoparabolic systems: a priori estimates and impulse-point controllability. *Proceedings of the Academy of Sciences of Ukraine*, 4, 43–49 [in Ukrainian]. [Гуляницький, А., & Семенов, В. (2012). Інтегро-диференціальні системи псевдопараболічного типу: априорні оцінки та імпульсно-точкова керованість. *Доповіді НАН України*, 4, 43–49].

- Jhangeer, A., Ibraheem, F., Jamal, T., Rahimzai, A., & Khan, I. (2024). Investigating pseudo parabolic dynamics through phase portraits, sensitivity, chaos and soliton behavior. *Scientific Reports*, 14(1), 15224. <https://doi.org/10.1038/s41598-024-64985-7>
- Jin, B., & Zhou, Z. (2024). *Numerical treatment and analysis of time-fractional evolution equations*. Springer Cham. <https://doi.org/10.1007/978-3-031-21050-1>
- Kubica, A., Ryszewska, K., & Yamamoto, M. (2020). *Time-fractional differential equations. a theoretical introduction*. Springer Singapore. <https://doi.org/10.1007/978-981-15-9066-5>
- Lakes, R. (1998). *Viscoelastic solids*. CRC Press Revivals. <https://doi.org/10.1201/9781315121369>
- Liu, H., Kim, E.-S., & Fang, Z. B. (2025). On a semilinear pseudo-parabolic equation with nonlinear convolution terms. *Nonlinear Analysis: Real World Applications*, 84, 104307. <https://doi.org/10.1016/j.nonrwa.2024.104307>
- Liu, X., Shi, J., Wang, C., & Fan, D. (2025). Global boundedness of solutions to a class of partial differential equations with time delay. *Journal of Differential Equations*, 435, 113232. <https://doi.org/10.1016/j.jde.2025.113232>
- Lyashko, S. (2002). *Generalized optimal control of linear systems with distributed parameters*. Kluwer Academic Publishers. <https://doi.org/10.1007/b130433>
- Lyashko, S., & Nomirovskii, D. (2003). Generalized solutions and optimal controls in systems describing the dynamics of a viscous stratified fluid. *Differential Equations*, 39(1), 90–98. <https://doi.org/10.1023/A:1025176109884>
- Lyashko, S., Nomirovskii, D., & Sergienko, T. (2001). Trajectory and final controllability in hyperbolic and pseudohyperbolic systems with generalized actions. *Cybernetics and Systems Analysis*, 37(5), 756–763. <https://doi.org/10.1023/A:1013871026026>
- Lyashko, S., & Nomirovskiy, D. (2003). The generalized solvability and optimization of parabolic systems in domains with thin low-permeable inclusions. *Cybernetics and Systems Analysis*, 39(5), 737–745. <https://doi.org/10.1023/B:CASA.0000012094.62199.de>
- Lyashko, S., & Semenov, V. (2001). Controllability of linear distributed systems in classes of generalized actions. *Cybernetics and Systems Analysis*, 37(1), 13–32. <https://doi.org/10.1023/A:1016607831284>
- Lyashko, S., Semenov, V., & Sergienko, T. (2002). Controllability and optimization of pseudohyperbolic systems. *Cybernetics and Systems Analysis*, 38(4), 586–596. <https://doi.org/10.1023/A:1021162303764>
- Mainardi, F. (1996). Fractional relaxation-oscillation and fractional diffusion-wave phenomena. *Chaos, Solitons & Fractals*, 7(9), 1461–1477. [https://doi.org/10.1016/0960-0779\(95\)00125-5](https://doi.org/10.1016/0960-0779(95)00125-5)
- Mainardi, F., Mura, A., Pagnini, G., & Goreflo, R. (2006). Fractional relaxation and time-fractional diffusion of distributed order. *IFAC Proceedings Volumes*, 39(11), 1–21. <https://doi.org/10.3182/20060719-3-PT-4902.00002>
- Maqbul, M., & Raheem, A. (2020). Time-discretization schema for a semilinear pseudo-parabolic equation with integral conditions. *Applied Numerical Mathematics*, 148, 18–27. <https://doi.org/10.1016/j.apnum.2019.09.002>
- Metzler, R., & Klafter, J. (2000). The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1), 1–77. [https://doi.org/10.1016/S0370-1573\(00\)00070-3](https://doi.org/10.1016/S0370-1573(00)00070-3)
- Miller, T., Tam, A. K., Marangell, R., Wechselberger, M., & Bradshaw-Hajek, B. H. (2024). Analytic shock-fronted solutions to a reaction–diffusion equation with negative diffusivity. *Studies in Applied Mathematics*, 153(1), e12685. <https://doi.org/10.1111/sapm.12685>
- Miller, T., Tam, A. K., Marangell, R., Wechselberger, M., & Bradshaw-Hajek, B. H. (2025). Shock selection in reaction–diffusion equations with partially negative diffusivity using nonlinear regularisation. *Physica D: Nonlinear Phenomena*, 474, 134561. <https://doi.org/10.1016/j.physd.2025.134561>
- Nomirovskii, D. (2007). Generalized solvability and optimization of a parabolic system with a discontinuous solution. *Journal of Differential Equations*, 233(1), 1–21. <https://doi.org/10.1016/j.jde.2006.09.025>
- Nomirovskiy, D. (2004). Generalized solvability of parabolic systems with nonhomogeneous transmission conditions of nonideal contact type. *Differential Equations*, 40(10), 1467–1477. <https://doi.org/10.1007/s10625-004-0013-1>
- Nomirovskiy, D. (2006). Unique solvability of pseudohyperbolic equations with singular right-hand sides. *Mathematical Notes*, 80(3), 550–562. <https://doi.org/10.1007/s11006-006-0174-8>
- Nomirovskiy, D., & Anikushyn, A. (2011). Generalized solutions for linear operators with weakened a priori inequalities. *Ukrainian Mathematical Journal*, 62(8), 1175–1186. <https://doi.org/10.1007/s11253-011-0435-x>
- Nomirovskiy, D., & Vostrikov, O. (2016). Generalized statements and properties of models of transport processes in domains with cuts. *Cybernetics and Systems Analysis*, 52(6), 931–942. <https://doi.org/10.1007/s10559-016-9895-1>
- Pokojoy, M., & Anikushyn, A. (2017). Multidimensional thermoelasticity for nonsimple materials—well-posedness and long-time behavior. *Applicable Analysis*, 96(9), 1561–1585. <https://doi.org/10.1080/00036811.2017.1295447>
- Pokojoy, M., & Anikushyn, A. (2019). Global well-posedness and exponential stability for heterogeneous anisotropic maxwell's equations under a nonlinear boundary feedback with delay. *Journal of Mathematical Analysis and Applications*, 475(1), 278–312. <https://doi.org/10.1016/j.jmaa.2019.02.042>
- Popescu, M. N., & Dietrich, S. (2004). Model for spreading of liquid monolayers. *Physical Review E*, 69, 061602. <https://doi.org/10.1103/PhysRevE.69.061602>
- Ragab, M., Abouelregal, A. E., AlShaibi, H. F., & Mansouri, R. A. (2021). Heat transfer in biological spherical tissues during hyperthermia of magnetoma. *Biology*, 10(12), 1259. <https://doi.org/10.3390/biology10121259>
- Renardy, M., Hrusa, W., & Nohel, J. (1987). *Mathematical problems in viscoelasticity*. Longman Science & Technology.
- Salhi, J. (2025). Exponential stability for degenerate/singular hyperbolic equations with delayed boundary feedback. *Acta Applicandae Mathematicae*, 198(1), 2. <https://doi.org/10.1007/s10440-025-00737-7>
- Tymchyshyn, I., & Nomirovskii, D. (2021). Generalized solvability of a parabolic model describing transfer processes in domains with thin inclusions. *Differential Equations*, 57(8), 1053–1062. <https://doi.org/10.1134/S0012266121080097>
- Vromans, A. J., van de Ven, F., & Muntean, A. (2019). Parameter delimitation of the weak solvability for a pseudo-parabolic system coupling chemical reactions, diffusion and momentum equations. *Advances in Mathematical Sciences and Applications*, 28(2), 273–311.
- Wang, K. (2024). Two-grid methods for nonlinear pseudo-parabolic integro-differential equations by finite element method. *Computers & Mathematics with Applications*, 168, 174–189. <https://doi.org/10.1016/j.camwa.2024.05.032>
- Wu, F., & Yin, G. (2025). Two-time-scale stochastic functional differential equations: Inclusion of infinite delay and coupled segment processes. *Journal of Differential Equations*, 435, 113238. <https://doi.org/10.1016/j.jde.2025.113238>
- Zhang, Y., & Li, X. (2025). A meshless finite point method for a class of parabolic differential equation of neutral delay. *Engineering Analysis with Boundary Elements*, 178, 106276. <https://doi.org/10.1016/j.enganabound.2025.106276>

Отримано редакцією журналу / Received: 21.08.25
 Прорецензовано / Revised: 12.02.26
 Схвалено до друку / Accepted: 16.04.26
 Опубліковано / Published: 05.06.26

Вікторія НАЗАРЧУК, магістр
ORCID ID: 0009-0007-2252-4584
e-mail: nazarchukviktoria@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

КЕРОВАНІСТЬ ПСЕВДОПАРАБОЛІЧНИМИ ІНТЕГРО-ДИФЕРЕНЦІАЛЬНИМИ СИСТЕМАМИ З ПАМ'ЯТТЮ

Досліджено лінійні псевдопараболічні інтегро-диференціальні рівняння з інтегральним членом типу Вольтерра, що забезпечують ефективний математичний апарат для опису динамічних процесів із пам'яттю та нелокальними взаємодіями. Такі математичні моделі є характерними для різних напрямів прикладних досліджень, де еволюція системи залежить не лише від поточного стану, але й від її попередніх станів. Типові приклади охоплюють поведінку в'язкопружних матеріалів (таких як полімери, біополімери та високоміцні метали за високих температур), процеси фільтрації підземних вод і перенесення домішок у пористих середовищах, а також задачі, пов'язані з перенесенням тепла й поживних речовин у тканинах.

Метою роботи є узагальнення відомих теоретичних результатів на ширший клас псевдопараболічних рівнянь шляхом введення додаткових членів як у диференціальну, так і в інтегральну частини операторів. Для досягнення цієї мети використано методику апріорних оцінок у негативних нормах, розроблену С. І. Ляшком та його колегами. Запропонований підхід дає змогу досліджувати коректність постановки задач, узагальнену розв'язність, а також питання керованості й оптимального керування для систем із розподіленими параметрами.

Основним результатом дослідження є встановлення апріорних оцінок у негативних нормах, що враховують нові члени, введені до псевдопараболічного інтегро-диференціального рівняння. Ключова відмінність цієї роботи від попередніх полягає в тому, що отримані оцінки доведено без накладання вимог невід'ємності на члени нижчого порядку диференціального оператора. Це суттєво розширює застосовність теорії, даючи можливість поширити теореми про коректність початково-крайових задач, а також результати щодо керованості й існування оптимального керування на ширший клас систем псевдопараболічного типу.

Ключові слова:

псевдопараболічне інтегро-диференціальне рівняння, оператор Вольтерра, ефект пам'яті, апріорні оцінки, коректність постановки, узагальнена розв'язність, керованість, оптимальне керування.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 539.3

DOI: <https://doi.org/10.17721/1812-5409.2026/1.21>

Igor ULITKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-6452-141X
email: igorulitko@knu.ua

Oleksandr BORYSEIKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0009-0001-3023-5332
email: boryseiko@knu.ua

Iryna LEBEDYEVA¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-7150-1310
email: lebedyevaiv@knu.ua

Oleksandr KURLKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-1711-7083
email: alexandr.kurylko@knu.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

RESONANCE PROPERTIES OF A PIEZOELECTRIC GYROSCOPIC SENSOR BASED ON BULK ACOUSTIC SHEAR WAVES

The article develops a mathematical model of a piezoelectric gyroscopic sensor based on bulk acoustic shear waves propagating through the thickness of a layer with piezoelectric anisotropy perpendicular to the electrical load. The layer rotates around an axis perpendicular to its lateral faces, and the excited shear waves have the properties of Coriolis dispersion. The obtained conjugate equations of motion are one-dimensional through the layer thickness only in the case of the absence of mechanical forces at the boundaries and homogeneous conditions of the electrical load.

An analytical solution to the problem is constructed, and exact formulas for the distributions of stresses and displacements through the layer thickness in traveling shear waves with elliptical polarization of coupled components in the median plane are obtained. In accordance with the exact solution for the thickness component of the electric displacement field, an expression for the conduction current function in the external circuit was obtained, and the amplitude-frequency characteristic of the gyroscopic sensor was constructed on the primary (excited) and secondary (sensor) modes. It was determined that on the first main resonance of the layer, the efficiency of the excited mode is about 24 percent, and the efficiency of the sensor mode is about 1.8 percent at a value of the normalized angular velocity of the order of 0.01.

Key words: shear waves, piezoelectric effect, resonance, Coriolis forces, wave gyroscopes.

AMS 2020 classification: 74F99, 74H05, 74J05, 74J10.

Introduction

In inertial navigation systems that are subjected to significant shock and vibration effects during operation, it is proposed to use solid-state gyroscopic sensors based on bulk acoustic waves (BAWs), built on the effects of Coriolis dispersion of elastic waves (Shoenberg, & Censor, 1973; Ulitko, 1995). One such effect is a change in the direction of the shear wave polarization vector over a finite length of the waveguide, which is proportional to the magnitude of the measured angular velocity (Sarapuloff, & Ulitko, 2001; Gribkova, Peregudov, & Shevelko, 2011; Sarapuloff, 2013; Liu et al., 2024). In the works mentioned above, additional piezoquartz emitters and receivers or complex spatial systems of sectioned electrodes for electrostatic excitation are used to excite and detect waves in fused quartz and silicon carbide elements.

Another way to create BAWs sensors is to use piezoelectric materials in their sensing elements and, accordingly, the phenomenon of the piezoelectric effect for excitation and detection of waves (Yang, Dejous, & Hallil, 2023). In this case, gyroscopes based on BAWs do not require any additional means for generating and receiving a signal, except the electrode coatings of the piezocrystal, which simplifies and significantly reduces the cost of mass production of such sensors. If we follow the methods of developing other types of piezoresonant sensors, the main factor that requires analysis is the function of the conductivity current or electrical impedance and its dependence on the resonant frequencies of the primary (exciting) and secondary (sensing) modes of the gyroscope, i.e., the amplitude-frequency characteristic (AFC). The relevant questions are: how far are the primary and secondary resonant frequencies of a certain resonance spaced from each other, and also how efficient energetically is the excitation of the secondary frequency compared to the primary?

The aim of the research was to establish the conduction current function and the energy electromechanical coupling coefficient (EMCC) (Ulitko, 1977) as consequences of the complete analytical solutions of the conjugate electroelasticity problem in the mathematical model of a layer with piezoelectric anisotropy perpendicular to the loading electric field.

1. Problem formulation for a piezoelectric layer under homogeneous electrical excitation

Let us suggest that the flat side faces of the layer $z = \pm h$ (see Fig. 1) are not mechanically loaded

$$\sigma_z|_{z=\pm h} = 0, \quad \tau_{zy}|_{z=\pm h} = 0, \quad \tau_{zx}|_{z=\pm h} = 0, \quad (1)$$

and let the uniform loading electric field be applied to the electrodes that completely cover these faces

$$E_z|_{z=\pm h} = \pm \frac{V_0}{h} \exp(i\omega t), \quad E_x = 0, \quad E_y = 0 \quad (2)$$

with the frequency ω and the electric potential difference $2V_0$. The piezoelectric material is assumed to be transversely isotropic and has an anisotropy axis, denoted by the dipole symbol in Fig. 1. The coordinate system $Oxyz$ is a relative reference system; the axis of rotation Oz does not change its position in space.

© Ulitko Igor, Boryseiko Oleksandr, Lebedyeva Iryna, Kurylko Oleksandr, 2026

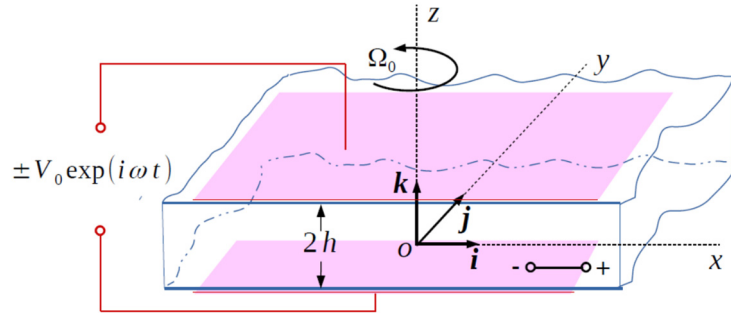


Fig. 1. To the formulation of the problem of harmonic oscillations of a uniformly rotating layer

Based on the analysis of the piezoelectric effect equations for this type of material anisotropy (Berlincourt, Curran, & Jaffe, 1964) under mechanical (1) and electrical (2) boundary conditions, it follows that the electroelastic state of the layer is determined by the shear stresses τ_{xz} , τ_{yz} , and displacements that are independent of the coordinates of the midplane x , y and given by

$$\mathbf{u}(z, t) = \mathbf{i}u_x(z, t) + \mathbf{j}u_y(z, t). \quad (3)$$

The component of displacement $u_z(z, t)$ and normal stresses σ_z , which are not affected by the inertia forces of the rotating motion, and the corresponding equations describing the thickness fluctuations of the layer, will be excluded from further consideration.

The equations for shear waves are derived from the linear equations of the piezoelectric effect, the equations of Newton's dynamics and Maxwell's electrostatics, and the linear kinematic formulas of Cauchy (Grinchenko, Ulitko, & Shulga, 1989). Taking into account the components of relative, translational, and Coriolis accelerations in the right-hand parts, we write them in projections on the axis Ox, Oy

$$\begin{aligned} Ox: \quad & c_{44}^E \frac{\partial^2 u_x}{\partial z^2} + e_{15} \frac{\partial^2 \Psi}{\partial z^2} = \rho(\ddot{u}_x - \Omega_0^2 u_x - 2\Omega_0 \dot{u}_y), \\ Oy: \quad & \frac{c_{11}^E - c_{12}^E}{2} \frac{\partial^2 u_y}{\partial z^2} = \rho(\ddot{u}_y - \Omega_0^2 u_y + 2\Omega_0 \dot{u}_x), \\ \Psi: \quad & \varepsilon_{11}^S \frac{\partial^2 \Psi}{\partial z^2} = e_{15} \frac{\partial^2 u_x}{\partial z^2}. \end{aligned} \quad (4)$$

Equations (4) form a conjugate system of electroelasticity equations for a piezoelectric layer that rotates uniformly with an angular velocity Ω_0 . The first and second equations are conjugate for components $-2\Omega_0 \dot{u}_y$ and $2\Omega_0 \dot{u}_x$ of Coriolis acceleration, as well as the first and third equations are conjugate for electric potential Ψ . The derivative $\partial^2 \Psi / \partial z^2$ can be expressed from the third equation in terms of $\partial^2 u_x / \partial z^2$ and substituted into the first equation. Consequently, we obtain the following:

$$\frac{\partial^2 u_x}{\partial z^2} = \frac{1}{c_{2,1}^2} (\ddot{u}_x - \Omega_0^2 u_x - 2\Omega_0 \dot{u}_y), \quad \frac{\partial^2 u_y}{\partial z^2} = \frac{1}{c_{2,2}^2} (\ddot{u}_y - \Omega_0^2 u_y + 2\Omega_0 \dot{u}_x), \quad (5)$$

where $c_{2,1} = [c_{44}^E(1 + e_{15}^2 / (\varepsilon_{11}^S c_{44}^E)) / \rho]^{1/2}$ is the propagation velocity of a coupled electroelastic shear wave in the direction of anisotropy (Ox), $c_{2,2} = [(c_{11}^E - c_{12}^E) / 2\rho]$ is the propagation velocity of an incoherent shear wave in the isotropy plane (Oy). In written formulas, $c_{11}^E, c_{12}^E, c_{44}^E$ are the elastic moduli, e_{15} is the piezoelectric constant, ε_{11}^S is the dielectric permittivity, ρ is the density of the piezoelectric material. Let us write down the formulas for shear stresses and the thickness component of the electrical displacement in the form:

$$\tau_{xz} = c_{44}^E \frac{\partial u_x}{\partial z} + e_{15} \frac{\partial \Psi}{\partial z}, \quad \tau_{yz} = \frac{c_{11}^E - c_{12}^E}{2} \frac{\partial u_y}{\partial z}, \quad D_z = -\varepsilon_{11}^S \frac{\partial \Psi}{\partial z} + e_{15} \frac{\partial u_x}{\partial z}. \quad (6)$$

2. Solution of the electroelasticity boundary value problem, AFC, energy efficiency of the secondary mode

Solutions for the displacements of Coriolis-dispersive plane waves in a layer are characterized by the fact that each component u_x and u_y is a complex superposition of electroelastically coupled and incoherent shear waves with wave numbers k_1, k_2 depending on the normalized angular velocity $\varepsilon = \Omega_0 / \omega$,

$$u_x(z, t) = \frac{e_{15}}{c_{44}^E} \frac{V_0}{h} \cdot \frac{1}{k_1^2 - k_2^2} \left\{ \left(k_1^2 - \frac{\omega^2}{c_{2,2}^2} (1 + \varepsilon^2) \right) \frac{\cos k_1 z}{k_1 \sin k_1 h} - \left(k_2^2 - \frac{\omega^2}{c_{2,2}^2} (1 + \varepsilon^2) \right) \frac{\cos k_2 z}{k_2 \sin k_2 h} \right\} \cos(\omega t), \quad (7)$$

$$u_y(z, t) = -\frac{e_{15}}{c_{44}^E} \frac{V_0}{h} 2i\varepsilon \frac{\omega^2}{c_{2,2}^2} \cdot \frac{1}{k_1^2 - k_2^2} \left\{ \frac{\cos k_1 z}{k_1 \sin k_1 h} - \frac{\cos k_2 z}{k_2 \sin k_2 h} \right\} \sin(\omega t).$$

Expressions for k_1, k_2 are given in (Ulitko, 1995). The trajectories of particle motion in a wave are ellipses, and the direction of the displacement vector $\mathbf{u}$ is formed by these components. When the wave propagates through the thickness of the layer $-h \leq z \leq h$, it is rotated by a certain angle, linearly dependent on the value of the angular velocity $\varepsilon = \Omega_0 / \omega$ (polarization effect).

The solutions for stresses τ_{xz}, τ_{yz} are written similarly using formulas (6).

For resonant oscillations with frequency ω , the magnitude of the conduction current in the external (observational) electrical circuit is determined by the following formula (Grinchenko, Ulitko, & Shulga, 1989)

$$\hat{I} \exp(i\omega t) = -\frac{d}{dt} \iiint_{S; -h \leq z \leq h} \hat{D}_z(z) \exp(i\omega t) dV, \quad (8)$$

where integration is performed over an arbitrary area S of the surface electrode on the front face of the layer $z = \pm h$, through which the electrical load is applied and detected, and by the thickness of the layer $-h \leq z \leq h$. In the case of solid layer electrodes, a unit area value can be set for S . Determining $\hat{D}_z$ by formulas (6), (7), after integration over the layer thickness $-h \leq z \leq h$, we find

$$\hat{I}(\lambda) = 2c_{2,1} C_0 V_0 \times \lambda \left\{ 1 + \left(\frac{e_{15}^2}{c_{44}^E \varepsilon_{11}^S} \right) \left[\frac{1 - A(\varepsilon)}{2} \frac{\cos \lambda}{\lambda \sin \lambda} + \frac{1 + A(\varepsilon)}{2} \frac{\cos \left(\lambda \frac{k_2}{k_1} \right)}{\left(\lambda \frac{k_2}{k_1} \right) \sin \left(\lambda \frac{k_2}{k_1} \right)} \right] \right\}, \quad (9)$$

where $C_0 = \varepsilon_{11}^S \frac{S}{h}$ is the electrostatic capacitance of the layer, $k_1 h = \lambda$, $k_2 h = \lambda k_2 / k_1$ are the wavenumbers, and $A(\varepsilon)$ is some monotonic decreasing function. The AFC graph (8) is shown in Fig. 2.

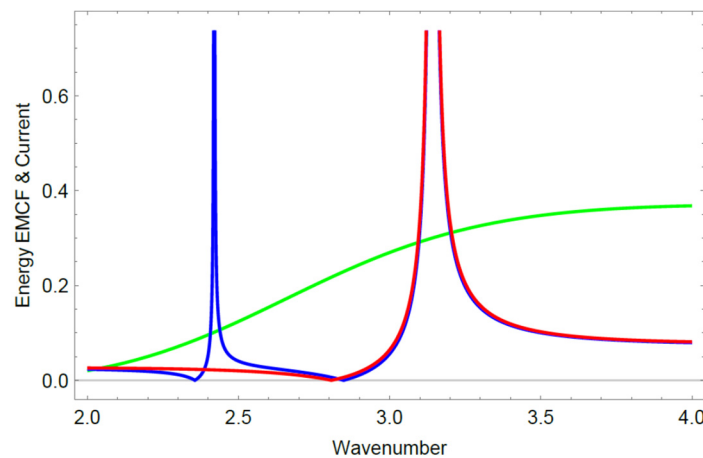


Fig. 2. Calculated amplitude-frequency characteristic of a gyroscopic sensor on the shear of BAWs

Discussion and conclusions

The calculated AFC (see Fig. 2) in the primary (red curve) and sensor (blue curve) modes demonstrates parameters of the gyroscopic sensor output signal that are satisfactory for practice. For a normalized angular velocity value of 0.01 at the first main resonance, the frequency of the sensor mode is 0.78 of the excited frequency, and the efficiency of the energy conversion according to the energy criterion (green curve) (Ulitko, 1977), is about 24% for the primary mode and about 1.5 % for the sensor mode compared to the static shear electromechanical coupling factor k_{15}^2 .

Authors' contribution: Igor Ulitko – basic research concept, problem statement, solutions for coupled displacements; Oleksandr Boryseiko – analysis of sources, preparation of literature review, energy analysis; Iryna Lebedyeva – AFC analysis; Oleksandr Kurylko – resonance and antiresonance frequency calculations, energy conversion.

Sources of funding. Funding is partially provided by Taras Shevchenko National University of Kyiv.

References

- Berlincourt, D. A., Curran, D. R., & Jaffe, H. (1964). *Piezoelectric and piezomagnetic materials and their function in transducers*. In: *Physical Acoustics*, Mason, W.P. Ed. Academic Press.
- Gribkova, E., Peregodov, A., & Shevelko, M. (2011). Theoretical and experimental researches of bulk acoustic wave's availability in angular motion sensors. *Proceedings of the IEEE-R. North West section*, 1, 69–72.

- Grinchenko, V. T., Ulitko, A. F., & Shulga, N. A. (1989). Electroelasticity. In *Mechanics of Coupled Fields in Structural Elements* (Vol. 5). Naukova Dumka [in Russian]. [Гринченко, В. Т., Улитко, А. Ф., & Шульга, Н. А. (1989). Электроупругость. В *Механика связанных полей в элементах конструкций* (Т. 5). Наукова думка].
- Liu, Z., Long, Y., Wehner, C., Wen, H., & Ayazi, F. (2024). 4H silicon carbide bulk acoustic wave gyroscope with ultra-high Q-factor for on-chip inertial navigation. *Communications Engineering*, 3(87), 1–9. <https://doi.org/10.1038/s44172-024-00234-z>
- Sarapuloff, S. A. (2013). Polarization precession effects for shear elastic waves in rotated solids. *Journal of Korean Society. for Noise and Vibration Engineering (KSNVE)*, 4, 842–848.
- Sarapuloff, S. A., & Ulitko, I. A. (2001). Rotation influence upon bulk waves in an elastic medium and their usage in solid-state gyroscopy. *Gyroscopy & Navigation*, 4(35), 62–70.
- Shoenberg, M., & Censor, D. (1973). Elastic waves in rotating media. *Quarterly Applied Mathematics*, 31(4), 115–125. <https://doi.org/10.1090/QAM/99708>
- Ulitko, A. F. (1977). Theory of electromechanical energy conversion in nonuniformly deformable piezoceramics. *Soviet Applied Mechanics*, 13(10), 1055–1062. <https://doi.org/10.1007/BF00883190>
- Ulitko, I. A. (1995). Dispersion of plane harmonic waves in uniformly rotating elastic space. *Reports NAS Ukraine*, 1, 54–57 [in Russian]. [Улитко, И. А. (1995). Дисперсия плоских гармонических волн в равномерно вращающемся упругом пространстве. *Доклады национальной академии наук Украины*, 1, 54–57].
- Yang, Y., Dejous, C., & Hallil, H. (2023). Trends and applications of surface and bulk acoustic wave devices. *Micromachines*, 14(43), 2–21. <https://doi.org/10.3390/mi14010043>

Отримано редакцією журналу / Received: 13.09.25

Прорецензовано / Revised: 26.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Ігор УЛІТКО¹, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0001-6452-141X

e-mail: igorulitko@knu.ua

Олександр БОРИСЕЙКО¹, канд. фіз.-мат. наук, доц.

ORCID ID: 0009-0001-3023-5332

e-mail: boryseiko@knu.ua

Ірина ЛЕБЕДЄВА¹, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0001-7150-1310

e-mail: lebedyevaiv@knu.ua

Олександр КУРИЛКО¹, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-1711-7083

e-mail: alexandr.kurylko@knu.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

РЕЗОНАНСНІ ВЛАСТИВОСТІ П'ЄЗОЕЛЕКТРИЧНОГО ГІРОСКОПІЧНОГО СЕНСОРА НА ОБ'ЄМНИХ ХВИЛЯХ ЗСУВУ

Розроблено математичну модель п'єзоелектричного гіроскопічного сенсора на об'ємних акустичних хвилях зсуву, що поширюються за товщиною шару із п'єзоелектричною анізотропією, перпендикулярною до навантажувального електричного поля. Шар обертається навколо перпендикулярної до його бічних граней осі, а збуджувані зсувні хвилі мають властивості Коріолісової дисперсії. Отримані спряжені рівняння руху є одновимірними за товщиною шару лише у разі відсутності механічних сил на границях та однорідних умов електричного навантаження.

Побудовано аналітичний розв'язок задачі, отримано точні формули розподілів напружень і переміщень за товщиною шару в біжучих зсувних хвилях, що мають еліптичну поляризацію зв'язаних компонентів у серединній площині. Відповідно до точного розв'язку для товщинної компоненти поля зміщених зарядів отримано вираз для функції струму провідності в зовнішньому колі та побудовано амплітудно-частотну характеристику гіроскопічного сенсора на первинній (збуджуваній) та вторинній (сенсорній) модах. Визначено, що на першому основному резонансі шару ефективність збуджуваної моди становить близько 24 %, а ефективність сенсорної моди – близько 1,8 % за значення нормованої кутової швидкості порядку 0,01.

Ключові слова: хвилі зсуву, п'єзоелектричний ефект, резонанс, сили Коріоліса, хвильові гіроскопи.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; або в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 539.3

DOI: <https://doi.org/10.17721/1812-5409.2026/1.22>

Олена ШИКУЛА^{1,2}, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-7385-2816
e-mail: ensh@ukr.net

Наталія ЖУКОВА², канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-2699-3771
e-mail: zhukova_n@ukr.net

Віктор ВИШНІВСЬКИЙ¹, д-р тех. наук, проф.
ORCID ID: 0000-0003-1923-4344
e-mail: vish_vv@ukr.net

Світлана БІЛОУСОВА¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-2557-8146
e-mail: svetlana_belousova@ukr.net

¹Державний університет інформаційно-комунікаційних технологій, Київ, Україна
²Інститут механіки ім. С. П. Тимошенка НАНУ, Київ, Україна

МОДЕЛЬ ДЕФОРМУВАННЯ АРМОВАНИХ У ПЛОЩИНІ ВОЛОКНИСТИХ МАТЕРІАЛІВ ІЗ ФІЗИЧНО НЕЛІНІЙНОЮ МАТРИЦЕЮ

У разі збільшення навантаження багато однорідних і композитних матеріалів проявляють нелінійний характер залежностей між макронапруженнями та макродеформаціями. Це може бути обумовлено фізичною нелінійністю деформування компонентів. Волокнисті матеріали використовують у спорудах і деталях, що функціонують в умовах високих силових і температурних навантажень. З огляду на це прогнозування їхніх нелінійних властивостей є актуальним.

Мета роботи – побудова моделі та дослідження нелінійного деформування армованих у площині волокнистих композитних матеріалів із фізично нелінійною матрицею.

Армовані у площині волокнисті матеріали розглянуто як багатокомпонентні матеріали з випадковим розташуванням волокон, визнаючи матрицю і волокна кожного напрямку окремими компонентами. Будемо вважати, що матеріал волокон – трансверсально-ізоотропний із відносною симетрією уздовж волокон, матеріал матриці – ізоотропний фізично нелінійний. Визначення ефективних деформативних властивостей і напружено-деформованого стану матеріалу відбувається у два етапи. На першому етапі визначено властивості підсистеми, що є односпрямованим волокнистим матеріалом, утвореним волокнами певного напрямку і частиною матриці. Розв'язання задачі першого етапу здійснюється на основі моделі односпрямованого волокнистого матеріалу стохастичної структури. Розв'язок побудовано із застосуванням методу умовних моментів Л. П. Хорошуна. На другому етапі за обчисленими властивостями підсистем визначено ефективні властивості всієї системи. Розв'язок побудовано на основі схеми Фойхта. Нелінійні рівняння, що враховують фізичну нелінійність матриці, розв'язано за ітераційним методом.

Розроблено алгоритм визначення ефективних властивостей та напружено-деформованого стану матеріалу із фізично нелінійною матрицею. Встановлені закони зв'язку між макронапруженнями і макродеформаціями в матеріалі та залежності середніх деформацій і напружень у матриці матеріалу від макродеформацій. Досліджено криві деформування матеріалу. З'ясовано, що нелінійність компонентів суттєво впливає на ефективні деформативні властивості та напружено-деформований стан матеріалу.

Ключові слова: волокнистий композитний матеріал, армовані у площині матеріали, нелінійна матриця, ефективні деформативні властивості, напружено-деформований стан, вплив нелінійності, комп'ютерне моделювання

Класифікація відповідно до AMS 2020: 74-10.

Вступ

Просторово армовані й армовані у площині волокнисті композити набули широкого застосування в різноманітних галузях техніки, де пред'являються вимоги до високої міцності за зниженої маси конструкцій. В умовах зростання навантаження багато однорідних і композитних матеріалів демонструють нелінійний характер залежностей між макронапруженнями та макродеформаціями. Це може бути обумовлене фізичною нелінійністю деформування компонентів. Такий вид нелінійності є типовим для композитів на основі пластичної металевої матриці, а також на основі полімерів за підвищених температур. Армовані у площині волокнисті матеріали використовують в оболонках, що входять у конструкції, які функціонують в умовах значних силових або температурних впливів. З огляду на це прогнозування їхніх нелінійних властивостей і деформування є актуальним.

Нелінійні волокнисті та просторово армовані волокнисті композитні матеріали вивчалися багатьма авторами. Визначенню ефективних властивостей волокнистих композитних матеріалів присвячені дослідження (Хорошун та ін., 1993; Khoroshun et al., 2008; Khoroshun, & Nazarenko, 2013; Maslov, 2017; 2022; Хорошун, & Шикун, 2011; 2016) та ін. Розв'язання лінійної задачі про деформування стохастично неоднорідних волокнистих матеріалів просторового армування було отримано в роботах (Хорошун та ін., 1993; Khoroshun, & Shikula, 2008; Khoroshun, & Nazarenko, 2013). Однак автори вивчали лише лінійне деформування. Не було побудовано моделі деформування в разі нелінійного деформування матриці, тож побудова такої моделі є на часі.

Об'єкт дослідження – армовані у площині волокнисті композитні матеріали.

Метою пропонованої роботи є побудова моделі та дослідження нелінійного деформування армованих у площині волокнистих матеріалів із фізично нелінійною матрицею.

У статті армовані у площині волокнисті матеріали розглянуто як багатокомпонентні матеріали з випадковим розташуванням волокон, визнаючи матрицю і волокна кожного напрямку окремими компонентами. Будемо вважати, що матеріал волокон – трансверсально-ізотропний із віссю симетрії уздовж волокон, матеріал матриці – ізотропний фізично нелінійний. Визначення ефективних деформативних властивостей і напружено-деформівного стану матеріалу проведено у два етапи. На першому етапі з'ясовуємо ефективні властивості підсистеми, що є односпрямованим волокнистим матеріалом, утвореним волокнами певного напрямку і частиною матриці. Розв'язання задачі першого етапу здійснюємо на основі моделі односпрямованого волокнистого матеріалу стохастичної структури. В основу покладено стохастичні диференційні рівняння теорії пружності, розв'язок побудовано із застосуванням методу умовних моментів Л. П. Хорошуна. На другому етапі за обчисленими властивостями підсистем визначаємо ефективні властивості всієї системи. Розв'язання задачі другого етапу побудовано на основі схеми Фойхта. Нелінійні рівняння, що враховують фізичну нелінійність матриці, розв'язано за ітераційним методом.

1. Матеріали та методи дослідження

Розглянемо просторово армований волокнистий матеріал стохастичною структури, що є системою волокон N напрямків, пов'язаних матрицею. Волокна k -го напрямку характеризуються напрямними косинусами (рис. 1)

$$\gamma_1^{(k)} = \sin \theta_k \cos \varphi_k; \gamma_2^{(k)} = \sin \theta_k \sin \varphi_k; \gamma_3^{(k)} = \cos \theta_k \quad (0 \leq \theta_k \leq \pi/2; 0 \leq \varphi_k \leq 2\pi; k = 1, 2, \dots, N). \quad (1)$$

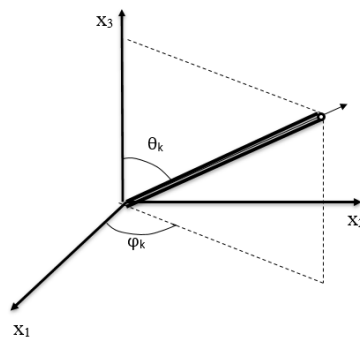


Рис. 1. Напрямні косинуси волокон

Нехай об'ємні вмісти волокон k -го напрямку і матриці дорівнюють відповідно $c_1^{(k)}$ і c_2 . Загальна концентрація волокон $c_1 = \sum_{k=1}^N c_1^{(k)}$, при цьому $c_1 + c_2 = 1$. Визначення ефективних деформативних характеристик матеріалу будемо виконувати у два етапи. На першому етапі з'ясовуємо ефективні властивості підсистеми, що є односпрямованим волокнистим матеріалом, утвореним волокнами k -го напрямку і частиною матриці з об'ємним вмістом $c_2^{(k)} = c_2 c_1^{-1} c_1^{(k)}$.

2. Визначення ефективних деформативних властивостей просторово армованих волокнистих матеріалів

Нехай макрооб'єм просторово армованого волокнистого матеріалу перебуває в умовах однорідного навантаження. Позначимо пружні сталі волокон k -го напрямку $\lambda_{ij}^{(k)}$, а деформативні характеристики матриці, що залежать від деформації, – $\lambda_2 = \lambda_2(\epsilon)$, $\mu_2 = \mu_2(\epsilon)$. Тоді вирази для ефективних характеристик підсистеми мають такий вигляд (Гузь та ін., 2003; Хорошун та ін., 1993):

$$\begin{aligned} \lambda_{11}^{*(k)} + \lambda_{12}^{*(k)} &= c_1^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)}) + 2c_2^{(k)}(\lambda_2 + \mu_2) - \frac{c_1^{(k)}c_2^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)} - 2\lambda_2 - 2\mu_2)^2}{2c_1^{(k)}(\lambda_2 + \mu_2) + c_2^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)}) + 2m^{(k)}}; \\ \lambda_{11}^{*(k)} - \lambda_{12}^{*(k)} &= c_1^{(k)}(\lambda_{11}^{(k)} - \lambda_{12}^{(k)}) + 2c_2^{(k)}\mu_2 - \frac{c_1^{(k)}c_2^{(k)}(\lambda_{11}^{(k)} - \lambda_{12}^{(k)} - 2\mu_2)^2}{2c_1^{(k)}\mu_2 + c_2^{(k)}(\lambda_{11}^{(k)} - \lambda_{12}^{(k)}) + \frac{2m^{(k)}n^{(k)}}{n^{(k)} + 2m^{(k)}}}; \\ \lambda_{13}^{*(k)} &= c_1^{(k)}\lambda_{13}^{(k)} + c_2^{(k)}\lambda_2 - \frac{c_1^{(k)}c_2^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)} - 2\lambda_2 - 2\mu_2)(\lambda_{13}^{(k)} - \lambda_2)}{2c_1^{(k)}(\lambda_2 + \mu_2) + c_2^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)}) + 2m^{(k)}}; \\ \lambda_{33}^{*(k)} &= c_1^{(k)}\lambda_{33}^{(k)} + c_2^{(k)}(\lambda_2 + 2\mu_2) - \frac{2c_1^{(k)}c_2^{(k)}(\lambda_{13}^{(k)} - \lambda_2)^2}{2c_1^{(k)}(\lambda_2 + \mu_2) + c_2^{(k)}(\lambda_{11}^{(k)} + \lambda_{12}^{(k)}) + 2m^{(k)}}; \\ \lambda_{44}^{*(k)} &= c_1^{(k)}\lambda_{44}^{(k)} + c_2^{(k)}\mu_2 - \frac{c_1^{(k)}c_2^{(k)}(\lambda_{44}^{(k)} - \mu_2)^2}{c_1^{(k)}\mu_2 + c_2^{(k)}(\lambda_{44}^{(k)} + s^{(k)})}. \end{aligned} \quad (2)$$

Ефективні характеристики (2) – компоненти тензорів ефективних пружних характеристик у локальній системі координат x'_i , де вісь x'_3 паралельна волокнам підсистеми. У загальній системі координат згідно з перетворенням

$$\begin{aligned} x'_i &= \alpha_{ij}x'_j; x'_i = \alpha_{ji}x_j; \alpha_{11} = \cos \theta \cos \varphi; \alpha_{12} = -\sin \varphi; \alpha_{13} = \sin \theta \cos \varphi; \\ \alpha_{21} &= \cos \theta \sin \varphi; \alpha_{22} = \cos \varphi; \alpha_{23} = \sin \theta \sin \varphi; \alpha_{31} = -\sin \theta; \alpha_{32} = 0; \alpha_{33} = \cos \theta. \end{aligned} \quad (3)$$

Тензори ефективних деформативних характеристик $\lambda_{ijlm}^{*(k)}$ у загальній системі координат виражаються через відповідні тензори $\lambda_{ijlm}^{*(k)}$ в локальній системі координат співвідношеннями

$$\lambda_{ijlm}^{*} = \alpha_{ip} \alpha_{jq} \alpha_{lr} \alpha_{ms} \lambda_{pqrs}^{*} \quad (4)$$

Ефективні характеристики всієї системи відповідно до схеми Фойхта визначаємо формулами

$$\lambda_{ijlm}^{**} = \frac{1}{c_1} \sum_{k=1}^N c_1^{(k)} \lambda_{ijlm}^{*(k)} \quad (5)$$

Підставляючи (4) у (5) і переходячи до матричних позначень, отримуємо

$$\begin{aligned} \lambda_{ijlm}^{**} = & \frac{1}{c_1} \sum_{k=1}^N c_1^{(k)} \{ [\alpha_{i1}^{(k)} \alpha_{j1}^{(k)} \alpha_{l1}^{(k)} \alpha_{m1}^{(k)} + \alpha_{i2}^{(k)} \alpha_{j2}^{(k)} \alpha_{l2}^{(k)} \alpha_{m2}^{(k)} + \frac{1}{2} (\alpha_{i1}^{(k)} \alpha_{j2}^{(k)} + \alpha_{i2}^{(k)} \alpha_{j1}^{(k)}) (\alpha_{l1}^{(k)} \alpha_{m2}^{(k)} + \alpha_{l2}^{(k)} \alpha_{m1}^{(k)})] \lambda_{1111}^{*(k)} + \\ & + [\alpha_{i1}^{(k)} \alpha_{j1}^{(k)} \alpha_{l2}^{(k)} \alpha_{m2}^{(k)} + \alpha_{i2}^{(k)} \alpha_{j2}^{(k)} \alpha_{l1}^{(k)} \alpha_{m1}^{(k)} - \frac{1}{2} (\alpha_{i1}^{(k)} \alpha_{j2}^{(k)} + \alpha_{i2}^{(k)} \alpha_{j1}^{(k)}) (\alpha_{l1}^{(k)} \alpha_{m2}^{(k)} + \alpha_{l2}^{(k)} \alpha_{m1}^{(k)})] \lambda_{1122}^{*(k)} + \\ & + \alpha_{i3}^{(k)} \alpha_{j3}^{(k)} \alpha_{l3}^{(k)} \alpha_{m3}^{(k)} \lambda_{3333}^{*(k)} + [(\alpha_{i1}^{(k)} \alpha_{j1}^{(k)} + \alpha_{i2}^{(k)} \alpha_{j2}^{(k)}) \alpha_{l3}^{(k)} \alpha_{m3}^{(k)} + \alpha_{i3}^{(k)} \alpha_{j3}^{(k)} (\alpha_{l1}^{(k)} \alpha_{m1}^{(k)} + \alpha_{l2}^{(k)} \alpha_{m2}^{(k)})] \lambda_{1313}^{*(k)} + \\ & + [(\alpha_{i1}^{(k)} \alpha_{j3}^{(k)} + \alpha_{i3}^{(k)} \alpha_{j1}^{(k)}) (\alpha_{l1}^{(k)} \alpha_{m3}^{(k)} + \alpha_{l3}^{(k)} \alpha_{m1}^{(k)}) + (\alpha_{i2}^{(k)} \alpha_{j3}^{(k)} + \alpha_{i3}^{(k)} \alpha_{j2}^{(k)}) (\alpha_{l2}^{(k)} \alpha_{m3}^{(k)} + \alpha_{l3}^{(k)} \alpha_{m2}^{(k)})] \lambda_{44}^{*(k)} \}. \end{aligned} \quad (6)$$

Зауважимо, що у випадку фізичної нелінійності матриці, оскільки $\lambda_2(<\varepsilon_{ij}^2>)$, $\mu_2(<\varepsilon_{ij}^2>)$, ефективні деформативні характеристики (6) також є функціями середніх у матриці деформацій $<\varepsilon_{ij}^2>$. Тому для визначення ефективних властивостей просторово армованого матеріалу потрібно встановити співвідношення між середніми деформаціями в матриці $<\varepsilon_{ij}^2>$ та макродеформаціями $<\varepsilon_{ij}>$, що задаються. Вони мають такий вигляд:

$$c_2 <\varepsilon_{ij}^2> = -[\lambda_{ijpq}^{1*} - \lambda_{ijpq}^{2*}]^{-1} [\lambda_{pqlm}^{**} - \lambda_{pqlm}^{1*}] <\varepsilon_{lm}>. \quad (7)$$

Формули (2)–(4), (6), (7) дають можливість написати ітераційний алгоритм визначення ефективних деформативних характеристик просторованого матеріалу за заданих макродеформацій.

3. Дослідження ефективних деформативних властивостей армованих у площині волокнистих матеріалів

Нехай волокна розташовані у площині x_1x_2 уздовж сторін рівностороннього трикутника. Концентрації волокон за різними напрямками в площині x_1x_2 вважаємо однаковими. У цьому випадку маємо матеріал із гексагональною симетрією. На рис. 2–5 побудовано відповідно графіки залежностей осьових і зсувних макронапружень від макродеформацій за різних об'ємних концентрацій волокон c_1 .

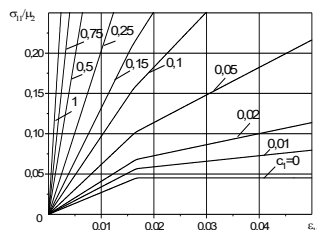


Рис. 2. Залежність макронапруження $<\sigma_{11}>$ від макродеформації $<\varepsilon_{11}>$

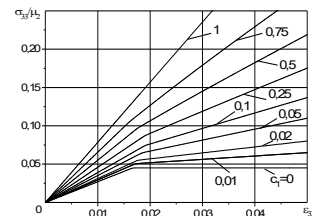


Рис. 3. Залежність макронапруження $<\sigma_{33}>$ від макродеформації $<\varepsilon_{33}>$

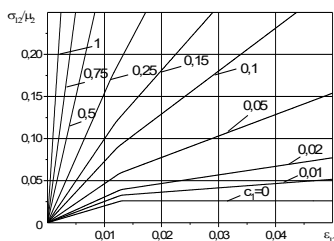


Рис. 4. Залежність макронапруження $<\sigma_{12}>$ від макродеформації $<\varepsilon_{12}>$

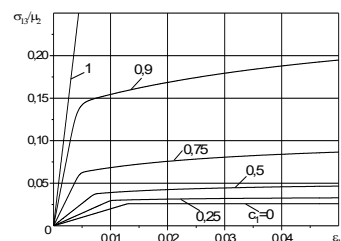


Рис. 5. Залежність макронапруження $<\sigma_{13}>$ від макродеформації $<\varepsilon_{13}>$

Матеріал добре витримує всі осьові навантаження, нелінійно деформується тільки за невеликого об'ємного вмісту волокон, а за зсувних навантажень у напрямку, перпендикулярному площині розташування волокон, нелінійність матриці дуже суттєво впливає на деформування композиту за всіх значень об'ємного вмісту волокон.

Дискусія і висновки

Побудовано модель і досліджено нелінійне деформування армованих у площині волокнистих композитних матеріалів із фізично нелінійною матрицею. Розроблено алгоритм визначення ефективних властивостей і напружено-деформованого стану матеріалу із фізично нелінійною матрицею. Встановлено закони зв'язку між макронапруженнями і макродеформаціями в матеріалі та залежності середніх деформацій і напружень у матриці матеріалу від макродеформацій. Досліджено криві деформування матеріалу. З'ясовано, що нелінійність компонентів суттєво впливає на ефективні деформативні властивості та напружено-деформований стан матеріалу. Матеріал добре витримує всі осьові навантаження, нелінійно деформується тільки за невеликого об'ємного вмісту волокон, а за зсувних навантажень у напрямку, перпендикулярному площині розташування волокон, нелінійність матриці дуже суттєво впливає на деформування композиту за всіх значень об'ємного вмісту волокон.

Внесок авторів: Олена Шикун – концептуалізація; методика; Наталія Жукова – створення алгоритму для розрахунків; Віктор Вишнівський – програмне забезпечення; Світлана Білоусова – графічна інтерпретація чисельних результатів.

Джерела фінансування. Роботу виконано в межах НДР "Дослідження поведінки композитних матеріалів різної структури (армованих короткими волокнами та шаруватих) та оболонкових конструкцій з них в умовах дії стискаючих навантажень", № держреєстрації 0119U103519, Інститут механіки НАНУ.

Список використаних джерел

- Гузь, А. Н., Хорошун, Л. П., Михайлова, М. И., Бабич, Д. В., & Шикун, Е. Н. (2003). *Механика композитов: Т. 12. Прикладные исследования* (А. Н. Гузь, Ред.). "А.С.К.".
- Хорошун, Л. П., Маслов, Б. П., Шикун, Е. Н., & Назаренко, Л. В. (1993). *Механика композитов: Т. 3. Статистическая механика и эффективные свойства материалов* (А. Н. Гузь, Ред.). Наукова думка. <https://librarygo.lpnu.ua/?elbook=mehanyka-kompozytov-v-12-tomah-tom-3-statysticheskaya-mehanyka-y-%D1%8Dffektyvn%D1%8Be-svoystva-materyalov>
- Хорошун, Л. П., & Шикун, Е. Н. (2011). Деформирование физически нелинейных стохастических композитных материалов. Деформирование и кратковременная повреждаемость физически нелинейных стохастических композитных материалов. *Успехи механики: Т. 6* (кн. 2) (А. Н. Гузь, Ред.). Литера ЛТД.
- Хорошун, Л. П., & Шикун, Е. Н. (2016). Ефективні деформовні властивості волокнистих композитних матеріалів при нелінійному деформуванні компонентів. *Доповіді Національної академії наук України*, 6, 47–55. <https://doi.org/10.15407/dopovidi2016.06.047>
- Khoroshun, L. P., Nazarenko, L. V., Müller, W. H., & Wille, R. (2008). Homogenization of Unidirectional and Arbitrarily Oriented Fiber-Reinforced Materials by the Method of Conditional Moments. *PAMM Proc. Appl. Math. Mech.*, 8, 10451–10452. <https://doi.org/10.1002/pamm.200810451>
- Khoroshun, L. P., & Nazarenko, L. V. (2013). Deformation and damage of composite materials with anisotropic components (Review). *Int. Appl. Mech.*, 49(4), 388–455. <https://doi.org/10.1007/s10778-013-0578-6>
- Khoroshun, L. P., & Shikula, E. N. (2008). Deformation of physically nonlinear stochastic composites. *Int. Appl. Mech.*, 44(12), 1325–1351. <https://doi.org/10.1007/s10778-009-0159-x>
- Maslov, B. P. (2017). Stress concentration in nonlinear viscoelastic composites. *J. Mech Adv Technol.*, 79(1), 5–10. <https://doi.org/10.20535/2521-1943.2017.79.66490>
- Maslov B. P. (2022). Nonlinear hereditary creep of isotropic composites of random structure. *Int. Appl. Mech.* 58(1), 75–90. <https://doi.org/10.1007/s10778-022-01136-3>

References

- Guz, A. N., Khoroshun, L. P., Mikhailova, M. I., Babich, D. V., & Shikula, E. N. (2003) *Mechanics of composites: T. 12. Applied research* (A. N. Guz, Ed.). "A.S.K." [in Russian].
- Khoroshun, L. P., & Nazarenko, L. V. (2013). Deformation and damage of composite materials with anisotropic components (Review). *Int. Appl. Mech.*, 49(4), 388–455. <https://doi.org/10.1007/s10778-013-0578-6>
- Khoroshun, L. P., & Shikula, E. N. (2008). Deformation of physically nonlinear stochastic composites. *Int. Appl. Mech.*, 44(12), 1325–1351. <https://doi.org/10.1007/s10778-009-0159-x>
- Khoroshun, L. P., & Shikula, E. N. (2011). *Deformation of physically nonlinear stochastic composite materials. Deformation and short-term damage of physically nonlinear stochastic composite materials*. Advances in Mechanics: Vol. 6 (book 2) (by A. N. Guz, Ed.). Litera LTD [in Russian].
- Khoroshun, L. P., & Shikula, E. N. (2016) Effective deformation properties of fibrous composite materials with nonlinear deformation of components. *Proceedings of the National Academy of Sciences of Ukraine*, 6, 47–55 [in Ukrainian]. <https://doi.org/10.15407/dopovidi2016.06.047>
- Khoroshun, L. P., Maslov, B. P., Shikula, E. N., & Nazarenko, L. V. (1993). *Mechanics of composites: Vol. 3. Statistical mechanics and effective properties of materials* (A. N. Guz, Ed.). Nauk. Dumka [in Russian]. <https://librarygo.lpnu.ua/?elbook=mehanyka-kompozytov-v-12-tomah-tom-3-statysticheskaya-mehanyka-y-%D1%8Dffektyvn%D1%8Be-svoystva-materyalov>
- Khoroshun, L. P., Nazarenko, L. V., Müller W. H., & Wille, R. (2008). Homogenization of Unidirectional and Arbitrarily Oriented Fiber-Reinforced Materials by the Method of Conditional Moments. *PAMM Proc. Appl. Math. Mech.*, 8, 10451–10452. <https://doi.org/10.1002/pamm.200810451>
- Maslov, B. P. (2017). Stress concentration in nonlinear viscoelastic composites. *J. Mech Adv Technol.*, 79(1), 5–10. <https://doi.org/10.20535/2521-1943.2017.79.66490>
- Maslov, B. P. (2022). Nonlinear hereditary creep of isotropic composites of random structure. *Int. Appl. Mech.* 58(1), 75–90. <https://doi.org/10.1007/s10778-022-01136-3>

Отримано редакцією журналу / Received: 25.09.25
Прорецензовано / Revised: 26.02.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Elena SHIKULA^{1,2}, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-7385-2816
e-mail: ensh@ukr.net

Nataliia ZHUKOVA², PhD (Phys. & Math.), Senior Researcher
ORCID ID: 0000-0002-2699-3771
e-mail: zhukova_n@ukr.net

Victor VYSHNIVSKYI¹, DSc (Engin.), Prof.
ORCID ID: 0000-0003-1923-4344
e-mail: vish_vv@ukr.net

Svitlana BILOUSOVA¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-2557-8146
e-mail: svetlana_belousova@ukr.net

¹State University of Information and Communication Technologies, Kyiv, Ukraine

²S. P. Tymoshenko Institute of Mechanics, National Academy of Science of Ukraine, Kyiv, Ukraine

MODEL OF DEFORMATION OF IN-PLANE REINFORCED FIBROUS MATERIALS WITH PHYSICALLY NONLINEAR MATRIX

With increasing load, many homogeneous and composite materials exhibit nonlinear relationships between macrostresses and macrostrains. This may be due to the physical nonlinearity of component deformation. Fibrous materials are used in structures and parts operating under high force and temperature loads. Therefore, predicting their nonlinear properties is relevant.

The aim of the work is to build a model and study the nonlinear deformation of in-plane reinforced fibrous composite materials with a physically nonlinear matrix.

In-plane reinforced fibrous materials are considered as multicomponent materials with a random arrangement of fibers, considering the matrix and fibers of each direction as separate components. We will assume that the fiber material is transversely isotropic with an axis of symmetry along the fibers, the matrix material is isotropic physically nonlinear. The determination of the effective deformation properties and the stress-strain state of the material is carried out in two stages. At the first stage, the properties of the subsystem, which is a unidirectional fibrous material formed by fibers of a certain direction and part of the matrix, are determined. The solution of the first stage problem is carried out on the basis of a model of a unidirectional fibrous material of a stochastic structure. The solution is built using the method of conditional moments of L.P. Khoroshun. At the second stage, the effective properties of the entire system are determined based on the calculated properties of the subsystems. The solution is built on the basis of the Voight scheme. Nonlinear equations that take into account the physical nonlinearity of the matrix are solved by the iterative method.

An algorithm for determining the effective properties and stress-strain state of a material with a physically nonlinear matrix has been developed. The laws of the relationship between macrostresses and macrostrains in the material and the dependence of the average strains and stresses in the material matrix on macrostrains have been established. The deformation curves of the material have been studied. It has been established that the nonlinearity of the components significantly affects the effective deformation properties and the stress-strain state of the material.

K e y w o r d s : *fibrous composite material, in-plane reinforced materials, nonlinear matrix, effective deformation properties, stress-strain state, influence of nonlinearity, computer modeling*

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

КОМП'ЮТЕРНІ НАУКИ ТА ІНФОРМАТИКА

UDC 519.175+532.546

DOI: <https://doi.org/10.17721/1812-5409.2026/1.23>

Oleg AFANASIEV, PhD Student

ORCID ID: 0009-0007-9491-6456

e-mail: oleg.afanasiev@karazin.ua

V. N. Karazin Kharkiv National University, Kharkiv, Ukraine

Natalya KIZILOVA, DSc. (Phys. & Math.), Prof.

ORCID ID: 0000-0001-9981-7616

e-mail: kizilova@karazin.ua

V. N. Karazin Kharkiv National University, Kharkiv, Ukraine

Sergii POSLAVSKYI, PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0002-1049-9947

e-mail: s.poslavskyi@karazin.ua

V. N. Karazin Kharkiv National University, Kharkiv, Ukraine

RANDOM NETWORKS IN MODELING HEAT AND MASS TRANSFER
IN POROUS MATERIALS

A detailed literature review on mathematical models of mass and heat transfer processes in porous media based on random graphs (networks) is presented. In these models, individual pores (graph nodes) are considered to be connected with a given probability by narrow channels (graph edges) through which a liquid or gas can move. A classification of the main types of random networks used in mathematical modeling of porous media is given, as well as corresponding discrete and continuous mathematical models for single-component Newtonian liquids and mixtures of immiscible liquids. In coupled heat and mass transfer problems, double random networks are usually considered, one of which corresponds to heat-mass transfer with a liquid, and another is associated with heat transfer through a solid phase. The available algorithms for constructing random graphs, mathematical models of heat and mass transfer on networks, and the results of modeling of liquid flow, several miscible or immiscible liquids of Newtonian and non-Newtonian liquids are presented. It is shown that, in addition to the possibility of the disappearance of a network edge due to the suffusion and/or aggregation of liquid/solid phase particles, it is also possible to restore the connection between neighboring nodes due to the destruction and washout of the solid phase, as well as disaggregation/disintegration of particle conglomerates. The results of the review analysis of the literature can be used to select the best random graph (network) model, discrete or continuous mathematical model, and numerical method for calculations on a specific case of a porous medium.

Key words: flows in porous media, random graphs, machine learning methods, mathematical modeling.

MS 2020 classification: 76S05, 05C80.

Introduction

Random graphs (or random networks) are mathematical objects that arise by taking n vertices (vertex set) and placing an edge (edge set) between any pair of distinct vertices with a fixed probability, independently for all pairs (Hofstad, 2016). Edges may appear or disappear during the evolution of the corresponding dynamical system (for example, due to the erosion of solid particles in a porous medium or, conversely, pore clogging); they may also be directed and weighted. This approach is referred to as neuron-like, as it corresponds to the structure of neural networks used in artificial intelligence approaches (Zhou et al., 2020). Random graphs are widely used in applied mathematics and mechanics to model fluid flow and heat transfer in porous media, multiphase systems with phase transitions, and related processes. The most well-known examples are the random graph models of Erdős-Rényi and Gilbert: $G(n, m)$, where a random graph has n vertices and m edges and all such graphs are equally likely, and $G(n, p)$, where each edge is present independently with probability p . The number of edges k incident to a vertex is called the degree of that vertex. For large n , the degree distribution of a random graph approaches the Poisson distribution $p_k = \langle k \rangle^k e^{-\langle k \rangle} / k!$, where $\langle k \rangle$ is the average over all vertices (Li et al., 2021). In the limit $n \rightarrow \infty$, with $m = [C_n^2 p]$ (where square brackets denote the integer part), the models of random graphs $G(n, m)$ and $G(n, p)$ become equivalent.

When applied to heat and mass transfer in porous media, graph models are referred to as pore-network models. Mass transfer processes may be accompanied by heat transfer. For example, evaporation of water from soil leads to a decrease in its surface temperature. Spontaneous fluid percolation may lead to local temperature jumps of several degrees due to surface chemical interaction between the fluid and the solid phase. Heat transfer occurs not only through the fluid in the pore space, but also through the solid phase. Thus, the geometry of the "thermal network" differs from that of the network associated with mass transfer, and therefore, they are usually specified separately. Both networks are treated as random, since flow stagnation with local pore blocking may occur, causing abrupt changes in both mass transfer between network nodes and convective heat transfer. Similar problems arise for flow through porous media involving immiscible fluids, where the concentrations, and even the presence of each fluid in a given pore, are random (Joekar-Niasar, & Hassanizadeh, 2012; Fyhn et al., 2021; Savani et al., 2017a).

Pore-network models consist of pore bodies (nodes) and connections between them (edges). Pore bodies are relatively large cavities of the pore space, whereas edges represent local constrictions separating these cavities. This approach is known as the watershed algorithm (Koch et al., 2021). Modern percolation theory on random graphs belongs to the class of models describing the geometrical properties of random media (Li et al., 2021). In statistical physics, percolation theory is often accompanied by scaling laws, fractal properties (self-similarity), self-organization criticality, and renormalization, which

© Afanasiev Oleg, Kizilova Natalya, Poslavskiy Sergii, 2026

are all of immense importance theoretically in many diverse fields of physics. The term percolation, as well as the geometrical and probabilistic concepts of transport processes in porous media, first appeared as mathematical formulations in the work (Broadbent, & Hammersley, 1957). According to this concept, the connection of any two adjacent pores is probabilistic, supposing the probability is p . If most pores are connected into a single large cluster, the fluid can reach the center of this cluster from the surface of the porous medium. The percolation problem reduces to determining whether there exists a connected path of pores from the surface to the center of the medium for a given probability p . Such problems have most often been studied on square lattices. It can be shown that for $p > 0.6$, large clusters appear in the pore space that connect the surface of the medium with its center. This ensures percolation throughout most of the medium (Li et al., 2021).

Random graph models are also used to describe resource consumption and the eventual resource depletion during fluid flow in porous media. Directed edges forming shortest paths that connect pairs of nodes in the direction of fluid flow are removed if the flow rate through them falls below a certain threshold. As a result of this edge-removal procedure, together with the elimination of disconnected nodes, the graph becomes disconnected, i.e., it undergoes a percolation transition. In (Kim, & Radicchi, 2024), it was shown that under limited flow capacity, this transition is identical to that of standard percolation, where a single giant cluster shrinks as edges are removed from the graph. However, for an infinite flow budget, the transition becomes sharper than in standard percolation and is accompanied by an abrupt fragmentation of the graph into a set of clusters of equal size.

The object of this study is various types of random graphs used to model heat and mass transfer processes in porous media. The aim of this work is to identify the most promising combinations of random network type, mathematical model, and numerical algorithms for simulating transport processes in porous, fractured, and porous-fractured media.

1. Classification of models

1.1. Geometry of random networks

1.1.1. Regular lattices with random edges

In two-dimensional problems, regular triangular, square (Fig. 1a), rhombic, hexagonal, or bi-hexagonal lattices (Ewing, & Gupta, 1993) can be used. Connections (edges) between certain pairs of neighboring nodes are randomly removed (with a given probability), or their conductance (hydraulic permeability) is reduced. In three-dimensional cases, tetrahedral or cubic lattices are most commonly used. Tree-like graphs are also employed in the form of binary trees or Bethe lattices (Fig. 1b). These are infinite connected graphs without cycles (infinite trees) in which all vertices have the same degree (Li et al., 2021).

1.1.2. Fully random networks

These networks have edges assigned with a given probability according to the Gilbert model $G(n, p)$, the Erdős-Rényi model $G(n, m)$, and several others, while the degree distribution of the vertices follows a Poisson distribution (Fig. 1c, d).

1.1.3. Small-world graphs

A small-world network is a graph in which two arbitrary vertices a and b are not adjacent with high probability. However, one can reach b from a in a small number of steps via other vertices. As a result, such a graph is characterized by a path length (i.e., the number of steps) between two vertices $L \sim \log(n)$. In such graphs, clusters arise – relatively small subsets of vertices that are connected to their hub (i.e., a node with a degree significantly higher than the average in the graph) by short paths of length $L \sim 2 - 4$ (Fig. 1e). Such properties are exhibited by various porous media, including fractured soils.

1.1.4. Scale-free networks

Scale-free networks are networks in which the degree distribution follows a power-law, at least asymptotically. That is, the fraction $P(k)$ of nodes in the network that have k connections to other nodes depends on k as $P(k) \sim k^{-q}$, $2 < q < 3$ (Fig. 1f). These models are based on observational data on the number of citations of scientific publications, the topology of the Internet, and others, which have shown that some nodes (hubs) have significantly more connections than others, and that the network as a whole exhibits a power-law distribution of node degrees. Similar properties are observed in some porous media.

1.1.5. Other types of networks

In addition, the following types of random networks are distinguished: fully connected, star-shaped (centroid), ring, bus, hybrid, partial, and incomplete networks (Xiong, Baychev, & Jivkov, 2016).

1.1.6. Multiscale random networks

In real porous media, there exist large (mesoscale), small (microscale), and nanoscale pores, as well as corresponding fractures at different scales. To model heat and mass transfer processes in such complex media, multiscale random networks with different types and degrees of connectivity have been deployed (Rabbani, Babaei, & Javadpour, 2020).

1.2. Percolation models on random networks

1.2.1. Percolation on regular lattices with prescribed random conductivity

In this model, nodes and/or edges are randomly considered either occupied (with probability p) or unoccupied (with probability $1 - p$). For convenience, each unoccupied node/connection with probability $p = 0$ can often be removed. However, in models with deformation and erosion of the solid phase, such connections may reappear (with probability $p \neq 0$), and therefore they must be retained to allow for switching. In the case of flow of a mixture of immiscible fluids through a cross-section of a porous medium (Savani et al., 2017a), channels (graph edges) through which a given fluid flows at a given time are selected randomly (Fig. 2a). This is because such flow most often occurs in the form of intermittent motion of fluid slugs (Fig. 2b) with different probabilities, i.e., with different hydraulic conductivities of the edges (Fig. 2c). Monte Carlo methods are also used to model percolation on random graphs (Savani et al., 2017b; Gjennestad, Winkler, & Hansen, 2020).

1.2.2. Discrete percolation models

Discrete percolation models are typically studied using probability theory. One of the simplest examples is the Bernoulli lattice, where each cell of the lattice is open with probability p or closed with probability $1 - p$. The study of percolation on such a lattice makes it possible to determine the values of p at which large clusters of open cells appear, forming pathways for the percolation of a fluid or gas.

One of the key problems in percolation theory is the determination of the percolation threshold p_c , which indicates the critical probability at which the transition from fragmented clusters to large connected clusters capable of transmitting a fluid or gas through the entire material begins.

At the same time, there exist discrete models designed, for example, to evaluate the efficiency of filters, in which the clogging process is described by a deterministic sequence of channels becoming blocked (Goldsztein, 2009).

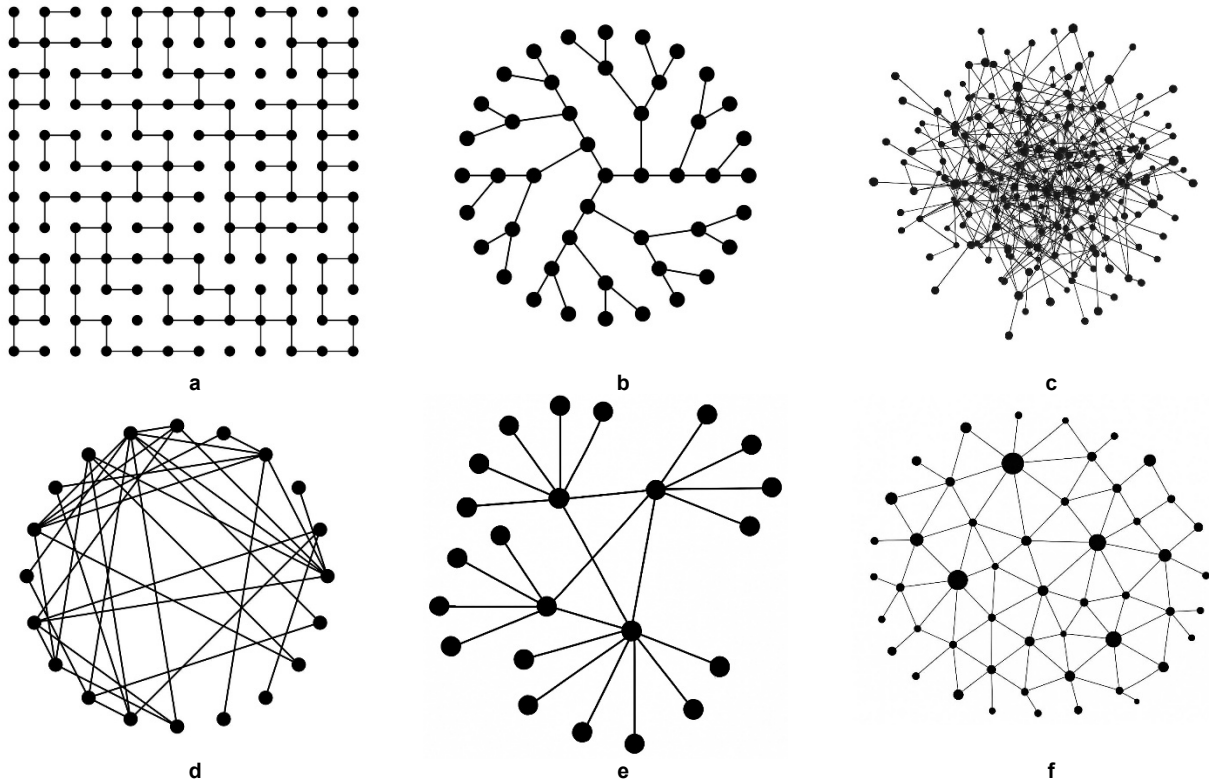


Fig. 1. Pore-network geometries: regular lattice (a), Bethe lattice (b), 3D random network with $p = 0.5$ (c), 2D random network with $p = 0.2$ (d), small-world network (e), and scale-free network (f)

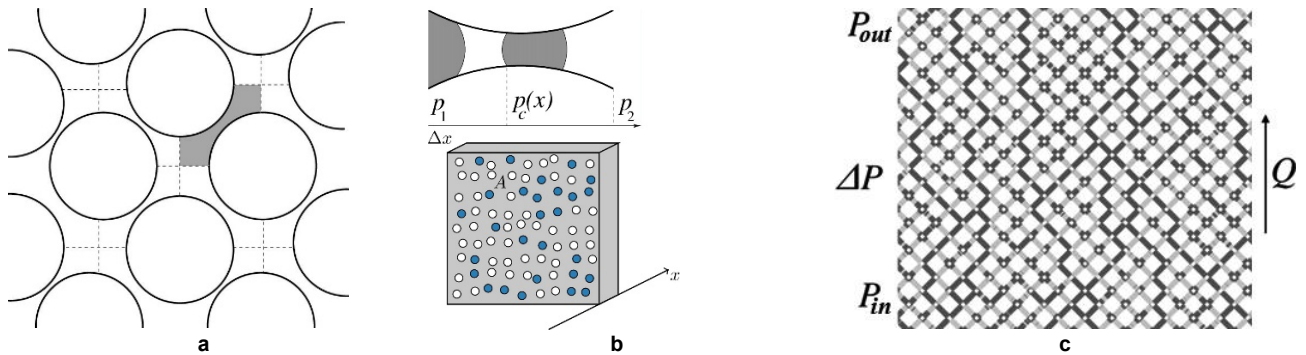


Fig. 2. Pore-space geometry: (a) Kozeny model, with the link between two vertices shown in gray; (b) a single link with two immiscible fluids flowing through the pore space; (c) a random rhomboidal network of nodes and channels with different randomly assigned hydraulic conductances, indicated by color

1.2.3. Continuum percolation models on random networks

In the case of two immiscible fluids with relative concentrations φ and $1 - \varphi$, the average flow velocity $\bar{v} = \varphi \bar{v}_1 + (1 - \varphi) \bar{v}_2$ is determined by the pressure difference at the inlet (P_{in}) and outlet (P_{out}) nodes of the network (Fig. 2c), where $\bar{v}_1$ and $\bar{v}_2$ are the velocities of fluids 1 and 2 (e. g., hydrophilic and hydrophobic), which satisfy Darcy's law:

$$\bar{v}_1 = -\frac{K_1(p, \varphi)}{\mu_1} \bar{\nabla} P_1, \quad \bar{v}_2 = -\frac{K_2(p, \varphi)}{\mu_2} \bar{\nabla} P_2, \quad (1)$$

where μ_1 , μ_2 are the fluid viscosities, and K_1 , K_2 are the permeabilities of the network edges corresponding to each fluid and dependent on the probability p . The local hydrostatic pressures P_1 , P_2 are related through the capillary pressure $P_c = P_1 - P_2$ (Savani et al., 2017a).

Darcy's law establishes a linear relationship between the pressure gradient and the filtration velocity of a fluid through a porous medium. This linear relationship holds only for slow flows with Reynolds number $Re < 10$. The filtration velocity at which Darcy's law ceases to be valid is referred to as the critical filtration velocity. The pore size distribution significantly affects the critical characteristics of the percolation process. Variability in material properties (network heterogeneity) also influences the velocity parameters of filtration.

For the systematic study of traditional problems within the framework of the hydrodynamic theory of filtration, the stochastic theory of filtration processes is employed.

2. Machine learning in percolation models

2.1. Machine learning methods

2.1.1. Supervised learning

Machine learning methods have rapidly become pervasive instruments due to better fitting quality and predictive quality in comparison with traditional models in terms of phase transitions and critical phenomena. Usually, machine learning can be divided into supervised and unsupervised learning. In the former, the machine receives a set of inputs and labels. Supervised learning models are trained with high accuracy to predict labels.

For the analysis of a two-dimensional percolation model, we use datasets obtained via Monte Carlo simulations and a convolutional neural network (CNN). The CNN architecture consists of four layers, including two convolutional layers and two fully connected layers (Fig. 3). Percolation configurations X are used as input data:

$$X = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,N} \end{pmatrix} \quad (2)$$

The CNN receives the permeability p as an output.

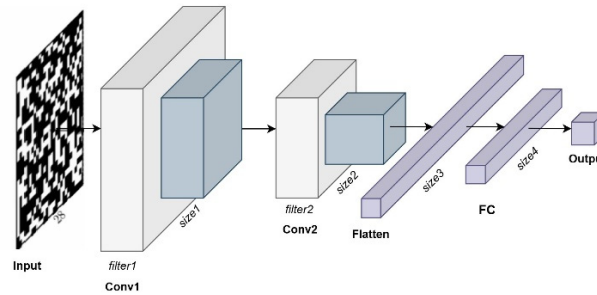


Fig. 3. The structure of the CNN with four layers, including "Conv1", "Conv2", "FC" та "Output"

2.1.2. Unsupervised learning

On the other hand, in unsupervised learning models, there are no labels. Unsupervised learning can be used as a meaningful analysis tool, such as sample generation, feature extraction, and cluster analysis.

The following unsupervised models are used to generate new initial configurations: a variational autoencoder (VAE) and a convolutional variational autoencoder (cVAE). A VAE is an architecture introduced by Diederik P. Kingma and Max Welling and belongs to the class of variational Bayesian methods (Kingma, & Welling, 2014). The VAE is composed of an encoder and a decoder. The VAE uses two different CNNs as two probability density distributions. The encoder in the VAE, called the inference network $P_{encoder}(Z|X)$, can generate the latent variables Z . The decoder in the VAE, called the generative network $P_{decoder}(\hat{X}|Z)$, reconstructs the raw configuration X .

To better reconstruct the raw two-dimensional configuration X , the fully connected layer in the VAE is replaced by the convolutional layer. Now we have got the cVAE. The architecture of the cVAE is shown in Fig. 4. Usually, the performance of the cVAE is better than the VAE due to the configuration X with spatial attribute (Cheng et al., 2021).

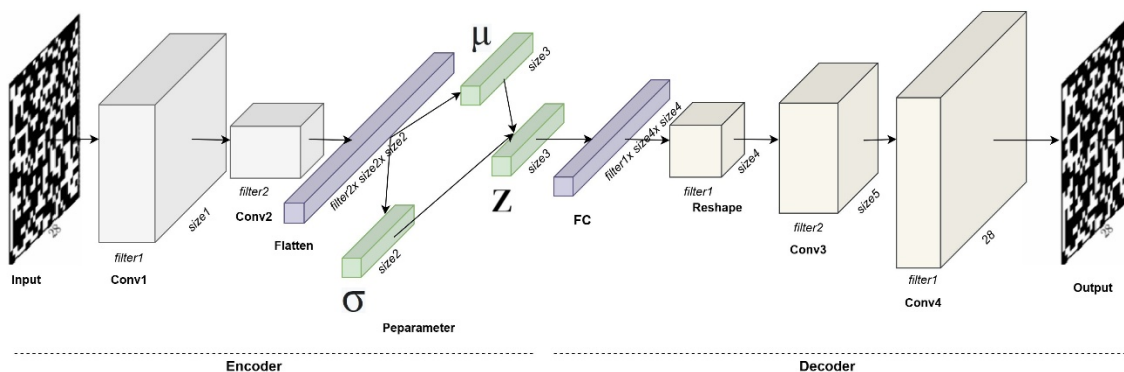


Fig. 4. The structure of the cVAE with an encoder and a decoder. The left part refers to the encoder with three layers, including the input layer, two convolutional layers, a flatten layer, and a fully connected layer with mean value " μ " and the standard deviation " σ ". The right part is a decoder with four layers, including the layer " Z ", a fully connected layer "FC", a reshape layer, two transposed convolutional layers, and an output layer

2.1.3. Training and optimization

For training, we make use of the TensorFlow library to perform the CNN with four layers. Let the first two layers of this CNN consist of two convolutional layers ("Conv1" and "Conv2"), both of which possess "filter1"=32 and "filter2"=64 filters of size 3x3, and a stride of 1. Each convolutional layer is followed by a max-pooling layer of size 2x2. The final convolutional layer ("Conv2") is strongly interlinked to a fully connected layer "FC" with 128 variables. The output layer ("Output"), followed by "FC", is a fully connected layer. For the two convolution layers and the fully connected layer "FC", a rectified linear unit (ReLU) $a = \max(0, x)$ is chosen as activation function due to its reliability and validity (Agarap, 2018). However, the output layer has no activation function.

After determining the framework of CNN, here we mention how to train CNN for deducing p . First, we carry out the Adam algorithm (Kingma, & Ba, 2014) as an optimizer to update parameters, i.e., weights and biases. Then, a mini batch size of 256 and a learning rate of 10^{-4} are selected for its timesaving. Following this treatment, CNN is trained for 1000 epochs on 40000 uncorrelated and shuffled configurations, respectively. Before training, we split X configurations (2) into a 32 000 training set and an 8000 testing set.

While training, we monitor using the loss function, i.e., mean squared error (MSE) (3) for training and testing sets (Cheng et al., 2021). y_i^{raw} and y_i^{pred} denote the true and predicted values of p in Eq. (3).

$$MSE = \frac{1}{M} \sum_{i=1}^M (y_i^{pred} - y_i^{raw})^2. \quad (3)$$

If the loss function on the test set reaches a minimum, the optimal CNN is obtained. As the number of epochs increases, the MSE continues to decrease until it reaches a plateau. Another performance metric of the CNN is the difference between the initial permeability p and its predicted value $\hat{p}$. The dark gray points in Fig. 5 indicate a strong positive correlation between the actual permeability p and its predicted value $\hat{p}$. Additional statistical tests show that most deviations between p and $\hat{p}$ are less than 0.1. An additional optimization step, principal component analysis (PCA) can be used to reduce the dimensionality of the input data before training, facilitating data analysis, noise reduction, and the analysis of latent variables (Yu, & Lyu, 2020).

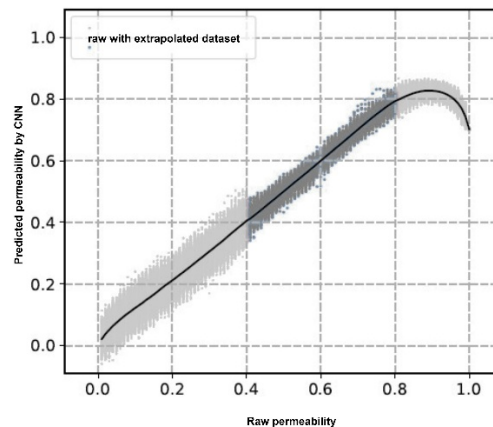


Fig. 5. Correlation between raw permeability and CNN-predicted permeability. Dark gray points denote data from the training range of the original dataset (0.4–0.8). Light gray points denote data from the extrapolated ranges (0–0.4 and 0.8–1.0). The black curve shows the mean predicted permeability for raw permeability values ranging from 0.01 to 1.0 with a step size of 0.01

Discussion and conclusions

As machine learning approaches are becoming increasingly popular and effective, they can be used to study percolation in porous-media models. In this paper, a CNN was used to model the permeability of porous media. VAE and cVAE were applied to generate new configurations – the generated variants preserve the main structural features but differ from the initial configurations, which confirms the robustness and generalization ability of the CNN. In addition, PCA was used for optimization in order to reduce the dimensionality of the initial configurations.

Neural networks provide fairly accurate predictions but require further cross-validation. The machine learning methods considered here are effective for studying physical characteristics in the percolation model. However, it should be acknowledged that they are based on a two-dimensional percolation model and therefore have certain limitations.

Authors' contribution: Oleg Afanasiev – machine learning, analysis of methods; Natalya Kizilova – conceptualization, theoretical studies (writing an original draft); Sergii Poslavskyi – random graphs, methodology.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Agarap, A. F. (2018). *Deep learning using rectified linear units (ReLU)*, *arXiv preprint*. <https://doi.org/10.48550/arXiv.1803.08375>
- Broadbent, S. R., & Hammersley, J. M. (1957). Percolation processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(03), 629. <https://doi.org/10.1017/s0305004100032680>
- Cheng, S., He, F., Zhang, H., Zhu, K., & Shi, Y. (2021). *Machine Learning Percolation Model*, *arXiv preprint*. <https://doi.org/10.48550/arXiv.2101.08928>
- Ewing, R. P., & Gupta, S. C. (1993). Modeling percolation properties of random media using a domain network. *Water Resources Research*, 29(9), 3169–3178. <https://doi.org/10.1029/93wr01496>
- Fyhn, H., Sinha, S., Roy, S., & Hansen, A. (2021). Rheology of immiscible two-phase flow in mixed wet porous media: dynamic pore network model and capillary fiber bundle model results. *Transport in Porous Media*, 139(3), 491–512. <https://doi.org/10.1007/s11242-021-01674-3>

- Gjennestad, M. A., Winkler, M., & Hansen, A. (2020). Pore network modeling of the effects of viscosity ratio and pressure gradient on steady-state incompressible two-phase flow in porous media. *Transport in Porous Media*, 132(2), 355–379. <https://doi.org/10.1007/s11242-020-01395-z>
- Goldsztajn, G. H. (2009). Clogging of Multigraphs as Toy Models of Filters. *SIAM J. Appl. Math.*, 70, 1078–1096. <https://doi.org/10.1137/090747579>
- Heydenreich, M., & Hofstad van der, R. (2017). *Progress in High-Dimensional Percolation and Random Graphs*. CRM Short Courses (CRMSC). Springer Cham. <https://doi.org/10.1007/978-3-319-62473-0>
- Hofstad van der, R. (2016). *Random Graphs and Complex Networks*. Cambridge University Press. <https://doi.org/10.1017/9781316779422>
- Joekar-Niasar, V., & Hassanizadeh, S. M. (2012). Analysis of fundamentals of two-phase flow in porous media using dynamic pore-network models: a review. *Critical Reviews in Environmental Science and Technology*, 42(18), 1895–1976. <https://doi.org/10.1080/10643389.2011.574101>
- Kim, M., & Radicchi, F. (2024). Shortest-path percolation on random networks. *Phys. Rev. Lett.*, 133, 047402. <https://doi.org/10.1103/PhysRevLett.133.047402>
- Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*, *arXiv preprint*. <https://doi.org/10.48550/arXiv.1412.6980>
- Kingma, D. P., & Welling, M. (2014). *Auto-encoding variational bayes*, *CoRR abs/1312.6114*. <https://doi.org/10.48550/arXiv.1312.6114>
- Koch, T., Weishaupt, K., Müller, J., Weigand, B., & Helmig, R. (2021). A (dual) network model for heat transfer in porous media. *Transport in Porous Media*, 140(1), 107–141. <https://doi.org/10.1007/s11242-021-01602-5>
- Li, M., Liu, R.-R., Lü, L., Hu, M.-B., Xu, S., & Zhang, Y.-C. (2021). Percolation on complex networks: Theory and application. *Physics Reports*, 907, 1–68. <https://doi.org/10.1016/j.physrep.2020.12.003>
- Rabbani, A., Babaei, M., & Javadpour, F. (2020). A triple pore network model (t-pnm) for gas flow simulation in fractured, micro-porous and meso-porous media. *Transport in Porous Media*. <https://doi.org/10.1007/s11242-020-01409-w>
- Savani, I., Bedeaux, D., Kjelstrup, S., Vassvik, M., Sinha, S., & Hansen, A. (2017a). Ensemble distribution for immiscible two-phase flow in porous media. *Physical Review E*, 95(2). <https://doi.org/10.1103/physreve.95.023116>
- Savani, I., Sinha, S., Hansen, A., Bedeaux, D., Kjelstrup, S., & Vassvik, M. (2017b). A Monte Carlo Algorithm for Immiscible Two-Phase Flow in Porous Media. *Transport in Porous Media*, 116(2), 869–888. <https://doi.org/10.1007/s11242-016-0804-x>
- Xiong, Q., Baychev, T. G., & Jivkov, A. P. (2016). Review of pore network modelling of porous media: Experimental characterisations, network constructions and applications to reactive transport. *Journal of Contaminant Hydrology*, 192, 101–117. <https://doi.org/10.1016/j.jconhyd.2016.07.00>
- Yu, W. & Lyu, P. (2020). Unsupervised machine learning of phase transition in percolation, *Physica A: Statistical Mechanics and its Applications*, 559, 125065. <https://doi.org/10.1016/j.physa.2020.125065>
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1, 57–81. <https://doi.org/10.1016/j.aiopen.2021.01.001>

Отримано редакцією журналу / Received: 18.07.25
Прорецензовано / Revised: 04.09.25
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Олег АФАНАСЬЄВ, асп.
ORCID ID: 0009-0007-9491-6456
e-mail: oleg.afanasiev@karazin.ua
Харківський національний університет ім. В. Н. Каразіна, Харків, Україна

Наталія КІЗІЛОВА, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-9981-7616
e-mail: kizilova@karazin.ua
Харківський національний університет ім. В. Н. Каразіна, Харків, Україна

Сергій ПОСЛАВСЬКИЙ, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-1049-9947
e-mail: s.poslavsky@karazin.ua
Харківський національний університет ім. В. Н. Каразіна, Харків, Україна

ВИПАДКОВІ МЕРЕЖІ У МОДЕЛЮВАННІ ТЕПЛОМАСОПЕРЕНОСУ В ПОРИСТИХ МАТЕРІАЛАХ

Представлено детальний огляд літератури з математичних моделей процесів тепломасопереоносу в пористих середовищах на основі випадкових графів (випадкових мереж). У цих моделях окремі пори (вузли графа) вважаються з даною ймовірністю зв'язаними між собою вузькими каналами (ребрами графа), крізь які рухається рідина або газ. Наведено класифікацію головних типів випадкових мереж, що використовуються в математичному моделюванні пористих середовищ, а також відповідних дискретних і континуальних математичних моделей для однокомпонентних ньютоновських і неньютоновських рідин. У зв'язаних задачах теплопереоносу розглянуто подвійні випадкові мережі, з яких одна відповідає тепломасопереоносу з рідиною, а друга – переоносу крізь тверду фазу. Наведено наявні алгоритми побудови випадкових графів, математичні моделі тепломасопереоносу на мережах і результати моделювання фільтрації рідини, кількох змішуваних або незмішуваних ньютоновських і неньютоновських рідин. Крім можливого зникнення ребер мережі внаслідок суфозії та/або агрегації частинок рідини / твердої фази, можливим також є відновлення зв'язків між сусідніми порами шляхом руйнування і вимивання твердої фази, а також дезагрегації / розпаду конгломератів частинок. Результати оглядового аналізу літератури можуть бути використані для вибору найкращої для конкретного випадку пористого середовища моделі випадкового графа (мережі), дискретної або континуальної математичної моделі, а також чисельного методу розрахунків. Розглянуто застосування методів машинного навчання для дослідження фізичних характеристик пористого середовища в моделі перколяції.

Ключові слова: течії в пористих середовищах, випадкові графи, методи машинного навчання, математичне моделювання.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 004.056.5

DOI: <https://doi.org/10.17721/1812-5409.2026/1.24>Олексій БАШУК, асп.
ORCID ID: 0009-0002-6778-2943
e-mail: a.bashuk@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПОСТКВАНТОВИХ ПРОТОКОЛІВ ЦИФРОВОГО ПІДПISУ, БАЗОВАНИХ НА ГЕШ-ФУНКЦІЯХ

В епоху активного розвитку інформаційних технологій і пристроїв поява портативних та загальнозастосованих квантових комп'ютерів є лише питанням часу. І хоч користь від їхньої в рази більшої ефективності, порівняно зі звичайними бітовими комп'ютерами, буде очевидною, водночас вони принесуть і багато вразливостей до сучасних алгоритмів криптографії та кібербезпеки. Щоб запобігти цьому, фахівці вже розпочали роботу над дослідженнями постквантових протоколів замість сучасних.

Не стали винятком і протоколи цифрового підпису. Наприклад, одним із підходів розв'язання окресленої проблеми є побудова постквантових протоколів цифрового підпису, базованих на геш-функціях. Першими варіантами стали схеми одноразових підписів, базованих на геш-функціях, які були захищені від можливостей квантових комп'ютерів, проте мали очевидну ваду одноразовості. Найвідомішим вважають так званий одноразовий підпис Лампорта. Однак цей недолік одноразовості усунули, запропонувавши різні схеми багаторазових цифрових підписів, базованих на створенні та використанні багатьох одноразових підписів. Були запропоновані ланцюгоподібні та деревоподібні схеми багаторазових підписів. Саме реалізації та порівнянню різних схем багаторазового підпису і присвячено пропонувану роботу.

У ході досліджень наведено теоретичні оцінки використання пам'яті та часу, необхідні для роботи різних схем. Подальша реалізація цих схем й остаточні заміри виявили відповідність теоретичних оцінок емпіричним результатам, а також явно продемонстрували переваги та недоліки різних схем. Найкращу продуктивність, зокрема, показали деревоподібні схеми, особливо ті, у яких дерево будується послідовно. Проте в певних випадках має сенс використовувати дерево з повним заглибленням під час побудови.

З огляду на відносну новизну та малу кількість досліджень у цій галузі, багаторазові цифрові підписи мають великий простір для подальших досліджень, наприклад, розгляд N-арних дерев та їхнього порівняння з бінарними.

Ключові слова: геш-функція, цифрові підписи, постквантові протоколи, одноразові цифрові підписи, багаторазові цифрові підписи, ланцюгоподібні схеми, деревоподібні схеми.

Класифікація відповідно до AMS 2020: 81P94, 94A60.

Вступ

Технологічний прогрес не стоїть на місці, і в майбутньому одним із ключових чинників у ньому стануть квантові комп'ютери. Проте до найбільших ризиків їхнього впровадження належить вразливість поточних криптографічних алгоритмів, в основі яких лежить розрахунок обмеженості звичайних бітових комп'ютерів. Традиційні криптосистеми спираються на задачі факторизації цілих чисел або задачі дискретного логарифмування, які легко можна розв'язати на достатньо великих квантових комп'ютерах. З огляду на це останнім часом стали дуже актуальними дослідження у сфері постквантових протоколів у криптографії, що є незалежними від квантових обчислень, а тому стійкими до квантових атак. Протоколи, засновані на геш-функціях, можуть бути стійкими до квантових обчислень.

Не є винятком і цифрові підписи. Найтиповішим прикладом постквантового цифрового підпису став одноразовий підпис Лампорта (Boneh, & Shoup, 2017). Іншими словами, надійність цього підпису безпосередньо впливає з його одноразовості. Ідейно алгоритм полягає у такому. Будується приватний ключ, до якого входять пари випадкових чисел на кожен біт вхідного повідомлення. Для побудови ж публічного ключа обирають певну геш-функцію H , та з її допомогою знаходять геш-значення кожного елемента приватного ключа. Підписом повідомлення за допомогою такого ключа буде множина значень в обсязі кількості бітів вхідного повідомлення, що обираються з кожної пари приватного ключа залежно від значення відповідного біта вхідного повідомлення. За такої побудови легко переконатися, що це дійсно є підписом повідомлення, прогешувавши значення за допомогою функції H , та звіривши з відповідними значеннями публічного ключа. Водночас, не знаючи другої половини приватного ключа, ніхто інший не зможе в майбутньому скористатися ним для підпису повідомлень, відмінних від уже підписаного. Є й інші одноразові підписи, такі як WOTS (Merkle, 1989), WOTS+ (Hülsing, 2013), Bleichenbacher-Maurer OTS (Bleichenbacher, & Maurer, 1994), BiBa OTS (Perrig, 2001) та HORS (Reyzin, L., & Reyzin, N., 2002).

Проте в цих підписах все ще є недолік – вони одноразові. Напрошується логічне бажання – об'єднати багато ключів одноразових підписів в один. З огляду на це був запропонований підпис Меркеля (Boneh, & Shoup, 2017) на основі дерева Меркеля (Katz, & Lindell, 2014, с. 183–184). Створюється певна кількість одноразових ключів і на їхніх публічних ключах, як на листках, будується бінарне дерево, де кожна вершина містить геш-значення її дітей, а значення в корені – публічний ключ такого підпису. Тоді для підписання щоразу використовують черговий одноразовий підпис, що ще не був використаний, а також з ним надають усі необхідні значення з дерева на шляху від відповідного листка, яким було підписано повідомлення, до кореня дерева. Тоді для верифікації перевіряють безпосередньо сам одноразовий підпис, а також те, що надані значення дійсно є елементами дерева, що відповідає публічному ключу кореня.

І хоч такий підхід забезпечує багаторазовість, він все ще є обмеженим. Одним із виходів стали так звані ланцюгові підписи (Chain-Based Signature) (Katz, & Lindell, 2014, р. 465–468). У них теж за основу беруть певний одноразовий підпис. Під час підписання повідомлення підписують не тільки саме повідомлення, але також публічний ключ нового свіжоствореного одноразового підпису. Під час наступного підписання вже використовують цей новий підпис, за допомогою якого підписують нове повідомлення, а також публічний ключ ще одного створеного підпису. Таким чином вибудовується ланцюг із підписів. Щоб можна було верифікувати будь-який з них, потрібно надати весь ланцюг разом

© Башук Олексій, 2026

з відповідними публічними ключами та вхідними повідомленнями, що були підписані від початкового підпису до підпису поточного повідомлення. Разом з початковим публічним ключем, легко переконалися в дійсності кожного з підписів в ланцюзі, зокрема і в тому, що цікавить перевіряльника. Тобто фактично "підписом" повідомлення вважають не безпосередньо зроблений у ланцюзі останній підпис, а весь поточний стан ланцюга разом з останнім підписом. У такого підходу є очевидний недолік – з кожним новим підписом зростає як розмір самого підпису, так і час, необхідний на його верифікацію.

Однак ідея такого підходу давала можливість використання багаторазових цифрових підписів на основі геш-функцій і, як наслідок, привела до реалізації різного роду комбінацій і надбудов. Наприклад, був запропонований підпис на основі HMAC (Hash-based Message Authentication Code) (Lizama-Pérez, 2022). У протоколі цього підпису було використано одноразовий цифровий підпис WOTS, де, попри його одноразовість, він певним чином використовувався двічі – для підписання поточного основного повідомлення та верифікації наступного, а поверх цього був використаний HMAC – для перевірки цілісності інформації під час комунікації.

Щоб скоротити розмір підпису, було запропоновано поєднання ідеї ланцюгового підпису та дерева Меркеля. Наприклад, для побудови деревоподібного підпису (Tree-Based Signature) (Katz, & Lindell, 2014, p. 468–473) пропонується будувати бінарне дерево, аналогічне дереву Меркеля, тільки кожна вершина містить у собі не геш-значення своїх дочірніх вершин, а їхній підпис. Для цього для кожної дочірньої вершини створюють пари ключів для підпису, якщо вони ще не були створені раніше, і після цього підписують публічні ключі цих дочірніх вершин. Крім того, кожній вершині відповідає деякий рядок "a" з 0 та 1 такий, що в дочірніх вершин поточної вершини будуть рядки "a0" та "a1", а кореню при цьому відповідає порожній рядок. Вершини, у яких довжина рядка дорівнює довжині бітового значення вхідних повідомлень, вважають листками. Однак це дерево не будується повністю одразу, а розбудовується з часом. Публічним ключем такого дерева підписів буде публічний ключ вершини дерева. Під час підписання ж чергового вхідного повідомлення йде спуск по дереву від кореня так, щоб рядок кожної вершини був префіксом бітового запису вхідного повідомлення. Якщо в якійсь вершині ще немає дітей і вона не є листком, як уже зазначалося, для неї створюються ключі одноразових підписів та підписуються їхні публічні ключі за допомогою приватного ключа поточної вершини. Якщо ж вона є листком, то за допомогою її приватного ключа підписується відповідне вхідне повідомлення. Якщо ж таке вхідне повідомлення вже було підписане раніше – використовується збережене значення. У підсумку, підписом у такому дереві підписів буде безпосередньо сам підпис вхідного повідомлення у листку, усі підписи у вершинах на шляху від листка до кореня, а також публічні ключі відповідних дочірніх вершин. Для верифікації ж просто перевіряють, що всі підписи є дійсними (а отже, і публічні ключі дочірніх вершин теж), а також сам підпис вхідного повідомлення.

За такої побудови, всі підписи будуть мати фіксовану довжину, що відповідатиме довжині бітового запису вхідних повідомлень, а тому, починаючи з деякого моменту, довжина підпису та час, необхідний на його верифікацію, будуть кращими, ніж у підпису з ланцюговою структурою. Проте для зберігання цього дерева буде потрібно більше пам'яті, ніж для зберігання підпису з ланцюговою структурою. Крім цього, на відміну від дерева Меркеля, немає необхідності в побудові одразу всього дерева, через що і зникає обмеженість явного ліміту кількості одноразових підписів (залишається лише обмеження пристроїв на зберігання повного такого дерева). Також варто зауважити, що за такої структури в підписі вхідних повідомлень не фігурують раніше підписані вхідні повідомлення.

Якщо ж фігурування попередніх вхідних повідомлень у поточному підписі повідомлення не є проблемою, можна будувати подібне альтернативне дерево. Якщо запропоноване вище дерево будується за аналогією до алгоритму пошуку в глибину, добираючись до листків від кореня дерева фіксованого розміру, то можна будувати дерево за аналогією до алгоритму пошуку в ширину, а саме будувати вершини послідовно залежно від їхньої відстані до кореня дерева (Sequential Tree-Based Signature). При кожному новому підписанні вхідного повідомлення обирається наступна в черзі вершина, для неї створюються дві дочірні вершини з новими одноразовими підписами кожна та додаються в кінець черги, і, за допомогою приватного ключа поточної вершини, підписуються публічні ключі новостворених дочірніх вершин разом із вхідним повідомленням. Підписом у такому дереві, відповідно, вважають усі підписи на шляху від поточної вершини до кореня, разом із публічними ключами дочірніх вершин та відповідними минулими вхідними повідомленнями. За такої побудови кількість вершин, що фігурують в одному підписі, буде суттєво меншою, а отже, і кількість верифікацій одноразових підписів буде меншою, що має вплинути як на час підпису та верифікації, так і на розмір одного підпису, порівняно з попереднім запропонованим деревом. З іншого боку, в підписі будуть також фігурувати попередньо підписані вхідні повідомлення, що збільшить розмір однієї ланки підпису.

Ще однією альтернативою буде аналогічний підхід до попереднього, але в якому в підписах не будуть використовувати попередні підписані повідомлення. Натомість кожна вершина матиме не дві дочірні вершини, а три (Wide Sequential Tree-Based Signature). Ліва та права – зберігають свій звичайний функціонал і використовуються в подальшій побудові дерева, тоді як центральна вершина є листочком дерева і використовується лише для підписання поточного повідомлення. Завдяки такій додатковій генерації листка для підписання повідомлень у кожній звичайній вершині дерева, для перевірки проміжних підписів більше не потрібно знати відповідні проміжні повідомлення. Замість них для перевірки проміжних підписів надають відповідні публічні ключі листків. Як наслідок, у підписах не використовуються раніше підписані повідомлення, проте генерується в півтора рази більше одноразових підписів, що впливає як на час, так і на розмір дерева.

Ця робота спрямована на розроблення реалізацій різних видів багаторазових підписів на основі одноразових, а також порівнянню їхньої продуктивності.

1. Базові заміри

Усі вимірювання та дослідження проводилися на базі ноутбуку MacBook Pro 2021 року з процесором Apple M1 Pro та 32 ГБ оперативної пам'яті. Повний код реалізації можна знайти за посиланням: <https://github.com/albashuk/Dissertation/tree/master/Experiments/Python>.

В основі всіх названих одноразових підписів лежить використання геш-функцій Hash, що повертають HashSize-бітові значення, від вибору яких напряму залежать результати. Для проведення досліджень було обрано та використано одну з найпопулярніших геш-функцій SHA-256 (NIST, 2015). Заміри продуктивності наведені у табл. 1.

Таблиця 1

Заміри функції SHA-256		
Кількість ітерацій	Середній час гешування, с	Середній розмір гешу, Байт
1	0.00000715256	152.000
10	0.00000162125	196.000
100	0.00000109196	192.560
1000	0.00000088501	192.792
10000	0.00000094020	192.428
100000	0.00000079224	192.000
1000000	0.00000072344	192.447
10000000	0.00000071610	192.909

2. Теоретична оцінка цифрових підписів

Кожна схема цифрового підпису має три ключові функції: GEN – генерація публічного та приватного ключів підпису, SIGN – генерація цифрового підпису повідомлення за допомогою приватного ключа, VRFY – верифікація цифрового підпису за допомогою відповідного публічного ключа. Саме на часових замірах цих трьох функцій і буде сконцентрована увага. Проте в реальних умовах, у ході підписання повідомлення та верифікації підпису, частину часу займає пересилання підписів через мережу. Ця величина дуже залежить від умов з'єднання та може значно різнитися. Водночас за конкретно заданих умов ця величина лінійно залежить від розміру даних, що передаються. Тому як найбільш оптимальне рішення буде вимірюватися час виконання суто алгоритмічних дій кожної з трьох ключових функцій, а також розмір даних, що пересилаються.

2.1. Лампорт

Для генерації приватного ключа треба створити HashSize пар випадкових г-значень, а для публічного ключа – набір гешів до кожного із геш-значень у приватному ключі. Відповідно часова складність буде:

$$O(T_{GEN}) = HashSize * 2 * O(T_{Hash}) + HashSize * 2 * O(T_{Hash}), \quad (1)$$

а обсяг пам'яті публічного та приватного ключів:

$$O(S_{sk}) = O(S_{pk}) = HashSize * 2 * O(S_{Hash}). \quad (2)$$

Для підписання повідомлення потрібно спершу перевести його в HashSize-бітове значення. Після цього обирають HashSize геш-значень приватного ключа відповідного до бітів значення вхідного повідомлення, а тоді:

$$O(T_{SIGN}) = O(T_{Hash}) + 256 * O(T_1), \quad (3)$$

$$O(S_{SIGN}) = 256 * O(S_{Hash}). \quad (4)$$

Для верифікації підпису потрібно перевести вхідне повідомлення в HashSize-бітове значення, пройти по всіх геш-значеннях підпису, знайти геші від цих геш-значень і порівняти їх із відповідними значеннями публічного ключа відповідно до вхідного повідомлення. Тобто

$$O(T_{VRFY}) = O(T_{Hash}) + HashSize * O(T_{Hash}). \quad (5)$$

Для подальших розрахунків уведемо позначення OTS (One-Time Signature), як схеми одноразового підпису.

2.2. CBS (Chain-Based Signature)

Генерація приватного та публічного ключів повністю відповідають генерації приватного та публічного ключів відповідної схеми OTS. Для підписання нового повідомлення потрібно створити нову пару публічного та приватного ключів OTS. За допомогою останнього приватного ключа в списку необхідно підписати повідомлення разом із новоствореним публічним ключем та додати новостворені ключі разом з останнім підписом і повідомленням у кінці відповідних списків. Підписом повідомлення в цій схемі будуть поточні стани списків повідомлень, підписів і публічних ключів OTS, що були використані. Тоді необхідно знайти:

$$O(T_{SIGN}) = O(T_{OTS_GEN}) + O(T_{OTS_SIGN}), \quad (6)$$

$$O(S_{SIGN}) = N * (O(S_{OTS_msg}) + O(S_{OTS_pk}) + O(S_{OTS_SIGN})), \quad (7)$$

де N – кількість підписаних повідомлень, включно з поточним.

Для верифікації підпису потрібно пройти по всіх OTS-підписах із загального підпису та перевірити їхню дійсність, тобто:

$$O(T_{VRFY}) = N * O(T_{OTS_VRFY}), \quad (8)$$

де N – кількість OTS підписів у наданому CBS підписі.

Як бачимо, розмір підпису та час його верифікації з кожним новим підписом зростає лінійно.

2.3. TBS (Tree-Based Signature)

Генерація приватного та публічного ключів повністю відповідають генерації приватного та публічного ключів відповідної схеми OTS. Для підписання нового повідомлення, треба спершу перевести його в HashSize-бітове значення, потім, відповідно до бітів повідомлення, спуститися по дереву від кореня до листка, паралельно створюючи нові вершини дерева, та підписуючи їхні публічні ключі, за потреби. За допомогою ж приватного ключа листка підписується безпосередньо саме повідомлення. Підписом повідомлення в цій схемі будемо вважати списки підписів у вершинах на шляху від відповідного листка до кореня, а також відповідні публічні ключі з дочірніх вершин, що відповідають цим підписам. З огляду на це

$$O(T_{SIGN}) = HashSize * (O(T_{OTS_GEN}) + O(T_{OTS_GEN}) + O(T_{OTS_SIGN})) + O(T_{OTS_SIGN}), \quad (9)$$

$$O(S_{SIGN}) = HashSize * (O(S_{(OTS_pk)}) + O(S_{(OTS_pk)}) + O(S_{(OTS_SIGN)})) + O(S_{(OTS_SIGN)}). \quad (10)$$

А для верифікації підпису потрібно пройти по всіх OTS-підписах із загального підпису та перевірити їхню дійсність, тобто:

$$O(T_{VRFY}) = HashSize * O(T_{OTS_VRFY}). \quad (11)$$

Як бачимо, розмір підпису та час його верифікації з кожним новим підписом залишається сталим.

2.4. STBS (Sequential Tree-Based Signature)

Генерація приватного та публічного ключів повністю відповідають генерації приватного та публічного ключів відповідної схеми OTS. Для підписання нового повідомлення, необхідно взяти наступну в черзі обходу дерева в ширину вершину (можна легко відслідковувати за допомогою вказівників), створити дві дочірні вершини з новими парами ключів для них і за допомогою приватного ключа поточної вершини підписати повідомлення разом із публічними ключами двох її дочірніх вершин. Підписом повідомлення в цій схемі будуть вважатися списки підписів у вершинах на шляху від відповідного листка до кореня, а також відповідні публічні ключі з дочірніх вершин і підписані раніше повідомлення, що відповідають цим підписам. Зважаючи на це,

$$O(T_{SIGN}) = O(T_{OTS_GEN}) + O(T_{OTS_GEN}) + O(T_{OTS_SIGN}), \quad (12)$$

$$O(S_{SIGN}) = \log_2(N) * (O(S_{OTS_msg}) + O(S_{OTS_pk}) + O(S_{OTS_pk}) + O(S_{OTS_SIGN})), \quad (13)$$

де N – кількість підписаних повідомлень.

Для верифікації підпису треба пройти по всіх OTS підписах із загального підпису та перевірити їхню дійсність, тобто:

$$O(T_{VRFY}) = M * O(T_{OTS_VRFY}), \quad (14)$$

де M – кількість OTS підписів у наданому STBS-підписі.

2.5. WSTBS (Wide Sequential Tree-Based Signature)

Ця схема ідею повністю аналогічна попередній з тою відмінністю, що звичайні вершини дерева генерують не дві, а три вершини, тому:

$$O(T_{SIGN}) = O(T_{OTS_GEN}) + O(T_{OTS_GEN}) + O(T_{OTS_GEN}) + O(T_{OTS_SIGN}), \quad (15)$$

$$O(S_{SIGN}) = \log_2(N) * (O(S_{OTS_pk}) + O(S_{OTS_pk}) + O(S_{OTS_pk}) + O(S_{OTS_SIGN})) + O(S_{OTS_msg}), \quad (16)$$

$$O(T_{VRFY}) = M * O(T_{OTS_VRFY}), \quad (17)$$

де N – кількість підписаних повідомлень, а M – кількість OTS підписів у наданому WSTBS-підписі.

На перший погляд останні дві схеми переважають над схемою TBS як за часом, необхідним на підписання і верифікацію, так і за розмірами самого підпису. Натомість, як зазначається в книзі (Katz, & Lindell, 2014, р. 472–473), такий підхід теоретично дає можливість втілити схему без зберігання стану підписів. Отже, конкретизуємо: оскільки в основі всіх генерацій ключів лежать псевдовипадкові функції, за правильної реалізації, якщо знати корінне значення (сід) цих псевдовипадкових функцій, то будь-яку заздалегідь обрану пару публічного та приватного ключів одноразового підпису можна легко відтворити. Однак, урахувавши, що всі значення, які підписуються цими приватними ключами, за фіксованого значення сіда повністю детерміновані, а від підписуваного повідомлення просто обирається, яка гілка дерева стане підписом, то в постійному підтриманні стану підписів немає потреби, оскільки, знаючи сід, можна легко відтворити весь підпис на ходу. І хоча надійність такого підходу значно спирається на правильний вибір псевдовипадкової функції, натомість зникає потреба в постійній підтримці та зберіганні стану підписів.

3. Опис реалізації основних функцій

У цьому підрозділі будуть розглянуті реалізації основних функцій GEN, SIGN та VRFY для кожного реалізованого протоколу.

3.1. Лампорт

Для генерації приватного ключа створюється масив розмірності $2 \times \text{HashSize}$, що заповнюється гешами від випадково згенерованих значень. Після цього з нього формується публічний ключ, шляхом створення масиву такої самої розмірності, де кожне значення – геш від відповідного значення приватного ключа:

```
def gen() -> sk, pk:
    sk = [[SHA256(randomValue()) for j in 0..HashSize-1] for i in 0..1])
    pk = [[SHA256(sk[i][j]) for j in 0..HashSize-1] for i in 0..1])
    return sk, pk
```

Щоб підписати повідомлення, достатньо створити масив розмірності HashSize, який заповнити значеннями приватного ключа, відповідно до бітів підписуваного повідомлення:

```
def sign(sk, msg) -> sign:
    sign = [sk[((msg >> j) & 0b1)][j] for j in 0..HashSize-1])
    return sign
```

Для верифікації підпису достатньо переконаватися, що геш-значення від елементів з підпису відповідають значенням публічного ключа, відповідно до бітів підписуваного повідомлення:

```
def vrfy(pk, sign, msg) -> bool:
    for j in 0.. HashSize-1:
        if SHA256(sign[j]) != pk[((msg >> j) & 0b1)][j]:
            return False
    return True
```

3.2. CBS (Chain-Based Signature)

Генерація пари ключів CBS повністю відповідає генерації пари ключів, обраної OTS, із тою відмінністю, що приватний ключ CBS насправді являє собою структуру, яка зберігає списками усі проміжні повідомлення, їх підписи, а також згенеровані приватні та публічні ключі від одноразових підписів:

```
def gen() -> sk, pk:
    otsk, otpk = OTS.gen()
    sk = {otsks: [otsk], otpks: [None], otsigns: [], msgsgs: []}
    pk = otpk
    return sk, pk
```

Під час підписання генерується нова пара ключів одноразового підпису, публічний ключ якої разом із повідомленням підписується поточним приватним ключем одноразового підпису. Після чого сформований підпис, разом із повідомленням і новоствореними приватним та публічним ключами додаються в кінці відповідних списків приватного ключа підпису CBS. У кінці формується підпис CBS, що складається зі списків одноразових публічних ключів, одноразових підписів і повідомлень із секретного ключа CBS:

```
def sign(sk, msg) -> sign:
    otsk, otpk = OTS.gen()
    otsign = OTS.sign(sk.otsks.last(), msg || otpk)
    sk.add(otsk, otpk, otsign, msg)
    sign = {otpk: sk.otpk, otsigns: sk.otsigns, msg: sk.msgs}
    return sign
```

У процесі верифікації підпису відбувається перевірка всіх наданих одноразових підписів із підпису CBS. Для простоти перевірки, перший елемент зі списку одноразових публічних ключів замінюється на публічний ключ CBS:

```
def vrfy(pk, sign, msg) -> bool:
    if sign.msgs.last() != msg:
        return False
    sign.otpk[0] = pk
    for otpk, next_otpk, otsign, msg in sign:
        if not OTS.vrfy(otpk, otsign, msg || otpk):
            return False
    return True
```

3.3. TBS (Tree-Based Signature)

Генерація пари ключів TBS повністю відповідає генерації пари ключів обраної OTS, із тією відмінністю, що приватний ключ TBS насправді являє собою структуру, яка зберігає список вершини дерева, кожна з яких або містить у собі певне повідомлення і його одноразовий підпис та відповідну пару одноразових ключів, або посилання на дві дочірні вершини, одноразовий підпис їхніх публічних ключів та відповідну пару одноразових ключів:

```
def gen() -> sk, pk:
    otsk, otpk = OTS.gen()
    sk = [{otsk: otsk, otpk: otpk, left: None, right: None, sign: None, msg: None}]
    pk = otpk
    return sk, pk
```

Під час підписання повідомлення відбувається спуск униз по дереву до листка, значенню якого відповідає підписуване повідомлення. Якщо на шляху у вершини не існує дочірніх вершин, – вони створюються разом із парами одноразових ключів підписів, і новостворені публічні ключі підписуються одноразовим приватним ключем поточної вершини. Після цього відбувається заглиблення в одну з дочірніх вершин, відповідно до біту вхідного повідомлення. Коли ж досягається листкова вершина, відбувається підписання самого повідомлення. Підпис при цьому має аналогічну структуру до підпису CBS, але будується з елементів на шляху до листка і також містить публічні ключі дочірніх вершин до всіх, що є на цьому шляху:

```
def sign(sk, msg) -> sign:
    cur = 0
    sign = {otpk: [], lotpk: [], rotpk: [], otsigns: [], msg: []}
    for height in 0..HashSize-1:
        if sk[cur].sign == None:
            sk[cur].left = sk.size()
            lotsk, lotpk = OTS.gen()
            sk.add({otsk: lotsk, otpk: lotpk, left: None, right: None, sign: None, msg: None})

            sk[cur].right = sk.size()
            rotsk, rotpk = OTS.gen()
            sk.add({otsk: rotsk, otpk: rotpk, left: None, right: None, sign: None, msg: None})

            sk[cur].sign = OTS.sign(sk[cur].otsk, lotpk || rotpk)
            sk[cur].msg = None
        else:
            lotpk, rotpk = sk[sk[cur].left].otpk, sk[sk[cur].right].otpk

            sign.add(sk[cur].otpk, lotpk, rotpk, sk[cur].sign, sk[cur].msg)
            cur = ((msg >> height) & 0b1) ? sk[cur].right : sk[cur].left

    sk[cur].sign = OTS.sign(sk[cur].otsk, msg)
    sk[cur].msg = msg
    sign.add(sk[cur].pk, sk[cur].sign, sk[cur].msg)
    sign.otpk[0] = None
    return sign
```

Зважаючи на аналогічність структури підпису зі схемою CBS, код верифікації ідейно ідентичний:

```
def vrfy(pk, sign, msg) -> bool:
    if sign.msgs.last() != msg:
        return False
    sign.otps[0] = pk
    for otpk, lotpk, rotpk, otsign, msg in pk:
        if not OTS.vrfy(otpk, otsign, msg || lotpk || rotpk):
            return False
    return True
```

3.4. STBS (Sequential Tree-Based Signature)

Генерація пари ключів STBS повністю дублює генерацію пари ключів TBS для обраної OTS із тією відмінністю, що кожна вершина дерева також має посилання на батьківську, а разом із самим деревом також зберігається вказівник на наступну вершину для підпису нового повідомлення:

```
def gen() -> sk, pk:
    otsk, otpk = OTS.gen()
    sk = {tree: [{otsk: otsk, otpk: otpk, prev: None, left: None, right: None, sign: None, msg: None}], head: 0}
    pk = otpk
    return sk, pk
```

Під час підписання повідомлення спершу береться наступна вершина в дереві. Для неї генеруються дві дочірні вершини, разом із парами одноразових публічних ключів кожна, і новостворені публічні ключі, разом із новим повідомленням, підписуються одноразовим приватним ключем поточної вершини. Після цього будується підпис STBS за принципом проходження від поточної вершини до кореня, у який також додаються публічні ключі дочірніх вершин до всіх, що є на цьому шляху:

```
def sign(sk, msg) -> sign:
    sign = {otps: [], lotps: [], rotps: [], otsigns: [], msgs: []}
    cur = sk.head

    sk.tree[cur].left = sk.tree.size()
    lotsk, lotpk = OTS.gen()
    sk.tree.add({otsk: lotsk, otpk: lotpk, prev: cur, left: None, right: None, sign: None, msg: None})

    sk.tree[cur].right = sk.tree.size()
    rotsk, rotpk = OTS.gen()
    sk.tree.add({otsk: rotsk, otpk: rotpk, prev: cur, left: None, right: None, sign: None, msg: None})

    sk.tree[cur].sign = OTS.sign(sk.tree[cur].otsk, msg || lotpk || rotpk)
    sk.tree[cur].msg = msg

    while cur != None:
        lotsk, rotpk = sk[sk[cur].left].otpk, sk[sk[cur].right].otpk
        sign.addFromBegin(sk.tree[cur].pk, lotpk, rotpk, sk.tree[cur].sign, sk.tree[cur].msg)
        cur = sk.tree[cur].prev

    sk.head = sk.head + 1
    return sign
```

З огляду на аналогічність структури підпису зі схемою TBS, код верифікації ідейно ідентичний.

3.5. WSTBS (Wide Sequential Tree-Based Signature)

Генерація пари ключів WSTBS повністю дублює генерацію пари ключів STBS для обраної OTS із тією відмінністю, що кожна вершина дерева, окрім листків, також має посилання на листок дерева, яким уже підписується повідомлення:

```
def gen() -> sk, pk:
    otsk, otpk = OTS.gen()
    sk = {tree: [{otsk: otsk, otpk: otpk, prev: None, left: None, mid: None, right: None, sign: None, msg: None}], head: 0}
    pk = otpk
    return sk, pk
```

Під час підписання повідомлення спершу береться наступна вершина в дереві. Для неї генеруються три дочірні вершини разом із парами одноразових публічних ключів кожна, і новостворені публічні ключі підписуються одноразовим приватним ключем поточної вершини. Після цього будується підпис WSTBS за принципом проходження від поточної вершини до кореня:

```
def sign(sk, msg) -> sign:
    sign = {otps: [], lotps: [], motps: [], rotps: [], otsigns: [], msgs: []}
    cur = sk.head

    sk.tree[cur].left = sk.tree.size()
```

```

lotsk, lotpk = OTS.gen()
sk.tree.add({otsk: lotsk, otpk: lotpk, prev: cur, left: None, mid: None, right: None, sign: None, msg: None})

sk.tree[cur].mid = sk.tree.size()
motsk, motpk = OTS.gen()
sk.tree.add({otsk: motsk, otpk: motpk, prev: cur, left: None, mid: None, right: None, sign: None, msg: None})

sk.tree[cur].right = sk.tree.size()
rotsk, rotpk = OTS.gen()
sk.tree.add({otsk: rotsk, otpk: rotpk, prev: cur, left: None, mid: None, right: None, sign: None, msg: None})

sk.tree[cur].sign = OTS.sign(sk.tree[cur].otsk, lotpk || motpk || rotpk)
sk.tree[cur].msg = None

cur = sk.tree[cur].mid
sk.tree[cur].sign = OTS.sign(sk.tree[cur].otsk, msg)
sk.tree[cur].msg = msg

while cur != None:
    lotsk, motpk, rotpk = sk[sk[cur].left].otpk, sk[sk[cur].mid].otpk, sk[sk[cur].right].otpk
    sign.addFromBegin(sk.tree[cur].pk, lotpk, motpk, rotpk, sk.tree[cur].sign, sk.tree[cur].msg)
    cur = sk.tree[cur].prev

sk.head = sk.head mod 3 == 0 ? sk.head + 1 : sk.head + 2
return sign

```

Зважаючи на аналогічність структури підпису зі схемою TBS, код верифікації ідейно ідентичний.

4. Заміри реалізованих цифрових підписів

Було виміряно час генерації ключів підпису, час підписання повідомлення та час верифікації у схемі Лампорта (табл. 2):

Таблиця 2

Час виконання методів схеми Лампорта			
Кількість ітерацій	Середній час GEN, с	Середній час SIGN, с	Середній час VRFY, с
1	0.003451	0.000114	0.000749
100	0.003304	0.000166	0.000745
10000	0.003319	0.000172	0.000745
164187	0.003498	0.000175	0.000742

а також розмір згенерованих пар ключів та розмір підпису (табл. 3):

Таблиця 3

Розміри згенерованих пар ключів і підпису в схемі Лампорта		
Кількість ітерацій	Середній розмір пари ключів, МБ	Середній розмір підпису, МБ
1	0.1956	0.0439
100	0.1839	0.0424
10000	0.1837	0.0424
164187	0.1837	0.0424

Розміри пар ключів і підпису повністю відповідають очікуванням, але ось час виконання функцій GEN, SIGN та VRFY перевищує їх орієнтовно в 4-5 разів. Проте в цьому випадку це зумовлено певними аспектами реалізації й обмеженнями мови.

Перейдімо тепер до багаторазових схем підписів на основі підписів Лампорта. Для замірів поведінки різних схем багаторазових підписів генерували повідомлення розміром 2560 Біт і послідовно їх підписували, а потім перевіряли відповідні підписи. Заміряли час підписування, час підписування з копіюванням, час верифікації, розмір підпису та загальне використання оперативної пам'яті. З огляду на різну асимптотичну складність функцій різних схем, кількість повідомлень, на яких проводили заміри, різна. Наприклад, CBS була перевірена на 1000 повідомленнях, TBS – на 10259, STBS – на 83550, а WSTBS – на 59052. Усі зібрані дані можна побачити на графіках, де CBS червоного кольору, TBS – помаранчевого, STBS – зеленого та WSTBS – синього (рис. 1).

Показники загального використання пам'яті переважно очікувані, оскільки з кожним підписом структура, що зберігає поточний стан підписів (список або дерево), просто лінійно зростає на однакову величину. Проте у випадку TBS заміри постійно показували константну величину. Це пояснюємо тим, що внаслідок псевдовипадковості геш-функцій більшість вершин на шляху від кореня до листочка буде зустрічатися нам уперше та генеруватись, тоді як раніше створені вершини, за невикористання, автоматично переносяться з оперативної пам'яті на жорсткий диск. Розглянемо інші заміри детальніше.

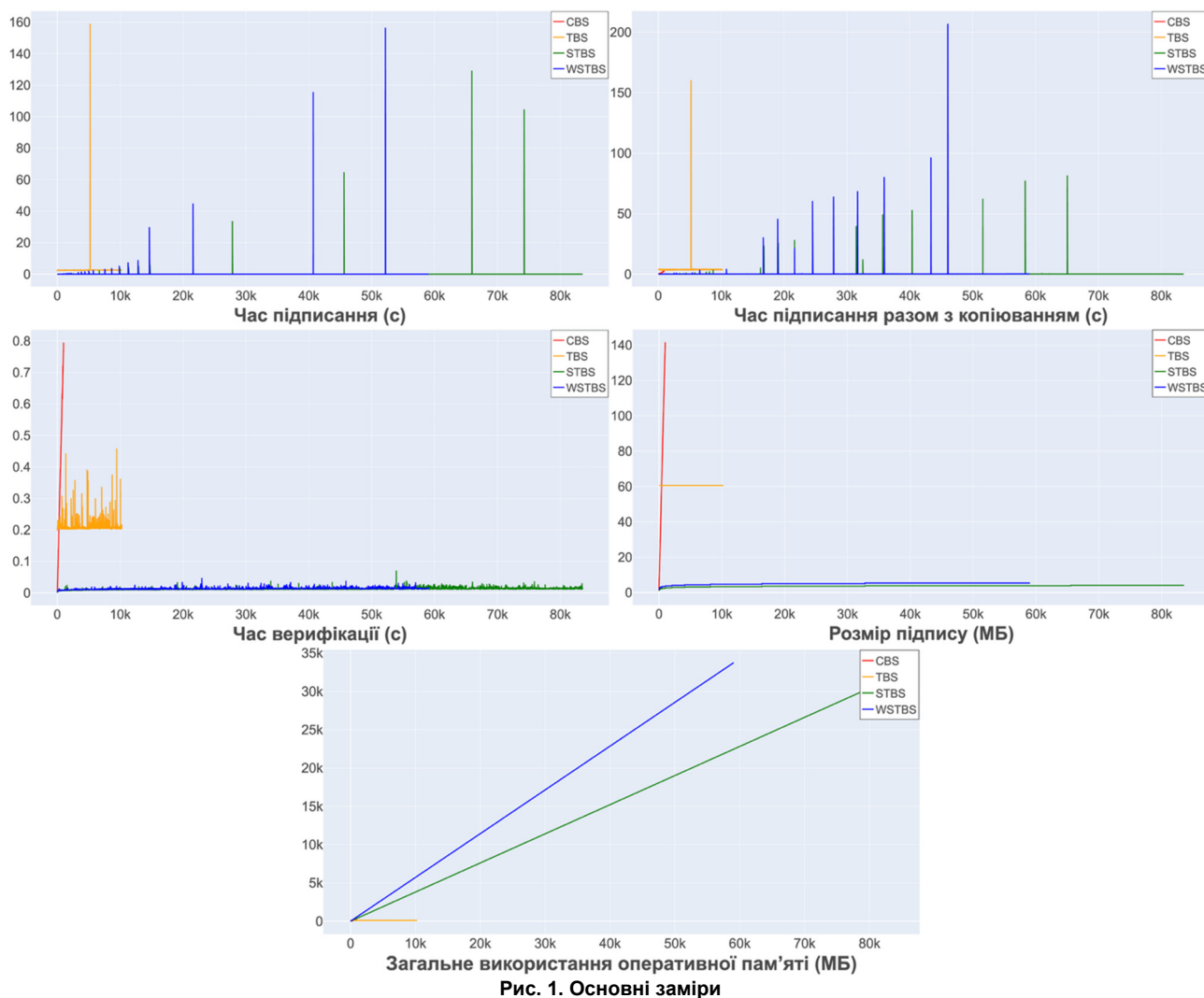


Рис. 1. Основні заміри

4.1. Час підписання

На обох графіках часу підписання видно послідовні сплески, що зростають. Проте єдина відмінність між графіками – це додаткові витрати часу на повне копіювання підпису, але на графіку з копіюванням сплесків помітно більше. Це легко пояснити внутрішніми процесами комп'ютера, на якому проводили заміри, що й викликали такий "шум". Конкретніше, більшість із них – це звільнення оперативної пам'яті під структуру, що зберігає підписи. Усунувши шум, матимемо (рис. 2):

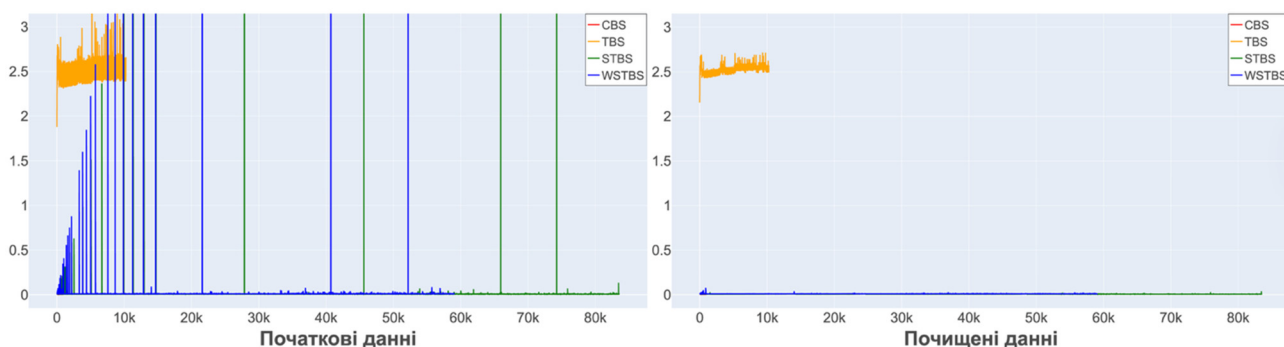


Рис. 2. Очищення даних часу підписання, с

Час, необхідний на підписання у схемі TBS, очікувано наближений до константного і практично відповідає теоретичним розрахункам, згідно з якими на підписання має йти 1.83 с. Якщо ж розглядати детальніше всі інші схеми та трохи усереднити дані, матимемо (рис. 3).

На графіках ще спостерігається "шум", але в іншому час теж наближений до константного та зберігає співвідношення відповідно до кількості генерацій ключів одноразового підпису за одного підписання.

4.2. Час верифікації

Розглянемо час верифікації (рис. 4).

У схемі CBS час верифікації росте лінійно, тоді як у схемі TBS час зберігається наближений до константного, що повністю відповідає асимптотичним оцінкам. Якщо ж виокремити схеми STBS та WSTBS, матимемо (рис. 5).

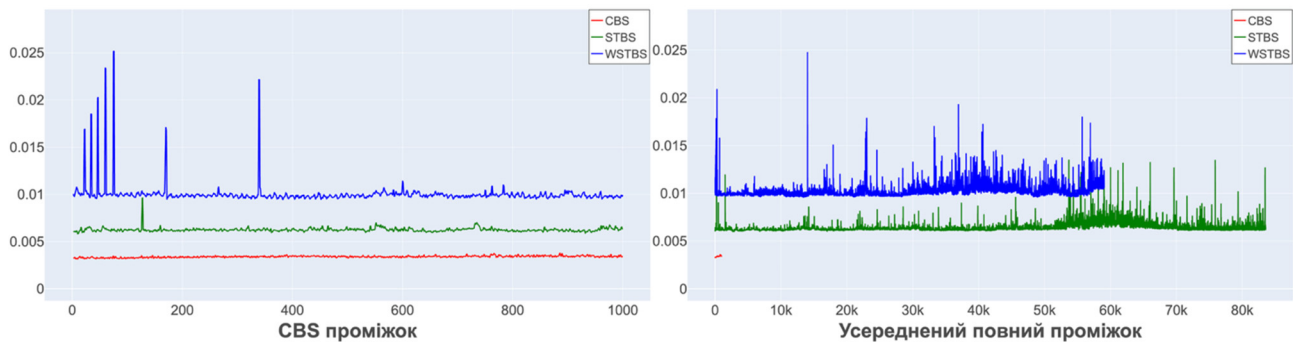


Рис. 3. Усереднені дані часу підписання, с

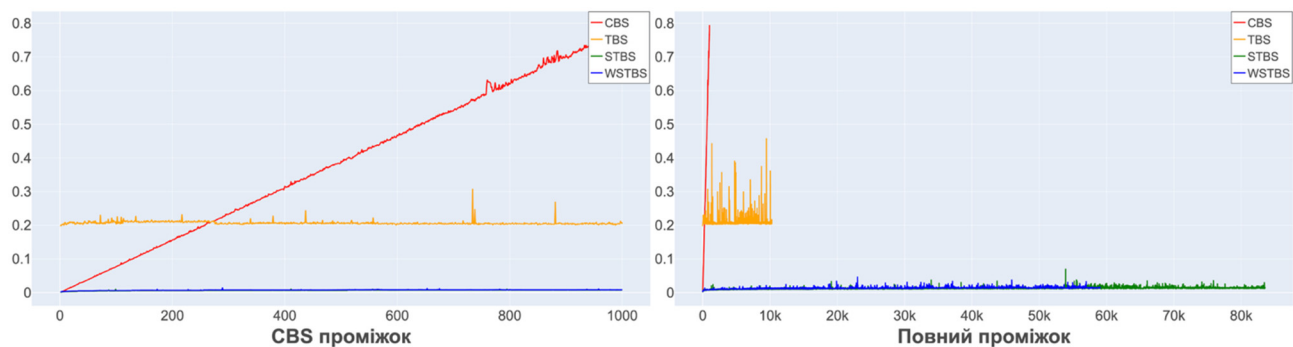


Рис. 4. Час верифікації, с

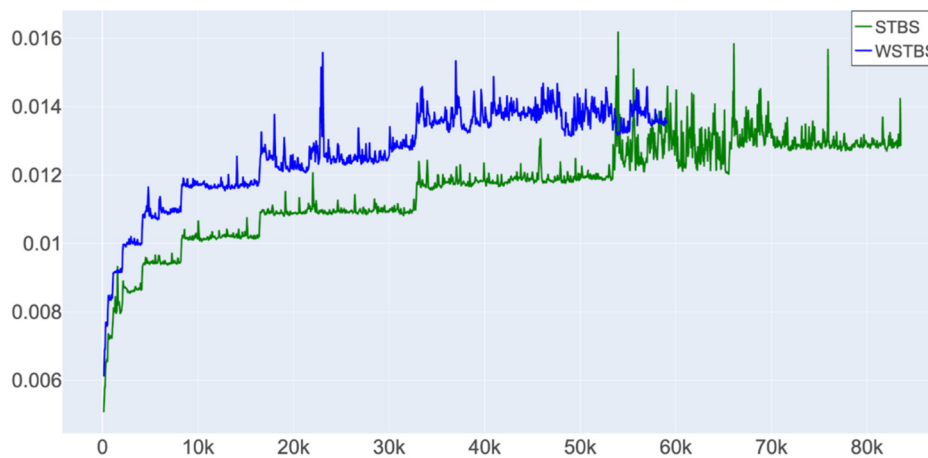


Рис. 5. Час верифікації STBS і WSTBS, с

Спостерігається логарифмічна асимптотика обох графіків з яскраво вираженими "сходінками" однакової висоти, що відповідають зростанню довжини підпису на одну ланку.

4.3. Розмір підпису

Розглянемо графіки розміру підпису (рис. 6).

Знову бачимо повну відповідність асимптотичним оцінкам. На останньому графіку також спостерігаємо невеликий "шум" та "сходінки" однакової висоти, що відповідають зростанню довжини підпису на одну ланку.

4.4. Загальний аналіз

З огляду на графіки можна побачити явну перевагу підписів на основі дерев над підписами на основі ланцюгів. Практично одразу спостерігаємо стрімке зростання як часу верифікації, так і розміру підпису, чого не відбувається в інших структурах. Водночас час підписання не сильно відрізняється від інших схем, окрім TBS. Як результат маємо абсолютну неконкурентоспроможність.

Якщо ж розглядати суто деревоподібні схеми, то STBS та WSTBS явно переважають над схемою TBS на декілька порядків за всіма заміряними показниками, окрім пам'яті, необхідної для підтримання стану структури даних. Якщо ж останнє, навпаки, є важливим, то схема TBS за правильної реалізації є фаворитом.

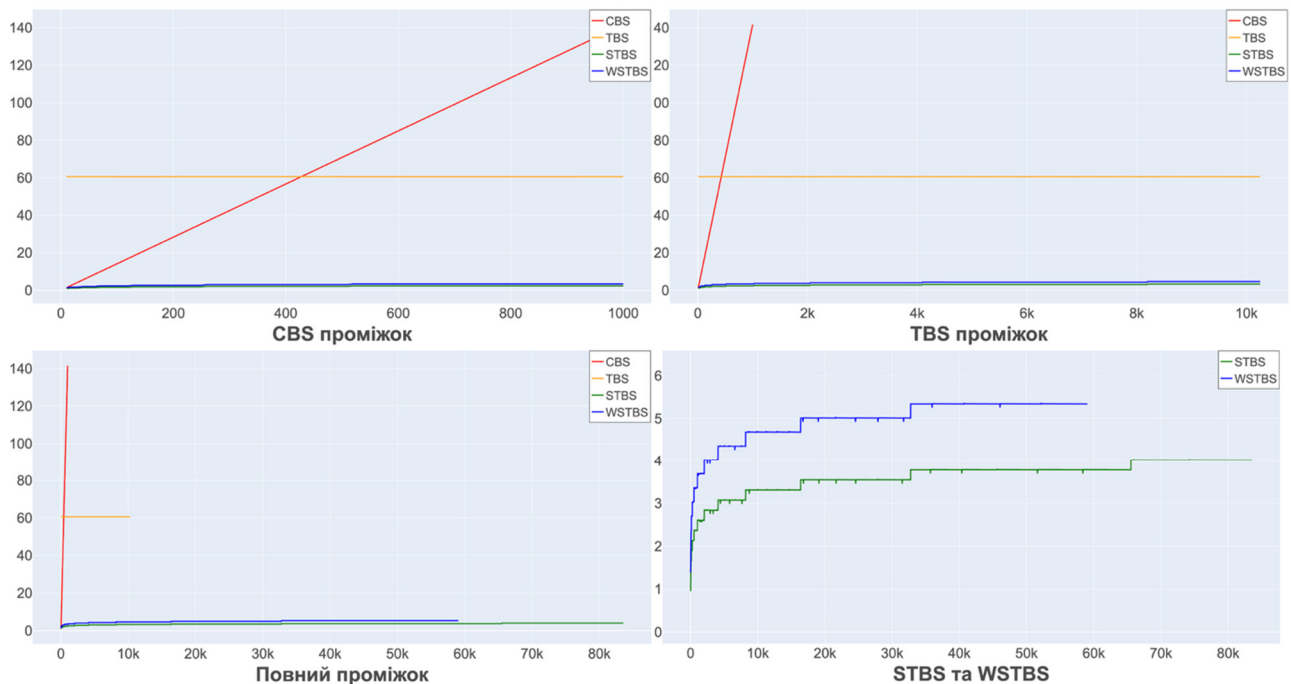


Рис. 6. Розмір підпису, МБ

Дискусія і висновки

У пропонованій роботі досліджені постквантові методи цифрових підписів, зокрема увагу було приділено схемі одноразового підпису Лампорта. Крім цього, були розглянуті та зіставлені й різні методи побудови схем багаторазового підпису на основі одноразових підписів.

Отримані результати показали, що найкращими з розглянутих методів побудови багаторазових підписів є схеми, що ґрунтуються на побудові бінарних дерев підпису. Проте вони мають як свої переваги, так і недоліки. Наприклад, якщо швидкодія за часом і малий розмір підпису не є основними цілями, то вигідніше скористатися схемою TBS, тоді як в іншому випадку доцільніше використовувати схеми STBS або WSTBS.

Варто зауважити, що в розглянутих схемах використовувалися лише бінарні дерева пошуку. З їхньою допомогою асимптотичну складність вдалося зменшити з лінійної до логарифмічної. Напрошується логічне припущення, що інші подібні структури збереження даних, зокрема й тернарні та N-арні дерева підписів, можуть дати хоча й асимптотично аналогічні оцінки, але порівняно кращі остаточні результати. З огляду на це наведені раніше структури мають бути розглянуті та досліджені в подальшому.

Джерела фінансування. Фінансування частково забезпечено Київським національним університетом імені Тараса Шевченка.

Список використаних джерел

- Bleichenbacher, D., & Maurer, U. M. (1994). Directed Acyclic Graphs, One-way Functions and Digital Signatures. In Y. G. Desmedt (Eds.), *Lecture Notes in Computer Science: Vol. 839. Advances in Cryptology – CRYPTO '94* (pp. 75–82). Springer. https://doi.org/10.1007/3-540-48658-5_9
- Boneh D., & Shoup, V. (2017). *A Graduate Course in Applied Cryptography*. https://crypto.stanford.edu/~dabo/cryptobook/BonehShoup_0_6.pdf
- Hülsing, A. (2013). W-OTS+ – Shorter Signatures for Hash-Based Signature Schemes. In A. Youssef, A. Nitaj, & A. E. Hassanien, (Eds.), *Lecture Notes in Computer Science: Vol. 7918. Progress in Cryptology – AFRICACRYPT 2013* (pp. 173–188). Springer. https://doi.org/10.1007/978-3-642-38553-7_10
- Katz, J., & Lindell, Y. (2014). *Introduction to Modern Cryptography, Second Edition*. Chapman and Hall/CRC. <https://doi.org/10.1201/b17668>
- Lizama-Pérez, L. A. (2022). Digital signatures over HMAC entangled chains. *Engineering Science and Technology, an International Journal*, 32(101076). <https://doi.org/10.1016/j.jestch.2021.11.002>
- Merkle, R. C. (1989). A Certified Digital Signature. In G. Brassard (Eds.), *Lecture Notes in Computer Science: Vol. 435. Advances in Cryptology – CRYPTO' 89 Proceedings* (pp. 218–238). Springer. https://doi.org/10.1007/0-387-34805-0_21
- NIST (2015). *FIPS PUB 180-4: Secure Hash Standard (SHS)*. <https://doi.org/10.6028/NIST.FIPS.180-4>
- Perrig, A. (2001). The BiBa one-time signature and broadcast authentication protocol. In P. Samarati (Eds.), *CCS '01: Proceedings of the 8th ACM conference on Computer and Communications Security* (pp. 28–37). Association for Computing Machinery. <https://doi.org/10.1145/501983.501988>
- Reyzin, L., & Reyzin, N. (2002). Better than BiBa: Short One-Time Signatures with Fast Signing and Verifying. In L. Batten, & J. Seberry (Eds.), *Lecture Notes in Computer Science: Vol. 2384. Information Security and Privacy* (pp. 144–153). Springer. https://doi.org/10.1007/3-540-45450-0_11

References

- Bleichenbacher, D., & Maurer, U. M. (1994). Directed Acyclic Graphs, One-way Functions and Digital Signatures. In Y. G. Desmedt (Eds.), *Lecture Notes in Computer Science: Vol. 839. Advances in Cryptology – CRYPTO '94* (pp. 75–82). Springer. https://doi.org/10.1007/3-540-48658-5_9

- Boneh D., & Shoup, V. (2017). *A Graduate Course in Applied Cryptography*. https://crypto.stanford.edu/~dabo/cryptobook/BonehShoup_0_6.pdf
- Hülsing, A. (2013). W-OTS+ – Shorter Signatures for Hash-Based Signature Schemes. In A. Youssef, A. Nitaj, & A. E. Hassanien, (Eds.), *Lecture Notes in Computer Science: Vol. 7918. Progress in Cryptology – AFRICACRYPT 2013* (pp. 173–188). Springer. https://doi.org/10.1007/978-3-642-38553-7_10
- Katz, J., & Lindell, Y. (2014). *Introduction to Modern Cryptography* (Second Edition). Chapman and Hall/CRC. <https://doi.org/10.1201/b17668>
- Lizama-Pérez, L. A. (2022). Digital signatures over HMAC entangled chains. *Engineering Science and Technology, an International Journal*, 32(101076). <https://doi.org/10.1016/j.jestch.2021.11.002>
- Merkle, R. C. (1989). A Certified Digital Signature. In G. Brassard (Eds.), *Lecture Notes in Computer Science: Vol. 435. Advances in Cryptology – CRYPTO' 89 Proceedings* (pp. 218–238). Springer. https://doi.org/10.1007/0-387-34805-0_21
- NIST (2015). *FIPS PUB 180-4: Secure Hash Standard (SHS)*. <https://doi.org/10.6028/NIST.FIPS.180-4>
- Perrig, A. (2001). The BiBa one-time signature and broadcast authentication protocol. In P. Samarati (Eds.), *CCS '01: Proceedings of the 8th ACM conference on Computer and Communications Security* (pp. 28–37). Association for Computing Machinery. <https://doi.org/10.1145/501983.501988>
- Reyzin, L., & Reyzin, N. (2002). Better than BiBa: Short One-Time Signatures with Fast Signing and Verifying. In L. Batten, & J. Seberry (Eds.), *Lecture Notes in Computer Science: Vol. 2384. Information Security and Privacy* (pp. 144–153). Springer. https://doi.org/10.1007/3-540-45450-0_11

Отримано редакцією журналу / Received: 25.06.24

Прорецензовано / Revised: 26.11.25

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Oleksii BASHUK, PhD Student

ORCID ID: 0009-0002-6778-2943

e-mail: a.bashuk@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

COMPARATIVE ANALYSIS OF POST-QUANTUM DIGITAL SIGNATURE PROTOCOLS BASED ON HASH FUNCTIONS

In the era of active information technologies and device development, the emergence of portable and general-purpose quantum computers is only a matter of time. And although the benefit of their many times greater efficiency, compared to ordinary bit computers, will be obvious, at the same time, they will bring many vulnerabilities to modern algorithms of cryptography and cybersecurity. To prevent this from happening, society has already started researching post-quantum protocols to replace modern ones.

Digital signature protocols were no exception. Thus, one of the approaches was constructing post-quantum digital signature protocols based on hash functions. The first representatives were schemes of one-time signatures based on hash functions, which were protected from the capabilities of quantum computers but had the obvious disadvantage of one-time use. The most famous is the so-called Lamport one-time signature. However, this disadvantage of one-time use has been solved by proposing different schemes of multi-time digital signatures based on the creation and use of many one-time signatures. Thus, chain-based and tree-based schemes of multi-time signatures were proposed. This work is devoted to the implementation and comparison of various multi-time signature schemes.

During the research, theoretical estimates of memory usage and time required for the operations of various schemes were given. Further implementation of these schemes and final measurements showed the correspondence of theoretical estimates with empirical results and clearly demonstrated the advantages and disadvantages of different schemes. Thus, tree-based schemes showed the best performance, especially those where the tree is built sequentially. However, in certain cases, it makes sense to use a tree with full-depth construction.

Due to the relative novelty and small amount of research in this field, multi-time digital signatures have much room for further research, for example, considering N-ary trees and comparing them with binary ones.

Keywords: hash function, digital signatures, post-quantum protocols, one-time digital signatures, multi-time digital signatures, chain-based schemes, tree-based schemes.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.25>

Олександр ГАЛУШКО, асп.
ORCID ID: 0009-0009-0275-6542
e-mail: health24@ukr.net

Київський національний університет імені Тараса Шевченка, Київ, Україна

Інтегрована модель факторного відбору та категоріальної регресії для прогнозування рівня ризику імплантації діагностично-терапевтичних пристроїв при патології шлунково-стравохідного з'єднання

Представлено інтегрований метод прогнозування рівня ризику імплантації діагностично-терапевтичних пристроїв у ділянці шлунково-стравохідного з'єднання на основі факторного аналізу та категоріальної регресії. На вибірці із 558 пацієнтів було виділено 15 клінічних та ендоскопічних показників, що характеризують стан слизової, анатомічні особливості та симптоматику рефлюксу. Факторний аналіз методом головних осей з обертанням Oblimin дає можливість звести початковий набір змінних до трьох латентних факторів: F_1 – "Клінічні та морфологічні ознаки гастроєзофагеальної рефлюксної хвороби", F_2 – "Анатомо-функціональні показники стравоходу" та F_3 – "Ерозивні ураження слизової". Ці фактори були використані як предиктори в категоріальній регресії для прогнозування п'яти рівнів ризику імплантації. Модель продемонструвала високу якість прогнозу ($R^2 = 0,858$), усі фактори були статистично значущими ($p < 0,001$), а показники мультиколінеарності предикторів $> 0,8$ підтвердили відсутність критичної мультиколінеарності. Отримана модель забезпечує стандартизовану стратифікацію ризику, допомагаючи клініцистам об'єктивніше оцінювати доцільність імплантації та визначати пацієнтів, які потребують додаткової оптимізації стану перед процедурою.

Ключові слова: імплантація пристроїв, факторний аналіз, категоріальна регресія, ризик, шлунково-стравохідне з'єднання, стратифікація ризику.

Класифікація відповідно до AMS 2020: 62J12, 62P10.

Вступ

Імплантацію діагностичних і терапевтичних пристроїв у ділянці шлунково-стравохідного з'єднання (зокрема систем бездротового рН-моніторингу, стимуляторів нижнього стравохідного сфінктера, капсульних сенсорів) дедалі частіше застосовують у клінічній практиці для діагностики та лікування гастроєзофагеальної рефлюксної хвороби (ГЕРХ) та пов'язаних із нею патологій. Водночас ці втручання пов'язані з ризиком розвитку ускладнень, що зумовлює необхідність попередньої об'єктивної оцінки локального стану слизової оболонки, анатомо-функціональних особливостей стравоходу та клінічної симптоматики пацієнта.

На практиці рішення щодо доцільності та безпечності імплантації часто ґрунтується на експертній оцінці лікаря з урахуванням великої кількості клінічних та ендоскопічних показників різного типу (бінарних, порядкових і кількісних). Це ускладнює стандартизацію підходу, підвищує суб'єктивність рішень й обмежує можливості формалізованої стратифікації ризику. Крім того, велика кількість взаємопов'язаних змінних і наявність рідкісних категорій створюють проблеми для класичних багатовимірних моделей через мультиколінеарність і нестійкість оцінок.

У зв'язку із цим актуальним є розроблення математично обґрунтованих моделей, що дають змогу інтегрувати різномірні клінічні показники, зменшувати розмірність даних без втрати інформації та забезпечувати об'єктивне прогнозування рівня ризику імплантації.

Латентні фактори у медичних дослідженнях застосовують для виявлення прихованих (небезпосередньо невимірюваних) структур у складних переліках клінічних, лабораторних або анкетних даних. Вони допомагають зменшити розмірність даних й об'єднати кілька взаємопов'язаних показників у спільні концептуальні компоненти (напр., "метаболічний синдром", "психоемоційний стан", "анатомо-функціональні порушення"). Подальше використання цих факторів у регресійному аналізі дає змогу будувати прогностичні моделі, які є більш стабільними, інтерпретованими та менш чутливими до мультиколінеарності між первинними змінними. Такий підхід широко застосовують для: оцінки ризику розвитку захворювань (Smith et al., 2017); побудови моделей латентних класів або підтипів пацієнтів (Guo, Wall, & Amemiya, 2006); аналізу психометричних і клінічних шкал (Viana, Gonçalves, & Paez, 2024); розробки методів факторної регресії для інтеграції змінних і латентних показників у єдину модель (Devlieger, 2019) та ін.

Завдяки цьому факторно-регресійний підхід вважають ефективним інструментом для побудови персоналізованих прогностичних систем у клінічній медицині.

Об'єктом пропонованого дослідження є клінічні та ендоскопічні характеристики пацієнтів із патологією шлунково-стравохідного з'єднання, що використовуються для оцінки ризику імплантації діагностичних і терапевтичних пристроїв.

Мета роботи – розроблення й обґрунтування інтегрованої математичної моделі прогнозування рівня ризику імплантації пристроїв у ділянці шлунково-стравохідного з'єднання на основі поєднання факторного аналізу та категоріальної регресії.

Для досягнення цієї мети поставлено такі завдання:

- відібрати релевантні клінічні й ендоскопічні показники, що характеризують стан пацієнтів;
- здійснити факторний аналіз для зменшення розмірності та виділення латентних клінічно інтерпретованих факторів;
- побудувати модель категоріальної регресії з оптимальним шкалюванням для прогнозування порядкової змінної "рівень ризику";
- оцінити якість, стабільність та інтерпретованість отриманої моделі.

У дослідженні спочатку виконано факторний аналіз за метод головних осей із обертанням Oblimin для виділення клінічно інтерпретованих латентних факторів із 15 ендоскопічних і клінічних показників, а потім отримані фактори постають як предиктори в категоріальній регресії для прогнозування п'яти рівнів ризику імплантації.

1. Формування вибірки, змінних і шкали ризику

У представленій статті побудовано математичну модель, що інтегрує клінічні й ендоскопічні дані для прогнозування рівня ризику до імплантації діагностичних та/або терапевтичних пристроїв, зокрема систем Bravo, JSPH-1, BEST Capsule, EndoStim для стимуляції нижнього стравохідного сфінктера, капсульних датчиків Smart nanowire та інших (Lawenko, & Lee, 2016; Li, Liu, & Tao, 2013), у пацієнтів із патологією шлунково-стравохідного з'єднання, з метою підвищення точності й об'єктивності клінічного ухвалення рішень.

Показники для моделі: гіперемія слизової в субкардіальній ділянці (бінарна змінна: 1 – наявна; 0 – відсутня) – X_1 ; скарги на печію (бінарна змінна: 1 – наявна; 0 – відсутня) – X_2 ; скарги на відчуття регургітації (бінарна змінна: 1 – наявна; 0 – відсутня) – X_3 ; наявність дефектів слизової класифікованих за Лос-Анджелеською класифікацією оцінки рефлюкс-езофагіту (порядкова змінна: дефекти відсутні – 0; тип А – 1; тип В – 2; тип С – 3; тип D – 4) – X_4 ; наявність ерозій, покритих фібрином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_5 ; наявність ерозій, покритих гематином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_6 ; наявність виразок, покритих фібрином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_7 ; наявність виразок, покритих гематином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_8 ; показник розташування Z-лінії (кількісна змінна, см) – X_9 ; висота верхнього стравохідного сфінктера (кількісна змінна, см) – X_{10} ; показник діафрагмального звуження (кількісна змінна, см) – X_{11} ; показник висоти пролабування шлунка у стравохід або кардіальну ділянку шлунка (кількісна змінна, см) – X_{12} ; довжина пролабування (кількісна змінна, см) – X_{13} ; гіперемія в нижній третині стравоходу (бінарна змінна: 1 – наявна; 0 – відсутня) – X_{14} ; куріння (бінарна змінна: 1 – наявне; 0 – відсутнє) – X_{15} .

На основі вибірки із 558 пацієнтів ендоскопічного відділення комунального некомерційного підприємства "Кіровоградська обласна лікарня Кіровоградської обласної ради" та їхніх клінічних характеристик (X_1 – X_{15}) лікарі змогли встановити рівні ризику для лікувальної імплантації. Було виділено 5 рівнів ризику. Проаналізуємо основні характеристики для градації на групи за рівнями ризику.

0 (відсутній ризик). Ймовірність ускладнень мінімальна. Наслідки незначні та легко контролювані. Процедура імплантації вважається безпечною. Базові критерії: може бути гіперемія в нижній третині стравоходу або гіперемія в субкардіальній ділянці шлунка; довжина пролабування менше 2 см; відсутність дефектів слизової.

1 (низький ризик). Ймовірність ускладнень мінімальна. Наслідки незначні, добре контролювані. Процедура імплантації безпечна, потрібне лише стандартне спостереження. Базові критерії: може бути гіперемія в нижній третині стравоходу і/або гіперемія в субкардіальній ділянці шлунка; довжина пролабування не перевищує 2 см; можуть бути дефекти слизової у вигляді ерозій, покритих фібрином.

2 (середній ризик). Ймовірність ускладнень помірна. Наслідки контролювані, однак рекомендоване дотримання дієти, режиму харчування та посилене клінічне спостереження. Базові критерії: може бути гіперемія в нижній третині стравоходу і/або гіперемія в субкардіальній ділянці шлунка; довжина пролабування не перевищує 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих фібрином.

3 (високий ризик). Ймовірність ускладнень підвищена. Наслідки клінічно значущі, потребують посиленого медичного нагляду. Процедура імплантації допускається лише після ретельної оцінки. Базові критерії: довжина пролабування не перевищує 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих гематином. Комбінації "гематин" піднімають ризик до високого незалежно від гіперемії чи набряку.

4 (критичний ризик). Ймовірність ускладнень висока. Наслідки можуть бути небезпечними для життя. Процедура імплантації не рекомендована, поки ризики не будуть усунені або зменшені. Базові критерії: довжина пролабування більше 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих фібрином і гематином.

Зауважимо, що тут наведені ключові маркери щодо градації за групами ризику. Лікарі враховували весь спектр зазначених показників X_1 – X_{15} .

2. Виділення латентних факторів методом головних осей

Оскільки кількість змінних дуже велика, а також деяка частина змінних містить поодинокі значення (є проблема розрізнення даних), то для знаходження моделі, що буде відображати рівні ризику, проведемо факторний аналіз змінних X_1 – X_{15} і визначимо латентні фактори, які пояснюють основні закономірності в даних і зменшують кількість змінних, залишаючи максимально можливу інформацію.

Отже, для спостережуваних p змінних X_1 – X_{15} припускаємо, що існує m ($m < p$) латентних факторів $F_1, F_2, \dots, F_m$, такі, що кожна змінна може бути представлена як:

$$X_j = \lambda_{j1}F_1 + \lambda_{j2}F_2 + \dots + \lambda_{jm}F_m + \varepsilon_j, \quad j = 1, \dots, 15, \quad (1)$$

де λ_{jk} – факторні навантаження, що показують, наскільки X_j пов'язана з фактором F_j ; ε_j – унікальна частина змінної X_j , яку фактори не пояснюють.

Будемо використовувати метод головних осей (Principal Axis Factoring, PAF), що не вимагає припущення про нормальний розподіл даних (Finch, & French, 2019; Yong, & Pearce, 2013). Це робить його привабливим вибором, коли дані не відповідають умовам нормальності. Порівняно з методом максимальної правдоподібності, що потребує нормальності даних, PAF є більш гнучким і стійким до порушень цієї умови.

Першим кроком зазначеного методу є стандартизація змінних через середні значення та стандартні відхилення змінних

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, \quad i = 1, \dots, 558, \quad j = 1, \dots, 15,$$

де Z_{ij} – стандартизоване значення змінної j для i -го пацієнта; $\bar{X}_j$ – середнє значення змінної X_j ; s_j – стандартне відхилення змінної X_j .

Це забезпечує, що всі змінні мають $E(Z_j) = 0$, $Var(Z_j) = 1$.

Далі будемо матрицю кореляцій $R = [r_{jk}]_{p \times p}$, елементи якої визначаємо так:

$$r_{jk} = \frac{\text{Cov}(X_j, X_k)}{s_j s_k}, \quad j, k = 1, \dots, 15.$$

Для бінарних змінних коефіцієнт кореляції Пірсона чисельно збігається з коефіцієнтом контингенції ϕ , який використовуємо для оцінювання зв'язку між бінарними змінними (Akoglu, 2018), тому застосування стандартної матриці кореляцій у SPSS (Statistical Package for the Social Sciences – статистичне програмне забезпечення для соціальних наук) коректне навіть за змішаних типів змінних.

У матричній формі основну факторну модель для всієї вибірки записуємо як:

$$Z = F\Lambda' + E,$$

де Z – матриця стандартизованих змінних розмірності $n \times p$; F – матриця факторних балів (латентних факторів) розмірності $n \times m$; Λ' – транспонована матриця факторних навантажень розмірності $m \times p$; E – матриця унікальних залишків розмірності $n \times p$.

Для кожного окремого спостереження (пацієнта) це рівняння має вигляд:

$$z_i = \Lambda f_i + \varepsilon_i, \quad i = 1, \dots, n,$$

де z_i – вектор стандартизованих змінних для i -го пацієнта; f_i – вектор значень латентних факторів; ε_i – вектор залишків похибок.

PAF розраховує такі факторні навантаження $\Lambda = [\lambda_{jk}]$, щоб коваріаційна структура змінних задовольняла рівність:

$$R = \Lambda\Lambda' + \Psi,$$

де $\Lambda\Lambda'$ – дисперсія, пояснена латентними факторами; $\Psi = \text{diag}(\psi_1, \dots, \psi_p)$ – діагональна матриця унікальних дисперсій, тобто залишків, що не пояснюються факторами.

Для кожної змінної X_j обчислюємо загальну дисперсію змінної, пояснену факторами h_j^2 , тобто частку дисперсії змінної, поясненої виділеними факторами:

$$h_j^2 = \sum_{k=1}^m \lambda_{jk}^2.$$

У пакеті SPSS розраховуємо Initial (початкові) і Extraction (виділені) h_j^2 . Високі значення h_j^2 ($h_j^2 \rightarrow 1$) означають, що фактори добре пояснюють змінну X_j .

Метод PAF відшукує такі факторні навантаження Λ , щоб звести до мінімуму різницю між емпіричною матрицею кореляцій та її факторною апроксимацією ($\Lambda\Lambda' + \Psi$). На практиці це досягається ітераційно. Початкові значення ψ_j задають як $1 - h_j^2$. З матриці $R - \Psi$ обчислюють власні значення та власні вектори. Формуються нові навантаження Λ із власних векторів, що відповідають найбільшим власним значенням. Процес повторюється доти, доки зміни h_j^2 стануть незначними.

Для полегшення інтерпретації результатів скористаємося методом обертання Oblimin із нормалізацією Кайзера. Цей метод дає факторам можливість бути корельованими, що важливо для клінічних ознак, які можуть бути взаємопов'язані. Також факторні навантаження після обертання легше інтерпретувати, вони чітко показують, які змінні входять до кожного латентного фактору.

Після обертання

$$\Lambda^* = \Lambda T,$$

де T – матриця обертання, що спрощує структуру факторних навантажень таким чином, що кожна змінна має суттєве навантаження на один фактор.

Критерієм придатності даних для факторного аналізу є тест Кайзера – Мейєра – Олкіна (KMO), що вимірює пропорцію дисперсії змінних, яку можна пояснити спільними факторами. KMO обчислюємо за формулою:

$$KMO = \frac{\sum_{j \neq k} r_{jk}^2}{\sum_{j \neq k} r_{jk}^2 + \sum_{j \neq k} p_{jk}^2}, \quad p_{jk} = -\frac{r^{jk}}{\sqrt{r^{jj} \cdot r^{kk}}},$$

де r^{jk} – елемент оберненої матриці кореляцій R^{-1} ; r^{jj} , r^{kk} – діагональні елементи R^{-1} , що відповідають змінним X_j , X_k .

Значення KMO близьке до 1 вказує на високу придатність даних для факторного аналізу.

Тест сферичності Бартлетта перевіряє гіпотезу про те, що кореляційна матриця змінних є одиничною, тобто змінні не корельовані між собою. Тест є важливим перед застосуванням факторного аналізу, оскільки значна кореляція між змінними – необхідна умова для виділення латентних факторів. Статистику критерію Бартлетта розраховуємо так:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \ln(\det(R)).$$

Якщо в SPSS p -значення тесту Бартлетта менше обраного рівня значущості, то гіпотезу про одиничність кореляційної матриці відхиляють, що свідчить про наявність значущих кореляцій між ними.

Проаналізуємо детально результати факторного аналізу для нашої моделі. З табл. 1 видно, що оптимальна кількість факторів дорівнює 3. Згідно з критерієм Кайзера, у моделі залишають лише ті фактори, власні значення яких перевищують 1 (Statistics Solutions, n. d.).

З табл. 1 видно, що перші три фактори мають власні значення: $\lambda_1 = 6,143$; $\lambda_2 = 2,862$; $\lambda_3 = 1,516$, що разом пояснюють приблизно 70,1 % загальної дисперсії вихідних змінних до обертання. Після обертання Oblimin із нормалізацією Кайзера сумарна пояснена дисперсія становить 64,6 %, що є прийнятним результатом для медико-біологічних досліджень, де змінні мають складну багатовимірну природу.

Вибір саме трьох факторів додатково підтверджується графіком кам'янистого осипу (scree plot) (рис. 1), на якому чітко спостерігаємо "перелом" між третім і четвертим факторами. Після третього фактору власні значення різко зменшуються та наближаються до нуля, що свідчить про незначний внесок наступних факторів у пояснення дисперсії. Отже, рішення про виокремлення трьох латентних факторів є теоретично та статистично обґрунтованим.

Таблиця 1

Пояснена дисперсія факторів за результатами факторного аналізу методом головних осей

Фактор	Початкові власні числа			Після виділення: сума квадратів навантажень			Після обертання: сума квадратів навантажень*
	Загальне	% дисперсії	Кумулятивно, %	Загальне	% дисперсії	Кумулятивно, %	Загальне
1	6,143	40,952	40,952	5,864	39,093	39,093	5,390
2	2,862	19,079	60,030	2,629	17,527	56,619	2,833
3	1,516	10,106	70,136	1,196	7,972	64,592	2,663
4	0,900	6,001	76,138				
5	0,760	5,064	81,202				
6	0,711	4,741	85,944				
7	0,641	4,274	90,218				
8	0,427	2,849	93,067				
9	0,338	2,253	95,320				
10	0,240	1,602	96,921				
11	0,183	1,221	98,142				
12	0,129	0,860	99,003				
13	0,077	0,516	99,519				
14	0,067	0,449	99,969				
15	0,005	0,031	100,000				

Примітки: 1. **Метод виділення** – метод головних осей (PAF). 2. Позначка (*) означає, що коли фактори корельовані, то суми квадратів навантажень не можна додавати для отримання загальної дисперсії.

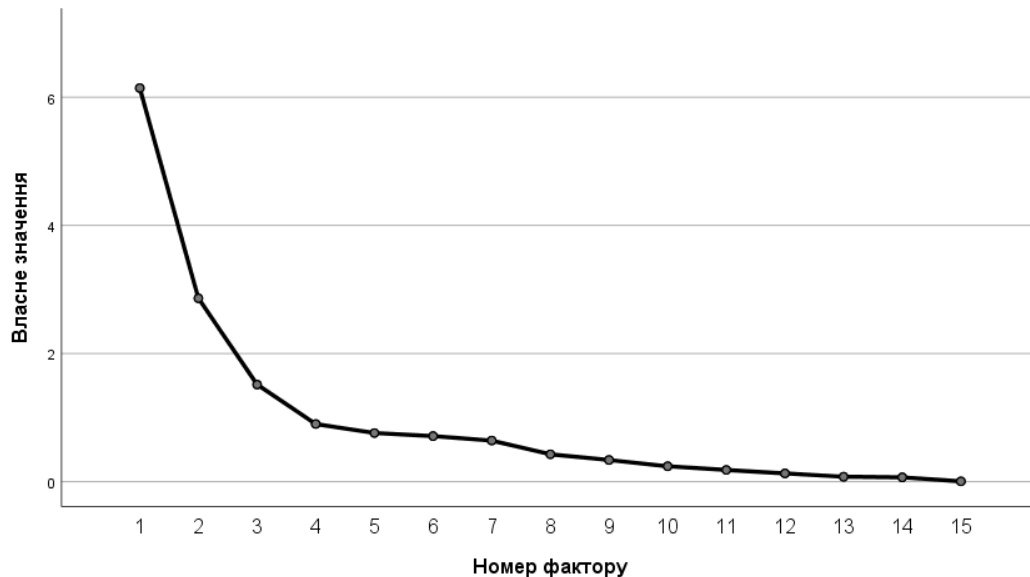


Рис. 1. Графік кам'янистого осипу (scree plot)

На основі кореляційної матриці та таблиці h_j^2 (табл. 2) можемо здійснити клінічну інтерпретацію важливості початкових змінних X_1-X_{15} на визначення факторних змінних F_1-F_3 .

Таблиця 2

Загальні дисперсії змінних, пояснені факторами, та клінічна інтерпретація змінних у факторному аналізі

Предиктори	Початкові h_j^2	Виділені h_j^2	Клінічна інтерпретація	Рівень впливу
X_1	0,732	0,454	Локальне запалення, пов'язане з анатомічними змінами (висота сфінктера, довжина пролабування)	Середній
X_2	0,712	0,678	Класичний симптом ГЕРХ, корелює з дефектами слизової	Високий
X_3	0,787	0,649	Ключовий симптом ГЕРХ, пов'язаний із тяжкими ураженнями	Високий
X_4	0,910	0,922	Ключовий ендоскопічний маркер, корелює із симптомами	Високий
X_5	0,763	0,915	Активне ураження, пов'язане із симптомами, анатомією та курінням	Високий
X_6	0,175	0,083	Рідкісне ураження, клінічно малозначуще	Низький
X_7	0,715	0,639	Тяжкі ураження, пов'язані із симптомами та структурними змінами	Високий
X_8	0,255	0,199	Малоспецифічне, рідкісне ураження	Низький
X_9	0,795	0,711	Анатомічний маркер ГЕРХ, відображає тяжкість рефлюксу	Високий
X_{10}	0,878	0,897	Механічний бар'єр проти рефлюксу, пов'язаний із довжиною пролабування	Високий
X_{11}	0,987	0,858	Основний анатомічний маркер тяжкого рефлюксу	Високий
X_{12}	0,991	0,977	Маркер тяжкого рефлюксу, пов'язаний із діафрагмальним звуженням	Високий
X_{13}	0,966	0,623	Механічне значення, вторинний маркер	Високий
X_{14}	0,533	0,519	Показник запалення, асоційований із симптомами і структурою	Середній
X_{15}	0,608	0,564	Фактор ризику, посилює симптоми й ураження слизової	Середній

Згідно з нашим результатом, значення $KMO = 0,740$ (табл. 3). Це свідчить про значне перекриття інформації між змінними та наявність середньої часткової кореляції. Отже, проведення факторного аналізу є цілком обґрунтованим. Іншими словами, вибірка і структура даних дають змогу виокремити латентні фактори. За тестом сферичності Бартлетта $p = 0,000 < 0,05$, тож гіпотезу про одиничну кореляційну матрицю відкидаємо. Це означає, що змінні взаємопов'язані, і факторний аналіз доцільний.

Таблиця 3

Критерії придатності даних для факторного аналізу (KMO та тест сферичності Бартлетта)		
Статистика критерію KMO		0,740
Тест сферичності Бартлетта	Хі-квадрат статистика	8597,565
	Ступені свободи	105
	Рівень значущості (p)	0,000

На основі таблиці факторних навантажень для 15 клінічних змінних методом головних осей з обертанням Oblimin виділимо латентні фактори та надамо їх змістовну інтерпретацію (в дужках зазначатимемо факторні навантаження клінічних показників).

Фактор 1 – F_1 "Клінічні та морфологічні ознаки ГЕРХ" охоплює такі показники: показник діафрагмального звуження (0,830); наявність виразок, покритих фібрином (0,780); показник розташування Z-лінії (0,748); скарги на відчуття регургітації (0,730); наявність дефектів слизової за Лос-Анджелеською класифікацією (0,729); куріння (0,598); скарги на печію (0,592); гіперемія в нижній третині стравоходу (0,572); наявність виразок, покритих гематином (0,438).

Фактор 2 – F_2 "Анатомо-функціональні показники стравоходу" охоплює такі показники: висота верхнього стравохідного сфінктера (– 0,870); показник довжини пролабування (– 0,783); показник висоти пролабування шлунка у стравохід або кардіальну ділянку (0,698); наявність гіперемії слизової в субкардіальній ділянці (0,635).

Фактор 3 – F_3 "Ерозивні ураження слизової" охоплює такі показники: наявність ерозій, покритих фібрином (0,973); наявність ерозій, покритих гематином (0,256).

Після обертання розраховуємо факторні бали для кожного спостереження. У нашому випадку це буде:

$$f_{i1}, f_{i2}, f_{i3}, \quad i = 1, \dots, 558,$$

де f_{i1} – факторний бал для фактора 1 "Клінічні та морфологічні ознаки ГЕРХ"; f_{i2} – факторний бал для фактора 2 "Анатомо-функціональні показники стравоходу"; f_{i3} – факторний бал для фактора 3 "Ерозивні ураження слизової".

Факторні бали f_{ik} для кожного спостереження i на факторі k розраховуємо як зважену комбінацію стандартизованих змінних:

$$f_{ik} = \omega_{1k}Z_{i1} + \omega_{2k}Z_{i2} + \dots + \omega_{pk}Z_{ip},$$

де ω_{jk} – коефіцієнт ваги для змінної Z_j на факторі F_k . Вагу в пакеті SPSS обчислюємо методом регресії, де факторні бали – це найкраща оцінка факторів за мінімумом середньоквадратичної похибки.

Далі обчислюємо кореляційну матрицю факторних балів:

$$r_{F_j F_k} = \frac{Cov(F_j, F_k)}{\sigma_{F_j} \sigma_{F_k}}, \quad j, k = 1, 2, 3,$$

де $Cov(F_j, F_k)$ – коваріація між факторами F_j та F_k ; σ_{F_j} , σ_{F_k} – стандартні відхилення факторних балів відповідно.

У табл. 4 наведено матрицю відповідних факторних кореляцій.

Таблиця 4

Матриця факторних кореляцій			
Фактор	1	2	3
1	1,000	– 0,149	0,288
2	– 0,149	1,000	– 0,001
3	0,288	– 0,001	1,000

Позадіагональні елементи в табл. 4 показують взаємну кореляцію між факторами: F_1 і F_2 ($r = -0,149$) – слабка негативна кореляція; майже незалежні; F_1 і F_3 ($r = 0,288$) – слабка позитивна кореляція, деяке перетинання ознак; F_2 і F_3 ($r = -0,001$) – практично відсутня кореляція, фактори незалежні.

Кінцева форма факторної моделі (1), що показує, як кожна змінна розкладається на вплив факторів плюс залишкову частину, має вигляд:

$$X_1 = -0,130 \cdot F_1 + 0,635 \cdot F_2 + 0 \cdot F_3 + \varepsilon_1;$$

$$X_3 = 0,730 \cdot F_1 - 0,217 \cdot F_2 + 0 \cdot F_3 + \varepsilon_3;$$

$$X_5 = 0 \cdot F_1 + 0 \cdot F_2 + 0,973 \cdot F_3 + \varepsilon_5;$$

$$X_7 = 0,780 \cdot F_1 - 0,153 \cdot F_2 - 0,373 \cdot F_3 + \varepsilon_7;$$

$$X_9 = 0,748 \cdot F_1 + 0,303 \cdot F_2 + 0,201 \cdot F_3 + \varepsilon_9;$$

$$X_{11} = 0,830 \cdot F_1 + 0,359 \cdot F_2 + 0,192 \cdot F_3 + \varepsilon_{11};$$

$$X_{13} = 0 \cdot F_1 - 0,783 \cdot F_2 + 0 \cdot F_3 + \varepsilon_{13};$$

$$X_{15} = 0,598 \cdot F_1 + 0 \cdot F_2 + 0,301 \cdot F_3 + \varepsilon_{15}.$$

$$X_2 = 0,592 \cdot F_1 - 0,182 \cdot F_2 + 0,369 \cdot F_3 + \varepsilon_2;$$

$$X_4 = 0,729 \cdot F_1 - 0,258 \cdot F_2 + 0,347 \cdot F_3 + \varepsilon_4;$$

$$X_6 = 0 \cdot F_1 + 0 \cdot F_2 + 0,256 \cdot F_3 + \varepsilon_6;$$

$$X_8 = 0,438 \cdot F_1 + 0 \cdot F_2 - 0,202 \cdot F_3 + \varepsilon_8;$$

$$X_{10} = 0,133 \cdot F_1 - 0,870 \cdot F_2 + 0,260 \cdot F_3 + \varepsilon_{10};$$

$$X_{12} = 0,698 \cdot F_1 + 0,729 \cdot F_2 + 0,188 \cdot F_3 + \varepsilon_{12};$$

$$X_{14} = 0,572 \cdot F_1 - 0,186 \cdot F_2 + 0,226 \cdot F_3 + \varepsilon_{14};$$

Отже, три виділені фактори переважно незалежні і відображають різні аспекти клінічних змін у пацієнтів.

3. Побудова й оцінка моделі категоріальної регресії ризику

Перейдемо від 15 клінічних показників різного типу і природи до 3 факторних змінних, що повно та значущо охоплюють дані показники. На основі отриманих трьох факторних змінних побудуємо модель категоріальної регресії, яка дасть можливість класифікувати пацієнтів за рівнями ризику для лікувальної імплантації.

Категоріальна регресія (в SPSS: Optimal Scaling Regression / CATREG) – це узагальнення лінійної регресії, що дає змогу використовувати змішані типи змінних: кількісні, порядкові, номінальні. Перед побудовою регресійної моделі виконується оптимальне шкалювання: кожній номінальній або порядковій змінній присвоюють оптимальні числові значення, які максимізують множинну кореляцію між предикторами та залежною змінною.

Зауважимо, що залежна змінна у нас порядкова (можливо її також визначати як номінальну) й у неї наявні такі категорії ризику: 0 (відсутній ризик); 1 (низький ризик); 2 (середній ризик); 3 (високий ризик); 4 (критичний ризик). Інші змінні – кількісні – це факторні навантаження для трьох факторів.

Підсумуємо: $Y \in \{0,1,2,3,4\}$ – це порядкова змінна; незалежні змінні – кількісні факторні бали (латентні фактори), які ми отримали із факторного аналізу: F_1, F_2, F_3 .

Тоді категоріальна регресійна модель у нашому випадку матиме вигляд:

$$\psi(Y_i) = \beta_0 + \beta_1 \phi_1(f_{1i}) + \beta_2 \phi_2(f_{2i}) + \beta_3 \phi_3(f_{3i}) + \varepsilon_i, \quad i = 1, \dots, n,$$

де $\psi: \{0,1,2,3,4\} \mapsto \mathbb{R}$ – монотонна функція оптимального шкалювання залежної змінної (зберігає порядок): $\psi(0) < \psi(1) < \psi(2) < \psi(3) < \psi(4)$; $\phi_j: \mathbb{R} \mapsto \mathbb{R}$ – функція оптимального шкалювання j -го предиктора ($j = 1,2,3$); $\beta_0, \beta_1, \beta_2, \beta_3$ – коефіцієнти регресії; ε_i – випадкова похибка з $E(\varepsilon_i) = 0$.

Коефіцієнти $\beta_0, \beta_1, \beta_2, \beta_3$ та трансформації $\psi, \phi_1, \phi_2, \phi_3$ визначаються шляхом мінімізації суми квадратів залишків:

$$\min_{\beta_0, \beta_1, \beta_2, \beta_3, \psi, \phi_1, \phi_2, \phi_3} \sum_{i=1}^n \left[\psi(Y_i) - \beta_0 - \sum_{j=1}^3 \beta_j \phi_j(f_{ji}) \right]^2.$$

В алгоритмі використовуємо наближення через дискретизацію (IBM, n. d.):

$$\phi_j(f_{ji}) = \sum_{k=1}^{K_j} c_{jk} \mathbf{1}(f_{ji} \in I_{jk}),$$

де I_{jk} – інтервали (градації) значень фактора F_j ; c_{jk} – оптимальні числові коефіцієнти шкали для кожної градації; $\mathbf{1}(f_{ji} \in I_{jk})$ – індикаторна функція (1, якщо f_{ji} належить інтервалу I_{jk}).

Зрештою формуються нові змінні:

$$Y_i^* = \psi(Y_i), \quad f_{ji}^* = \phi_j(f_{ji}), \quad j = 1,2,3.$$

Модель у просторі шкалованих змінних має вигляд лінійної регресії:

$$Y_i^* = \beta_0 + \beta_1 f_{1i}^* + \beta_2 f_{2i}^* + \beta_3 f_{3i}^* + \varepsilon_i. \quad (2)$$

Відповідно, β_j – приріст очікуваного шкализованого індексу ризику Y^* за одиничного приросту F_j^* в умовах фіксованих інших предикторів. Знак β_j визначає напрям впливу фактора на шкалований ризик, а модуль – відносну силу впливу.

Показники якості моделі розраховуємо аналогічно, як і у випадку лінійної регресії, їх наведено у табл. 5 та 6.

Таблиця 5

Параметри якості моделі категоріальної регресії для залежної змінної "Група ризику"

Множинний R	R ²	Скоригований R ²	Середня квадратична помилка прогнозу
0,926	0,858	0,857	0,142

Множинний коефіцієнт кореляції між фактичними і прогнозованими значеннями залежної змінної "Група ризику" $R = 0,926$ свідчить про дуже сильний зв'язок. Коефіцієнт детермінації $R^2 = 0,858$ означає, що 85,8 % варіації "Група ризику" пояснюється моделлю. Скоригований $R^2 = 0,857$ з поправкою на кількість предикторів і розмір вибірки майже такий самий, як R^2 , тобто модель не надто ускладнена.

Таблиця 6

Результати аналізу дисперсії (ANOVA) для моделі категоріальної регресії залежної змінної "Група ризику"

Джерело	Сума квадратів	Ступені свободи	Середній квадрат	F-критерій	Рівень значущості (p – value)
Регресія	476,376	5	95,275	665,264	0,000
Залишки	78,624	549	0,143		
Загальна	555,000	554			

З табл. 6 видно, що модель значущо пояснює змінну Y – "Група ризику", оскільки p – value = 0,000. Отже, ймовірність того, що модель випадкова, менша 0,001.

Стандартизовані коефіцієнти (табл. 7) показують силу та напрямок впливу предикторів: F_1 має невеликий позитивний вплив (+ 0,126); F_2 – дуже сильний негативний вплив (– 0,857); F_3 – невеликий позитивний вплив (+ 0,164). Кожен коефіцієнт статистично значущий (p-value < 0,001). Відповідно найсильніший предиктор F_2 , що свідчить про те, що змінні, які початково сформували зазначений фактор, найсильніше впливають на виявлення рівня ризику для лікувальної імплантації. Проте зауважимо, що всі фактори здійснюють суттєвий вплив на залежну змінну навіть на рівні значущості 0,001.

Таблиця 7

Стандартизовані коефіцієнти та статистична значущість предикторів моделі категоріальної регресії для залежної змінної "Група ризику"

Предиктор	Стандартизовані коефіцієнти		Ступені свободи	F-критерій	Рівень значущості (p – value)
	Beta	Оцінка стандартної похибки			
F_1	0,126	0,028	2	20,869	0,000
F_2	-0,857	0,021	1	1681,605	0,000
F_3	0,164	0,033	2	25,502	0,000

Відповідно нашу модель категоріальної регресії можна записати у вигляді (2):

$$Y_i^* = \beta_0 + 0,126 \cdot f_{1i}^* - 0,857 \cdot f_{2i}^* + 0,164 \cdot f_{3i}^* + \varepsilon_i.$$

У табл. 8 наведені важливі характеристики категоріальної регресії. Нульова кореляція – це проста кореляція кожного предиктора із залежною змінною, без контролю за іншими предикторами. Часткова кореляція – кореляція предиктора із залежною змінною, урахувавши вплив інших предикторів. Напівчасткова кореляція: у цьому випадку вплив інших предикторів вилучається лише із предиктора, а не із залежної змінної. Внесок ознаки – це показник, що відображає "вагу" предиктора після оптимального шкалювання у категоріальній регресії. Показник мультиколінеарності зі значенням $< 0,1$ свідчить про сильну мультиколінеарність (Liao, 2019).

Таблиця 8

Кореляції, внесок предикторів та показники мультиколінеарності для моделі категоріальної регресії (залежна змінна "Група ризику")

Предиктор	Кореляція			Внесок предиктора	Показник мультиколінеарності	
	Нульова кореляція	Часткова кореляція	Напівчасткова кореляція		Після трансформації	До трансформації
F_1	0,255	0,311	0,123	0,037	0,961	0,823
F_2	-0,906	-0,911	-0,831	0,905	0,939	0,886
F_3	0,300	0,393	0,161	0,057	0,960	0,920

Отже, найвищі абсолютні значення усіх типів кореляцій має фактор F_2 (нульова -0,906, часткова -0,911, напівчасткова -0,831). Це знову підтверджує, що саме F_2 є основним фактором, що визначає зміну залежної змінної "Група ризику". Високе значення внеску предиктора після трансформації (0,905) також підтверджує його провідну роль у моделі. F_1 і F_3 демонструють слабші зв'язки із залежною змінною (кореляції 0,255 та 0,39). Їхній внесок після трансформації становить 0,037 і 0,057 відповідно, що вказує на другорядний внесок у модель.

Значення показників мультиколінеарності до та після трансформації для всіх факторів суттєво не відрізняються. Це означає, що оптимальне шкалювання не спотворило структуру зв'язків, а лише уточнило їх.

Усі значення показника мультиколінеарності $> 0,8$, що свідчить про високу незалежність факторів між собою, тобто про відсутність мультиколінеарності.

Отже, модель стабільна, і жоден фактор не дублює інформацію іншого.

Дискусія і висновки

Методологічний підхід "факторний аналіз у поєднанні з категоріальною регресією" ефективно скоротив багатовимірну початкову інформацію (15 змінних) до трьох латентних факторів (F_1 – F_3) зі збереженням клінічної інтерпретованості і статистичної інформативності.

Індикатори придатності даних (статистика КМО = 0,740; тест сферичності Бартлетта $p < 0,001$) підтверджують доцільність факторного аналізу для цієї вибірки ($n = 558$). Перші три фактори пояснюють приблизно 70 % дисперсії до обертання (після обертання сумарна інтерпретована дисперсія становить 64,6 %), що є прийнятним результатом для медико-біологічних даних.

Коротко проаналізуємо клінічне значення виділених факторів. Фактор F_1 "Клінічні та морфологічні ознаки ГЕРХ" охоплює симптоми та структурні маркери рефлюксу, а також має помірний позитивний вплив на ризик. Фактор F_2 "Анатомо-функціональні показники стравоходу" є найсильнішим предиктором і має значний (негативний у межах моделі шкалювання) вплив на категорію ризику, що підкреслює роль анатомічних характеристик (висота сфінктера, довжина пролабування, висота пролабування) у визначенні придатності до імплантації. Фактор F_3 "Ерозивні ураження слизової" додає інформацію про активність ерозивного процесу і підвищує прогностичну здатність моделі.

Категоріальна регресія забезпечила високу якість прогнозу: множинний $R = 0,926$ і $R^2 = 0,858$ (скоригований $R^2 = 0,857$), що вказує на те, що запропонована комбінація факторів пояснює значну частину варіації рівнів ризику. Усі три фактори статистично значущі ($p < 0,001$). Показники мультиколінеарності предикторів $> 0,8$ свідчать про відсутність критичної мультиколінеарності.

Отримана модель може слугувати допоміжним інструментом для стандартизованої стратифікації ризику перед імплантацією пристроїв у ділянці шлунково-стравохідного з'єднання, сприяючи об'єктивнішому ухваленню рішень і визначенню пацієнтів, яким потрібна додаткова оптимізація стану перед процедурою.

Зауважимо, що дослідження обмежене однорегієюною вибіркою та ретроспективною / обсерваційною природою аналізу; модель потребує зовнішньої валідації на незалежних когортах, перевірки робастності для різних підгруп пацієнтів (напр., за віком, супутніми хворобами) й оцінки клінічного впливу за проспективного застосування. Подальші дослідження можуть розглянути інтеграцію додаткових маркерів (функціональні тести, тривалі моніторинги) і використання інших алгоритмів класифікації для порівняння продуктивності.

Джерела фінансування. Фінансування частково забезпечено Київським національним університетом імені Тараса Шевченка.

Список використаних джерел

- Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Devlieger, I. (2019). *Factor score regression: Models for the analysis of variables and factors* [Doctoral dissertation, Ghent University]. Ghent University Academic Repository. <https://backoffice.biblio.ugent.be/download/8655145/8655146>
- Finch, W. H., & French, B. F. (2018). *Educational and Psychological Measurement* (1st ed.). Routledge. <https://doi.org/10.4324/9781315650951>
- Guo, J., Wall, M., & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1), 145–163. <https://doi.org/10.1093/biostatistics/kxi046>
- IBM. (n. d.). *CATREG*. IBM SPSS Statistics Algorithms. <https://public.dhe.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/14.0/catreg.pdf>

- Lawenko, R. M. A., & Lee, Y. Y. (2016). Evaluation of gastroesophageal reflux disease using the Bravo capsule pH system. *Journal of Neurogastroenterology and Motility*, 22(1), 25–30. <https://pubmed.ncbi.nlm.nih.gov/26717929/>
- Li, J. N., Liu, C. L., & Tao, X. H. (2013). Clinical utility and tolerability of JSPH-1 wireless esophageal pH monitoring system. *BMC Gastroenterology*, 13(10). <https://doi.org/10.1186/1471-230X-13-10>
- Liao, Y. (2019). *Feature Selection: Removing Multicollinearity – Python Automation and Machine Learning for EM and ICs*. An Online Book, First Edition. https://www.globalsino.com/ICs/page3268.html?utm_source=chatgpt.com
- Smith, L. E., Smith, D. K., Blume, J. D., Siew, E. D., & Billings IV, F. T. (2017). Latent variable modeling improves AKI risk factor identification and AKI prediction compared to traditional methods. *BMC Nephrology*, 18(1), 55. <https://doi.org/10.1186/s12882-017-0465-1>
- Statistics Solutions. (n. d.). *Comprehensive guide to factor analysis*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/?utm_source=chatgpt.com
- Viana, A. T., Gonçalves, K. C. M., & Paez, M. S. (2024). Factor model for ordinal categorical data with latent factors explained by auxiliary variables applied to the Major Depression Inventory. *Journal of Applied Statistics*, 51(14), 2866–2893. <https://doi.org/10.1080/02664763.2024.2321913>
- Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>

References

- Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Devlieger, I. (2019). *Factor score regression: Models for the analysis of variables and factors* [Doctoral dissertation, Ghent University]. Ghent University Academic Repository. <https://backoffice.biblio.ugent.be/download/8655145/8655146>
- Finch, W. H., & French, B. F. (2018). *Educational and Psychological Measurement* (1st ed.). Routledge. <https://doi.org/10.4324/9781315650951>
- Guo, J., Wall, M., & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1), 145–163. <https://doi.org/10.1093/biostatistics/kxi046>
- IBM. (n. d.). *CATREG*. IBM SPSS Statistics Algorithms. <https://public.dhe.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/14.0/catreg.pdf>
- Lawenko, R. M. A., & Lee, Y. Y. (2016). Evaluation of gastroesophageal reflux disease using the Bravo capsule pH system. *Journal of Neurogastroenterology and Motility*, 22(1), 25–30. <https://pubmed.ncbi.nlm.nih.gov/26717929/>
- Li, J. N., Liu, C. L., & Tao, X. H. (2013). Clinical utility and tolerability of JSPH-1 wireless esophageal pH monitoring system. *BMC Gastroenterology*, 13(10). <https://doi.org/10.1186/1471-230X-13-10>
- Liao, Y. (2019). *Feature Selection: Removing Multicollinearity – Python Automation and Machine Learning for EM and ICs*. An Online Book, First Edition. https://www.globalsino.com/ICs/page3268.html?utm_source=chatgpt.com
- Smith, L. E., Smith, D. K., Blume, J. D., Siew, E. D., & Billings IV, F. T. (2017). Latent variable modeling improves AKI risk factor identification and AKI prediction compared to traditional methods. *BMC Nephrology*, 18(1), 55. <https://doi.org/10.1186/s12882-017-0465-1>
- Statistics Solutions. (n. d.). *Comprehensive guide to factor analysis*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/?utm_source=chatgpt.com
- Viana, A. T., Gonçalves, K. C. M., & Paez, M. S. (2024). Factor model for ordinal categorical data with latent factors explained by auxiliary variables applied to the Major Depression Inventory. *Journal of Applied Statistics*, 51(14), 2866–2893. <https://doi.org/10.1080/02664763.2024.2321913>
- Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>

Отримано редакцією журналу / Received: 30.11.25

Прорецензовано / Revised: 03.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Oleksandr HALUSHKO, PhD Student

ORCID ID: 0009-0009-0275-6542

e-mail: health24@ukr.net

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

INTEGRATED MODEL OF FACTOR SELECTION AND CATEGORICAL REGRESSION FOR PREDICTING THE IMPLANTATION RISK OF DIAGNOSTIC-THERAPEUTIC DEVICES IN GASTROESOPHAGEAL JUNCTION PATHOLOGY

The article presents an integrated method for predicting the implantation risk of diagnostic-therapeutic devices in the gastroesophageal junction based on factor analysis and categorical regression. In a sample of 558 patients, 15 clinical and endoscopic indicators were selected, characterizing the mucosal condition, anatomical features, and reflux symptoms. Factor analysis using principal axis factoring with oblimin rotation reduced the initial set of variables to three latent factors: F_1 – "clinical and morphological features of gastroesophageal reflux disease", F_2 – "anatomical and functional indicators of the esophagus", and F_3 – "erosive mucosal lesions". These factors were used as predictors in categorical regression to forecast five levels of implantation risk. The model demonstrated high predictive quality ($R^2 = 0,858$), all factors were statistically significant ($p < 0,001$) and multicollinearity indicators of predictors $> 0,8$ confirmed the absence of critical multicollinearity. The resulting model provides standardized risk stratification, enabling clinicians to more objectively assess the appropriateness of device implantation and identify patients who require additional optimization prior to the procedure.

Key words: device implantation, factor analysis, categorical regression, risk, gastroesophageal junction, risk stratification.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 004.912

DOI: <https://doi.org/10.17721/1812-5409.2026/1.26>

Микола ГЛИБОВЕЦЬ, д-р фіз.-мат. наук, проф.

ORCID ID: 0009-0005-6942-8026

e-mail: glib@ukma.edu.ua

Національний університет "Києво-Могилянська академія", Київ, Україна

Данило ВАНИН, магістр

ORCID ID: 0009-0007-4510-8209

e-mail: danylo.vanin@gmail.com

Національний університет "Києво-Могилянська академія", Київ, Україна

ВИКОРИСТАННЯ ОДНО- ТА БАГАТОМОВНИХ МОДЕЛЕЙ НА БАЗІ BERT ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ АВТОМАТИЧНОЇ ОБРОБКИ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

Об'єктом дослідження пропонованої статті є одно- та багатомовні моделі на основі BERT. Предметом дослідження – порівняння продуктивності таких моделей на завданнях обробки природної мови (ОПМ) з акцентом на їхньому застосуванні для української мови. Методологічну основу порівняльного аналізу становило використання стандартних підходів до навчання й оцінки моделей. У дослідженні використовували доступні джерела інформації. Експерименти виконувалися лише на відкритих або власноруч донаведених моделях й охоплювали наявні українські бенчмарки.

Результати дослідження свідчать про те, що як одномовні, так і багатомовні моделі на основі BERT можуть бути ефективними для розв'язання завдань ОПМ залежно від конкретної мови, завдання та доступних ресурсів. Хоча одномовні моделі часто перевершують багатомовні в завданнях своєї конкретної мови, багатомовні моделі можуть мати перевагу, коли ресурси для навчання одномовних моделей обмежені.

Результати порівняння демонструють те, що моделі, натреновані "з нуля" на українських текстах або дотреновані на них, загалом мають кращі результати на завданнях ОПМ української мови, ніж багатомовні. Це вдалося показати на завданнях розпізнавання іменованих сутностей, класифікації текстів і заповнення пропусків. Багатомовні моделі демонструють конкурентні або кращі результати у "легших" завданнях, наприклад, у задачі розрізнення значень слів, особливо коли ресурси для повноцінного тренування одномовної моделі обмежені.

Таке зіставлення підтверджує необхідність адаптації або повного тренування україномовних моделей для спеціалізованих задач, водночас залишаючи простір для економічно доцільного використання багатомовних систем у типових сценаріях.

Практична цінність роботи полягає у формуванні узагальненої карти продуктивності сучасних моделей та у підґрунті для створення повномасштабного українського GLUE-подібного бенчмарку, що стимулюватиме подальші дослідження та підвищить точність застосувань ОПМ для української мови.

Ключові слова: обробка природної мови, великі мовні моделі, одно- та багатомовні моделі, BERT.

Класифікація відповідно до AMS 2020: 68T50, 68T07.

Вступ

Обробка природної мови (ОПМ) охоплює безліч завдань, спрямованих на те, щоб інтелектуальні програмні системи могли розуміти, інтерпретувати та генерувати людську мову. Останні дослідження у сфері глибокого машинного навчання, зокрема створення та використання моделей на основі архітектури трансформера (Vaswani et al., 2017; Глибовець, & Бікчентаєв, 2022), як BERT (Devlin et al., 2019), значно розширили можливості для розв'язання задач обробки природної мови (ОПМ). Та їх використання потребує значних ресурсів.

Досить перспективною альтернативою для мов з обмеженими ресурсами стали багатомовні (multilingual) моделі, що навчаються на злитих корпусах багатьох мов і показують високі результати на багатьох із них. Для мов з обмеженою кількістю ресурсів для навчання багатомовні моделі одразу показували передові на той час результати. Причина цьому – крос-лінгвістичний трансфер, або "перевикористання" знань про одну мову для розуміння іншої. Багатомовні моделі можуть "перевикористовувати" певні "знання" про мови з великими ресурсами на мовах з обмеженими ресурсами.

Нині багато досліджень присвячені вивченню можливостей одномовних і багатомовних моделей. Деякі дослідження, як-от (Virtanen et al., 2019), демонструють, що одномовні моделі перевершують багатомовні на багатьох окремих завданнях для певної мови. Інші дослідження доводять, що багатомовні моделі можуть одночасно бути ефективними для великоресурсних мов і показувати кращі результати, ніж одномовні для тих мов, де ресурси обмежені (напр. у випадку української мови).

Об'єктом дослідження цієї роботи є одно- та багатомовні моделі на основі BERT. Предметом дослідження – порівняння продуктивності таких моделей на окремих типах завдань ОПМ (української мови): розпізнавання іменованих сутностей, розрізнення значень, класифікація текстів, заповнення пропусків і розмічування частин мови. Мета дослідження – оцінити ефективність різних типів моделей BERT для застосування до виконання зазначених завдань. Для досягнення цієї мети було вирішено оцінити: результати одномовних моделей, які були натреновані на переважно українських текстах; адаптованих до української багатомовних моделей (наприклад, за методом WECHSEL); використання "чистих" багатомовних моделей. Зрозуміло, що потрібно було провести і порівняльний аналіз отриманих результатів застосування різних типів моделей та визначити особливості їхнього використання.

Методологічною основою дослідження було використання стандартних підходів до навчання й оцінки моделей. Ми використовували доступні дотреновані на завданні моделі, або базові моделі, які були дотреновані для завдання у межах дослідження. Оцінювали моделі за допомогою таких метрик, як F1-score (для розпізнавання іменованих сутностей) або точність (для розмічування частин мови). Завдання заповнення пропусків оцінювали за допомогою опитування, а також через ручне порівняння й аналіз якості відповідей різних моделей.

У цьому дослідженні для порівняння одно- та багатомовних моделей були обрані ті завдання, для яких існують уже готові набори даних чи бенчмарки українською мовою, є результати оцінки україномовних моделей або публічно доступні моделі, які дотреновані або можна дотренувати для певного завдання.

Дослідження має як наукове, так і практичне значення. З наукової позиції, воно допомагає краще розуміти компроміс між використанням одно й багатомовних моделей і потенційні особливості кожного з підходів. З практичного боку – результати дослідження можна використовувати для зваженого обрання найоптимальнішої моделі для конкретного завдання. Зібрані результати порівняння та перелік моделей можна також використати, як основу для створення комплексного бенчмарку оцінки застосування моделей для української мови.

1. Основні означення та результати

Обробка природної мови (ОПМ) передбачає безліч завдань. Серед них можна виділити: класифікація тексту (Text Classification); виявлення спаму (Spam Detection); визначення наміру (Intent Detection); маркування послідовностей (Sequence Labeling); генерація тексту (Text Generation); машинний переклад (Machine Translation); порівняння тексту (Text Comparison); інформаційний пошук (Information Retrieval); аналіз тексту (Text Analysis); розуміння мови (Language Understanding).

Вибір архітектури мовної моделі залежить від конкретного завдання ОПМ. Наприклад, для нескладних завдань типу "послідовність – послідовність" (машинний переклад), можуть бути достатніми рекурентні нейронні мережі (RNN). Завдання класифікації спаму може бути розв'язано і наївним баєсовим класифікатором. Поширена зараз трансформерна архітектура стала дуже популярною завдяки універсальності та здатності навчатися виявляти зв'язки між елементами тексту, розташованими на різних відстанях. Вони розв'язують проблему згасання важливості контексту та регулярного перерахунку контекстних ваг у рекурентних архітектурах.

Часто завданням на етапі базового тренування є просте передбачення слова в контексті. Наприклад, модель навчається передбачувати "замасковане" слово у реченні. Для цього завдання не потрібні розмічені дані і для навчання потенційно можна скористатися будь-яким якісним текстом на цільовій мові.

Незважаючи на те, що BERT уже вміють на загальному рівні розуміти мови (вміння передбачати слова у контексті), усе ж цього недостатньо для більш специфічних завдань. Базова модель може лише передбачати масковані слова, а для складнішого завдання, відповідно, як відповідь на питання чи переказ тексту – модель має бути донавчена на розмічених даних. Замість того, щоб навчати модель "з нуля", процес тонкого налаштування починається з використання вже наявної попередньо навченої моделі, яку адаптують під конкретне завдання. Це дає змогу "перевикористати" знання, отримані з великого неструктурованого набору даних під час попереднього навчання. Додавання нових даних, що близькі до кінцевого завдання, допомагає налаштувати модель для розв'язання конкретних задач.

Тонке налаштування передбачає оптимізацію гіперпараметрів, таких як швидкість навчання, розмір пакету даних і кількість епох навчання. Це допомагає досягти найкращої продуктивності на специфічних для завдання даних і також визначається емпірично.

Одномовний BERT (*Bidirectional Encoder Representations from Transformers* – двонаправлене кодування представлень із трансформерів) – це мовна модель, що використовує трансформерну архітектуру. Вона складається із шарів самоуваги (self-attention) і нейронних мереж прямого поширення (Devlin et al., 2019). Модель попередньо навчається на великих одномовних корпусах через передбачення замаскованого токена у нерозміченому тексті. У цьому методі частина вхідних токенів випадково маскується, а модель навчається прогнозувати їх на основі контексту.

Такий підхід допомагає BERT запам'ятовувати інформацію про мову з неструктурованих даних, як-от значення окремих слів, порядок слів у реченні та граматичні категорії. Завдяки цьому модель можна легко налаштувати для різних завдань із мінімальними модифікаціями, специфічними для завдання (Devlin et al., 2019). Розроблені багатомовні варіанти BERT, такі як mBERT та XLM-RoBERTa (Shaham et al., 2024), розширюють можливості моделі для роботи з кількома мовами.

Ці моделі навчаються на об'єднаних корпусах до 100 мов, використовуючи спільну інформацію між різними мовами для поліпшення роботи на окремих мовах, а особливо для мов з низьким рівнем ресурсів, таких як українська (Liu et al., 2019). Основна ідея полягає в тому, що літери та фрагменти слів часто збігаються у схожих мовах, що спрощує тренування токенизатора, а також багато мов мають подібну семантичну структуру, логіку побудови слів чи навіть спільні слова. Використовуючи багатомовні корпуси, вдається значно збільшити розмір тренувальних даних, а модель може скористатися зі схожості різних мов.

Проте ефективність багатомовних моделей BERT може варіюватися залежно від конкретної мови та завдання. Питання того, які з двох моделей є ефективнішими для конкретної мови, можна визначити лише емпірично. Це дослідження присвячене такому емпіричному порівнянню цих моделей на різних завданнях ОПМ для української. Багатомовні моделі можуть досягати кращих результатів у широкому діапазоні мов з великими ресурсами. Такий "крослінгвістичний трансфер" дає моделям змогу використовувати знання, отримані в одній мові, для інших мов, що особливо корисно для мов з обмеженими ресурсами (Sanh et al., 2019).

Одномовні моделі часто можуть перевершувати свої багатомовні аналоги в деяких завданнях, включно з виявленням мови ворожнечі та генеруванням речень (Pires et al., 2019; Sido et al., 2021). Ця вища продуктивність може бути пов'язана зі здатністю одномовної моделі глибше розуміти нюанси і тонкощі однієї мови, а не розподіляти свої ресурси між кількома мовами (Ruder, 2022). Важливим також є тип даних, на яких тренувалася модель.

Порівняння результатів одномовних і багатомовних моделей для української мови може вказати, у яких напрямках дослідники повинні приділити більше уваги розробленню якісних україномовних наборів даних і моделей.

Українська мова має багату морфологічну структуру, характерну для синтетичних мов (Haltiuik, & Smywiński-Pohl, 2024). Вона стикається зі значними викликами в області обробки природної мови, насамперед через нестачу даних для тренування. У 2020 р. українську мову визначили як "rising star" – мову, що страждає від нестачі розмічених наборів даних (Joshi et al., 2020). Основною перешкодою для української мови є нестача загальнодоступних, орієнтованих на конкретні завдання корпусів (Haltiuik, & Smywiński-Pohl, 2024). Це заважало розвитку надійних одномовних моделей

для української мови та змушувало дослідників покладатися на крос-мовне трансферне навчання з інших мов (Hamotskyi et al., 2024; Torge et al., 2023).

У представленому дослідженні буде також перевірена гіпотеза, що модель, натренована на корпусі слов'янських мов, може показати кращі результати, ніж одномовна модель для конкретних завдань.

2. Порівняння одно та багатомовних моделей на завданнях обробки української мови

2.1. Бенчмарки для оцінки результатів

Оскільки метою роботи є порівняння результатів моделей на різних завданнях обробки природної мови, важливо визначити потенційні бенчмарки та завдання, які можна використати для такого порівняння. Попри дедалі більший інтерес до ОПМ для української, комплексних бенчмарків для оцінки мовних моделей досі бракує. І хоча для конкретних завдань існують готові набори даних, стандартизованого та повного бенчмарку, подібного до GLUE (Wang et al., 2019) або SuperGLUE (Wang et al., 2020) для англійської чи KLEJ (Rybak et al., 2020) для польської, який би покривав широке коло завдань, наразі немає.

Нам були доступні комплексні бенчмарки: UA-datasets (Panchenko et al., 2022), Eval-UA-tion (Chaplynskyi, & Romanushyn, 2024). Перша колекція містить набори даних для різних завдань, таких як розмітка частин мови (POS tagging) і класифікація новин (Panchenko et al., 2022). Другий набір охоплює дані для оцінки продуктивності мовних моделей у таких завданнях, як розуміння наративів оповідань (UA-CBT), асоціація статей із їхніми заголовками (UP-Titles) й елементарні мовні завдання (LMES).

Через відсутність готових комплексних бенчмарків для порівняння моделей, у пропонованій роботі використано поєднання результатів на декількох окремих завданнях. Серед них розпізнавання іменованих сутностей, розрізнення значень слів, класифікація текстів, завдання бенчмарку Eval-UA-tion, заповнення пропусків і розмічення частин мови.

2.2. Розпізнавання іменованих сутностей (NER)

Розпізнавання іменованих сутностей полягає у виявленні та класифікації імен, назв організацій, місць, топографічних назв тощо у текстах.

Для української мови дослідники нещодавно випустили оновлений корпус NER-UK 2.0 (Chaplynskyi, & Romanushyn, 2024), що містить 323 200 слів і 21 993 іменовані сутності. У дослідженні автори також долучили початкові результати для моделі robertalarge (Hugging Face, 2024). У табл. 1 наведено порівняння результатів різних моделей на попередній версії корпусу NER-UK 1.0 (GitHub, 2025). Тут об'єднані результати з (Haltiuik, & Smywiński-Pohl, 2024); Hugging Face, 2024); GitHub, 2025) та власного дослідження. Значення в дужках позначають стандартне відхилення метрики.

Таблиця 1

Результати оцінки моделей на NER-UK 1.0

Тип і мова моделі	Розмір моделі	NER-UK 1.0 F1 оцінка
RoBERTa WECHSEL Ukrainian	base	90.81 (1.51)
RoBERTa WECHSEL Ukrainian	large	91.24 (1.16)
RoBERTa Scratch Ukrainian	base	89.57 (1.01)
RoBERTa Scratch Ukrainian	large	89.96 (0.89)
ELECTRA Ukrainian	base	90.43 (1.29)
XLM RoBERTa Multilingual	base	90.86 (0.81)
XLM RoBERTa Multilingual	large	90.16 (2.98)
LiBERTa Ukrainian (pretrained)	large	91.27 (1.22)
RoBERTa Slavic	large	90.89
uk_ner_web_trf_13class RoBERTa	large	91.3

Таблиця 1 містить результати моделі, натренованої для української мови методом WECHSEL (Hugging Face, 2025), багатомовної моделі XLM-R та нещодавно опублікованої моделі LiBERTa (Haltiuik, & Smywiński-Pohl, 2024). Остання була натренована з "нуля" на наборах даних української мови. Порівняння показує, що моделі, адаптовані для української мови, ефективніші за багатомовні в розпізнаванні іменованих сутностей.

Як зазначено в (Haltiuik, & Smywiński-Pohl, 2024), друга велика модель XLM-R демонструє гірші результати, ніж усі базові моделі. Вона також має найбільшу варіативність результатів. Це підкреслює необхідність тренування моделей, специфічних для певної мови, оскільки як WECHSEL-RoBERTa, так і LiBERTa мають меншу варіативність результатів.

Таблиця також містить результати моделі на корпусі слов'янських мов (болгарська, чеська, польська, словенська, російська й українська) (Piskorski et al., 2024). Ця модель була додатково дотренована на частині набору даних у ході дослідження. Можна побачити, що результати моделі, натренованої на слов'янських мовах, кращі, ніж у багатомовних моделях, але все ж гірші, ніж у моделей, адаптованих для української мови.

На бенчмарку WikiANN (Pan et al., 2017), який також передбачає розпізнавання іменованих сутностей, LiBERTa, яка повністю натренована на українській мові, показує гірші результати. Результати порівняння моделей наведені у табл. 2. Значення взято з (Haltiuik, & Smywiński-Pohl, 2024).

Таблиця 2

Результати оцінки моделей на WikiANN

Тип і мова моделі	Розмір моделі	WikiANN micro f1
RoBERTa WECHSEL Ukrainian	base	92.98 (0.12)
RoBERTa WECHSEL Ukrainian	large	93.22 (0.17)
ELECTRA Ukrainian	base	92.99 (0.11)
XLM RoBERTa Multilingual	base	92.27 (0.09)
XLM RoBERTa Multilingual	large	92.92 (0.19)
LiBERTa Ukrainian (pretrained)	large	92.50 (0.07)

Автори зазначають, що такі результати можуть бути спричинені відмінностями між датасетами та навчальними даними для моделі. Модель і токенизатор були натреновані переважно на українських текстах, тоді як багатомовні моделі часто тренуються на даних, зібраних із Wikipedia (Haltiuk, & Smywiński-Pohl, 2024).

Отже, у завданні NER моделі, адаптовані для української мови, показують вищу продуктивність порівняно з багатомовними моделями. Для цього завдання створення та використання моделей, орієнтованих на українську мову, важливе для досягнення кращих результатів. WECHSEL-RoBERTa та LiBERTa, наприклад, показують високу точність і невелику варіативність. Проте різниця між результатами моделі, навченої на українських даних, та RoBERTa, навченої за методом WECHSEL, незначна. Це може означати, що попереднє тренування на українських даних не дало значної переваги порівняно з методом WECHSEL.

2.3. Розрізнення значень слів (Word Sense Disambiguation)

Розрізнення значень слів (WSD) допомагає розв'язати проблему неоднозначності мови (наявність омонімів "ключ журавлів" – "ключ від дверей"). Оскільки багато слів мають кілька значень, натренована модель намагається визначити правильний сенс слова залежно від контексту. Розв'язок цього завдання може бути частиною ширшого конвеєру з моделей, ціль якого переказати текст чи відповісти на питання про текст.

У дослідженні (Laba et al., 2023) автори порівнюють точність розрізнення значень слів (WSD) для української мови, використовуючи різні стратегії об'єднання та попередньо навчені моделі без додаткового тренування.

Таблиця 3 (значення запозичено з (Laba et al., 2023)) демонструє, що модель paraphrasemultilingual-mpnet-base-v2 (PMMBv2) з (Reimers, & Gurevych, 2019) показує кращі результати WSD, ніж модель RoBERTa, попередньо навчена на українській мові у 2020 р. (Radchenko, 2020). У роботі (Laba et al., 2023) підвищили продуктивність PMMBv2 до 0.779, тонко налаштувавши її на основі датасету зі Словника української мови (СУМ).

Таблиця 3

Середня точність розрізнення значень слів без тонкого налаштування

Тип і мова моделі	Розмір моделі	Середня точність WSD
mBERT cased multilingual	base	0.602
XLM RoBERTa Multilingual	base	0.529
XLM RoBERTa Multilingual	large	0.547
XLM RoBERTa Ukrainian	base	0.528
RoBERTa Ukrainian	large	0.580
Paraphrase MPNET Multilingual	base	0.735

Утім, щоб отримати релевантне порівняння між одномовними та багатомовними моделями в задачі WSD, необхідно протестувати сучаснішу українську модель, таку як LiBERTa (Chaplynskyi, & Romanyshyn, 2024), з додатковим налаштуванням на тому ж наборі даних.

2.4. Класифікація текстів

У дослідженні (Haltiuk, & Smywiński-Pohl, 2024) автори представили модель LiBERTa, що була попередньо навчена з нуля на українських даних. Вони надають результати порівняння її продуктивності на завданні класифікації текстів з Ukrainian News Classification benchmark (Panchenko et al., 2022). Цей бенчмарк показує, наскільки точно модель може визначити, до якого новинного джерела належить певна стаття (багатокласова класифікація).

Як видно з табл. 4 (значення запозичено з (Haltiuk, & Smywiński-Pohl, 2024)), модель, попередньо навчена на українських даних, перевершує багатомовну модель XLM-R за продуктивністю. Однак модель, адаптована до української мови за допомогою методу WECHSEL, показує ще кращі результати.

Таблиця 4

Порівняння результатів моделей на задачі класифікації українських новин

Тип і мова моделі	Розмір моделі	Класифікація новин macro F1
RoBERTa WECHSEL Ukrainian	large	96.48 (0.09)
XLM RoBERTa Multilingual	large	95.13 (0.49)
LiBERTa Ukrainian (pretrained)	large	95.44 (0.04)

2.5. Eval-UA-tion

Eval-UA-tion (Hamotskyi et al., 2024) – це спеціальний бенчмарк, розроблений для оцінки продуктивності мовних моделей на українській мові. Він складається з трьох основних завдань: UA-CBT, UP-Titles, LMentry-static-UA (LMES).

Перше завдання створене за аналогією з тестом Children's Book Test. Воно передбачає заповнення пропусків у розповідях правильними словами з наданого набору варіантів. Завдання оцінює здатність моделі розуміти текст і передбачати пропущені слова на основі контексту.

У завданні UP-Titles потрібно зіставити статті з онлайн-газети "Українська Правда" з їхніми правильними заголовками з наданого набору схожих варіантів. Це оцінює здатність моделі зрозуміти основну ідею статті та відрізнити її від схожих тем.

LMentry-static-UA (LMES) – набір завдань, розроблений на основі бенчмарку LMentry, спеціально для оцінки можливостей мовних моделей. Ці завдання, хоч і прості для людей, часто є складними для штучного інтелекту – наприклад, визначення найдовшого слова з пари або виділення N-го слова в реченні.

Eval-UA-tion містить базові результати для кожного завдання: показник людських відповідей і випадкової генерації. Мета Eval-UA-tion полягає у всебічній оцінці можливостей мовних моделей для української мови та стимулюванні подальших досліджень у цій сфері. Він був представлений на конференції UNLP 2024 і містить оцінки для різних багатомовних й одномовних моделей. Незважаючи на використання сучасніших моделей замість BERT, порівнянню яких присвячено цю роботу, автори показують результати, що є корисними для порівняння продуктивності багатомовних й одномовних моделей.

Згідно з результатами, представленими на рис. 1, модель SherlockAssistant/Mistral-7B-Instruct-Ukrainian, яка пройшла попереднє налаштування на українських даних, виявилася ефективнішою для завдання UP-Titles (Hamotskyi et al., 2024), ніж багатомовні моделі, що не є OpenAI, та навіть перевершила GPT-3. Хоча GPT-4, багатомовна модель, показала найкращі результати, варто зауважити, що пряме порівняння ускладнюється через значні відмінності у розмірі моделі й обсязі набору даних для попереднього навчання порівняно з іншими моделями.

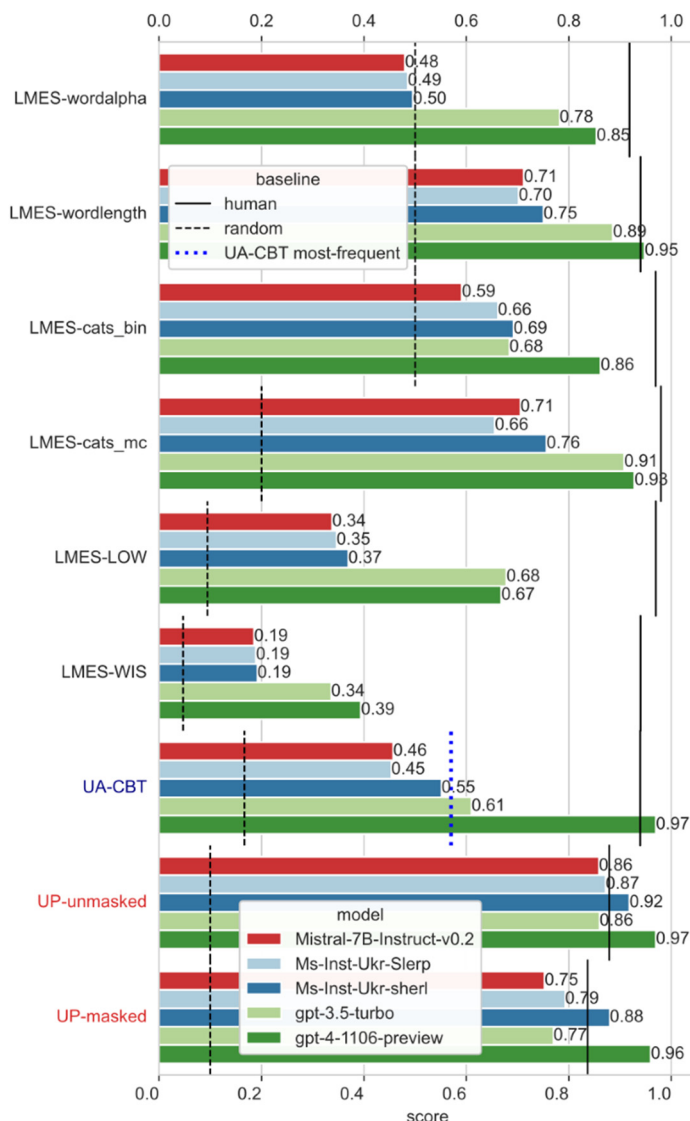


Рис. 1. Результати порівняння моделей на завданнях Eval-UA-tion (Hamotskyi et al., 2024). Позначення: baseline – базова модель; human – оцінка людиною; random – випадковий вибір; UA-CBT most-frequent – найчастіший клас. Вісь X – значення метрики (score)

2.6. Заповнення пропусків (Mask filling)

На етапі попереднього навчання моделей на основі BERT зазвичай їх тренують передбачати випадково замасковані слова у великому корпусі даних без розмітки. Це допомагає мовній моделі вивчати фундаментальні представлення мови або мовні структури, такі як порядок слів у реченнях і значення слів (через навчання токенизатора та перетворення векторів). Ці попередньо навчені моделі потім донавчаються на специфічних наборах даних для конкретних завдань. Тому оцінювання продуктивності моделі на завданні заповнення пропусків потенційно може показати, наскільки добре модель "розуміє" українську мову.

Для оцінювання моделей нами створено набір речень із замаскованими словами, що вимагають не лише знання української мови, але й "розуміння" українського контексту (культура, поезія, історія, кухня, традиції). Якість заповнення пропусків різними моделями дає змогу визначити якість даних, використаних під час навчання. Наприклад, відсутність оригінальних українських текстів може призвести до гірших результатів у реченнях, пов'язаних з українською культурою.

Для дослідження якості заповнення пропусків ми створили набір питань, що генерувалися моделлю Gemini, а також дописувалися та перевірялися вручну. Для заповнення пропусків на кожному питанні використовували кілька моделей. Оцінювані моделі охоплювали багатомовні моделі, моделі, які були додатково навчені на українських текстах, й одномовні моделі. Проводили два типи оцінювання: ручне оцінювання для висновків і порівняння продуктивності, та загальне оцінювання.

Загальне оцінювання проводили за допомогою вебсайту, де респонденти бачили випадкове запитання з пропуском і відповіді від кількох моделей (інформація про назву моделі, що згенерувала конкретну відповідь, була прихована). Учасникам опитування пропонували оцінити кожну відповідь за шкалою від 1 до 5 за такими критеріями: 5 (точно та контекстно релевантне заповнення пропуску), 4 (переважно точне, але з незначними помилками), 3 (деяка релевантність, але не зовсім точне), 2 (погана релевантність і точність), 1 (повністю нерелевантне або граматично неправильне).

Оскільки метою опитування була оцінка, наскільки якісно заповнені пропуски моделями, а не результатом якої моделі нададуть перевагу люди (Human Feedback), то для опитування були запрошені люди. Для всіх учасників українська мова є основною мовою спілкування. Усі учасники мали вищу освіту.

Оцінювані моделі охоплювали багатомовний XLM-RoBERTa (базовий і великий), GPT4o (як еталон), RoBERTa WECHSEL український (базовий і великий) та YouScan RoBERTa (базовий, попередньо навчений на українському наборі даних). Результати ручного та загального оцінювання наведено нижче.

Результати ручного оцінювання

Багатомовні моделі зазвичай показують гірше знання українського контексту (табл. 6).

Моделі, попередньо навчені або перенесені на українську, часто показують краще знання українського контексту (напр., у табл. 7 модель WECHSEL додала нішеві українські топоніми).

Таблиця 6

Приклад результатів заповнення пропусків різними моделями

Видатні українські письменники, такі як [MASK] і [MASK], зробили великий внесок у літературу. (Питання згенероване ШІ)	
Назва моделі	Результат заповнення пропусків
XLM RoBERTa base	Видатні українські письменники, такі як Шевченко і Шевченка , зробили великий внесок у літературу.
RoBERTa WECHSEL base	Видатні українські письменники, такі як Франко і Шевченко , зробили великий внесок у літературу.
YouScan RoBERTa Uk base	Видатні українські письменники, такі як Шевченко і інші , зробили великий внесок у літературу.

Таблиця 7

Приклад результату заповнення пропуску моделлю WECHSEL

Українські Карпати славляться своїми курортами, такими як [MASK] і [MASK], де можна насолоджуватися [MASK]. (Питання згенероване ШІ)	
Назва моделі	Результат заповнення пропусків
RoBERTa WECHSEL large	Українські Карпати славляться своїми курортами, такими як Трускавець і Яремче, де можна насолоджуватися природою.

Моделі, попередньо навчені на українських даних, також зазвичай генерують несподіваніші відповіді порівняно з іншими моделями. Вони також більш послідовні із закінченнями слів, що є важким завданням для багатомовних моделей, оскільки українська мова є морфологічно багатою.

Однак усі моделі послідовно показують брак якості міркувань і знання українського культурного контексту порівняно з набагато більшою та зрілою моделлю GPT4o. Відповіді схильні бути більш загальними, особливо для запитань, пов'язаних із науковими фактами (табл. 8, 9).

Таблиця 8

Приклад результатів заповнення пропусків, що вимагають знання фактів, різними моделями

Софія Київська – це [MASK], яка була побудована у [MASK] році і є прикладом [MASK] архітектури. (Питання згенероване ШІ)	
Назва моделі	Результат заповнення пропусків
GPT4	Софія Київська – це собор , яка була побудована у 1037 році і є прикладом візантійської архітектури.
XLM RoBERTa base	Софія Київська – це церква , яка була побудована у 1905 році і є прикладом сучасної архітектури.
RoBERTa WECHSEL base	Софія Київська – це церква , яка була побудована у 2007 році і є прикладом української архітектури.
RoBERTa WECHSEL large	Софія Київська – це будівля , яка була побудована у 1991 році і є прикладом сучасної архітектури.
YouScan RoBERTa Uk base	Софія Київська – це церква , яка була побудована у 2001 році і є прикладом народної архітектури.
XLM RoBERTa large	Софія Київська – це церква , яка була побудована у 1929 році і є прикладом української архітектури.

Таблиця 9

Приклад результатів заповнення пропусків, що вимагають знання фактів про історію України чи біологію, згенеровані різними моделями

Людська пам'ять функціонує завдяки [MASK] і [MASK], що дозволяє запам'ятовувати [MASK]. (Питання згенероване ШІ)	
Назва моделі	Результат заповнення пропусків
GPT4	Людська пам'ять функціонує завдяки мозковій діяльності і нейронним зв'язкам , що дозволяє запам'ятовувати інформацію .
XLM RoBERTa base	Людська пам'ять функціонує завдяки енергії і GPS , що дозволяє запам'ятовувати інформацію .
RoBERTa WECHSEL base	Людська пам'ять функціонує завдяки мозку й відео , що дозволяє запам'ятовувати інформацію .
RoBERTa WECHSEL large	Людська пам'ять функціонує завдяки науці й розуму , що дозволяє запам'ятовувати людей .
YouScan RoBERTa Uk base	Людська пам'ять функціонує завдяки технології й енергії , що дозволяє запам'ятовувати інформацію .
XLM RoBERTa large	Людська пам'ять функціонує завдяки руху і часу , що дозволяє запам'ятовувати інформацію .

Це ще раз підкреслює значну перевагу великих комерційних моделей над меншими безкоштовними моделями з відкритим кодом. Така відмінність у продуктивності, ймовірно, пов'язана з тим, що комерційні моделі мають більшу кількість параметрів і навчаються на значно ширших наборах даних.

Результати загального оцінювання

Під час загального оцінювання всі відповіді моделей оцінювали від 5 до 1. GPT4o використовували як еталон. Результати загального оцінювання, наведені на рис. 2, показують, що моделі, перенесені на українську за допомогою методу WECHSEL, перевершують модель YouScan RoBERTa, попередньо навчену на українській мові. Однак найнижчі бали отримали багатомовні моделі XLM-R.

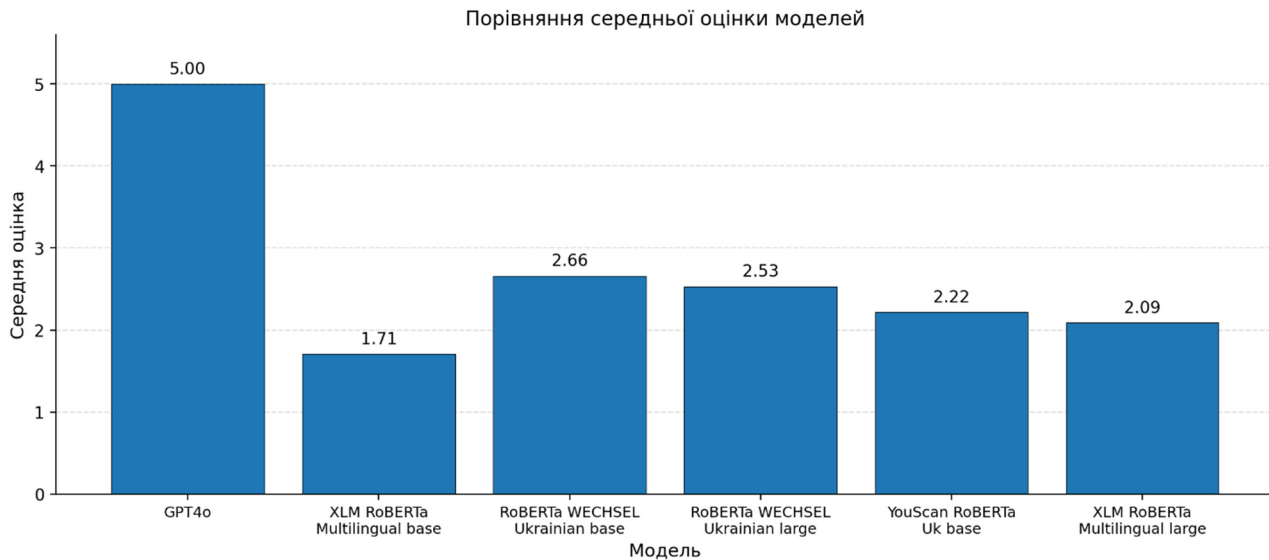


Рис. 2. Результати загального оцінювання моделей (онлайн-опитування)

Водночас жодна модель не досягла якості комерційної моделі GPT4o.

Для завдання заповнення пропусків моделі, попередньо навчені або перенесені на українську, зазвичай працюють краще, ніж багатомовні моделі. Проте адаптовані до української мови за допомогою методу WECHSEL багатомовні моделі, несподівано перевершують моделі, попередньо навчені виключно на українських текстах. Можливим поясненням цього може бути те, що моделі WECHSEL отримують перевагу від глибшого розуміння мови та величезних масивів даних, що використовуються для навчання базових моделей. Однак залишається незрозумілим, чи покажуть вони вищі результати, ніж остання українська модель LiBERTa, друга версія якої показала кращу продуктивність, ніж WECHSEL на інших завданнях (Haltiuk, & Smywiński-Pohl, 2024).

2.7. Розмічування частин мови (POS tagging)

Автори (Haltiuk, & Smywiński-Pohl, 2024) порівнюють продуктивність різних багатомовних й одномовних моделей на українській частині набору даних Universal Dependencies (Nivre et al., 2017). Розмічування частин мови (POS tagging) – це процес призначення граматичної категорії (або "тегу") кожному слову в тексті. Ці теги вказують на частину мови слова, наприклад, іменник, дієслово, прикметник, прислівник, займенник, прийменник, сполучник або вигук.

Згідно з (Haltiuk, & Smywiński-Pohl, 2024), продуктивність усіх моделей у цьому завданні є досить високою та схожою, що може свідчити про те, що завдання POS tagging є відносно простим для української мови, а різниця в результатах може бути випадковою. Результати порівняння наведені в табл. 10.

Таблиця 10

Результати моделей на задачі розмічування частин мови

Тип і мова моделі	Розмір моделі	UD POS (точність)
RoBERTa WECHSEL Ukrainian	base	98.57 (0.03)
RoBERTa WECHSEL Ukrainian	large	98.74 (0.06)
ELECTRA Ukrainian	base	98.59 (0.06)
XLM RoBERTa Multilingual	base	98.45 (0.07)
XLM RoBERTa Multilingual	large	98.71 (0.04)
LiBERTa Ukrainian (pretrained)	large	98.62 (0.08)

Отже, це порівняння показує, що для розмічування частин мови різниця у продуктивності між одномовними й багатомовними моделями є мінімальною.

2.8. Результати порівняння

Результати порівняння одно- та багатомовних моделей наведено в табл. 11.

Можемо зробити висновок, що адаптовані до української одномовні моделі часто демонструють кращі результати у спеціалізованих завданнях. Прикладом цього є показники порівняння моделей для розпізнавання іменованих сутностей і класифікації текстів. Це часто пов'язано з тим, що моделі, спеціально треновані на українських даних, краще розуміють контекст й особливості мови, а також орієнтуються в контексті української культури, від чого можуть страждати багатомовні моделі.

Таблиця 11

Порівняння одно- та багатомовних моделей на українських завданнях NLP

Завдання ОПМ для української мови	Порівняння одно- та багатомовних моделей	Додаткові коментарі
Розпізнавання іменованих сутностей (NER)	Одномовна модель тренувана переважно на українських текстах показує найкращий результат	
Розрізнення значень слів (Word Sense Disambiguation)	Багатомовна модель дотренована українською показує кращий результат	Є потреба порівняти з результатом більш сучасної української моделі LiBERTa v2
Класифікація текстів	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська LiBERTa	Друга версія моделі LiBERTa перевершує WECHSEL, але результати ще не опубліковані
UA-CBT (частина EvalUA-tion)	Модель Mistral-7B-Sherlock дотренована на українських даних показує кращі результати ніж багатомовна модель	Mistral-7B-Sherlock не досягає рівня gpt-41106-preview
UP-Titles		
LMentry-static-UA (LMES):		
Заповнення пропусків	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська YouScan/Roberta	Жодна модель не досягає рівня GPT4
Розмічування частин мови (POS Tagging)	Незначна різниця у результатах одно- та багатомовних моделей	Можливо, пов'язано з легкістю задачі

Водночас багатомовні моделі, дотреновані українською мовою, також можуть бути ефективними для завдань, що не потребують глибокого розуміння специфіки мови. Прикладом таких задач є розрізнення значень. Потенційно відмінність між моделями, натренованими з "нуля" українською та дотренованими українською, може зменшитись, залежно від якості та кількості тренувальних даних. Для деяких простих задач, як розмічування частин мови, різниця між різними типами моделей незначна. У практиці для таких задач вибір правильної моделі може не особливо впливати на ефективність.

Одномовна модель, попередньо навчена на українських даних, хоча і не перевершує WECHSEL-модель для української мови, однак уже показує кращі результати, ніж попередні повністю українські моделі. Це може свідчити про те, що час, коли модель, натренована лише на українських даних, показуватиме кращі результати, ніж адаптована модель, вже близько. (Згідно з виступом (Haltiuik, & Smywiński-Pohl, 2024) на конференції UNLP, друга версія моделі LiBERTa вже показує кращі результати, ніж WECHSEL модель)).

Отже, вибір між одномовними та багатомовними моделями залежить від конкретного завдання: для більш спеціалізованих і складних задач доцільніше використовувати одномовні моделі, тоді як для загальніших задач багатомовні можуть бути аналогічно ефективними. Ця відмінність потенційно може нівелюватися більше, якщо моделі були адаптовані до української мови.

3. Обмеженість і майбутнє дослідження

Описавши отримані результати, важливо відзначити певні обмеження цього дослідження, які варто враховувати під час інтерпретації та подальшого використання висновків.

Обмеженість обчислювальних ресурсів й обсягів даних. Повноцінне навчання "з нуля" великих мовних моделей, таких як BERT, потребує значних обчислювальних потужностей і величезних обсягів даних. Попередня оцінка вартості та часу навчання моделі (64 GPU протягом 4 діб за ціною \$4.5/год) становить приблизно \$7000. Окрім цього, доступні корпуси українських текстів обмежені (оригінальний набір даних BERT містить 3.3 млрд слів). Через ці обмеження в цьому дослідженні переважно використовували тонке налаштування вже навчених моделей. Хоча воно і дає змогу досягти непоганих результатів з меншими витратами грошей і часових ресурсів, це не допомогло повністю оцінити потенціал моделі, оскільки вона вже адаптована до певних даних на етапі базового тренування та містить упередження стосовно цих даних.

Для кращого порівняння роботи моделей потрібно мати ресурси для тренування моделі на корпусі слов'янських мов. До виходу LiBERTa (Haltiuik, & Smywiński-Pohl, 2024) існувала потреба також навчити з "нуля" україномовну модель, що, попри спроби, не вдалося зробити в межах поточного дослідження через брак обчислювальних ресурсів.

Відсутність комплексного бенчмарку. На відміну від англійської мови, для якої існує стандартизований набір бенчмарків GLUE, що охоплює різні задачі ОПМ, для української мови такого інструменту бракує. Це суттєво ускладнює порівняння різних моделей і підходів, оскільки доводиться використовувати окремі набори даних і метрики, які можуть бути не повністю репрезентативними для всіх аспектів мови.

Відсутність універсального стандарту оцінки робить порівняння різних моделей менш об'єктивним й ускладнює процес перевірки для кожного завдання.

Використання результатів конференції UNLP 2024. Конференція UNLP 2024, що відбулася 25 травня 2024 р., стала важливою подією для розвитку ОПМ в Україні. На конференції представили нові моделі, натреновані "з нуля" чи дотреновані на українських текстах (Haltiuik, & Smywiński-Pohl, 2024; Kiulian et al., 2024). Крім того, конференція відзначилася розширенням бенчмарків для оцінки моделей. У межах нашої роботи ми здійснили спробу інтегрувати деякі із цих результатів, проте обмежений час не дав змоги ґрунтовніше використати нові моделі та бенчмарки.

Майбутня робота та напрямки досліджень. Результати дослідження показали, що спеціалізовані україномовні моделі мають значні переваги в більшості завдань з обробки української мови. Однак з'явилася нагальна потреба в створенні комплексного україномовного бенчмарку, аналогічного до GLUE або KLEJ, для всебічної оцінки моделей за різними завданнями. Ця потреба стає ще актуальнішою з огляду на активний розвиток моделей, натренованих переважно на українських даних, та появу невеликих комплексних бенчмарків.

Основними напрямками майбутньої роботи є: розробка комплексного бенчмарку, який буде українським аналогом GLUE; розгляд завдань на глибоке розуміння української мови; розширення та диверсифікація набору даних, що використовуються для навчання й оцінки моделей. Це сприятиме підвищенню точності й адекватності моделей.

При розробці бенчмарку та моделей необхідно приділяти значну увагу етичним аспектам, таким як упередженість, дискримінація та дезінформація. Важливо забезпечити, щоб моделі були справедливими, прозорими та не завдавали шкоди суспільству. Це передбачає ретельний аналіз даних, алгоритмів і результатів роботи моделей, а також розробку механізмів контролю та корекції.

Загалом реалізація запропонованих напрямків досліджень має значний потенціал для розвитку української спільноти ОПМ та створення потужних інструментів для виконання різноманітних завдань у різних сферах життя.

Дискусія і висновки

В межах представленого дослідження вдалося порівняти результати одно- та багатомовних моделей типу BERT на окремих завданнях обробки української мови. Результати порівняння демонструють те, що моделі, натреновані "з нуля" на українських текстах або дотреновані на них, загалом мають кращі результати, ніж багатомовні. Це вдалося показати на завданнях розпізнавання іменованих сутностей, класифікації текстів та заповнення пропусків. Водночас багатомовні моделі показують наблизений або кращий результат на легших завданнях (розрізнення значень слів).

Українські моделі показали кращі можливості захоплювати український контекст і враховувати культурні нюанси, які можуть втрачатися в багатомовних моделях. Хоча модель LiBERTa натренована "з нуля" на українських наборах даних показала гірші результати, ніж просто перенесена на українську методом WECHSEL багатомовна модель – згідно з виступом авторів LiBERTa (Haltiuk, & Smywiński-Pohl, 2024), друга модель уже перевершує WECHSEL-RoBERTa. Це може позначати, що кількість нерозмічених якісних даних українською доходить до того моменту, коли тренувані лише на них моделі вже можуть показати кращі результати, ніж адаптовані багатомовні.

Дослідження показало, що вибір між одно- та багатомовними моделями значно залежить від завдання.

У роботі були зібрані та порівняні одно- та багатомовні моделі для різних мов, що додатково підкреслило важливість проведення окремого порівняння для української мови. Ефективність певного підходу залежить від мови, завдання та доступних ресурсів. Наприклад, одномовна фінська модель BERT перевершила багатомовну mBERT, але результати порівняння для португальської мови були кардинально протилежними.

Для порівняння одно- та багатомовних моделей були зібрані результати з різних джерел, серед яких і наукові статті та моделі. Деякі з використаних моделей були додатково налаштовані. Для завдань, де потрібно було заповнити пропуски, ми використовували готові попередньо тренувані моделі.

Більшість моделей, використаних у дослідженні, не досягають рівня комерційних моделей, таких як ChatGPT4 через обмеженість кількості параметрів моделі, обсягу даних для тренування та можливостей отримувати зворотний зв'язок щодо роботи моделі.

Процес порівняння моделей значно ускладнювала відсутність комплексного бенчмарку для української мови, який був би аналогом GLUE або KLEJ.

Результати пропонованого дослідження можуть допомогти у визначенні оптимальної моделі для завдань, що були представлені в порівнянні. Крім того, зібрані показники порівняння моделей і перелік завдань можуть слугувати основою для розробки комплексного бенчмарку для української мови. Загалом результати роботи також підкреслюють важливість розробки та вдосконалення одномовних моделей для української мови, що допоможе забезпечити їхню ефективність у широкому спектрі завдань обробки мови.

Внесок авторів: Микола Глибовець – концептуалізація, методика; Данило Ванін – програмне забезпечення; Микола Глибовець – аналіз і підготовка теоретичних основ дослідження; Данило Ванін – аналіз результатів, збір і перевірка емпіричних даних, аналіз джерел, підготовка огляду літератури.

Джерела фінансування. Це дослідження не отримало жодного гранту від фінансової установи в державному, комерційному або некомерційному секторах.

Список використаних джерел

- Глибовець, А. М., & Бікчентаєв, М. (2022). Програмна система перевірки на плагіат українських текстів. *Наукові записки НаУКМА. Комп'ютерні науки*, 5, 16–25. <https://doi.org/10.18523/2617-3808.2022.5.16-25>
- Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 23–29). ELRA and ICCL.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- GitHub. (2025). *Lang-Uk/ner-uk*. <https://github.com/lang-uk/ner-uk>
- Haltiuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch (Presentation). In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 120–128). ELRA and ICCL.
- Hamotskyi, S., Levbarg, A.-I., & Hänig, C. (2024). Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 109–119). ELRA and ICCL.
- Hugging Face. (2024). *Uk_ner_web_trf_13class*. https://huggingface.co/dchaplynsky/uk_ner_web_trf_13class
- Hugging Face. (2025). *Roberta-large-wechsel-ukrainian*. <https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Kiulian, A., Polishko, A., Khandoga, M., Chubych, O., Connor, J., Ravishankar, R., & Shirawalmath, A. (2024). *From bytes to borsch: Fine-tuning Gemma and Mistral for the Ukrainian language representation*. arXiv preprint arXiv:2404.09138. <https://doi.org/10.48550/arXiv.2404.09138>
- Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.unlp-1.2>

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692. <https://doi.org/10.48550/arXiv.1907.11692>
- Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 1–8). Association for Computational Linguistics. <https://aclanthology.org/E17-5001>
- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 1946–1958). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1178>
- Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., & Turuta, O. (2022). Ukrainian news corpus as text classification benchmark. In O. Ignatenko, V. Kharchenko, V. Kobets, H. Kravtsov, Y. Tarasich, V. Ermolayev, D. Esteban, V. Yakovyna, & A. Spivakovsky (Eds.), *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science* (Vol. 1635, pp. 550–559). Springer, Cham. https://doi.org/10.1007/978-3-031-14841-5_37
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1493>
- Piskorski, J., Marcinczuk, M., & Yangarber, R. (2024). Cross-lingual named entity corpus for Slavic languages. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4143–4157). ELRA and ICCL.
- Radchenko, V. (2020). *We trained the Ukrainian language model*. Youscan. <https://youscan.io/blog/ukrainian-language-model/>
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv:1908.10084. <https://doi.org/10.48550/arXiv.1908.10084>
- Ruder, S. (2022). *The state of multilingual AI*. [ruder.io](https://www.ruder.io/stateof-multilingual-ai/). <https://www.ruder.io/stateof-multilingual-ai/>
- Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1191–1201). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.111>
- Sanh, V., Wolf, T., Ly, J., & Rush, A. M. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108. <https://api.semanticscholar.org/CorpusID:203626972>
- Shaham, U., Aharoni, R., Rappoport, A., & Goldberg, Y. (2024). *Multilingual instruction tuning with just a pinch of multilinguality*. arXiv preprint arXiv:2401.01854. <https://doi.org/10.48550/arXiv.2401.01854>
- Sido, J., Švec, J., Pecina, P., & Konopík, M. (2021). *Czert: Czech BERT-like model for language representation*. arXiv preprint arXiv:2103.13031. <https://doi.org/10.48550/arXiv.2103.13031>
- Torge, S., Politov, A., Lehmann, C., Saffar, B., & Tao, Z. (2023). Named entity recognition for low-resource languages - Profiting from language families. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)* (pp. 1–10). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bsnlp-1.1>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish*. arXiv preprint arXiv:1912.07076. <https://doi.org/10.48550/arXiv.1912.07076>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019, May). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019) (poster session)* (pp. 353–355). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-5446>
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *SuperGLUE: A stickier benchmark for general-purpose language understanding systems*. arXiv preprint arXiv:1905.00537. <https://doi.org/10.48550/arXiv.1905.00537>

References

- Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 23–29). ELRA and ICCL.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- GitHub. (2025). *Lang-Uk/ner-uk*. <https://github.com/lang-uk/ner-uk>
- Haltuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch (Presentation). In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 120–128). ELRA and ICCL.
- Hamotskyi, S., Levbarg, A.-I., & Hänig, C. (2024). Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 109–119). ELRA and ICCL.
- Hlybovets, A. M., & Bikchentaiev, M. (2022). Software system for checking plagiarism of Ukrainian texts [Prohramna systema perevirky na plahiat ukrainskykh tekstiv]. *Naukovi zapysky NaUKMA. Kompiuterni nauky*, 5, 16–25 [in Ukrainian]. <https://doi.org/10.18523/2617-3808.2022.5.16-25>
- Hugging Face. (2024). *Uk_ner_web_trf_13class*. https://huggingface.co/dchaplynsky/uk_ner_web_trf_13class
- Hugging Face. (2025). *Roberta-large-wechsel-ukrainian*. <https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Kiulian, A., Polishko, A., Khandoga, M., Chubych, O., Connor, J., Ravishankar, R., & Shirawalmath, A. (2024). *From bytes to borsch: Fine-tuning Gemma and Mistral for the Ukrainian language representation*. arXiv preprint arXiv:2404.09138. <https://doi.org/10.48550/arXiv.2404.09138>
- Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.unlp-1.2>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692. <https://doi.org/10.48550/arXiv.1907.11692>
- Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 1–8). Association for Computational Linguistics. <https://aclanthology.org/E17-5001>

- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 1946–1958). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1178>
- Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., & Turuta, O. (2022). Ukrainian news corpus as text classification benchmark. In O. Ignatenko, V. Kharchenko, V. Kobets, H. Kravtsov, Y. Tarasich, V. Ermolayev, D. Esteban, V. Yakovyna, & A. Spivakovsky (Eds.), *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science* (Vol. 1635, pp. 550–559). Springer, Cham. https://doi.org/10.1007/978-3-031-14841-5_37
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1493>
- Piskorski, J., Marcinczuk, M., & Yangarber, R. (2024). Cross-lingual named entity corpus for Slavic languages. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4143–4157). ELRA and ICCL.
- Radchenko, V. (2020). *We trained the Ukrainian language model*. Youscan. <https://youscan.io/blog/ukrainian-language-model/>
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv:1908.10084. <https://doi.org/10.48550/arXiv.1908.10084>
- Ruder, S. (2022). *The state of multilingual AI*. [ruder.io](https://www.ruder.io/stateof-multilingual-ai/). <https://www.ruder.io/stateof-multilingual-ai/>
- Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1191–1201). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.111>
- Sanh, V., Wolf, T., Ly, J., & Rush, A. M. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108. <https://api.semanticscholar.org/CorpusID:203626972>
- Shaham, U., Aharoni, R., Rappoport, A., & Goldberg, Y. (2024). *Multilingual instruction tuning with just a pinch of multilinguality*. arXiv preprint arXiv:2401.01854. <https://doi.org/10.48550/arXiv.2401.01854>
- Sido, J., Švec, J., Pecina, P., & Konopík, M. (2021). *Czert: Czech BERT-like model for language representation*. arXiv preprint arXiv:2103.13031. <https://doi.org/10.48550/arXiv.2103.13031>
- Torge, S., Politov, A., Lehmann, C., Saffar, B., & Tao, Z. (2023). Named entity recognition for low-resource languages – Profiting from language families. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)* (pp. 1–10), Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bsnlp-1.1>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish*. arXiv preprint arXiv:1912.07076. <https://doi.org/10.48550/arXiv.1912.07076>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019, May). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019) (poster session)* (pp. 353–355). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-5446>
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *SuperGLUE: A stickier benchmark for general-purpose language understanding systems*. arXiv preprint arXiv:1905.00537. <https://doi.org/10.48550/arXiv.1905.00537>

Отримано редакцією журналу / Received: 19.08.25

Прорецензовано / Revised: 24.01.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Mykola GLYBOVETS, DSc (Phys. & Math.), Prof.
 ORCID ID: 0009-0005-6942-8026
 e-mail: glib@ukma.edu.ua
 National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Danylo VANIN, Master of Arts
 ORCID ID: 0009-0007-4510-8209
 e-mail: danylo.vanin@gmail.com
 National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

THE APPLICATION OF MONOLINGUAL AND MULTILINGUAL BERT-BASED MODELS FOR TEXT AUTOMATION TASKS IN UKRAINIAN

The object of this study is mono- and multilingual models based on the BERT architecture. The research focuses on comparing their performance in natural language processing (NLP) tasks, with an emphasis on their application to the Ukrainian language. The methodological framework follows standard approaches to model training and evaluation, relying on publicly available information sources. Experiments were conducted solely on open-source or self-fine-tuned models and covered all existing Ukrainian benchmarks.

Overall, the findings show that both monolingual and multilingual BERT models can be effective for NLP tasks, depending on the target language, specific task, and available resources. While monolingual models often outperform multilingual ones for tasks in their own language, multilingual models can prove advantageous when resources for training language-specific models are limited.

The comparison demonstrates that models trained from scratch on Ukrainian texts—or further fine-tuned on them—generally achieve higher results on Ukrainian NLP tasks than multilingual counterparts. This advantage was evident in named-entity recognition, text classification, and masked-token prediction. Multilingual models, however, achieve competitive or superior performance in "lighter" tasks such as word-sense disambiguation, especially when full training of a monolingual model is not feasible.

These findings underscore the need to adapt or fully train Ukrainian-language models for specialized tasks while still allowing for the cost-effective use of multilingual systems in everyday scenarios. The practical significance of this work lies in providing a consolidated performance map of modern models and laying the groundwork for a full-scale Ukrainian GLUE-like benchmark, which will stimulate further research and improve the accuracy of Ukrainian NLP applications.

Keywords: natural language processing, large language models, monolingual and multilingual models, BERT.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

УДК 519.83

DOI: <https://doi.org/10.17721/1812-5409.2026/1.27>

Serhii DOTSENKO¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0003-2913-3790
e-mail: sergei204@knu.ua

Georgi DIMITROV², PhD, Prof.
ORCID ID: 0000-0002-4785-0702
e-mail: g.dimitrov@unibit.bg

Anastasiya VECHERKOVSKAYA¹, PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0003-2913-3790
e-mail: a.vecherkovska@knu.ua

Igor MAKUSHENKO¹, PhD (Phys. & Math.)
ORCID ID: 0000-0002-4313-4936
e-mail: igormakushenko@knu.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²University of Library Studies and Information Technologies, Sofia, Bulgaria

ALGORITHMIC APPROACHES TO MODELING AND ANALYZING ENDGAME POSITIONS IN BACKGAMMON

The final phase of the backgammon game, known as "races", is considered. This phase is characterized by minimizing the interaction of opponent's checkers and reducing the game process mainly to assessing the rate of moving checkers to the "home" and their subsequent withdrawal from the board. The delivered analysis covers the specifics of this stage in various backgammon variants, in particular in classic backgammon, using mathematical models and probability theory. The paper proposes methods for optimizing strategies in the final phase of the game, which are based on a comprehensive analysis of possible game positions and calculating the probabilities of the dices values. Special attention is paid to algorithmic approaches to decision-making, which allow calculating the most effective game actions to achieve the optimal result.

The results of the study are valuable both for improving the game strategies of practicing players and beginners, and for developing theoretical approaches to game analysis based on modeling and assessing the effectiveness of decision-making in conditions of limited interaction between participants.

Key words: backgammon, "races" stage, probability theory, total probability formula, simulation, optimal strategies, doubling cube, pip-count, game optimization.

AMS 2020 classification: 62H30, 34C60.

Introduction

Backgammon is one of the oldest board games, gaining popularity both in Europe and the Middle East. Despite its centuries-old history, this game remains relevant due to the variety of its variations, among which the most common are long nardy, backgammon, tavli, etc. A special place among them is occupied by backgammon, which is the subject of numerous studies in the field of game theory and probability theory, and also serves as the basis for holding international championships and online tournaments. The main strategy of the game is to move checkers to the "home", or final area, with their subsequent withdrawal from the board. However, the most intensive mathematical and algorithmic analysis concerns the final stage of the game, which is known as "races" or "speed game". In this phase, the opponent's checkers cease to interact, and success is determined by the player's ability to bring checkers correctly into the house for their subsequent withdrawal as quickly as possible.

The complexity of analyzing strategies in backgammon has repeatedly attracted the attention of researchers. In particular, in the work (Witter, 2021) it was proven that the problem of determining the possibility of winning from an arbitrary position belongs to the class of computationally complex ones, which explains the relevance of the search for approximate methods. A significant contribution to the understanding of the stochastic aspects of the game was made by models of optimal strategies for doubling bets, presented in the work (Blath, & Morters, 2001), which showed the effectiveness of decision-making under uncertainty. A separate area of research is devoted to the "races" phase: in the article (Ross et al., 2007) mathematical models for estimating the probabilities of winning were built and compared with the results of simulations. The work of Tesauro played a key role in the development of the algorithmic approach, who used temporal difference learning methods and neural networks to evaluate positions and improve strategies in the TD-Gammon project (Tesauro, 1995). In addition, an analysis of the effectiveness of search algorithms, such as minimax and its variations, was presented in the works (Bell, 1997; Olsen, 2019), which confirms the promising application of optimization techniques in strategy simulation.

Thus, the analysis of "races", as a key phase of the backgammon game, has not only mathematical but also applied significance, as it allows to construct highly efficient strategies. This paper examines in detail the methods of optimizing strategies for this stage of the game with an emphasis on the use of modern analytical tools based on probability theory and decision-making algorithms.

1. Mathematical model. The main concept

Pip-count (PC). Let the fields of the board be numbered from 1 to 24, so that the nearest field in the house is numbered 1, etc., and the field farthest from the house is numbered 24. Then the sum of the products of the number of checkers by the numbers of the fields they occupy is called the pip-count or score. The opponent, to calculate his score, numbers the fields in reverse order. Thus, in backgammon, before the start of the game, each player's score equals to

$$2 \cdot 24 + 5 \cdot 13 + 3 \cdot 8 + 5 \cdot 6 = 167.$$

In fact, the score is the minimum number of points that a player must roll (over a number of moves) to bring the checkers into the house and then to withdraw them. It is said that a player who has a lower/higher score than the opponent is ahead/behind in speed, respectively. The skill of quickly calculating the score is an important component of a player's skill, the technique of such calculations can be found in (Kissane, 1992).

The roll. Each player's move is preceded by the roll of two dice. In this case, if the number of points on the dice matches (for example, 3-3), then it is said that a "double" has been rolled, and the number of points rolled is doubled (i. e. in this case the player has not two, but four moves of three points). The probability distribution of the number of points in one roll, taking into account doubles, is given in Tab. 1.

Table 1

Probability distribution of the number of points in one roll (two dice)													
j	3	4	5	6	7	8	9	10	11	12	16	20	24
π_j	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{4}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

This probability distribution has the following characteristics:

$$M(j) = \mu = \frac{49}{6} \approx 8.167, \sigma(j) = \sigma = \frac{\sqrt{665}}{6} \approx 4.297.$$

Doubling cube. In backgammon there is a rule of doubling bets. Let the game be played at a single bet at start. A player who has sufficient (in his subjective opinion) advantages in the position can say "double" before his roll. In this case, his opponent must make one of two decisions - to surrender without further play (and pay a single bet) or to continue the game at a doubled bet. There can be several doublings during the game. A double cube is a cube with degrees of 2 on its sides - from 2 to 64 and which is an indicator of the doubling status. At the beginning, when the game is played at a single bet, the cube is in the middle and this situation is called "our cube". If the player has accepted the offer to double, then the cube moves closer to him on the board, and the upper face shows the current (doubled) bet value. In this case, he is called the "owner of the cube", and the situation that has arisen is "my cube" (for the one who accepted the doubling) and "his cube" (for the opponent), and only the "owner of the cube" has the right to further double.

2. Optimal doubling strategies

It turns out that the correct managing of the double cube (i. e., an adequate evaluation of situations in which the player needs to double the bet himself, and either to accept to surrender as opponent doubles) is no less important than the ability to make the right moves. This problem was first considered in (Keeler, & Spencer, 1975). Let $W(t)$ be a standard Wiener process and $W(0) = 0.5$, the observation of the process ends at time t_0 , when the value of the Wiener process trajectory takes the value 1 (then it is considered that the 1st player won) or 0 (then the 2nd player won). Then at each time t the conditional probability of the 1st player winning is equal to the value $W(t)$, and the conditional probability of the 2nd player winning is equal to the additional probability. For this situation, it is proved that the 1st player should double at the moment when $W(t) = 0.8$. The second player should accept the double if $W(t) < 0.8$ and surrender if $W(t) > 0.8$. If the first player doubled at $W(t) = 0.8$, then the second player is indifferent to whether to accept the double or surrender, in any case his average loss is equal to the current bet. Due to the symmetry of the model, the critical doubling point for the second player is obviously 0.2. But at the end of the game, situations may arise that are not adequate to the continuous game model. The situations where this inadequacy is most evident are the so-called one-step and two-step games considered in (Tuck, 1980).

In subsequent works, it was shown that the "continuous game" model, although it plays a significant role in estimating the threshold values of doubling and doubling acceptance, is not fully adequate, since the players' chances can change dramatically during a single roll (Tuck, 1980).

One step game. Let the first player after his throw win with probability p and lose with additional probability. Such a model is fully adequate to the real game situation in which the 2nd player has one or two checkers, which he will definitely throw away if it comes to his turn, i. e. one checker in the cell 1,2 or 3 or two checkers 1-1 or 1-2. It is worth doubling for the 1st player when his expected gain at doubling is greater than the expected gain if he does not double, i. e. $4p - 2 >> 2p - 1$. Then the first player in the situations "my cube" and "our cube" should double if $p > 0.5$. In this case, the 2nd player should surrender if, having accepted the doubling, he loses more than by surrendering, i. e. $4p - 2 > 1$ or $p > 0.75$.

Two steps game. Let the first player after his roll win with probability p and with additional probability the move goes to the second player. If the move has reached the 2nd player, he also makes a roll, after which he wins with (conditional) probability q , otherwise the 1st player wins. Then the following result occurs. Consider the square $[0; 1] \times [0; 1]$, where the values of p and q are plotted along the abscissa/ordinate axes, respectively. Let the square be divided into 4 regions by such curves.

$$P_D(q) = \begin{cases} 0.5, q \geq 0.75; \\ \frac{4q-2}{4q-1}, 0.5 \leq q \leq 0.75; \\ 0, q \leq 0.5. \end{cases} \quad (1)$$

$$P_R(q) = \begin{cases} \frac{3-2q}{4-2q}, q \geq 0.75; \\ \frac{6q-3}{6q-2}, 0.5 \leq q \leq 0.75; \\ 0, q \leq 0.5. \end{cases} \quad (2)$$

$$P_Q(q) = \begin{cases} 0.75, q \geq 0.75; \\ \frac{8q-3}{8q-2}, 0.5 \leq q \leq 0.75; \\ 1 - \frac{1}{4q}, 0.25 \leq q \leq 0.5; \\ 0, q \leq 0.25. \end{cases} \quad (3)$$

Then the correct behavior of the 1st player will be as follows. In the region "left and above" $P_D(q)$ it should not be double in any case. In the region enclosed between $P_D(q)$ and $P_R(q)$, it should be double in the situation "our cube" and not double in the situation "my cube". In the region "right and below" $P_R(q)$ it should be doubled in any case (see Fig. 1).

The correct behavior of the 2nd player is to accept doubling if the point (p, q) lies "to the left and above" and to surrender if "to the right and below" the curve $P_Q(q)$. In more detail, how the inequalities, with respect to expected winnings, give rise to the above-described curves $P_D(q)$, $P_R(q)$, $P_Q(q)$, was considered in (Tuck, 1980).

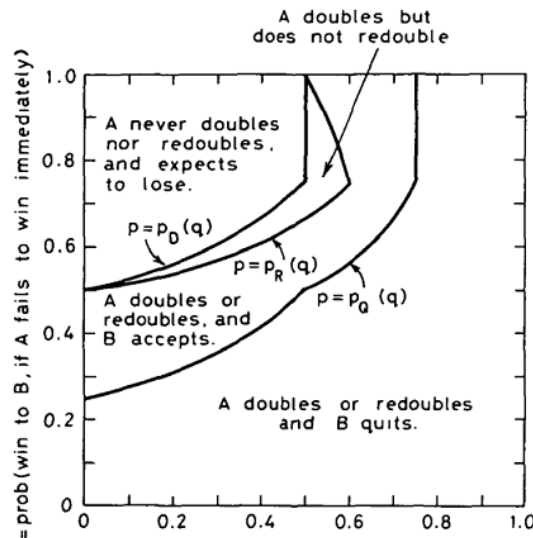


Fig. 1. Diagram of reasonable doubling and accepting doubling zones

Applying the results to backgammon gives the following practical rules. In a one-step game, player 1 should double if he has one checker left in any position or two checkers in positions 2 – 5 or better. Player 2 should surrender if player 1 has one checker left in any position or two checkers in positions 1 – 4 or better. Otherwise, player 2 should accept the double. For example, in a one-step game in position 2 – 5, although player 1's probability of winning is $\frac{19}{36} \approx 0.528$, player 1 should double, and player 2 should accept the double. This "increased aggressiveness" of player 1 is explained only by the fact that his chances of winning, although only slightly exceed $\frac{1}{2}$, are in a "double now or never" situation, i. e. he will not have a chance to wait for a more favorable situation, as in the continuous game model (Keeler, & Spencer, 1975).

3. Single checker model

In (Ross et al., 2007) the "one-checker model" in the tempo game was considered. This means the following. Let the players' checkers not interfere with each other, the players have scores S_A and S_B , respectively, and after the roll, each player subtracts the number of points that were rolled from his score. The player who first withdraw all his points wins. Such a model would be adequate if the players had one checker each, and these checkers moved to their homes on a "large board". But, as noted above, the score is the minimum number of points that a player must throw to remove the checkers from the game. In reality, it may be necessary to throw more points. For example, let a player have two checkers in positions 2 and 3 (totally 5 points), and throws 1 – 6 (7 points). Although the number of points thrown out formally exceeds the score, it is not enough to throw both checkers off the board.

As a result, one checker remains on the board, which will be withdrawn on the next move. If the player has a large number of checkers in positions 1 and 2, this will also lead to the fact that the real amount of points that will be needed to withdraw all the checkers will exceed the score, since in one throw you can remove at best 2 or 4 checkers (if a double did not fall out/fell out, respectively). For example, let there be 3 checkers left on 1 and 2 on 2 (i. e. the total score is 7). Then it is not important for the player whether large numbers will be present in the throw. For example, let 6 – 5 be thrown out, but only 2 checkers can be removed. The next throw 5 – 4 can remove 2 more checkers. Then 5 – 3 the last checker is removed. So for throwing 5 checkers with a score of 7, throwing $11 + 9 = 20$ points turned out to be not enough. The thrown points are obviously used inefficiently. Another factor that leads to inefficient use of discarded points at the stage of discarding checkers is "holes in positions". For example, let a player have 5 checkers in positions 5, 3 and 1: formally the score is 45, but there is a hole in position 4 and, having discarded 4, the player can only move checkers from 5 to 1 (see Fig. 2).

A clear explanation of the inefficiency of using the points on the rolled dices can be the following situation. Suppose you have a sum of money in notes of different denominations. Suppose you periodically need to buy goods for different amounts, but the seller never gives change. If at first you may have no problems finding the required amount without change, then later you have to overpay, and the actual amount of money spent exceeds the total cost of the purchased goods.

summed up until a sum equal to T is reached, then the number of terms has an asymptotically normal distribution with mean $\frac{T}{\mu}$ and variance $\frac{\sigma^2 T}{\mu^3}$.

Thus, to withdraw k points, the following number of throws will be needed:

$$n_k \sim N\left(\frac{k}{\mu}, \frac{\sigma^2 k}{\mu^3}\right),$$

$$F(k, s) = P(n_s - n_k \geq 0) \sim \Phi\left(\frac{s - k + \mu/2}{\sqrt{s + k}} \cdot \frac{\sqrt{\mu}}{\sigma}\right). \quad (6)$$

However, even at the assumption, that a player has phenomenal computational abilities and can calculate the value of $F(k, s)$ using one of the formulas (or study Table 1 and perform linear interpolation), the main question still remains open – to double or not to double?

4. Tips for doubling

Instead of calculating the probabilities of winning, the authors calculated the "advice to double" and "advice to accept doubling" matrices using dynamic programming and presented the data in a form convenient for practical use.

In (Buro, 1999) the concept of "equity" (E) is considered, which means a dimensionless quantity equal to the ratio of the expectation of the win to the current bet, which varies in the range from 1 (if the player is sure to win) to -1 (if he loses). Let us assume that both players behave rationally, that is, they double their bets, fold, or accept the doubling in such a way as to maximize the mathematical expectation of their winning.

Let's take a look at these functions: $E_{my}^0(k, s), E_{my}^1(k, s), E_{my}(k, s), E_{our}^0(k, s), E_{our}^1(k, s), E_{our}(k, s)$.

Here the subscript my /our corresponds to the situation my cube/our cube respectively. The superscript is 0 if the player did not double, and 1 if he doubled in this situation (assuming that in all subsequent situations the players will be rational). Let

$$E_{my}(k, s) = \max(E_{my}^0(k, s), E_{my}^1(k, s)), \quad (7)$$

$$E_{our}(k, s) = \max(E_{our}^0(k, s), E_{our}^1(k, s)) \quad (8)$$

rational choice of the player at this step. For all introduced functions, we define initial values corresponding to the final (trivial) game situations:

$$E(0, j) = 1, \quad (9)$$

$$E(i, 0) = -1, \quad (10)$$

$$E(1, j) = E(2, j) = E(3, j) = 1, \quad (11)$$

where $i, j > 0$.

$$E^*(i, j) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ -1 & 1 & \dots & 1 \\ -1 & 1 & \dots & 1 \\ -1 & \text{elements} & \text{that} & \text{are} \\ & \text{not} & \text{calculated} & \text{yet} \end{pmatrix}$$

First, let's calculate the elements of the matrices E_{my}^* , then E_{our}^* .

$$E_{my}^0(k, s) = \sum_{i=3}^{24} \sum_{j=3}^{24} \pi_i \pi_j E_{my}((k-i)_+, (s-j)_+), \quad (12)$$

$$E_{my}^1(k, s) = \min\left(1, -2 \sum_{i=3}^{24} \pi_i E_{my}(s, (k-i)_+)\right), \quad (13)$$

$$E_{our}^0(k, s) = -\sum_{j=3}^{24} \pi_j E_{our}(s, (k-j)_+), \quad (14)$$

$$E_{our}^1(k, s) = E_{my}^1(k, s). \quad (15)$$

These formulas may be explained as follows. Consider (12). Let in the situation (k, s) the 1st player not double. Then the first player withdraw some (random, according to his roll) number of points i from his k . The second player has no opportunity to double, he makes a throw and withdraw (random, according to his roll) number j , and so, the score becomes equal to $((s-i)_+, (k-j)_+)$, and the move again passes to the first player. By virtue of the assumption of optimality of all subsequent steps E becomes equal to $E_{my}((k-i)_+, (s-j)_+)$, averaged over the probabilities of possible throws of both players. If the 1st or 2nd arguments of the function become zero, this means that the 1st or 2nd player wins. If both arguments become zero, this means that the 1st player wins, since it was previously determined that, although formally, after the 1st argument becomes zero, the 2nd player makes a "useless" throw.

Consider (13). Let in the situation (k, s) "my cube" the 1st player doubled. If the 2nd player surrenders, then $E = 1$. If the 2nd player accepts the double, then the 1st player withdraw (random, according to his roll) points, the bet will double, the move and the right to double will pass to the 2nd player. Thus, for the 2nd player the situation will be $(s, (k-i)_+)$, my cube, i. e. E will be equal to $E_{my}((k-i)_+, (s-j)_+)$. Averaging over all possible throws and taking into account that the values of E are always equal in magnitude and opposite in sign for the opponents, we will obtain for the 1st player $E = -2E_{my}(s, (k-i)_+)$.

Based on the assumption of rationality of the players, the 2nd player will choose the one of the two actions (surrender, accept) that corresponds to the minimum of the two terms (13).

Consider (14). Let in the situation (k, s) and "our cube" the 1st player did not double. Then similarly to (12) the 1st player withdraw some (random, according to his roll) number of points i from his k , while the 2nd player gets into the situation, "our cube". Therefore, for the first player the value of E is taken with the opposite sign.

Consider (15). The consequences of doubling in the situations "my cube" and "our cube" are the same, all considerations are similar to (13), so $E_{our}^1(k, s) = E_{my}^1(k, s)$, so both matrices may be denoted as $E^1(k, s)$.

Based on the calculated matrices E , we construct matrices of advice regarding doubling $Db_{my}(k, s)$ та $Db_{our}(k, s)$ and accepting doubling $Ac(k, s)$. If in this situation it is advisable to double, then $Db_*(k, s) = 1$, if it is advisable not to double, then $Db_*(k, s) = 0$. In a similar way, $Ac(k, s) = 1$ or 0, depending on accepting double is advisable or not. Then

$$Db_{my}(k, s) = \chi(E^1(k, s) \geq E_{my}^0(k, s)), \quad (16)$$

$$Db_{our}(k, s) = \chi(E^1(k, s) \geq E_{our}^0(k, s)).$$

The sign of the non-strict inequality in these formulas is explained as follows. Let the 2nd player be hopelessly behind, for example, $k = 10, s = 100$ (in this case they say "there is no theory"). Then, obviously, all values in (16) are equal to each other and equal to 1. Obviously, in this position one should double and this is the end the game, the outcome of which is predetermined. If the inequality were strict, then the game would continue until the very end. When calculating the matrix $Ac(k, s)$ one should take into account that the optimal behavior of the 2nd player has already been taken into account when calculating $E^1(k, s)$, therefore

$$Ac(k, s) = \chi(E^1(k, s) < 1). \quad (17)$$

It turns out that in each row of the matrices $Db_{my}(k, s), Db_{our}(k, s)$ there are zeros at the beginning, and starting from some number - ones. Accordingly, in $Ac(k, s)$ there are first ones, then zeros. Let us construct the vectors $db_{my}(k)$ and $db_{our}(k)$, in which the k -th component is equal to the index of the first one in the row of the matrices $Db_{my}(k, s)$ and $Db_{our}(k, s)$ respectively. Let us also set the k -th component of the vector $ac(k)$ equal to the last one in the k -th row of the matrix $Ac(k, s)$. Then let

$$\begin{aligned} db_{my}^*(k) &= db_{my}(k) - k, \\ db_{our}^*(k) &= db_{our}(k) - k, \\ ac^*(k) &= ac(k) - k \end{aligned}$$

vectors characterizing the minimum advantage for doubling and the maximum possible lag for accepting it. It turns out that the components of these vectors form non-decreasing numerical sequences that can be compactly presented in the form of a table (see Tab. 3). It makes no sense to put the data for $k < 22$ into the table, since in this case the required number of throws depends strongly on the position

Table 3

Doubling Tips Table					
Minimal advantage for doubling at our cube		Minimal advantage for doubling at my cube		Maximum lag for accepting a doubling	
k	$s - k$	k	$s - k$	k	$s - k$
22-27	0	23-27	1	23-26	3
28-32	1	28-33	2	27-31	4
33-38	2	34-38	3	32-37	5
39-45	3	39-44	4	38-44	6
46-53	4	45-51	5	45-51	7
54-62	5	52-60	6	52-59	8
63-72	6	61-69	7	60-68	9
73-82	7	70-79	8	69-78	10
83-93	8	80-89	9	79-88	11
94-105	9	90-100	10	89-98	12
106-117	10	101-111	11	99-110	13
118-129	11	112-124	12	111-122	14
130-143	12	125-136	13	123-134	15
144-157	13	137-150	14	135-148	16
158-170	14	151-163	15	149-161	17
		164-170	16	162-170	18

Let us consider the analysis of another possible scenario of the end of the game using dynamic programming. Let opponents A and B have m and n checkers respectively in position 1. With each move, the players discard 2 or 4 checkers with probabilities $5/6$ and $1/6$ respectively (if a double doesn't rolled or rolled respectively). Let us consider the following equivalent model.

Let $k = \left\lceil \frac{m+1}{2} \right\rceil, s = \left\lceil \frac{n+1}{2} \right\rceil$ be conditional points, so that one or two conditional points can be deducted (if double is rolled on dices or not respectively) for one roll. Since $m, n \leq 15$, then $k, s \leq 8$. Next, similarly to the previous case, the matrices E, Db, Ac are calculated, but now $j = 1$ or 2 with probabilities $5/6$ та $1/6$ respectively.

The results of the calculations can be formulated as easy-to-remember advice: in the "my cube" situation, double, if $k < s$ або $k = s \leq 4$, in the "our cube" situation, if $k < s$ or $k = s \leq 8$. However, the advice to double in situations (7,7) and (8,8) gives a small benefit on average, but increases the risk, since

$$E_{our}^1(8,8) - E_{our}^0(8,8) = 1.6 \cdot 10^{-3}, E_{our}^1(7,7) - E_{our}^0(7,7) = 7.2 \cdot 10^{-3}.$$

So whether to double or not in this situation is a matter of risk appetite. It should also be noted that for these results to apply, it is not necessary for all checkers to be on 1. For example, the position "1 checker on 4, 3 on 3, 6 on 2 and 5 on 1" is not significantly worse than 15 on 1", so all the above considerations can be applied to it.

Discussion and conclusions

This paper provides a comprehensive analysis of the final stage of backgammon game ("races"), a phase characterized by minimal interaction between players' checkers and the outcome depending first of all on the speed of the checkers' movement. During the study:

1. A mathematical model of the game process is proposed to assess the effectiveness of decisions, based on the pip-count indicator, which allows quantitatively assessing the player's advantage.
2. Optimal strategies for using the double cube are created based on probabilistic calculations and dynamic programming principles. It is shown that the critical points of doubling and accepting doubling depend on the probabilities of winning, modeled as Wiener processes.
3. Practical algorithms for evaluating game positions, in particular for one-step and two-step games, have been found. The results confirm the importance of accurate probability calculations in decision-making.
4. Recommendations for doubling down and accepting doubling down in various game situations have been developed, allowing players to maximize their advantage in the game.

The obtained results demonstrated the possibility of increased accuracy of modeling of "races" phases by reducing the state space through the localization of critical moments and the use of the central limit theorem. The practical consequences of the work are the simplification of strategic decision-making in competitive conditions and the creation of a basis for improving training methods for players.

The results of the study confirmed the high importance of using probabilistic and mathematical models for effective evaluation of positions in the "races" phase. The proposed pip-count calculation method and its adaptation to backgammon allow for accurate prediction of the winning chances. Such calculations are especially relevant for sophisticated positions.

Future research could be aimed at expanding the model to adapt to individual playing styles of opponents and take into account psychological aspects of decision-making. Another promising direction is the integration of the developed methods into artificial intelligence to automate learning and analysis of game data.

Authors' contribution: Serhii Dotsenko – conceptualization, methodology; Georgi Dimitrov – methodology, theoretical foundations of the study; Anastasiya Vecherkovskaya – methodology, source analysis, preparation of a literature review; Igor Makushenko – methodology, data collection and verification, empirical research.

Sources of funding. This study was conducted at the author's own expense and partially funded by Taras Shevchenko National University of Kyiv.

References

- Bell, P. (1997). *Backgammon: Winning with the doubling cube*. The Gammon Press.
- Blath, J., & Mörters, P. (2001). *Backgammon, doubling the stakes, and Brownian motion*. <https://plus.maths.org/content/backgammon-doubling-stakes-and-brownian-motion>
- Buro, M. (1999). Efficient approximation of backgammon race equities. *International Computer Chess Association Journal*, 22(3), 133–142. <https://doi.org/10.3233/icg-1999-22303>
- Keeler, E. B., & Spencer, J. (1975). Optimal doubling in backgammon. *Operations Research*, 23(6), 1063–1071. <https://doi.org/10.1287/opre.23.6.1063>
- Kissane, J. (1992). *Cluster counting*. <https://bkgm.com/articles/McCool/cluster.html>
- Lamford, P. (2002). *Improve your backgammon*. Gloucesters Publishers plc.
- Olsen, M. (2019). *Cube like a boss – Patterns, intuition & strategy*. Backgammon Galaxy.
- Ross, A., Benjamin, A., & Munson, M. (2007). Estimating winning probabilities in backgammon races. In S. Ethier, & W. Eadington (Eds), *Optimal Play: Mathematical Studies of Games and Gambling* (pp. 269–291). Reno: Institute for the Study of Gambling and Commercial Gaming.
- Tesauro, G. (1995). TD-Gammon, a self-teaching backgammon program. In Alan F. Murray (Ed.), *Applications of Neural Networks* (pp. 267–285). https://doi.org/10.1007/978-1-4757-2379-3_11
- Tuck, E. O. (1980). Doubling strategies for backgammon-like games. *Journal of the Australian Mathematical Society Series B*, 21(4), 440–451. <https://doi.org/10.1017/s0334270000002137>
- Witter, R. T. (2021). Backgammon Is Hard. In D. Z. Du, D. Du, C. Wu, & D. Xu (Eds), *Lecture Notes in Computer Science: Vol. 13135. Combinatorial Optimization and Applications* (pp. 484–496). Springer. https://doi.org/10.1007/978-3-030-92681-6_38
- Zadeh, N., & Kobliska, G. (1977). On optimal doubling in backgammon. *Management Science*, 23(8), 853–858. <https://doi.org/10.1287/mnsc.23.8.853>

Отримано редакцією журналу / Received: 29.10.25
Прорецензовано / Revised: 17.03.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Сергій ДОЦЕНКО¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0003-2913-3790
e-mail: sergei204@knu.ua

Георгій ДІМИТРОВ², д-р філософії
ORCID ID: 0000-0002-4785-0702
e-mail: g.dimitrov@unibit.bg

Анастасія ВЕЧЕРКОВСЬКА¹, канд. тех. наук, доц.
ORCID ID: 0000-0003-2054-2715
e-mail: a.vecherkovska@knu.ua

Ігор МАКУШЕНКО¹, канд. фіз.-мат. наук
ORCID ID: 0000-0002-4313-4936
e-mail: igormakushenko@knu.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Університет бібліотекознавства та інформаційних технологій, Софія, Болгарія

АЛГОРИТМІЧНІ ПІДХОДИ ДО МОДЕЛЮВАННЯ ТА АНАЛІЗУ ФІНАЛЬНИХ ПОЗИЦІЙ У ГРІ В НАРДИ

Розглянуто кінцеву фазу гри в нарди, відому як "races" або "гра за швидкістю", що характеризується мінімізацією взаємодії шашок супротивників та зведенням ігрового процесу переважно до оцінки швидкості переміщення власних шашок у "дім" і їхнього подальшого виведення з дошки. Представлений аналіз описує специфіку цієї стадії в різних варіантах нрд, зокрема й у класичних коротких нардах (backgammon), із застосуванням математичних моделей і теорії ймовірностей. Запропоновано методи оптимізації стратегій у кінцевій фазі гри, що базуються на комплексному аналізі можливих ігрових траєкторій та розрахунку ймовірностей випадіння значень гральних кубиків. Особливу увагу приділено алгоритмічним підходам до ухвалення рішень, що дають змогу обчислювати найефективніші ігрові дії для досягнення оптимального результату.

Результати дослідження є цінними як для вдосконалення ігрових стратегій гравців-практиків і новачків, так і для розвитку теоретичних підходів до аналізу ігор на основі моделювання й оцінки ефективності прийняття рішень в умовах обмеженої взаємодії між учасниками.

Ключові слова: нарди (backgammon), фаза "races", короткі нарди, теорія ймовірностей, математичне моделювання, оптимальні стратегії, алгоритмічний аналіз, дабл-куб (doubling cube), pip-count, ігрова оптимізація.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 004.932.2:004.89

DOI: <https://doi.org/10.17721/1812-5409.2026/1.28>

Volodymyr KUBYTSKYI, PhD Student

ORCID ID: 0000-0002-1529-8677

e-mail: vova.kubytskyi@gmail.com

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Artem BOZHOK, PhD Student

ORCID ID: 0009-0009-9572-6501

e-mail: artembozh@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

INTERMEDIATE LAYER CONFIGURATIONS FOR IMAGE SIMILARITY: EVALUATING MULTI-LEVEL FEATURE EXTRACTION ACROSS CNN AND TRANSFORMER ARCHITECTURES

Evaluating visual similarity between images requires vector representations that preserve information across multiple levels of abstraction. Multi-level feature aggregation from intermediate CNN layers has shown good results in near-duplicate detection tasks, yet no structured analysis exists of how architectural properties of different backbones affect the quality of such multi-level representations. This study addresses the question of which structural characteristics make an architecture suitable for intermediate layer aggregation.

We formalize the multi-level aggregation pipeline as Global Average Pooling applied to intermediate activations, followed by concatenation and L2-normalization, producing a fixed-length descriptor. A compact MLP classifier is trained on the resulting pairwise representations. We evaluate six backbone architectures – VGG-16, Inception v3, DenseNet-121, MobileNetV2, ResNet-50, and ViT-B/16 – analyzing their stage structure, spatial resolution progression, intermediate feature dimensionality, and information flow properties. Quantitative evaluation is performed on the INRIA Holidays dataset using F1-score for pairwise similarity classification. ResNet-50 with layers C2+C3+C5 achieves $F1 = 0.77$, outperforming all other configurations. VGG-16 multi-level aggregation yields $F1 = 0.72$ but suffers from parameter inefficiency (138M vs. 23M). DenseNet-121 reaches $F1 = 0.74$ despite feature redundancy from dense connections. ViT-B/16 intermediate blocks produce $F1 = 0.70$, limited by homogeneous block structure without clear spatial hierarchy. MobileNetV2 and Inception v3 show $F1 = 0.68$ and $F1 = 0.71$ respectively.

Clear stage hierarchy with residual connections, progressive spatial-to-semantic transition, and moderate intermediate dimensionality are the key architectural requirements for effective multi-level feature aggregation.

Keywords: multi-level feature aggregation, image similarity, convolutional neural networks, vision transformers, intermediate representations, near-duplicate detection, architectural comparison, ResNet-50.

AMS 2020 classification: 68T45, 68T07.

Introduction

Automated detection of visually similar images across large collections remains an open challenge in applied computer vision. When two photographs of the same room differ only by a watermark or a slight crop, a human recognizes them as near-duplicates instantly. Automated systems must solve the same problem at web scale, and their accuracy depends heavily on how images are encoded as vectors (Thyagarajan, & Kalaiarasi, 2021). The standard approach – extracting a fixed-length descriptor from a pre-trained deep network and comparing via a distance metric (Razavian et al., 2014) – works well in many settings, yet the choice of network architecture is rarely justified beyond convention. The same challenge appears across many visual domains – from detecting reposted content on social platforms to matching product images in e-commerce and verifying document authenticity.

A known weakness of single-layer representations from final network layers is the loss of low- and mid-level visual information. These features encode high-level semantic content but suppress textures, edges, and local structures that matter for near-duplicate detection and structural correspondence (Kordopatis-Zilos et al., 2017). This trade-off motivates multi-level feature aggregation – combining activations from several intermediate layers to preserve information across the full abstraction hierarchy (Kubytskyi, & Panchenko, 2023; Panchenko, Bozhok, & Kubytskyi, 2026).

Combining features from different depths has precedents in HyperColumns (Hariharan et al., 2015) and Feature Pyramid Networks (Lin et al., 2017), but both target dense prediction and require spatially aligned outputs. Image similarity needs a single compact descriptor per image. The pipeline in (Kubytskyi, & Panchenko, 2023) solves this by applying Global Average Pooling to each intermediate layer, concatenating the results, and L2-normalizing the combined vector. While effective on real estate image similarity tasks (Kubytskyi, & Panchenko, 2023; Panchenko, Bozhok, & Kubytskyi, 2026), the backbone architecture has not been compared against alternatives. Vision transformers (ViT) (Dosovitskiy et al., 2021) and contrastive vision-language models (CLIP) (Radford et al., 2021) offer architecturally distinct alternatives to CNNs, yet whether their intermediate features are useful for multi-level aggregation has not been studied.

The research object is the multi-level intermediate feature aggregation pipeline and its dependence on the structural properties of the backbone architecture.

The aim of this research is to evaluate the suitability of six deep learning architectures – VGG-16 (Simonyan, & Zisserman, 2014), Inception v3 (Szegedy et al., 2015), DenseNet-121 (Huang et al., 2017), MobileNetV2 (Sandler et al., 2018), ResNet-50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2021) – for multi-level intermediate feature aggregation in image similarity tasks, and to identify the architectural properties that determine aggregation quality.

The structure of the paper is as follows. Section 1 describes the aggregation pipeline and evaluation protocol. Section 2 presents the architectural analysis and experimental results. The paper concludes with a discussion of findings and practical recommendations for backbone selection.

© Kubytskyi Volodymyr, Bozhok Artem, 2026

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

1. Multi-level aggregation pipeline and evaluation protocol

The aggregation pipeline, originally proposed in (Kubytzkyi, & Panchenko, 2023), operates as follows. Given a backbone network, we select K intermediate layers at distinct depth levels. For each selected layer k , the activation tensor is reduced to a fixed-length vector via Global Average Pooling (GAP). The resulting vectors are concatenated and L2-normalized:

$$e = [GAP(F_1); GAP(F_2); \dots; GAP(F_K)], \quad e_{norm} = e / \|e\|_2, \quad (1)$$

where F_k denotes the activation tensor at layer k . The normalized descriptor (1) is architecture-agnostic and can be applied to any backbone that produces spatially-structured intermediate activations.

For pairwise similarity classification, we construct a symmetric feature vector from two image descriptors $x^{(a)}$, $x^{(b)}$ and train a compact MLP (2048→1024→512→256, ELU activation, batch normalization, dropout) with binary cross-entropy loss (Panchenko, Bozhok, & Kubytzkyi, 2026). The base CNN remains frozen; only the MLP is trained.

Evaluation is performed on the INRIA Holidays dataset (Jégou, Douze, & Schmid, 2008), containing 1,491 images grouped into 500 scene classes with variations in viewpoint, scale, illumination, and partial occlusions. Positive pairs are drawn from the same class; negative pairs are randomly sampled across classes in a balanced 1:1 ratio. The MLP is trained on 100 labeled pairs. All architectures use ImageNet-pretrained weights with frozen parameters, ensuring that differences in performance reflect the quality of intermediate representations rather than training procedures.

All experiments are implemented in Python 3.10 using PyTorch 2.1 with torchvision model implementations. Feature extraction and MLP training are performed on a single NVIDIA RTX 3090 GPU. Each configuration is evaluated over 10 independent runs with different random pair splits (seed values 0–9); we report the mean F1-score and standard deviation across runs. Input images are resized to 224×224 pixels with standard ImageNet normalization.

For each architecture, we select intermediate layers that correspond to distinct processing stages, aiming to capture features at maximally different levels of abstraction. The specific layer selections are: VGG-16 – after conv2_2, conv3_3, conv5_3 (128+256+512 = 896-D); Inception v3 – outputs of mixed stages at spatial resolutions 35×35, 17×17, 8×8 (concatenated branch outputs, 1280-D); DenseNet-121 – after denseblock1, denseblock2, denseblock4 (128+256+1024 = 1664-D, skipping denseblock3 to reduce redundancy); MobileNetV2 – after bottleneck stages 2, 4, 7 (32+64+160 = 256-D); ResNet-50 – C2, C3, C5 outputs (256+512+2048 = 2816-D, excluding C4 due to overlap with neighboring stages); ViT-B/16 – outputs of attention blocks 3, 6, 9, 12 (768×4 = 3072-D).

2. Architectural analysis and experimental results

We analyze six architectures along four structural criteria that determine their suitability for multi-level feature aggregation: (i) stage clarity – whether the architecture has well-defined, separable processing stages; (ii) spatial progression – whether spatial resolution decreases monotonically across stages; (iii) feature independence – whether features at different stages capture genuinely distinct information; and (iv) residual information flow – whether architectural mechanisms preserve input information through the network depth. Table 1 summarizes the structural properties.

Table 1

Structural properties of backbone architectures for multi-level feature aggregation

Architecture	Params (M)	Stage structure	Intermediate dims	Residual flow	Stage clarity	Inf. time (ms)	Aggregation suitability
VGG-16	138	Sequential	128/256/512	No	Med	~120	Limited by params
Inception v3	24	Branched	Mixed/varied	Partial	Low	~95	Ambiguous layers
DenseNet-121	8	Dense	128/256/512/1024	Dense	Low	~110	Feature redundancy
MobileNetV2	3.4	Inv. residual	32/64/160	Yes	Med	~35	Low-capacity features
ResNet-50	23	4-stage residual	256/512/2048	Yes	High	~75	Optimal
ViT-B/16	86	Uniform blocks	768/768/768	No	Low	~130	No spatial hierarchy

Note: inference times measured on NVIDIA RTX 3090, batch size 1, 224×224 input.

VGG-16 (Simonyan, & Zisserman, 2014) has a purely sequential structure of 3×3 convolutional blocks, making intermediate layer selection straightforward (e.g., conv2_2, conv3_3, conv5_3). The spatial resolution decreases by factor 2 at each pooling layer, creating a clear hierarchy from fine-grained (128-D at 56×56) to semantic (512-D at 7×7) features. However, VGG-16 lacks residual connections, meaning that low-level information is progressively diluted through the depth – early layer details must be re-learned at each subsequent stage rather than being preserved via skip connections. The fully-connected layers alone contain 123M of the total 138M parameters, making the architecture computationally inefficient for feature extraction when used as a frozen backbone. Multi-level aggregation produces a 896-dimensional descriptor (128 + 256 + 512), achieving F1 = 0.72 – lower than ResNet-50 despite having 6× more parameters.

Inception v3 (Szegedy et al., 2015) processes each spatial resolution through parallel branches with different kernel sizes (1×1, 3×3, 5×5, pooling). This multi-branch structure creates an ambiguity problem for multi-level aggregation: at each stage, there is no single canonical feature tensor to extract. Selecting one branch discards information from others; concatenating all branches yields an inconsistent representation across stages since the number and configuration of branches varies. While the total parameter count is moderate (24M), the internal topology does not support transparent extraction of hierarchically organized features. Aggregation from mixed outputs yields F1 = 0.71, but the result is sensitive to branch selection choices, making the pipeline less reproducible.

DenseNet-121 (Huang et al., 2017) connects each layer to all subsequent layers within a dense block via concatenation. While this ensures strong gradient flow and feature reuse, it creates a problem for multi-level aggregation: features at different stages are not independent. Each dense block's output already contains information from all preceding layers within the block, and transition layers between blocks further compress this accumulated information. Concatenating features from multiple dense blocks therefore introduces substantial redundancy – the same edge and texture information may appear in multiple components of the aggregated descriptor. Despite this, DenseNet-121 achieves F1 = 0.74, benefiting from its compact architecture (8M parameters) and efficient internal feature reuse that partially compensates for the redundancy at the aggregation level.

MobileNetV2 (Sandler et al., 2018) uses depthwise separable convolutions and inverted residual blocks to minimize computational cost and parameter count (3.4M). While the architecture does include residual connections and has a stage-like structure, the resulting intermediate features have very low channel dimensionality (32/64/160 at the selected stages). This low capacity limits the representational richness of each component in the aggregated descriptor. The multi-level descriptor is only 256-dimensional – an order of magnitude smaller than ResNet-50's 2816-D – and the F1 of 0.68 confirms that extreme parameter efficiency comes at the cost of feature richness for similarity tasks requiring fine-grained discrimination.

ResNet-50 (He et al., 2016) provides the most suitable structure for multi-level aggregation due to three key properties. First, its four clearly defined stages (C2, C3, C4, C5) with progressively decreasing spatial resolution ($56 \times 56 \rightarrow 28 \times 28 \rightarrow 14 \times 14 \rightarrow 7 \times 7$) and increasing channel dimensionality ($256 \rightarrow 512 \rightarrow 1024 \rightarrow 2048$) enable transparent layer selection at distinct levels of abstraction. Second, residual connections within each stage preserve low-level information through the entire depth, ensuring that early-stage features retain discriminative details that would otherwise be lost in a feedforward architecture. Third, the moderate total parameter count (23M) and intermediate dimensionality ($256+512+2048 = 2816$ -D with C4 excluded) balance expressiveness and computational efficiency. The C2+C3+C5 configuration achieves $F1 = 0.77$, consistently outperforming all other architectures. C4 is excluded due to significant feature overlap with neighboring stages (Kubyskyi, & Panchenko, 2023).

ViT-B/16 (Dosovitskiy et al., 2021) consists of 12 identical transformer blocks, each producing a 768-dimensional output over a 14×14 grid of patch tokens. Unlike CNNs, ViT lacks a spatial hierarchy: all blocks operate at the same patch resolution, and the self-attention mechanism allows each block to aggregate information globally from the first layer. Intermediate features from blocks 3, 6, 9, 12 have identical dimensionality and similar structural properties, limiting the complementarity that drives multi-level aggregation gains in CNNs. The concatenated 3072-dimensional descriptor achieves $F1 = 0.70$ – only 2 percentage points above the single-block baseline, compared to the 9 p.p. gain observed with ResNet-50. ViT-B/16 also has 86M parameters – $3.7 \times$ more than ResNet-50 – and longer inference time.

CLIP ViT-B/32 (Radford et al., 2021) shares the architectural limitations of standard ViT but adds a further constraint: contrastive text-image training suppresses fine-grained visual details in favor of semantic alignment, making it poorly suited for near-duplicate detection tasks (Gkelios, Boutalis, & Chatzichristofis, 2021).

To quantify the practical impact of architectural properties, we evaluate each backbone in the multi-level aggregation pipeline described in Section 1. For each architecture, we report the F1-score achieved by the multi-level configuration alongside the single-layer baseline (final convolutional or attention block only) to measure the aggregation gain – the improvement attributable to intermediate layer inclusion. Results are presented in Tab. 2.

Table 2

Multi-level aggregation performance on INRIA Holidays (F1-score)

Architecture	Multi-level config	Dim.	Multi-level F1	Single-layer F1	Gain (p.p.)	Key limitation
ResNet-50	C2+C3+C5	2816	0.77 ± 0.02	0.68 ± 0.03	+9	–
DenseNet-121	dense1+2+4	1664	0.74 ± 0.02	0.68 ± 0.03	+6	Feature redundancy
VGG-16	conv2+3+5	896	0.72 ± 0.03	0.66 ± 0.03	+6	No residual flow
Inception v3	mixed stages	1280	0.71 ± 0.03	0.67 ± 0.03	+4	Ambiguous layers
ViT-B/16	blocks 3+6+9+12	3072	0.70 ± 0.03	0.68 ± 0.03	+2	No spatial hierarchy
MobileNetV2	stages 2+4+7	256	0.68 ± 0.03	0.63 ± 0.03	+5	Low capacity

The results in Tab. 2 reveal two key patterns. First, multi-level aggregation consistently improves over single-layer baselines across all architectures, confirming that intermediate features carry complementary information regardless of backbone design. Second, the magnitude of improvement varies widely and correlates with the structural criteria defined above. ResNet-50 shows the largest gain (+9 p.p.), followed by DenseNet-121 and VGG-16 (both +6 p.p.), MobileNetV2 (+5 p.p.), Inception v3 (+4 p.p.), and ViT-B/16 (+2 p.p.).

The ordering of aggregation gains is not random. Architectures with clear stage hierarchy and residual connections (ResNet-50) benefit most because their intermediate features are both diverse and well-preserved. Architectures where intermediate features are structurally similar (ViT-B/16) or already entangled through dense connections (DenseNet-121) benefit less, because concatenation adds redundant rather than complementary information.

The descriptor dimensionality does not predict performance: ViT-B/16 produces the largest descriptor (3072-D) yet achieves lower F1 than ResNet-50 (2816-D) and DenseNet-121 (1664-D). This confirms that the structure of intermediate features matters more than their quantity. From a practical standpoint, ResNet-50 achieves the best accuracy with moderate computational cost (23M parameters, ~75 ms/image), making it the most efficient backbone for production similarity systems.

Discussion and conclusions

This study identifies three architectural requirements for effective multi-level feature aggregation in image similarity tasks. First, clear stage hierarchy with distinct spatial resolutions ensures that features from different layers capture complementary information – from fine-grained edges and textures at early stages to semantic concepts at deep stages. Second, residual connections preserve low-level details through the network depth, preventing the information loss that degrades aggregation quality in architectures like VGG. Third, moderate intermediate dimensionality provides sufficient representational capacity without the excessive redundancy observed in DenseNet or the insufficient capacity of MobileNet.

ResNet-50 uniquely satisfies all three requirements, which explains its best performance across all tested configurations. Its four-stage structure provides a natural progression from spatial to semantic features, while skip connections ensure that each stage retains access to input-level information.

ViT-B/16 shows the smallest aggregation gain (+2 p.p.) despite having the largest descriptor. The uniform block structure – all 12 blocks produce 768-dimensional outputs at the same spatial resolution – means that intermediate features differ primarily in degree of abstraction rather than in spatial resolution or feature type. Hierarchical vision transformers such as Swin Transformer (Liu et al., 2021) may address this limitation by introducing a stage-based structure more analogous to CNNs. CLIP adds a further constraint: contrastive pre-training suppresses fine-grained details, making it poorly suited for near-duplicate detection.

For practitioners selecting a backbone for multi-level similarity systems, these results translate into concrete guidelines. If the task requires fine-grained discrimination, ResNet-50 with C2+C3+C5 is the recommended choice. If inference speed is the primary constraint, MobileNetV2 offers 2× faster processing at lower accuracy. DenseNet-121 provides a middle ground when memory is limited. VGG-16, Inception v3, and ViT-B/16 are not recommended for this pipeline due to the structural limitations identified above.

Limitations of this study include the use of a single benchmark dataset and frozen pre-trained weights. Future work should extend the evaluation to additional visual domains, investigate fine-tuning of intermediate layers, and test hierarchical vision transformers that combine the stage-based structure of CNNs with the self-attention mechanism. Another direction is learned aggregation weights, where the contribution of each layer is optimized jointly with the similarity classifier.

Authors' contribution: Volodymyr Kubytskyi – conceptualization, methodology, software, formal analysis, writing – original draft. Artem Bozhok – data validation, formal analysis, writing – review and editing.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of ICLR 2021*. arXiv:2010.11929
- Gkelios, S., Boutalis, Y., & Chatzichristofis, S. A. (2021). Investigating the vision transformer model for image retrieval tasks. In *Proceedings of IEEE MMSP 2021*. <https://doi.org/10.1109/MMSP53017.2021.9733553>
- Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object localization and fine-grained localization. In *Proceedings of CVPR 2015* (pp. 447–456).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR 2016* (pp. 770–778).
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of CVPR 2017* (pp. 4700–4708).
- Jégou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *ECCV 2008* (pp. 304–317). Springer.
- Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-duplicate video retrieval by aggregating intermediate CNN layers. In *Multimedia Modeling*. Springer.
- Kubytskyi, V., & Panchenko, T. (2023). Enriched image embeddings as a combined outputs from different layers of CNN for various image similarity problems. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 180, pp. 321–333). Springer. https://doi.org/10.1007/978-3-031-36115-9_30
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of CVPR 2017* (pp. 2117–2125).
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV 2021* (pp. 10012–10022).
- Panchenko, T., Bozhok, A., & Kubytskyi, V. (2026). Multi-level CNN feature fusion from ResNet50 for near-duplicate image detection in real estate imagery. *Informatica*, 50(9). <https://doi.org/10.31449/inf.v50i9.12111>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of ICML 2021* (Vol. 139, pp. 8748–8763). PMLR.
- Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of CVPRW 2014* (pp. 806–813).
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of CVPR 2018* (pp. 4510–4520).
- Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of CVPR 2015* (pp. 1–9).
- Thyagarajan, K. K., & Kalaiarasi, G. A. (2021). A review on near-duplicate detection of images using computer vision techniques. *Archives of Computational Methods in Engineering*, 28(3), 897–916. <https://doi.org/10.1007/s11831-020-09400-w>

Отримано редакцією журналу / Received: 01.12.25
Прорецензовано / Revised: 26.03.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Володимир КУБИЦЬКИЙ, асп.
ORCID ID: 0000-0002-1529-8677
e-mail: vova.kubytskyi@gmail.com
Київський національний університет імені Тараса Шевченка, Київ, Україна

Артем БОЖОК, асп.
ORCID ID: 0009-0009-9572-6501
e-mail: artembozh@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

КОНФІГУРАЦІЇ ПРОМІЖНИХ ШАРІВ ДЛЯ ЗАДАЧ ПОДІБНОСТІ ЗОБРАЖЕНЬ: ОЦІНКА БАГАТОРІВНЕВОГО ВЕКТОРНОГО ПРЕДСТАВЛЕННЯ, ОТРИМАНОГО З АРХІТЕКТУР CNN І ТРАНСФОРМЕРІВ

Оцінка візуальної подібності зображень потребує векторних представлень, що зберігають інформацію на різних рівнях абстракції. Багаторівнева агрегація ознак із проміжних шарів CNN демонструє високу ефективність у задачах виявлення майже дублікатів, однак систематичний аналіз впливу архітектурних властивостей різних backbone-мереж на якість таких представлень досі не проводився. У пропонуваній роботі формалізовано конвеєр багаторівневої агрегації: GAP, конкатенація та L2-нормалізація проміжних активацій з подальшою класифікацією компактним MLP. Проаналізовано шість архітектур – VGG-16, Inception v3, DenseNet-121, MobileNetV2, ResNet-50 та ViT-B/16 – за критеріями чіткості стадій, просторової прогресії, незалежності ознак і залишкового потоку інформації. Кількісну оцінку виконано на

наборі даних INRIA Holidays за $F1$. ResNet-50 із шарами C2 + C3 + C5 досягає $F1 = 0,77$, перевершуючи всі інші конфігурації. VGG-16 дає $F1 = 0,72$ за надлишкової кількості параметрів. DenseNet-121 досягає $F1 = 0,74$ попри надмірність ознак. ViT-B/16 показує $F1 = 0,70$, обмежений однорідною структурою блоків. Чітка ієрархія стадій із залишковими зв'язками, поступовий перехід від просторових до семантичних ознак і помірна проміжна розмірність є ключовими архітектурними вимогами для ефективної багаторівневої агрегації.

Ключові слова: багаторівнева агрегація ознак, подібність зображень, згорткові нейронні мережі, візуальні трансформери, проміжні представлення, виявлення майже дублікатів, порівняння архітектур, ResNet-50.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 004.4:510.6

DOI: <https://doi.org/10.17721/1812-5409.2026/1.29>

Andrii KRYVOLAP, PhD (Phys. & Math.)

ORCID ID: 0009-0002-0493-9587

e-mail: andriikryvolap@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Olena SHYSHATSKA, PhD (Phys. & Math.)

ORCID ID: 0000-0001-8791-8989

e-mail: shyshatska@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Nataliia RUSINA, PhD (Ped.), Assoc. Prof.

ORCID ID: 0000-0002-5595-9548

e-mail: rusina@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

FRAME RULE IN MONOTONE FLOYD – HOARE LOGIC EXTENSION

Separation logic is an extension of Floyd – Hoare logic, aimed at the specification and verification of programs with a main focus on the representation of dynamic memory handling. Its main tool is the frame rule, which enables local proofs. Applying the composition-nominative approach to classical Floyd – Hoare logic, its monotone extension was developed. The main advantages it offers are the use of nominative data to describe the state of the program and the consideration of partial predicates. The main purpose of this research is to introduce the frame rule and pointer capabilities for monotone Floyd – Hoare logic. This further extension adds support for local reasoning, which is crucial today when the complexity of software only rises. Local reasoning means that a property of a part of the program can be proved separately and combined with proofs for other components if specific conditions hold. For this purpose, the nominative data was enriched with the notion of an address. New special compositions for storing and loading from dynamic memory were introduced. The logical connective separating conjunction was redefined as a composition in program algebra. This composition was proved to be monotone. To extend the proof of the program's approximations to a complete proof, all compositions have to be monotone. Furthermore, rules for new operators were added to the derivation system. Rules for storing to and loading from dynamic memory are modifications of the rules for the assignment operator. The frame rule states that, with the help of a separating conjunction, a valid Hoare triple can be augmented with an assertion for a part of the memory that was not modified by the program. It was shown that the classic version of the rule, implemented within Separation logic, is not sound for a monotone extension. Conditions for a sound version were presented. Investigation of possible restrictions on the operations with pointers, which will lead to constructive frame rule conditions, is the aim of future research. Complex nominative data with dynamic memory is out of the scope of this paper and should be considered separately.

Keywords: formal methods, software verification, program logic, nominative data, separation logic.

AMS 2020 classification: 03B70, 68Q60.

Introduction

Relevance of research. Software verification is a crucial part of software engineering as its main goal is ensuring the correctness and quality of program systems. The main challenge within this domain remains the availability of efficient tools supported by formal methods with sufficient expressiveness. Floyd – Hoare logic, after being introduced, became a standard basis for formal verification. Different extensions were studied over the years (Apt, & Olderog, 2019), covering, among others, parallel programs, complex data structures, etc. One such extension is the Separation logic. The main advantages of the suggested approach were the operators introduced to handle assertions over dynamic memory and a frame rule aimed at local reasoning. The importance of local reasoning is reinforced by the growing complexity of modern software. Further development resulted in the introduction of numerous variants of the Separation logic (Reynolds, 2002), covering different aspects of program systems. Among such variants of significant interest are the following:

- Separation Incorrectness logic (Raad et al., 2020), which enriches the Incorrectness logic with tools for handling dynamic memory. Incorrectness logic is a dual approach to Hoare logic, aimed at finding errors rather than proving correctness in all states.
- Exact Separation Logic (Maksimović et al., 2023), which combines in a single formal system over-approximation of Separation logic, and under-approximation of Separation Incorrectness logic.
- Concurrent Separation Logic (O'Hearn, 2007), where disjoint memory areas are treated more in terms of parallel programs, and communication is performed with the help of common memory.
- Dynamic Separation logic (De Boer, Hiep, & de Gouw, 2023), which combines Separation logic with dynamic logic, that is, an approach of defining Floyd – Hoare logic as modal logic.

Monotone Floyd – Hoare logic extension (Nikitchenko et al., 2018) uses nominative data and partial mappings to properly describe the state of execution of a program and account for a possible non-terminating behavior. It is defined within the composition-nominative approach (Shkilnyak, 2025), which also allows definitions in a semantic-syntactical manner. Different types of complex nominative data were broadly studied for the monotone Floyd – Hoare logic extension, also covering such problems as the unambiguity of the complex names for such data. But at the same time, pointers and dynamic

memory pose different challenges, such as aliasing, pointer arithmetics, and proving the exclusiveness of the memory access. To handle such problems, notions of dynamic memory and disjoint areas of dynamic memory need to be adopted for the monotone Floyd – Hoare logic extension. Combined with the frame rule variant, this would allow a program logic that will have local reasoning and nominative data capabilities, together with monotone operators.

Lately, there has been a rise in interest in an under-approximating approach to program logic that is dual to an over-approximating approach of the Floyd – Hoare logic. Logics that follow such an approach are aimed at finding the execution that leads to an error, rather than at proving that executions starting in a proper initial state will be correct. Examples of such logics are the Incorrectness logic (O'Hearn, 2019) and its Separation version (Raad et al., 2020). Even though under-approximation is not considered for the monotone Floyd – Hoare logic extension, it is a topic for future investigation. This proves that the frame rule and dynamic memory handling can be beneficial for a broad range of program logics.

The object of research is program logics that allow reasoning about the dynamic memory aspect of the programs and complex data structures.

The aim and objectives of the research are to introduce a nominative data extension with pointers and dynamic memory, provide the implementation of the frame rule of Separation logic within monotone Floyd – Hoare logic, and investigate conditions for its soundness.

1. Main definitions and results

Monotone Floyd – Hoare logic is defined with a semantic-syntactic approach. This means that first semantics is introduced as the following three-sorted program algebra:

$$\langle Pr^{V,A}, Fn^{V,A}, FPr^{V,A}; \vee, \neg, \{S_P^{\bar{v}}\}_{\bar{v} \in U(V)}, \{S_F^{\bar{v}}\}_{\bar{v} \in U(V)}, \{x\}_{x \in V}, \{\exists x\}_{x \in V}, id, \{AS^x\}_{x \in V}, \bullet, IF, WH, FH \rangle.$$

Here $Pr^{V,A} : ND(V, A) \xrightarrow{p} Bool$ is a set of all partial mappings from nominative data $ND(V, A)$ to a set of boolean values $Bool = \{True, False\}$ that represent semantics of boolean expressions in the program, $Fn^{V,A} : ND(V, A) \xrightarrow{p} A$ is a set of all partial mappings from nominative data $ND(V, A)$ to a set of values A that represent semantics of arithmetic expressions in the program, $FPr^{V,A} : ND(V, A) \xrightarrow{p} ND(V, A)$ is a set of all partial mappings from nominative data $ND(V, A)$ to nominative data $ND(V, A)$ that represent the semantics of program operators where nominative data models the current state of the variables. At the same time, nominative data is defined as the set of all partial mappings from the set of the names of the variables V to a set of all possible values A : $ND(V, A) : V \xrightarrow{p} A$.

Operators of the algebra are called compositions and denote the main constructs of the programming language. Superposition operators $\{S_P^{\bar{v}}\}_{\bar{v} \in U(V)}, \{S_F^{\bar{v}}\}_{\bar{v} \in U(V)}$ with parameter $\bar{v} \in U(V)$ correspond to the function or predicate calls, where $U(V)$ is the set of all finite sequences of different variable names. $\{x\}_{x \in V}$ is a parametric denaming operator used to retrieve the value of a variable. Compositions $\vee, \neg, \{\exists x\}_{x \in V}$ represent respective logical connectives. On the program level, compositions $id, \{AS^x\}_{x \in V}, \bullet, IF, WH$ are used to denote program constructs skip (idle operator), assignment operator, sequential execution, if-then-else operator, and while-do operator accordingly. FH is a special composition representing the semantics of the Floyd – Hoare triple.

Nominative data has to be extended to allow reasoning about dynamic memory or the heap. We will use the same assumptions as in classical Separation logic, namely: the set of all possible values is the same set that is used as addresses for pointers; variables can be assigned either a value or an address. This leads us to the following definition.

Definition 1. Dynamic nominative data over a set of variable names V and a set of addresses A is a partial mapping of the type: $DND(V, A) : V \cup A \xrightarrow{p} A$.

For further convenience, the next notations will be used to denote the stack and heap parts of the dynamic nominative data $st \in DND(V, A)$: $stack(st) = st|_V$ and $heap(st) = st|_A$. Here $st|_V$ marks the restriction of the mapping st to a set V . As dynamic nominative data represents a state of the program, both terms will be further used interchangeably.

Substituting nominative data with dynamic nominative data, we can introduce dynamic versions of the partial mapping types: $Pr_D^{V,A} : DND(V, A) \xrightarrow{p} Bool$, $Fn_D^{V,A} : DND(V, A) \xrightarrow{p} A$, $FPr_D^{V,A} : DND(V, A) \xrightarrow{p} DND(V, A)$.

Extension of most program algebra compositions on dynamic nominative data is straightforward and is omitted in this paper. Only superpositions require consideration. To describe operations with a heap, new compositions of storing values to and retrieving values from a specific address are introduced. We assume that to operate with an address, it has to be first stored in some variable. This won't limit the expressiveness of a programming language, but it will help to make address arithmetic expressions clearer. Also, as in the case of nominative data, only equitone predicates and functions are considered for dynamic nominative data. Mapping f over dynamic nominative data is considered equitone if for any state st , $f(st) \downarrow \Rightarrow \forall st' \supseteq st : f(st') \downarrow = f(st)$. Here, $f(st) \downarrow$ denotes that mapping f is defined on state st ; $f(st) \uparrow$ will be used to indicate that mapping f is undefined on state st . Restriction of $Pr_D^{V,A}$ and $Fn_D^{V,A}$ to only equitone mapping is natural, as it corresponds to the intuition that if information within a current state is sufficient to calculate a value of the expression of a condition, this value should not change for any extension of the state.

Definition 2. Parametric store composition ST^x of a type $Fn_D^{V,A} \xrightarrow{t} FPr_D^{V,A}$ with a parameter $x \in V$ can be defined on arbitrary state as $ST^x(f)(st) \simeq st \nabla [st(x) \mapsto f(st)]$. Here $[v_1 \mapsto a_1, \dots, v_n \mapsto a_n, a'_1 \mapsto a''_1, \dots, a'_m \mapsto a''_m]$ is an explicit form of dynamic nominative data representation for $v_1, \dots, v_n \in V$ and $a_1, \dots, a_n, a'_1, \dots, a'_m, a''_1, \dots, a''_m \in A$, where each pair denotes either name-value or address-value in data. Overriding operator ∇ is defined as follows: $st_1 \nabla st_2 = [l \mapsto a]l \in \text{dom } st_2 \wedge st_2(l) = a \vee l \in \text{dom } st_1 \setminus \text{dom } st_2 \wedge st_1(l) = a]$. In other words, $st = st_1 \nabla st_2$, the result of the application of

the overriding operator will be the state (nominative data) in which variable and address mapping that was defined in the right operand st_2 is retained together with all pairs from the left operator st_1 , that are not defined in st_2 .

Definition 3. Parametric load function LD^x of a type $Fn_D^{V,A}$ with a parameter $x \in V$ can be defined on an arbitrary state in the following way: $LD^x(st) \simeq st(st(x))$. Applying nominative data st to a variable x will yield the address stored in it, but the second application will return the value stored at that address.

Definition 4. Parametric extended predicate superposition composition $\{S_P^{\bar{v}, \bar{u}'}\}_{\bar{v}, \bar{u}' \in U(V)}$ of a type $Pr_D^{V,A} \times (Fn_D^{V,A})^{n+m} \xrightarrow{t} \xrightarrow{t} Pr_D^{V,A}$ with parameters $v_1, v_2, \dots, v_n, u_1, u_2, \dots, u_m \in V$ can be defined on arbitrary state in the following way:

$$S_P^{\bar{v}, \bar{u}'}(p, f_1, \dots, f_n, g_1, \dots, g_m)(st) \simeq p(st \nabla [v_1 \mapsto f_1(st), \dots, v_n \mapsto f_n(st), st(u_1) \mapsto g_1(st), \dots, st(u_m) \mapsto g_m(st)]).$$

A common type of property that needs to be verified for programs that manipulate the heap deals with exclusive access of certain program components to their own part of the heap. Separation conjunction is a logical connective introduced in Separation logic to handle such assertions without the need for repeated explicit conditions describing parts of the heap each component works with. Intuitively, it can be defined in the following way: formula $P * Q$ holds for each configuration where there are two disjoint subheaps, on one of which P holds and on another – Q holds. Let us give a formal definition of two states having disjoint heaps with a shared stack.

Definition 5. Two states $st_1, st_2 \in DND(V, A)$ are considered disjoint configurations $st_1 ||_{V,A} st_2$ if $stack(st_1) = stack(st_2)$ and $dom heap(st_1) \cap dom heap(st_2) = \emptyset$

If two states are disjoint, the overall state can be obtained by applying the override operator. Disjoint union will be used as a name for the resulting operation.

Definition 6. State $st \in DND(V, A)$ is called a disjoint union of two states $st_1, st_2 \in DND(V, A)$ (denoted $st = st_1 \oplus st_2$) if $st_1 ||_{V,A} st_2$ and $st = st_1 \nabla st_2$.

Based on the definition, the next properties of disjoint union can be obtained: $st_1 \oplus st_2 = st_2 \oplus st_1$, $stack(st_1 \oplus st_2) = stack(st_1) = stack(st_2)$, $heap(st_1 \oplus st_2) = heap(st_1) \cup heap(st_2)$. Additionally, if $st = st_1 \oplus st_2$ we will use $\ominus$ to mark the remainder of the heap: $st_1 = st \ominus st_2$ and $st_2 = st \ominus st_1$.

With a formal means to describe states with the same stack but non-intersecting areas of dynamic memory, separation conjunction composition can be introduced for program algebra monotone Floyd – Hoare logic extension with heap. For convenience, it will be used in infix form.

Definition 7. Separation conjunction composition $*$ of a type $Pr_D^{V,A} \times Pr_D^{V,A} \xrightarrow{t} Pr_D^{V,A}$ can be defined on the arbitrary state in the following way:

$$P * Q(st) = \begin{cases} True, & \text{iff exist } st_1, st_2 \text{ such that } st = st_1 \oplus st_2 \wedge P(st_1) = True \wedge Q(st_2) = True; \\ False, & \text{iff for every } st_1, st_2 \text{ such that } st = st_1 \oplus st_2 \text{ it holds that } P(st_1) = False \vee Q(st_2) = False; \\ \perp, & \text{otherwise.} \end{cases}$$

Let us recall the definition of the monotone operator prior to presenting the results on whether the separation conjunction composition is monotone. Monotonicity is treated in the sense of definedness.

Definition 8. A composition $C : (Pr_D^{V,A})^k \times (Fn_D^{V,A})^l \times (FPr_D^{V,A})^m \xrightarrow{t} Pr_D^{V,A}$ is called monotone over the first argument if for any $p, p_2, \dots, p_k, p' \in Pr_D^{V,A}$, $f_1, f_2, \dots, f_l \in Fn_D^{V,A}$, and $pr_1, pr_2, \dots, pr_m \in FPr_D^{V,A}$ it holds that $p \subseteq p'$ implies $C(p, p_2, \dots, p_k, f_1, f_2, \dots, f_l, pr_1, pr_2, \dots, pr_m) \subseteq C(p', p_2, \dots, p_k, f_1, f_2, \dots, f_l, pr_1, pr_2, \dots, pr_m)$. Here $\subseteq$ is considered for mappings as a set-theoretical relation over the graphs of respective mappings. A composition is called monotone if it is monotone over every argument.

Lemma 1. Separation conjunction composition is monotone.

Proof. Monotonicity will be proved for the first argument of composition. Proof for another argument is similar.

Assume that property doesn't hold, this means that $\exists P, P', Q \in Pr_D^{V,A}$ such that $P \subseteq P' \wedge P * Q \not\subseteq P' * Q$. The latter is possible in the following cases:

- $\exists st \in DND : P * Q(st) \downarrow = True \wedge (P' * Q(st) \downarrow = False \vee P' * Q(st) \uparrow)$,
- $\exists st \in DND : P * Q(st) \downarrow = False \wedge (P' * Q(st) \downarrow = True \vee P' * Q(st) \uparrow)$.

Let us consider the first case. If $P * Q(st) \downarrow = True$ then by the definition of composition there should exist such st_1 and st_2 that $P(st_1) = True$ and $Q(st_2) = True$ and $st = st_1 \oplus st_2$. From $P(st_1) = True$ and $P \subseteq P'$ we obtain $P'(st_1) = True$ and from definition of composition using $P'(st_1) = True$, $Q(st_2) = True$ and $st = st_1 \oplus st_2$ we can conclude that $P' * Q(st) \downarrow = True = P * Q(st)$ which contradicts with our assumption.

The second case, where $P * Q(st) \downarrow = False$, proof is similar. Using the definition of composition for every st_1 and st_2 such that $st = st_1 \oplus st_2$, either $P(st_1) = False$ or $Q(st_2) = False$. If $P(st_1) = False$, using $P \subseteq P'$ we obtain

$P'(st_1) = False$. Finally, this proves we cannot find a pair of disjoint states such that at least one of $P'(st_1) = False$ and $Q(st_2) = False$ won't hold. Thus, $P' * Q(st) \downarrow = False = P * Q(st)$, and in the second case, we also achieve a contradiction.

Both cases were considered, which means that $P * Q \subseteq P' * Q$ and composition is monotone over the first argument. $\square$

Combining the given definition, we will obtain a three-sorted algebra that will define the semantics of the monotone Floyd – Hoare logic extension with dynamic memory.

$$\begin{aligned} < Pr_D^{V,A}, Fn_D^{V,A}, FPrG_D^{V,A}; \vee, \neg, *, \{S_P^{\bar{v}, \bar{u}'}\}_{\bar{v}, \bar{u} \in U(V)}, \{S_F^{\bar{v}, \bar{u}'}\}_{\bar{v}, \bar{u} \in U(V)}, \{x\}_{x \in V}, \\ & \{\exists x\}_{x \in V}, id, \{AS^x\}_{x \in V}, \{ST^x\}_{x \in V}, \{LD^x\}_{x \in V}, \bullet, IF, WH, FH >. \end{aligned}$$

The next step, after new compositions were defined and their properties were investigated, is to provide sound derivation rules for a monotone Floyd – Hoare logic extension that operates with those compositions. Proof that existing rules remained sound when moving from nominative data to dynamic nominative data is straightforward and also omitted in this paper.

The rule for store composition can be viewed as the modification of the rule for assignment composition:

$$R_ST \frac{}{\{S_P^{x'}(p, f)\} ST^x(f) \{p\}}.$$

Lemma 2. Rule R_ST is sound.

Proof. The rule is sound if for valid premises – valid outcomes are derived. As in the rule, there are no premises – we need to prove that triple $\{S_P^{x'}(p, f)\} ST^x(f) \{p\}$ is valid. Triple $\{p\} r \{q\}$ is considered valid if for every interpretation and for every state st , such that $p(st) \downarrow = True$, either $r(st) \uparrow$ or $q(r(st)) \downarrow = True$. In our case, based on the definition of superposition composition, if $S_P^{x'}(p, f)(st) \downarrow = True$ then $p(st \nabla [st(x) \mapsto f(st)]) \downarrow = True$. If program $ST^x(f)$ terminates for starting state st , $p(ST^x(f)(st)) = p(st \nabla [st(x) \mapsto f(st)])$, which evaluates to True based on the previous assertion. This proves the soundness of the rule. $\square$

In the Separation logic frame, the rule for the derivation system is defined as follows:

$$R_FRAME_SEP \frac{\{p\} pr \{r\}}{\{p * q\} pr \{r * q\}}, ch(pr) \cap fv(q) = \emptyset.$$

Where $ch(pr)$ denotes variables that were changed by program pr and is defined inductively based on the structure of pr , $fv(q)$ is the list of free variables for predicate q . Such a variant of the rule is not sound for monotone Floyd – Hoare logic extension. The reason behind this fact is that the introduced condition doesn't account for complex pointer arithmetic available in the general case. We will investigate restrictions in the semantic form that will address this issue. The main idea behind condition $ch(pr) \cap fv(q) = \emptyset$, together with the use of the separation conjunction, was to describe that when there is an assertion about the part of the heap that the program doesn't affect – it can be combined with any valid assertion about the part of the heap that the program modifies.

Let us introduce tools to describe such aspects of the predicates and programs as the part of the heap affected by them. As pointer arithmetic is not limited in our case, part of the heap affected by the program in terms of updating stored values is a function over the state. It cannot be described by only considering the variables that the program updates. In the general case, the definition of the required function can be seen below.

Definition 9. Memory footprint function for a program $pr \in FPrG_D^{V,A}$ is a function of a type $DND(V, A) \xrightarrow{t} 2^A$ that for an arbitrary state st where $pr(st) \downarrow$ is equal $mem(pr)(st) = \{a | a \in A \wedge st(a) \downarrow \wedge pr(st)(a) \downarrow \neq st(a)\}$, and $mem(pr)(st) = \emptyset$ otherwise if program is not terminating on st .

The intuition behind the memory footprint function is the maximal heap area affected and changed by the program. It doesn't account for addresses that were updated in the process with the same value. In this case, a strict version of the function can be introduced that considers not only the addresses that were changed by the program but also the addresses that are required by the program to terminate.

Definition 10. A strict memory footprint function for a program $pr \in FPrG_D^{V,A}$ is such a function of a type $DND(V, A) \xrightarrow{t} 2^A$ that for an arbitrary state st , where $pr(st) \downarrow$, it can be defined as $mem_s(pr)(st) = \bigcup_i A_i$, where A_i are all such sets of addresses that $retain(pr, st, A_i) \wedge \neg \exists A'_i \subset A_i : retain(pr, st, A'_i)$ and $retain(pr, st, A_i)$ is a special notion for $pr(st)|_{V \cup A_i} = pr(st|_{V \cup A_i})|_{V \cup A_i} \wedge heap(st|_{V \cup (A \setminus A_i)}) = heap(pr(st)|_{V \cup (A \setminus A_i)})$, and $mem_s(pr)(st) = \emptyset$ if program is not terminating on st .

Remark 1. Based on the monotonicity of compositions it can be shown that the condition for A_i also holds for their union and any superset up to st , which means $retain(pr, st, mem_s(pr)(st))$ and for all A such that $mem_s(pr)(st) \subset A \subset \text{dom } heap(st)$ it holds that $retain(pr, st, A)$.

Remark 2. It is obvious that for arbitrary $st \in DND(V, A)$ and arbitrary $pr \in PrG_D^{V,A}$ $mem(pr)(st) \subseteq mem_s(pr)(st)$.

Restrictions on the memory modification capabilities for programs that are available within program algebra impact whether the memory footprint function can be defined inductively based on the program structure.

Similarly to the programs, a general definition of a memory function for predicates can be given based on the intuitive notion of an area of memory required for the predicate to be defined.

Definition 11. Memory footprint function for a predicate $p \in Pr_D^{V,A}$ is a function of a type $DND(V, A) \xrightarrow{t} 2^A$ that for an arbitrary state st where $p(st) \downarrow$ it is equal $mem_s(p)(st) = \bigcup_i A_i$, here A_i are all such sets of addresses such that $p(st|_{V \cup A_i}) \downarrow = p(st) \wedge \neg \exists A'_i \subset A_i : p(st|_{V \cup A'_i}) \downarrow = p(st)$, and $mem(p)(st) = \emptyset$ otherwise if predicate is not defined on st .

Using the notion of the memory footprint frame rule, modified semantical conditions can be introduced:

$$R_FRAME \frac{\{p\}pr\{r\}}{\{p * q\}pr\{r * q\}}, \forall st \in DND(V, A) : p(st) \rightarrow (mem_s(pr)(st) \cap mem(q)(st) = \emptyset), ch(pr) \cap fv(q) = \emptyset.$$

Theorem 1. Rule R_FRAME is sound.

Proof. Assume that triple $\{p * q\}pr\{r * q\}$ is not valid. This means that there exists such a state st that $p * q(st) = True$, $pr(st)$ is defined, and $r * q(pr(st)) = False$. If $p * q(st) = True$ then from the definition of separating conjunction exist st_1 and st_2 such that $st = st_1 \oplus st_2$, $p(st_1) = True$, $q(st_2) = True$.

Consider $st'_2 = st_2|_{V \cup mem(q)(st)}$, in this case we can construct $st'_1 \supseteq st_1$ such that $st'_1 = st \ominus st'_2$. This state will extend the heap of st_1 with the part of the heap of st_2 , not in st'_2 . Based on the properties of compositions of the predicate level $p(st'_1) = True$. $q(st'_2) = True$ holds based on the definition of memory footprint of a predicate if we show that $st_2|_{V \cup mem(q)(st)} = st_2|_{V \cup mem(q)(st_2)}$. The latter is equal to $mem(q)(st) \cap dom heap(st_2) = mem(q)(st_2)$ as heap part of the state can be restricted only up to the list of its addresses. This statement is true if we consider that $q(st_2) = True$, $st_2 = st|_{V \cup dom heap(st_2)}$, $q(st) = True$, and the definition of the memory footprint.

Next we shall prove that for any states st and st'_1 such that $st'_1 \subseteq st$, $stack(st) = stack(st'_1)$, $pr(st) \downarrow$ and also $dom heap(st'_1) \cap mem_s(pr)(st) = \emptyset$ there exist the states st'_2 , st'''_1 , such that $st = st'_1 \oplus st'_2$, $pr(st) = pr(st'_2) \oplus st'''_1$, $stack(st'''_1) = stack(pr(st))$ and $heap(st'''_1) = heap(st'_1)$. In this case, $st'_2 \supseteq st|_{V \cup mem_s(pr)(st)}$, which means that conditions from the definition of the strict memory footprint also hold for st'_2 and $pr(st) = pr(st'_2) \oplus st'''_1$.

It is easy to see that if $st'_1 = st'_2$ then all the required conditions hold and $st'_2 = st'_1$ as $st = st'_1 \oplus st'_2$. Thus, $pr(st) = pr(st'_1) \oplus st'''_1$, $stack(st'''_1) = stack(pr(st))$, and $heap(st'''_1) = heap(st'_2)$. Based on the definition of $ch(pr)$, $stack(st'''_1)|_{V \setminus ch(pr)} = stack(st)|_{V \setminus ch(pr)} = stack(st'_2)|_{V \setminus ch(pr)}$. As $ch(pr) \cap fv(q) = \emptyset$, we have $fv(q) \subseteq V \setminus ch(pr)$, together with $heap(st'_1) = heap(st'_2)$ and $stack(st'''_1)|_{V \setminus ch(pr)} = stack(st'_2)|_{V \setminus ch(pr)}$ this gives us $q(st'_1) = q(st'_2) = True$. To show that $r(pr(st'_1)) = True$ we need to use the fact that $\{p\}pr\{r\}$ holds for any state, and for state st'_1 , $p(st'_1) = True$ and $pr(st'_1) \downarrow$. From $pr(st) = pr(st'_1) \oplus st'''_1$, $q(st'''_1) = True$ and $r(pr(st'_1)) = True$ we can obtain $r * q(pr(st)) = True$, this contradicts with assumption, which proves that rule R_FRAME is sound. $\square$

2. Examples

To showcase the frame rule for monotone Floyd – Hoare logic extension, we will use a simple example with a specification of the stack data structure. This is a first-in-last-out data structure, assertions for two main operations – put a value on top of the stack and retrieve a value from the top of the stack. Memory allocation and deallocation are not considered in this example. The second stack and its properties are used to illustrate the frame rule and give another example of assertions about data structures.

Example. Without any restriction on generality, we will use the set of lowercase letters of the Roman alphabet as variable names $V = \{a, b, \dots, z\}$ and integer numbers as addresses $A = Int = \{\dots, -1, 0, 1, \dots\}$. Basic predicates and functions will include parametric predicates for relations greater, equals, and less, and parametric functions increment and decrement, denoted respectively $>_{x,y}$, $=_{x,y}$, $<_{x,y}$, $x++$, and $x--$. x and y are names of the variables that are used in mappings. The stack data structure is presented by its top and bottom addresses; all addresses that are greater than the bottom and less than the top are considered a part of the stack. To put a value on top of the stack – value is stored in the top address, and the variable holding the top is incremented. Retrieving value is exactly opposite – if top is greater than bottom, return the value at the top and decrement top. The stack is empty if the top address is equal to the bottom address. A stack is considered invalid if its top address is less than the bottom address.

The subprograms and predicates described above can be defined semantically with the first parameter being a variable storing the top address, and the second parameter being a variable storing the bottom address. $push_{x,y}(f) \simeq ST^x(f) \bullet AS^x(x++)$ pushes the value of the expression that represents function f to the top of the stack. $pop_{x,y,v} \simeq IF(>_{x,y}, AS^v(LD^x) \bullet AS^x(x--), id)$ pops the top value of the stack to the variable v ; if the stack is empty or invalid, it does nothing. $empty_{x,y} \simeq =_{x,y}$, $invalid_{x,y} \simeq <_{x,y}$.

Here are the properties of the stack data structure that can be proved: $\{empty_{x,y}\}push_{x,y}(f)\{\neg empty_{x,y}\}$, $\{empty_{x,y} \wedge \wedge =_{v,w}\}pop_{x,y,v}\{empty_{x,y} \wedge \wedge =_{v,w}\}$ and $\{empty_{x,y}\}push_{x,y}(f) \bullet pop_{x,y,v}\{empty_{x,y} \wedge S^w(=_{v,w}, f)\}$. It is easy to define memory footprints for $push$ and pop : $mem_s(push_{x,y}(f))(st) = \{st(x--(st))\}$, $mem_s(pop_{x,y}(f))(st) = \{st(x++(st))\}$. Nominative data $[x \mapsto 3, y \mapsto 1, 1 \mapsto 4, 2 \mapsto 2]$ representing a simple stack with two values in it.

If we have another stack with top a and bottom b condition that this stack contains values in the ascending order can be defined as $q = \forall e \forall f (>_{e,f} \wedge \geq_{a,e} \wedge \geq_{f,b} \rightarrow S^{x,y}(>_{x,y}, LD^e, LD^f))$. The frame rule cannot be used to prove that

$\{empty_{x,y} * q\} push_{x,y}(f) \bullet pop_{x,y,v}\{empty_{x,y} \wedge S^w(=_{v,w}, f)\} * q\}$. The memory footprint of a predicate q is $mem(q)(st) = \{i \in Int[b'(st) \leq i \leq a'(st)]\}$. There is no guarantee that operations on the initial stack won't overwrite values for addresses of the second stack. One example is $[x \mapsto 3, y \mapsto 1, 1 \mapsto 4, 2 \mapsto 2, a \mapsto 5, b \mapsto 3, 3 \mapsto 1, 4 \mapsto 5]$. This issue can be handled by adding to the precondition assertion that the top address of the first stack is less than the bottom address of the second, or vice versa.

The final Floyd – Hoare triple will be $\{(empty_{x,y} \wedge >_{b,x}) * q\} push_{x,y}(f) \bullet pop_{x,y,v}\{empty_{x,y} \wedge S^w(=_{v,w}, f)\} * q\}$. Shared variable values within the separating conjunction and the absence of the restriction on the free variables of p in R_FRAME made it possible to specify the needed condition. If we have another stack with top a and bottom b condition that the stack contains values in ascending order can be defined as $q = \forall e \forall f (>_{e,f} \wedge \geq_{a,e} \wedge \geq_{f,b} \rightarrow S^{x,y}(>_{x,y}, LD^e, LD^f)$. Frame rule cannot be used to prove $\{empty_{x,y} * q\} push_{x,y}(f) \bullet pop_{x,y,v}\{empty_{x,y} \wedge S^w(=_{v,w}, f)\} * q\}$. The memory footprint of the predicate q is $mem(q)(st) = \{i \in Int[b'(st) \leq i \leq a'(st)]\}$. There is no guarantee that operations on the initial stack won't overwrite values for addresses of the second stack. The example for such a state is $[x \mapsto 3, y \mapsto 1, 1 \mapsto 4, 2 \mapsto 2, a \mapsto 5, b \mapsto 3, 3 \mapsto 1, 4 \mapsto 5]$. This issue can be handled by adding to the precondition assertion that the top address of the first stack is less than the bottom address of the second, or vice versa. The final Floyd – Hoare triple will be $\{(empty_{x,y} \wedge >_{b,x}) * q\} push_{x,y}(f) \bullet pop_{x,y,v}\{empty_{x,y} \wedge S^w(=_{v,w}, f)\} * q\}$. Shared variable values within the separating conjunction and the absence of the restriction on the free variables of p in R_FRAME made it possible to specify the needed condition.

Discussion and conclusions

Dynamic nominative data was introduced, making it possible to define properties of the programs with pointers within the composition-nominative approach. To represent basic operations with the heap, new compositions were defined, and their monotonicity was proved. Similar results were achieved for the separation conjunction connective, which is the basis of the frame rule and assertions over the heap. Overall, the conditions of the frame rule were investigated in terms of monotone Floyd – Hoare logic extension, and a simple example of its application is provided. The updated conditions required new means that describe a part of the heap affected by the program and a part of the heap needed for the predicate to be defined. Basic properties of such memory mappings were presented.

Similar to what was achieved in already published papers, the introduction of the frame rule extended the capabilities of monotone Floyd – Hoare logic extension for local reasoning. But it is worth mentioning that the construction of memory footprints for special cases of logic, where address operations are restricted, is one of the directions of further research, together with a weaker version of the frame rule, which has a less complex condition. Other open questions are working with memory allocation and concurrency.

The general case of predicates over dynamic nominative data was considered. Further research is required to define the properties of predicates needed for efficient computation of the memory footprint of the predicate. This could also lead to the introduction of new logical connectives similar to the separating conjunction.

Authors' contribution: Olena Shyshatska — conceptualization; Nataliia Rusina — methodology; Andrii Kryvolap — analysis of sources, theoretical foundations of research.

Sources of funding. The study was partially financially supported by Taras Shevchenko National University of Kyiv.

References

- Apt, K. R., & Olderog, E.-R. (2019). Fifty years of Hoare's logic. *Formal Aspects of Computing*, 31(6), 751–807. <https://doi.org/10.1007/s00165-019-00501-3>
- De Boer, F. S., Hiep, H.-D. A., & de Gouw, S. (2023). Dynamic separation logic. *Electronic Notes in Theoretical Informatics and Computer Science*, 3, 5:1–5:19. <https://doi.org/10.46298/entics.12297>
- Maksimović, P., Cronjäger, C., Löb, A., Sutherland, J., & Gardner, P. (2023). Exact separation logic: Towards bridging the gap between verification and bug-finding. In Ali & Salvaneschi (Eds.), *37th European Conference on Object-Oriented Programming (ECOOP 2023)* (pp. 19:1–19:27). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/LIPICS.ECOOP.2023.19>
- Nikitchenko, M., Ivanov, I., Kornilowicz, A., & Kryvolap, A. (2018). Extended Floyd-Hoare logic over relational nominative data. *Communications in Computer and Information Science*, 826, 41–64. https://doi.org/10.1007/978-3-319-76168-8_3
- O'Hearn, P. W. (2019). Incorrectness logic. *Proceedings of the ACM on Programming Languages*, 4(POPL), 1–32. <https://doi.org/10.1145/3371078>
- O'Hearn, P. W. (2007). Resources, concurrency, and local reasoning. *Theoretical Computer Science*, 375(1), 271–307. <https://doi.org/10.1016/j.tcs.2006.12.035>
- Raad, A., Berdine, J., Dang, H.-H., Dreyer, D., O'Hearn, P., & Villard, J. (2020). Local reasoning about the presence of bugs: Incorrectness separation logic. In Lahiri & Wang (Eds.), *Computer Aided Verification* (pp. 225–252). Springer International Publishing. https://doi.org/10.1007/978-3-030-53291-8_14
- Reynolds, J. (2002). Separation logic: a logic for shared mutable data structures. In *Proceedings 17th Annual IEEE Symposium on Logic in Computer Science* (pp. 55–74). IEEE. <https://doi.org/10.1109/LICS.2002.1029817>
- Shkilnyak, S. (2025). Varieties of pure first-order logics of partial quasiary predicates. *Bulletin of Taras Shevchenko National University of Kyiv. Physical and Mathematical Sciences*, 79(2), 80–88. <https://doi.org/10.17721/1812-5409.2024/2.13>

Отримано редакцією журналу / Received: 10.04.25

Прорецензовано / Revised: 06.11.25

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Андрій КРИВОЛАП, канд. фіз.-мат. наук
ORCID ID: 0000-0003-4567-8901
e-mail: andriikryvolap@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Олена ШИШАЦЬКА, канд. фіз.-мат. наук
ORCID ID: 0000-0001-8791-8989
e-mail: shyshatska@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Наталія РУСІНА, канд. пед. наук, доц.
ORCID ID: 0000-0002-5595-9548
e-mail: rusina@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

РАМКОВЕ ПРАВИЛО В МОНОТОННОМУ РОЗШИРЕННІ ЛОГІКИ ФЛОЙДА – ХОАРА

Логіка розділення є розширенням логіки Флойда – Хоара, спрямованим на специфікацію та верифікацію програмних систем, в яких використовується динамічна пам'ять. Її основним засобом є рамкове правило, що дає змогу працювати з локальними доведеннями. Монотонне розширення логіки Флойда – Хоара було одержано, використовуючи композиційно-номінативний підхід. Основні переваги отриманої логіки полягають у застосуванні номінативних даних для опису стану програми та у роботі з частковими предикатами. Основна мета пропонованого дослідження – розроблення рамкового правила та засобів роботи з вказівниками для монотонної логіки Флойда – Хоара. З огляду на це, вона була збагачена інструментами для підтримки локальних доведень, що надзвичайно важливо сьогодні, коли складність програмного забезпечення лише зростає. Локальність означає, що властивість частини програми може бути доведена окремо та поєднана з властивостями інших компонентів, якщо визначені умови виконуються. Для цього номінативні дані було розширено поняттям адреси. Представлено нові спеціальні композиції збереження в динамічній пам'яті і завантаження з неї. Логічна зв'язка відокремлювальна кон'юнкція була перевизначена як композиція в програмній алгебрі. Доведено, що така композиція є монотонною. Щоб мати можливість розширити доведення для апроксимацій програми до доведення для програми загалом, усі композиції мають бути монотонними. Крім того, додано правила системи виводу для нових операторів. Правила для зберігання в динамічну пам'ять та завантаження з неї є модифікацією правила для оператора присвоєння. Рамкове правило – це правило, яке за допомогою відокремлювальної кон'юнкції описує, як можна розширити істинну трійку Хоара за допомогою твердження для частини пам'яті, яка не була змінена програмою. Було показано, що варіант правила, реалізований у логіці розділення, вже не є коректним для монотонного розширення. Були представлені модифіковані умови, за яких досягається коректність. Темою майбутніх досліджень є можливі обмеження на операції з вказівниками, які дають можливість запропонувати конструктивні умови для рамкового правила. Іншим важливим питанням, що виходить за межі цієї статті, є розгляд використання складних номінативних даних разом з динамічною пам'яттю.

Ключові слова: формальні методи, верифікація ПС, програмні логіки, номінативні дані, логіка розділення.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 004.75

DOI: <https://doi.org/10.17721/1812-5409.2026/1.30>

Anatolii PASHKO¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0001-6944-8477
e-mail: aap2011@ukr.net

Ihor PAPROTSKYI¹, PhD Student
ORCID ID: 0009-0006-1644-2833
e-mail: igorpapr@gmail.com

Andrii HLYBOVETS², DSc (Engin.), Prof.
ORCID ID: 0000-0003-4282-481X
e-mail: a.glybovets@ukma.edu.ua

Svitlana TERENCECHUK³, PhD (Phys. & Math.), Prof.
ORCID ID: 0000-0001-6527-4123
e-mail: terenchuksa@ukr.net

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

³Kyiv National University of Construction and Architecture, Kyiv, Ukraine

IMPROVEMENTS TO THE CIRCUIT BREAKER PREDICTION MODEL FOR INCREASING THE FAULT TOLERANCE OF MICROSERVICES

In high-load distributed systems, it is crucial to pay strict attention to the reliability of communication between architectural components, making fault tolerance a fundamental property of microservice architecture. While various fault tolerance patterns exist to improve microservice resilience, and their implementations are widely adopted, none are universal for the range of possible failures in service communication. Therefore, there is a continuous need to search for more effective approaches to achieving fault tolerance in distributed software systems.

This research is building upon our earlier study, which introduced the Circuit Breaker Prediction Model. The main principle of the Prediction model is to calculate the rating of communication between two services, defined as a convex combination of various performance metrics, and the rating value is then used to determine whether the interaction with the target service can be safely re-established after the period of blocking outgoing requests. Compared to standard implementations of this method, it was specifically designed to reduce the latency involved in blocking request transmission between microservices, especially under uncertain load conditions.

The purpose of this paper is to enhance this model and verify its effectiveness when combined with other resilience patterns, particularly Retry and Timeout. The study enhances the flexibility of the Circuit Breaker Prediction Model by describing an adaptable sliding window mechanism that dynamically adjusts the context size based on load conditions at runtime. Also, the algorithm of rating calculation is generalized to enable the usage of any substantial number of performance metrics. Finally, this research conducts simulated experiments with multiple service communication configurations and fault tolerance pattern combinations, using the standard Circuit Breaker implementation as a baseline for comparison. The findings show improved communication performance, mainly due to reduced request transmission delays under the same load conditions. The prospects of applying this model's principles to other fault tolerance patterns have been discussed, with particular emphasis on the potential of using machine-assisted optimization techniques in dynamic decision making.

Keywords: *microservices architecture, distributed systems, fault tolerance, resilience patterns, high load, circuit breaker, service communication, runtime adaptation.*

AMS 2020 classification: 68M15, 68M14.

Introduction

In software engineering, microservice architecture (MA) is considered a practical approach for managing large-scale enterprise applications. Breaking down the application structure into many functional parts not only allows for the division of various entity domains and the isolation of distinct parts of the business logic but also increases the autonomy of teams in research and development (R&D): this way, they can work in separate code repositories in parallel, making the development process more granular, faster, and isolated.

Recent surveys highlight the popularity of microservices in software development. According to the JetBrains 2023 Developer Ecosystem survey (JetBrains, 2024), 82 % of respondents use microservices in their daily work. Similarly, Stack Overflow's 2023 Developer Survey (Stack Overflow, 2024) found that nearly half of the participants said their companies rely on microservices. These statistics prove that more and more software development companies, teams, and developers are considering this approach suitable for their system design. This makes microservice architecture one of the leading system design trends in the 2020s.

Despite its popularity, the approach can sometimes be challenging in designing, developing, maintaining, securing, ensuring availability, and implementing correct application logic. Because the key feature of microservice architecture is decentralization, components rely heavily on communication, especially through events that may occur at different levels of the entire system. According to (Lewis, & Fowler, 2014) "the microservice preference towards choreography and event collaboration leads to emergent behavior," which may sometimes be negative and result in unforeseen consequences. If not handled properly, even a brief network failure can cause data inconsistencies between the services, potentially leading to

incorrect delivery of the intended functionality. To mitigate this risk, the developers must consider ensuring eventual message delivery between services. This is where the concept of "resilience" emerges.

Fault tolerance patterns, also known as resilience patterns, are commonly used to mitigate disruptions, network outages, security breaches, changes in application load, and system failures. They aim to recover from failures and operate effectively, minimizing the impact of incorrect system behavior on the application's availability.

While numerous fault tolerance patterns and techniques exist, and they can be classified into different categories, this research focuses on those that influence direct synchronous communication between services, one of the basic types of data transfer. Among such patterns, experts outline the following: Timeouts, Circuit Breaker (CB), Retries, Fallbacks, Bulkheads, Rate Limiters, and others (Nygard, 2018, pp. 110–146; Resilience4j, 2025). Even though they seem to be simple in their working principles, many software libraries and frameworks may provide different implementations of some of them. Mostly, those algorithms use the canonical logic of a pattern, and most software developers may use the existing solutions out of the box, changing only variable values in the configuration provided by a library. However, sometimes even a small optimization of the algorithm can lead to a meaningful performance increase.

In this article, we focus on the Circuit Breaker method and extend our previous research published in (Hlybovets, & Paprotskyi, 2024), where we proposed the Circuit Breaker Prediction Model as a modification to it. The model aimed at dynamic adaptation to conditions in synchronous communication between two microservices by predicting the performance of the interaction between two microservices and using a set of runtime metrics, reducing the Circuit Breaker's state-transition latency and improving the request success rate.

However, this model can still be enhanced to increase the method's ability to adapt to load changes. Also, the previous approach provided only a preliminary evaluation of the algorithm. Despite the promising results compared to the traditional Circuit Breaker, we want to further validate whether the implementation is more effective in more complex microservice setups. Apart from that, since the method used a particular set of performance metrics for the communication rating calculation, there is a need to generalize the algorithm so that it could work with any number or combination of metrics.

According to the background, **the object of this research** is microservices that operate under high-load conditions.

The goal of the study is to improve and validate the Circuit Breaker Prediction Model in a more practical environment and in conjunction with other resilience patterns, as well as to generalize the algorithm to support varying numbers and types of performance metrics.

In this work, we consider a "more practical environment" to involve different combinations of synchronous microservice-to-microservice communications, the simultaneous use of multiple resilience patterns, and larger workloads that change over time.

To achieve the goals of the research, **the following objectives were defined for this paper**:

- Enhance the model's ability to adapt to workload changes by introducing automatic adjustment of the sliding window size, which provides a more accurate data context for calculating service communication performance.
- Extend and unify the performance calculation formula introduced in our previous study, and present a generalized version that supports an arbitrary set of performance metrics.
- Validate the model under more practical conditions, including a series of experiments that involve its simultaneous use with other resilience patterns.

It makes sense to briefly mention the main concepts of the traditional Circuit Breaker pattern: it works analogously to a real-world electric fuse. In software programming, this concept is typically implemented as a state machine that wraps the logic of a given function, which makes a call to another service. Depending on the outcome of that function being executed repeatedly, the CB state machine can transition between three states: Closed, Open, and Half-Open (Fowler, 2014).

The flow where one microservice requests another may sound simple and appear easy to manage. However, since microservice architecture often distributes different domains across multiple services, the execution of application logic likely requires multiple cascading requests between them. Given a situation where there is one or more services between the two (as an example, service A makes a request to service B that makes a request to service C...), there is a possibility for some of them to have a failure in communication, which, if not handled properly, may result in cascading failures of the whole call chain. Not only will this break the application logic, but, more importantly, it will also overuse the available resources of all the services in the mentioned call chain.

The Circuit Breaker pattern is used to counteract the situations described above. It rejects the execution of requests that will almost certainly not result in a successful response, thus saving resources and reducing the network load (Fowler, 2014).

In the current study, other resilience patterns, which are often used in conjunction with the Circuit Breaker pattern, will be considered as the methodological basis for experimentation. The patterns of Retry and Timeout. Their role in the research is not isolated but rather complementary: they provide the context in which the proposed Circuit Breaker Prediction Model will be validated, allowing us to test its behavior under combined resilience strategies.

The mentioned mechanisms were described in detail in (Nygard, 2018, pp. 110–146; Resilience4j, 2025).

Since resilience in microservice architecture is a highly multifaceted topic, it has also become the subject of numerous research studies in the field of software engineering.

The study (De Souza Miranda et al., 2024) explores models, methods, and technologies for building resilient microservices. They rely on various strategies to maintain system stability and reliability. The authors examine and present the initial catalog of patterns for developing fault-tolerant microservices. They also provided a review of a wide range of literature and organized a survey in this field of study to determine the difficulties and solutions for them.

The paper (Valdivia et al., 2020) presents a multivocal systematic literature review for microservices, identifying the benefits and trade-offs of different patterns used in microservice architecture development and their association with quality attributes and metrics. The paper identified the metrics that help to measure a system's readiness to meet quality criteria that improve overall performance.

Recent research (Sun, Cui, & Cai, 2024) applied a deep reinforcement learning approach to fault tolerance in microservice environments. They use metrics gathered from a service mesh and cluster in their circuit breaker strategy. The authors aim

to dynamically adjust tripping thresholds and recovery times. This method learns policies from external telemetry data and focuses on environmental observations outside the service. In contrast, our work uses microservices' internal context data for calculating state predictions.

The research (Zhang et al., 2025) proposed a fault-tolerance framework using Spring Cloud and Docker. It utilizes techniques and features like service discovery, load balancing, and failover. The authors focused on fault tolerance at the architectural and orchestration levels, and focused on redundancy and quick service recovery. However, in our study, we focus on the individual pattern's level of optimization in the communication between individual pairs of services.

To sum up, the variety of different approaches in the recent studies shows that microservice resilience can be approached from various angles. It can be observed and improved at the architecture level by analyzing the whole system performance, or it can be optimized by changing the algorithms responsible for controlling service communication. Each perspective has different pros and cons depending on the size and complexity of the target informational system.

There are various implementations of the CB in different programming languages, various libraries, and frameworks. Many of them were mentioned in our previous study (Hlybovets, & Paprotskyi, 2024). The Resilience4j framework was selected for implementing the Prediction Model, and the choice was explained in the mentioned article.

Most frameworks provide the standard Circuit Breaker pattern implementation, but some of them add extra features and configuration variations to improve their customization capabilities. For example, (Resilience4j, 2025) offers highly customizable configuration, the possibility to combine multiple fault tolerance patterns, and advanced metrics (such as "failure rate" and "slow call rate"). The latter framework is used in this research as a base for experiments.

1. Methodology

Circuit Breaker Prediction Model. The main concepts of our model, introduced in (Hlybovets, & Paprotskyi, 2024), are briefly described below.

When using the Circuit Breaker pattern to resolve communication errors between microservices, it is crucial to understand how it can be configured most effectively. The most complex challenge in such a configuration is determining the optimal amount of time for the Circuit Breaker to wait in the Open state until the requests can be permitted.

The period between when a target service started failing to respond and when it became available is not static and may vary across different failure occurrences. As a result, there is almost always a point when the target service has fully recovered, but the Circuit Breaker in the consuming service remains in the Open/Half-open state, blocking requests from being sent. Moreover, the transitions from the Half-open state back to the Open state sometimes make the total delay even longer: after each such transition to the Open state, a cycle of waiting starts again, potentially resulting in the loss of those invaluable seconds of reestablished communication between services, which may cost a lot of business value in an enterprise application.

Since the problem of operational delays is related to the static time-guessing nature of pattern configuration, the Circuit Breaker Prediction Model was previously proposed. The idea of making the pattern to adapt to the communication performance between the two services and predict whether the target service has already recovered forms the expectation of working more efficiently than the original CB model.

Instead of using the canonical three states in the CB, the proposed model has only two: Open and Closed. Upon the target function being called, which is wrapped by the Circuit Breaker, a decision is made on whether to close the "circuit" by calculating the rating value of the communication. Hereinafter, the "rating value" represents the prediction of whether the target microservice is ready to receive requests. If it is higher than the Rating Threshold value configured in the application, the CB mechanism makes a transition to the Closed state, enabling it to make requests to the target service. There were no changes to the Closed state, its conditions for detecting failures, and transitioning to the Open state compared to the canonical CB. (Hlybovets, & Paprotskyi, 2024)

The Circuit Breaker Prediction Model implementation must gather a set of metrics to ensure its operation via the rating value calculation. The sliding window method of data collection is used, thereby taking into account only the last configured number of function invocations.

The particular metrics can vary depending on the future research of that algorithm, but the initial set of them was used in the previous study:

- **Failure rate.** The percentage of failed requests compared to the total number of requests observed in the sliding window. A request is considered failed if its invocation raises an exception from a preconfigured list.

- **Slow call rate.** The ratio of the number of requests that took more time than a configured threshold to the total number of requests recorded in the sliding window.

- **Success call rate.** The ratio of successful (not failed) requests to the total number of requests recorded in the sliding window.

- **Time in the Open state.** The current amount of time CB has been operating in the Open state, starting from the previous transition. It is also essential to check if the time in the Open state has not exceeded the preconfigured threshold, the maximum allowed time in that state, and close the CB if it does.

After some revision, we decided to use an enhanced list of metrics for our experiments in this research. The mentioned list is discussed in Section 1.4.

Generalizing the algorithm. Despite the model having been proven to work, there is a necessity to generalize the algorithm so that it can work with an arbitrary set of performance metrics.

Depending on its nature, we can logically categorize a specific metric as "positive" or "negative," based on how it influences the communication rating value. For example, the Failure rate and Slow call rate are negative metrics because the higher they are, the lower the CB rating value should be.

Therefore, formula (1) is used to represent the calculation logic for the Circuit Breaker Prediction Model:

$$rating = posM + negM, \quad (1)$$

where $posM$ and $negM$ are defined using (2) and (3):

$$posM = \sum_{i=1}^a (c_i * m_{pos^i}); \quad (2)$$

$$negM = \sum_{j=1}^b (c_j * (1 - m_{negj})), \quad (3)$$

where variable a holds the quantity of positive metrics in (2), variable b holds the quantity of negative metrics in (3), and both refer to the number of metrics gathered by the service in the runtime. The c_i and c_j are the preconfigured weight coefficients for each positive or negative metric, which define the level of influence of the corresponding metric on the whole rating value. The m_{posi} and m_{negj} are the values of each metric (also positive or negative). The weights of m_{negj} are inverted, are inverted, ensuring that all components contribute uniformly to communication quality.

To maintain the correctness of calculations, the following static configuration check, which verifies if the equality (4) is satisfied, must be performed when the CB is loaded into the context of an application:

$$\sum_{i=1}^a c_i + \sum_{j=1}^b c_j = 1. \quad (4)$$

The check ensures that all coefficient values for the metrics are properly distributed relative to one another, and therefore, the proportion of each metric in the overall rating calculation formula is accurately estimated.

Each metric in the mentioned equations is expressed as a ratio; the values are bound to the range from 0 to 1, so the final rating value is also in the range from 0 to 1:

$$c_i, m_{posi}, c_j, m_{negj}, rating \in [0; 1]. \quad (5)$$

Therefore, the whole rating calculation is derived from a normalized weighted aggregation of metrics, with negatively oriented ones inverted to ensure that higher values uniformly indicate better performance.

Finally, the rating value is compared with the preconfigured rating threshold. If the calculated value is bigger, the CB changes its state to Closed, resuming the communication.

It is also important to briefly describe the core part of the model's behaviour.

The algorithm was formalized in the following pseudocode listing:

When in **CLOSED** state:

```
* allow request *
result ← * execute call and gather metrics *
if (result is error AND error not in IGNORED_ERRORS_LIST):
    if (request duration OR failure counter exceeds thresholds):
        transitionToOpenState()
        * disallow further requests *
        * update metrics *
    else if (success):
        if (request duration OR failure counter exceeds thresholds):
            transitionToOpenState()
            * disallow further requests *
            * update metrics *
return
```

When in **OPEN** state:

```
On invocation attempt:
communicationRating ← calculateRatingValue(metrics)
if (communicationRating > RATING_TRANSITION_THRESHOLD):
    transitionToClosedState()
    * allow request *
    * update metrics *
else:
    * decline request and update metrics *
return
```

function **calculateRatingValue(metrics)**:

```
for each metric in metrics:
    * compute individual metric's component value *
    if metric is negatively oriented:
        * invert its contribution *
return sum of all rating components
```

According to what is described in this section, it is obvious that as long as both metric values and a sum of coefficients are normalized, any substantial number of performance metrics can be used in this calculation. Each metric contributes to the rating proportionally to its preconfigured weight, and negatively oriented metrics are inverted to ensure consistent interpretation. Also, each metric's component value can be calculated differently if it makes sense for the algorithm, and if any extra logic is required for the metric component to participate in the communication rating.

The Slow Call Rate metric value can be calculated by inverting the rate value and multiplying by the coefficient (Java code):

```
double slowCallRateMetric = countOfSlowCalls / totalCallsInAWindow;
```

```
double slowCallRating = (1 - slowCallRateMetric) * SLOW_CALL_RATE_COEFFICIENT;
```

However, such metrics as Consecutive Failure Streak can be calculated in the following way:

```
int maxStreak = Math.max(maxStreak, consecutiveFailures);
double failureStreakMetric = Math.min((float) maxStreak / S_MAX, 1.0);
double consecutiveFailureStreakComponent = (1 - failureStreakMetric) * FAILURE_STREAK_COEFFICIENT;
```

In this case, the consecutive failure streak metric is additionally normalized to [0, 1] using a saturation limit S_MAX , because there's a need to ensure that excessively long failure sequences do not disproportionately affect the final communication rating.

From an implementation perspective, such metric calculations can be provided by a fault-tolerance framework at the library level, allowing a software developer to potentially select from a predefined list of metric components available for the Circuit Breaker Prediction Model without the necessity to write their own calculation for it. The important detail is that all the reads and writes to the state of CB and its set of metrics must be implemented thread-safe, ensuring consistency and correctness.

Enhancing the adaptiveness to periodic load changes. It is common for a microservice to experience varying load levels over time, depending on the context within which the software system operates. For example, a Kafka worker instance subscribed to multiple user-related topics in the backend of a public business-related website might experience a higher average load during the working part of the day than at night. In this case, for each event processed by that worker instance, if a synchronous request function (wrapped in the Circuit Breaker instance) is invoked to another microservice through the processor's logic, the potential performance metrics increase in proportion to the number of such request invocations. This should provide a better context for the Circuit Breaker Prediction Model, if it is used in this setup.

Most modern Circuit Breaker implementations work based on a sliding window, which is a continuously updating time or call-based interval that collects recent request outcomes and performance metrics to evaluate the current reliability of service communication. Existing frameworks and libraries only provide the static window size, depending on the sliding window type, and only suggest using a fixed sampling duration; some of them also provide the mechanism of evicting the "bad instances" (e.g., Istio/Envoy – "Outlier Detection" (Istio, 2025)), but also without automatic window size adjustments.

In this study, we propose an enhancement to the Circuit Breaker Prediction Model for dynamically adjusting the window size based on recent load. Since our model analyzes performance metrics and is also based on a count-based sliding window, it makes sense to adjust the context to natural traffic oscillations: the lower the traffic, the less recent requests should influence the prediction about the state of communication between microservices, and vice versa.

To implement this idea, an adaptive sliding window controller was added to the existing model that scales the data window size according to the observed throughput.

First, the current request rate is calculated (6):

$$\hat{\lambda}_t = \frac{n_t}{\Delta}, \quad (6)$$

where n_t is the number of requests for the last interval Δ , and $\hat{\lambda}_t$ is the current request rate.

Then, we apply exponential smoothing to the observed request rate to stabilize short-term noise and produce a smoothed throughput (7):

$$\lambda_t = (1 - \gamma)\lambda_{t-1} + \gamma\hat{\lambda}_t, \quad \gamma \in (0,1], \quad (7)$$

where λ_t is the smoothed request rate at the moment t , λ_{t-1} is the previous smoothed request rate, $\hat{\lambda}_t$ is the current "raw" request rate from (6), and γ is the smoothing factor.

The smoothed rate is then used to compute the target window size at the moment t (8):

$$W_{target,t} = clip(\alpha\lambda_t, W_{min}, W_{max}), \quad (8)$$

where $W_{target,t}$ is the target sliding window size at the moment t , λ_t is the smoothed request rate at the moment t , α is the scaling coefficient that controls the proportional growth of the window size relative to the request rate, W_{min} , W_{max} are the lower and upper bounds for the allowable window size, preventing excessive shrinking or expansion, $clip$ is the bounding function that limits the resulting value within the specified range $[W_{min}, W_{max}]$.

To avoid oscillations due to minor traffic variations, the window size is updated only when its relative change exceeds the hysteresis threshold (9):

$$W_t = W_{target,t} \text{ if } \frac{|W_{target,t} - W_{t-1}|}{W_{t-1}} > \varepsilon; \\ \text{otherwise } W_t = W_{t-1}, \quad (9)$$

where W_t is the final window size applied at the current update step, $W_{target,t}$ is the newly computed target window size before stabilization, W_{t-1} is the sliding window size from the previous step (current value), and ε is the hysteresis threshold that defines the minimum relative change required to trigger an update.

The recalculation process is performed periodically (e.g., every couple of seconds, configured according to the application's needs) and runs as a separate thread from the main Circuit Breaker flows, or is initiated by a background scheduler via a dedicated timer. The window size and request count variables should be accessed and updated in a thread-safe way. The initial window size value W is meant to be configured at application startup.

As a result, the proposed algorithm conceptually operates as illustrated in Fig. 1:

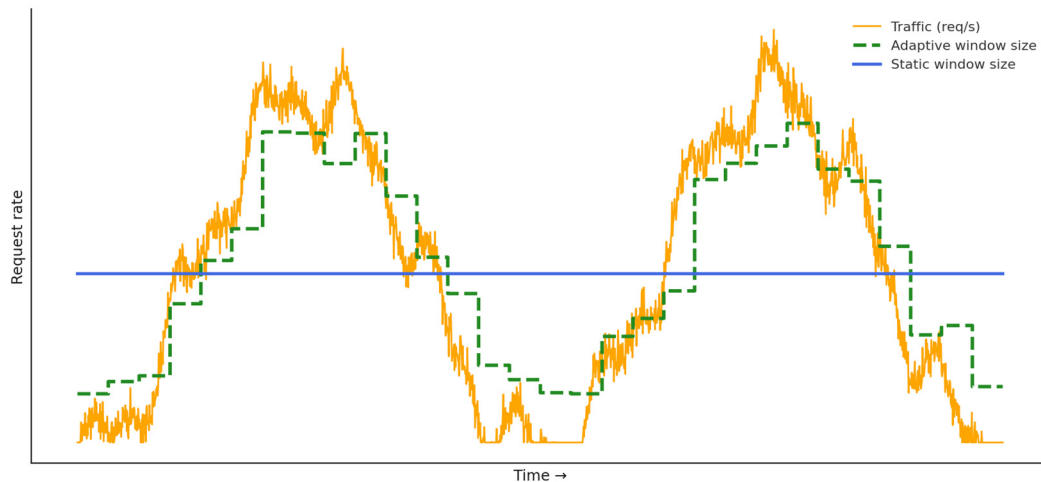


Fig. 1. Conceptual illustration of adaptive sliding window behavior compared to a static configuration under periodic load changes

The adaptive sliding window adjusts its size according to the current request rate, and uses exponential smoothing and a hysteresis threshold to avoid frequent value changes. In turn, the static window size remains insensitive to the load changes and may capture either too little or too much data, so the adaptive window size should provide a more balanced representation of system activity metrics.

We integrated this mechanism into the Circuit Breaker Prediction Model by default and used it in all scenarios in the experimental study.

Experiments. Unlike the earlier study that used a simplified service communication simulation, this work assesses the Circuit Breaker Prediction Model under more realistic and diverse experimental conditions.

In modern microservice systems, developers frequently apply multiple fault tolerance mechanisms together within the same communication flow. They typically combine Circuit Breaker with Retry, Timeout, and other resilience patterns. The experiments in this study were designed according to this idea, and the goal was to validate the CB Prediction Model under these combined conditions in different setups and in a controlled environment to allow measuring and observability.

The configuration of each resilience pattern was selected through extensive empirical analysis. Parameter values were chosen after a series of preliminary runs aimed at identifying stable and distinguishable behaviors between the compared CB implementations of traditional and the Prediction models. Although no formal optimization techniques were used during the preliminary tuning process, the configuration proved suitable for a comparison between the canonical and predictive versions of the Circuit Breaker.

To ensure repeatability, we used identical configuration values for all experiments, making the observed performance differences attributable to the model design itself rather than to random variation in settings.

In comparison to the previous study, we decided to enhance the set of performance metrics used for rating calculation in our experiments. To avoid redundancy, only one of the complementary success and failure metrics was kept, as the Success Call Rate and Failure Rate (with the inverted value) have the same contextual meaning when analyzing the state of communication between services. Additionally, a couple of new metrics were introduced to expand the behavior.

The Permitted Call Rate is updated after each invocation attempt and reflects the recent availability trend of the protected service.

The Consecutive Failure Streak metric accumulates failed requests until a successful one occurs, after which the counter resets to zero. Its value is then normalized to remain within a fixed $[0, 1]$ interval. To calculate it, we added a small addition to the pattern's base code, tracking consecutive failures through a simple internal counter.

The influence coefficients were slightly adjusted compared to previous work to reflect the updated metrics while maintaining the model's overall balance (see the previous set of metrics in Section 1.1). The main idea was to preserve the weight of reliability- and time-related indicators, but also allow some space to the new ones that capture how the Circuit Breaker behaves and how short-term instability appears.

The coefficients used in the experiments are shown in Tab. 1.

Table 1

Influence coefficients of metrics used for service communication rating calculation

Metric	Orientation	Coefficient	Description
Success call rate	+	0.3	The ratio of successfully completed requests to the total number of requests within the sliding window.
Slow call rate	-	0.15	The ratio of requests exceeding the preconfigured response time threshold. Adds the context of request processing speed.
Permitted call rate	+	0.2	The ratio of requests permitted by the Circuit Breaker to the total number of attempted requests. Shows communication invocation availability.
Consecutive failure streak	-	0.1	The normalized length of the current sequence of consecutive failed requests. Captures short-term spikes of instability.
Time in the Open state	-	0.25	The current duration of the Circuit Breaker operating in the Open state. Shows the context of the current waiting time denying the requests.

Since the goal of the new CB implementation is to reduce the potential time spent in the Open state while the target service has already recovered from the outage and thus increase the percentage of successful invocations under load, the experiments were designed as a simulation where one microservice (utilizing the CB) makes a large number of requests to another service. The target service, in turn, tends to fail with a random probability and then recover after a random time within a fixed interval. The response times are generated randomly, limited by a fixed timeframe.

To simulate load fluctuations, we designed the workload generation process to follow a precomputed trace of request intensities over a fixed period of experiment runtime. The load was computed using the intensity function, which was defined as a smooth oscillating signal with additive noise, representing natural load variations. We configured the duration of one full oscillation cycle to be 20 min, while the total duration of each experimental run was 30 min, resulting in 1.5 cycles per run.

Since the workload trace was precomputed, it determined the number of requests to be sent at each discrete tick.

For each time tick Δt , the expected request rate $\lambda(t)$ was calculated, and the final workload trace can be represented as (10):

$$\begin{aligned}\lambda(t) &= \lambda_{min} + A(1 + \sin(\frac{2\pi t}{T})) + \eta(t), \\ \eta(t) &= \text{Poisson}(\mu_n),\end{aligned}\quad (10)$$

where λ_{min} is the baseline request rate, A is the load amplitude, T is the period of oscillation, and $\eta(t)$ is a Poisson-distributed random noise with a mean value μ_n .

Each computed value $\lambda(t)$ was rounded to the nearest integer to obtain the number of requests issued during the time interval Δt , and these requests were distributed across a fixed-size worker pool of 16 threads. The purpose of the mentioned worker pool was to act like independent users that constantly initiate a varying number of invocations of the source service (where the CB is configured). The same workload trace was reused across all runs to ensure consistency and comparability between experiments.

For this study, the following values were used for the parameters mentioned above: $\lambda_{min} = 1000$ requests/s, $A = 800$ requests/s, $T = 1200$ s (20 min), $\Delta t = 0.5$ s, $\mu_n = 200$.

During a 30-minute run, this workload generated approximately:

- 3.5 million requests in total (220 thousand requests per worker);
- an average intensity of around 120 requests/s per worker;
- a minimal intensity of around 420 requests/s;
- a maximum intensity of around 3100 requests/s.

Such a configuration provided a stable yet sufficiently dynamic load for observing how both Circuit Breaker implementations process the requests.

To model the behavior of an external dependency, a simulated target service was implemented, where we defined internal pseudo-randomized logic. The only purpose of this service was to simulate realistic failure and recovery processing instead of performing a real business logic, so each API endpoint invocation was followed by responding with a delaying thread execution, instead of the real downstream call.

The target service was switching between three operational states:

- UP – the service is healthy, and returns a successful response after a pseudo-random delay (taken from a uniform distribution between T_{min} and T_{max} values).
- DEGRADED – the service is partially healthy, meaning it returns either a successful response after a delay, or fails with a preconfigured probability.
- DOWN – the service is not operational. All requests return a failed response after a maximum delay.

Transitions between the mentioned states were implemented by a semi-random schedule, controlled by a timer that recalculated the next transition interval based on semi-random values within a fixed range. The idea behind this design was to replicate the temporal instability and simulate a real-world microservice, where the service's availability may be dependent on load spikes, resource exhaustion, or network issues.

The main configuration parameter values used in the process described above (and fixed for the experiment) are presented in Tab. 2.

Table 2

State configuration parameter values used in the experiments

Parameter \ State	UP	DEGRADED	DOWN
Response time, ms	50–400	200–2500	3000
Failure probability	0.01	0.15	1.0
Probability of switching to the DOWN state	0.05	0.2	-
Probability of switching to the DEGRADED state	Load-dependent: • Base degradation probability: 0.04 • Overload coefficient (sensitivity): 0.3 • Overload threshold: $0.7\lambda_{max}$, where λ_{max} is the current maximum request rate	-	0.3
Probability of switching to the UP state	-	0.4	0.6
Random delay bounds before switching back to UP state, s	-	1-10	1-10

The interval between state recalculations was set to 10 s.

We used the same pseudo-random seeds in our experiments to ensure the controlled and deterministic behavior during the testing of two CB implementations. This made the pseudo-random parts of the environment's behavior controllable, allowing performance differences to be attributed solely to the use of different Circuit Breaker models.

We also used the following Resilience4j configuration parameters of the traditional model of Circuit Breaker:

- Fixed sliding window size: 1000 requests.
- Slow-call duration threshold: 3.75 s.
- Time to wait in the Open state: 4 s.
- Failure rate threshold: 50 %.
- Slow-call rate threshold: 50 %.
- Permitted calls in Half-Open state: 20.

The default Circuit Breaker was used as a base for implementation by following the instructions documented in (Resilience4j, 2025). The CB Prediction Model was implemented via overriding the basic CB classes of Resilience4j components. Because the library already monitors the metrics selected earlier, it was very useful for this implementation. For more specific implementation details, refer to our previous work (Hlybovets, & Paprotskyi, 2024).

For the CB Prediction Model, we decided to use the following adapting window configuration described in detail in Section 1.3: smoothing factor (γ) = 0.2, recomputation cycle = 30 seconds, the target count-based sliding window $W_{target,t}$ is computed (8) with the scaling coefficient $\alpha = 0.6$, and bounds $W_{min} = 300$ and $W_{max} = 2600$. To prevent oscillations, the window is only changed when the relative deviation from the current value exceeds 20% upward or 8% downward. The starting window size $W_0 = 1000$ requests, so the traditional and adaptive versions of CB start with the same window size.

Circuit Breaker. Given the setup specified above, the simplest way to compare the difference in pattern accuracy between the proposed model and the regular implementation is to run the simulation only using the Circuit Breaker.

Each request from the previously mentioned workers is targeted at the source service's API, which processes it by calling the invocation function wrapped in a particular Circuit Breaker implementation. In this case, either the default implementation from the Resilience4j library or our enhanced implementation of the Circuit Breaker Prediction Model was used. The invocation function, in turn, makes another synchronous request to the API of the target service. The effectiveness of this communication is measured by calculating the success rate statistics.

See Fig. 2 for a schematic representation of this setup in our experiments.

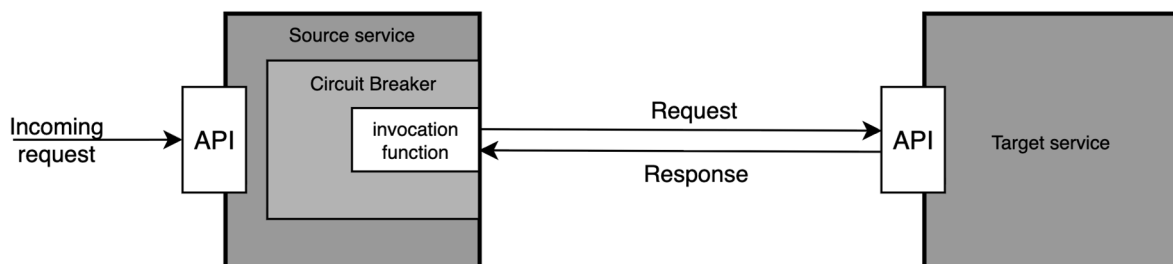


Fig. 2. The simplest Circuit Breaker usage in the experiment configuration

The results of applying the environment described in Section 1.4 to this and all the subsequent experimental configurations listed later in this section are shown in Section 2.

Circuit Breaker and Retry. The combination of the CB and Retry patterns was analyzed and recreated, since these patterns are often used together. The Resilience4j framework already provides a highly customizable implementation of the Retry pattern, which meets the needs of this research. It is important to mention that the Retry pattern context was wrapped by the Circuit Breaker, thus providing the following invocation flow: request, then a sequence of retries if a failure occurs, then the logic of the Circuit Breaker in case the number of retries is exceeded.

The schematic representation of this process is shown in Fig. 3.

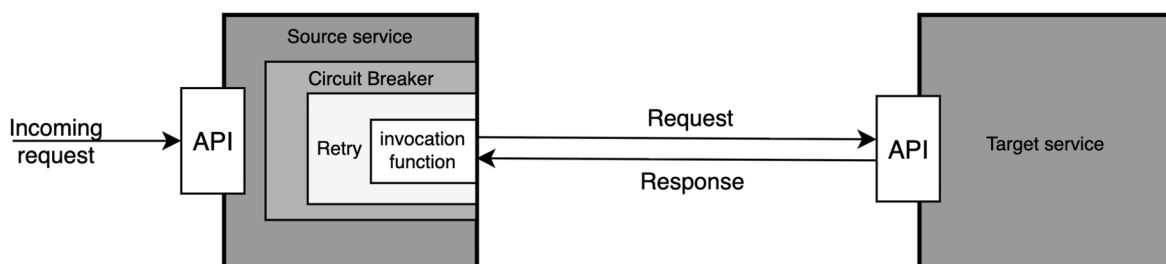


Fig. 3. The combination of Circuit Breaker and Retry in the experiment configuration

This way, most small failures, such as network blips, should be covered by the logic of the Retry mechanism, leaving the Circuit Breaker with the duty of handling more significant outages and issues.

A simple configuration with two retries and an exponential backoff was established, resulting in the following delay order between attempts: 0.5 seconds and 1 second.

Circuit Breaker, Retry, and Timeout. It is also a common use case when the Timeout is used together with the already mentioned patterns. In Resilience4j, this pattern's implementation is called a TimeLimiter, and its configuration is simple: a single duration is set, after which the request is considered failed. We put a Timeout to wrap the invocation function first (see Fig. 4), so the invocation flow results in the following: if a request takes too long to receive the response, the Timeout breaks this function's invocation, and the first retry is executed; in case this happens more times, and after all the retries were executed, the Circuit Breaker's logic considers this function's work as failed.

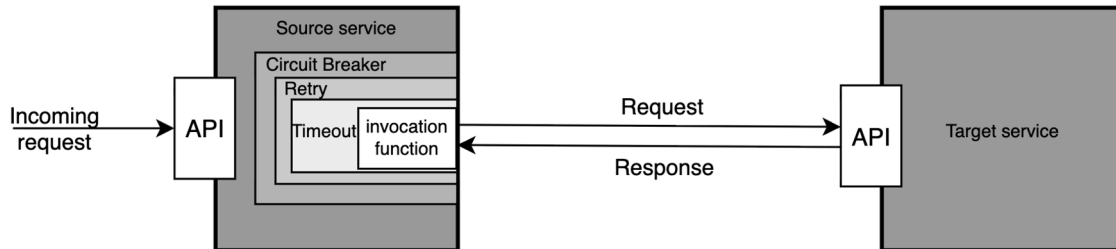


Fig. 4. The combination of Circuit Breaker, Retry, and Timeout in the experiment configuration

In our setup, we use a 2-second timeout. Since we previously configured the two retries with delays of 0.5 seconds and 1 second, this resulted in a worst-case response time of up to 7.5 seconds. Half of this duration (3.5 seconds) is already considered a slow request by both Circuit Breaker models.

Three Services: Circuit Breaker, Retry, and Timeout. Since one of the CB's main goals is to reduce the probability of cascading failures, the impact of using the different implementations should also be analyzed in a more complex setup.

This experiment involves adding an additional service to the call chain and configuring the Circuit Breaker accordingly (Fig. 5).

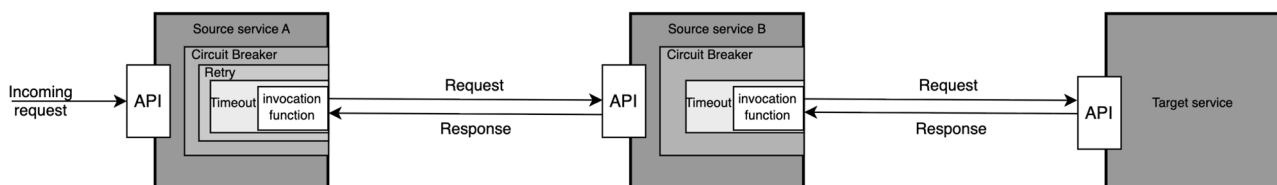


Fig. 5. Chaining the calls – adding an additional service with Circuit Breaker and Timeout in the middle of the flow

Service A sends a request to Service B, which in turn sends a request to the Target service synchronously.

In this two-hop configuration, retries were applied only at the edge (Service A). Service B used a Circuit Breaker with a per-attempt TimeLimiter but no retries. This choice avoids retry-storm amplification and better reflects production guidance for multi-hop RPC chains. We also had to increase the Timeout duration on Service A to 2.25 s (consequently, the worst-case time budget was increased to 8.25 s) and the slow call threshold on Service A to 4.0 s. All other parameters remained identical to the single-hop experiments to maintain comparability.

The percentage of successful requests in this setup only counts those that successfully invoke all three services in the chain and return a response, thereby processing the entire end-to-end flow.

2. Results

While running all the previously mentioned scenarios, we collected the following primary metrics to compare the traditional and predictive models of the Circuit Breaker: request success rate, end-to-end (e2e) 95th percentile (p95) latency, and the fraction of time in the unhealthy state (Half-Open + Open for the traditional model, and only Open for the Prediction model). To count the success rate, we used lock-free accumulators. Latency statistics and the CB state-based unhealthy time were accumulated from internal transition events using aggregated counters. All the aggregations were performed in memory to avoid the I/O overhead and any additional logging during the experiment runs.

Each experiment scenario was run in four iterations. The first was run using the traditional implementation of CB. The other three were run using the Circuit Breaker Prediction model, but with a different threshold value for the calculated rating, which triggered the state change from Open to Closed. We decided to use the following values for that parameter (based on empirical research): 0.60, 0.65, and 0.70.

In Tab. 3, we present the results gathered from the first experimental setup described in Section 1.4.1.

Table 3

Performance results: Circuit Breaker only configuration

Model	Rating Threshold	Success rate (%)	E2e p95 latency (ms)	Unhealthy time (%)
Traditional CB	—	53.4	2520	27.6
Prediction CB	0.6	60.8	2388	21.9
	0.65	59.7	2426	22.7
	0.70	55.1	2483	25.8

Under the basic configuration, both Circuit Breaker versions showed similar trends in stability and latency during load oscillations. However, the predictive model consistently maintained a higher success rate and shorter recovery times. Compared with the traditional CB implementation, it increased the overall success rate by approximately 7–9 percent, reduced the 95th-percentile end-to-end latency by around 6 %, and decreased the total unhealthy time by almost 5 %. These

improvements indicate that the adaptive window mechanism provided a more responsive adjustment to traffic variations, maintaining better throughput even under unstable service conditions.

After adding the Retry pattern usage to the given setup (Section 1.4.2), the following results were obtained (Tab. 4).

Table 4

Performance results: Circuit Breaker and Retry configuration

Model	Rating Threshold	Success rate (%)	E2e p95 latency (ms)	Unhealthy time (%)
Traditional CB	–	74.2	2260	12.8
Prediction CB	0.6	86.1	2142	10.1
	0.65	84.7	2170	10.6
	0.70	79.8	2229	12.1

After adding the retries, as shown in the table, overall service availability increased notably, with the success rate increasing from roughly 74 % to over 86% in the best-performing configuration of the predictive model. In parallel, the unhealthy time dropped from 12.8 % to nearly 10 %, suggesting that the model allowed faster reopening of the circuit once the target service resumed stable operation.

Adding the Timeout pattern to the previous configuration (Section 1.4.3) provided us with the following results (Tab. 5):

Table 5

Performance results: Circuit Breaker, Timeout, and Retry configuration

Model	Rating Threshold	Success rate (%)	E2e p95 latency (ms)	Unhealthy time (%)
Traditional CB	–	83.5	1764	8.8
Prediction CB	0.6	94.6	1639	6.9
	0.65	93.8	1660	7.1
	0.70	90.1	1712	7.9

As expected, the addition of the Timeout pattern reduced the total response time, so the difference between the CB Prediction Model and the traditional CB was almost 7 % in the p95 latency and about a 10–11 % in the success rate.

Finally, introducing the scenario discussed in Section 1.4.4, presented in Tab. 6:

Table 6

Performance results: two-hop chain – Service A to Service B to Target service, Retries only at A

Model	Rating Threshold (A / B)	Success rate (%)	E2e p95 latency (ms)	Unhealthy time (%) at A / B
Traditional CB	–	69.8	3010	19.8 / 17.5
Prediction CB	0.60 / 0.60	81.6	2790	15.7 / 13.2
	0.65 / 0.65	80.2	2831	16.2 / 13.6
	0.70 / 0.70	76.4	2905	17.4 / 14.9

In the two-service chain setup, overall latency and error accumulation increased as expected, but the general pattern remained consistent. It improved the end-to-end success rate by approximately 11-12%, reduced the 95th-percentile latency by about 8%, and shortened the unhealthy periods at both Service A and Service B by up to 3-4%. These results demonstrate that the approach continues to provide value in chained communication scenarios, where slower recovery recognition in any upstream service can lead to additional failures.

In all experiment configurations, different scenarios with varying Rating Threshold values showed that overall success depends on the correctly selected rating limit for transitioning from Open to Closed states of CB. However, even manual tuning of that parameter was intuitive: the higher the threshold, the more strict we want our CB to be towards communication instabilities.

Discussion and conclusions

The conducted study expanded the earlier work on the Circuit Breaker Prediction Model presented in (Hlybovets, & Paprotskyi, 2024) by improving it and validating its performance under more realistic and dynamic conditions.

The model was extended with an adaptive sliding window controller that scales the data window size according to the observed throughput. This increased the model's ability to adapt to workload fluctuations by providing a more accurate data context for calculating service communication performance.

The main concept of this model's calculation algorithm was generalized in the form of communication rating calculation formulas, allowing the use of any substantial number of performance metrics within the algorithm. The pseudocode instructions listing was also included to ensure reproducibility.

Through a series of simulation-based experiments combining the model with other fault-tolerance patterns, it was shown that adaptive estimation of communication performance can improve the CB recovery speed and reduce unnecessary downtime. Across all experiment configurations, the predictive Circuit Breaker implementation demonstrated an improvement in request success rate ranging from 7 % to 12 %, a reduction of 5–8 % in the 95th-percentile end-to-end latency, and a decrease of 4–6 % in total unhealthy operation time compared with the traditional model. The configurations involving two microservices demonstrated that integrating the prediction-driven rating of service communication calculation (1) increases the successful requests rate and reduces request latency, especially when combined with other stabilizing patterns like the Retry and the Timeout. When testing the three-service configuration, the experiment results remained positively stable: the percentage of successful requests remained better in the prediction-based Circuit Breaker, proving its better resilience capabilities in withstanding cascading failures.

As a result of the above-mentioned, we consider the study's goal to have been achieved.

The process of preparing the experiments revealed the role of tuning the parameters and thresholds to achieve optimal fault tolerance performance in this method. The tasks of adjusting thresholds and identifying perfect correlations between various performance metrics are complex for the human mind. It should be possible to do it more effectively by using automated data-driven analytical approaches and various optimization techniques.

The conclusions align with the study (Souza, Neves, & Kimura, 2024), which analyzed industry practices in research papers and surveyed microservice practitioners. It highlighted standard practices, gaps, supporting technologies, and the evolution of fault tolerance. The study confirms the benefits of continuously improving resilience patterns and methods, and emphasizes the growing integration of AI-based approaches into industry resilience strategies.

This points toward the future of prediction-based fault tolerance, where machine-assisted analysis may help reveal complex dependencies among runtime metrics and improve adaptive decision-making.

Authors' contribution: Anatolii Pashko – conceptualization, formal analysis, methodology, supervision, validation, writing – review & editing; Ihor Paprotskyi – conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review & editing; Andrii Hlybovets – conceptualization, methodology, writing – review & editing; Svitlana Terenchuk – conceptualization, methodology, validation, visualization, writing – original draft.

Sources of funding. Funding is partially provided by Taras Shevchenko National University of Kyiv.

References

- De Souza Miranda, F., Santos, D. S. D., Vilela, R. F., Assunção, W. K. G., Santos, R. C. D., & Pinto, V. H. S. C. (2024). A proposed catalog of development patterns for fault-tolerant microservices. In I. Machado, J. Maldonado, T. Conte, E. Canedo, J. Marques, B. B. de França, P. Matsubara, D. Viana, S. Soares, G. Santos, L. Rocha, B. Gadelha, R. Santos, A. Oran, & A. G. Seca Neto (Eds.), *SBQS 2024: XXIII Brazilian Symposium on Software Quality* (pp. 406–416). Association for Computing Machinery. <https://doi.org/10.1145/3701625.3701678>
- Zhang, P., Xiang, L., Song, Z., & Yang, Y. (2025). Adaptive load balancing and fault-tolerant microservices architecture for high-availability web systems using docker and spring cloud. *Discover Applied Sciences*, 7(7). <https://doi.org/10.1007/s42452-025-07320-7>
- Fowler, M. (2014). Circuit Breaker. [martinfowler.com](https://martinfowler.com/bliki/CircuitBreaker.html). <https://martinfowler.com/bliki/CircuitBreaker.html>
- Sun, X., Cui, B., & Cai, Z. (2024). Deep Q-learning based circuit breaking method for micro-services in cloud native systems. In Y. Sun, T. Lu, T. Wang, H. Fan, D. Liu, & B. Du (Eds.), *Communications in computer and information science: Vol. 2012. Computer supported cooperative work and social computing: ChineseCSCW 2023* (pp. 348–362). Springer. https://doi.org/10.1007/978-981-99-9637-7_26
- Souza, V. J. S., Neves, V. O., & Kimura, B. Y. L. (2024). Dependable Microservices in the Kubernetes era: A Practitioners survey. *Journal of Internet Services and Applications*, 15(1), 561–583. <https://doi.org/10.5753/jisa.2024.4000>
- Istio. (2025). *Documentation*. <https://istio.io/latest/docs>
- JetBrains. (2024). *General Development Trends –The State of Developer Ecosystem in 2023 Infographic*. *JetBrains: Developer Tools for Professionals and Teams*. https://www.jetbrains.com/lp/devecosystem-2023/development/#mcrsvc_design_approaches
- Hlybovets, A., & Paprotskyi, I. (2024). Increasing the fault tolerance in microservice architecture. *Cybernetics and Systems Analysis*, 60(3), 480–488. <https://doi.org/10.1007/s10559-024-00689-0>
- Resilience4j. (2025). Introduction. <https://resilience4j.readme.io/docs/getting-started>
- Lewis, J., & Fowler, M. (2014). Microservices. [martinfowler.com](https://martinfowler.com/articles/microservices.html). <https://martinfowler.com/articles/microservices.html>
- Valdivia, J. A., Lora-González, A., Limón, X., Cortes-Verdin, K., & Ocharán-Hernández, J. (2020). Patterns Related to Microservice Architecture: a Multivocal Literature Review. *Programming and Computer Software*, 46(8), 594–608. <https://doi.org/10.1134/s0361768820080253>
- Nygard, M. T. (2018). *Release it!: Design and Deploy Production-Ready Software* (2nd ed.). The Pragmatic Programmers LLC.
- Stack Overflow. (2024). Stack Overflow Developer Survey 2023. <https://survey.stackoverflow.co/2023/#section-developer-experience-developer-experience-processes-tools-and-programs-within-an-organization>

Отримано редакцією журналу / Received: 20.10.25
Прорецензовано / Revised: 18.03.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Анатолій ПАШКО¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-6944-8477
e-mail: aap2011@ukr.net

Ігор ПАПРОЦЬКИЙ¹, асп.
ORCID ID: 0009-0006-1644-2833
e-mail: igorpapr@gmail.com

Андрій ГЛИБОВЕЦЬ², д-р техн. наук, проф.
ORCID ID: 0000-0003-4282-481X
e-mail: a.glybovets@ukma.edu.ua

Світлана ТЕРЕНЧУК³, канд. фіз.-мат. наук, проф.
ORCID ID: 0000-0001-6527-4123
e-mail: terenchuksa@ukr.net

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Національний університет "Києво-Могилянська академія", Київ, Україна

³Київський національний університет будівництва і архітектури, Київ, Україна

УДОСКОНАЛЕННЯ ПРЕДИКТИВНОЇ МОДЕЛІ ПАТЕРНУ CIRCUIT BREAKER ДЛЯ ПІДВИЩЕННЯ ВІДМОВСТІЙКОСТІ МІКРОСЕРВІСІВ

У високонавантажених розподілених системах надзвичайно важливо приділяти особливу увагу надійності комунікації між архітектурними компонентами, що робить відмовостійкість однією з базових властивостей мікросервісної архітектури. Попри наявність різноманітних патернів підвищення відмовостійкості, спрямованих на поліпшення стійкості мікросервісів, жоден із них не є

універсальним для всього спектра можливих відмов у сервісній взаємодії. З огляду на це зберігається постійна потреба в пошуку ефективніших підходів до досягнення відмовостійкості в розподілених програмних системах.

Спираючись на наше попереднє дослідження, у якому було представлено предиктивну модель *Circuit Breaker*, запропонована робота детальніше досліджує її застосування для оцінювання ефективності в поєднанні з іншими поширеними патернами відмовостійкості. Центральним принципом предиктивної моделі є обчислення рейтингу комунікації між двома сервісами, визначеного як опукла комбінація різних метрик ефективності. Отримане значення рейтингу використовується для визначення того, чи може взаємодія із цільовим сервісом бути безпечно відновлена після періоду блокування. Подібно до традиційного патерна *Circuit Breaker*, запропонована модель спрямована на зменшення затримок під час передавання запитів в умовах невизначеного навантаження. Порівняно зі стандартними імплементаціями цього методу модель розроблено з метою скорочення часу блокування викликів між мікросервісами.

Метою статті є вдосконалення моделі *Circuit Breaker* та перевірка її ефективності в поєднанні з іншими патернами відмовостійкості, зокрема *Retry* та *Timeout*. У дослідженні розширено гнучкість *Circuit Breaker Prediction Model* шляхом опису адаптивного механізму ковзного вікна для метрик, що динамічно змінює розмір контексту залежно від навантаження під час роботи застосунку. Також алгоритм обчислення рейтингу було узагальнено, щоб забезпечити можливість використання довільної кількості метрик ефективності. У межах роботи проведено серію симульованих експериментів із різними конфігураціями комунікації між сервісами та комбінаціями патернів відмовостійкості, використовуючи стандартну реалізацію *Circuit Breaker* як базову для порівняння.

Отримані результати показують підвищення продуктивності комунікації насамперед шляхом зменшення затримок у передаванні запитів за однакових умов навантаження. Також розглянуто перспективи застосування принципів цієї моделі до інших патернів відмовостійкості, зокрема з акцентом на можливості використання різноманітних методів оптимізації для підтримки динамічного ухвалення рішень.

Ключові слова: мікросервісна архітектура, розподілені системи, відмовостійкість, патерни відмовостійкості, високе навантаження, *Circuit Breaker*, взаємодія сервісів, адаптація під час виконання.

Один з авторів (А. Пашко) є членом редколегії видання, тому він не брав участі у рецензуванні та прийнятті рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

One of the authors (A. Pashko) is a member of the journal editorial board, therefore did not take part in the peer-review process or in the decision to publish this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.31>

Mykola PYCHTAR, PhD (Ped.), Assoc. Prof.

ORCID ID: 0000-0003-4567-8901

e-mail: nppihtar@gmail.com

National Pedagogical Dragomanov University, Kyiv, Ukraine

Myroslava LYTVYN, Student

ORCID ID: 0009-0006-6792-6053

e-mail: myroslava.lytvyn.25@kse.org.ua

Kyiv School of Economics, Kyiv, Ukraine

STATISTICAL ANALYSIS OF FACTORS AFFECTING THE INCIDENCE OF DIABETES MELLITUS

The seriousness of the problem of diabetes mellitus is explained by the scale of its prevalence. In 2021, approximately 537 million people worldwide were living with diabetes, which accounts for about 6% of the global population. Therefore, the study and analysis of data on the dynamics of diabetes incidence are highly relevant, both for forecasting future trends and for supporting effective government regulation and prevention of the disease's spread.

The aim of this article is to conduct a statistical analysis of diabetes prevalence, to develop models of its dynamics for the city of Kyiv, and to analyze the influence of factors among women, such as the number of pregnancies, blood glucose levels, blood pressure, and body mass index (BMI), on the development of diabetes. Based on these variables, classification methods are applied to build predictive models using the R programming language.

Keywords: *time series, diabetes mellitus, forecasting and modeling, k-NN, Random Forest, logistic regression, algorithm accuracy.*

AMS 2020 classification: 60G10, 60G15, 91-10.

Introduction

Diabetes is a metabolic disorder that affects many people worldwide. Every year, the incidence rates are rising at an alarming rate. If treatment is delayed, diabetes-related complications in many vital organs of the body can lead to a fatal outcome. It is very important to detect the disease at an early stage, which can halt its progression before complications arise. The International Diabetes Federation (IDF) has stated that currently, over 537 million people worldwide have diabetes. Over the last few years, the number of people with diabetes has risen sharply, making it a global threat. Unfortunately, diabetes now consistently ranks third among the leading causes of death.

Because early detection is vital to halting the disease's progression before fatal complications arise, researchers have increasingly turned to machine learning as a viable solution for automated prediction and diagnosis. Various robust frameworks have been proposed to address the complexities of medical data. The use of data analysis and machine learning methods can significantly improve the diagnosis of diabetes and help to take timely measures for its prevention and treatment. Due to the relevance of this direction, machine learning methods have been actively developed in recent years (Aguilera-Venegas et al., 2023; Kangra, & Singh, 2023; Olisah, Smith, & Smith, 2022; Orlova et al., 2026). Works in this area demonstrate the high efficiency of the classification models for various data sets. Recent advancements also include specialized deep learning architectures like EchoceptionNet, which captures both temporal and spatial feature interactions, and the comparative use of Decision Tree Classifiers and Artificial Neural Networks (ANN), with the former often demonstrating superior accuracy in identifying diabetic conditions based on biochemical indicators like HbA1c (Iftikhar et al., 2025). In (Livinska, Skrypnik, & Galán-García, 2024), the accuracies of different machine learning algorithms were compared for diagnosing diabetes on binary data that do not include laboratory tests, and can be obtained by patient questionnaires.

This work examined data on people with diabetes in Kyiv and conducted a descriptive analysis to calculate key statistical characteristics; it also investigated two models based on ARIMA processes and linear regression for the percentage of children with diabetes. Using these, forecasts were constructed for the next 5 years. All this is described in more detail in the first chapter. The second chapter is devoted to the analysis of the diabetes incidence dataset in women, along with 9 other variables – factors that will be the focus of further work. The distribution of variables was analysed using histograms and box-and-whisker plots, and the relationships between them were investigated using a correlation matrix. The Student's t-test was used to compare the mean values of variables in the two groups (healthy, diabetic). Three machine learning methods were applied to classify patients into groups: k-nearest neighbours, logistic regression, and random forest. To compare these methods, the following quality measures were calculated: accuracy, sensitivity, specificity, and other indicators.

1. Statistical analysis and forecasting of the incidence of diabetes mellitus in Kyiv

This section is devoted to the analysis of data on the incidence of diabetes in Kyiv. Based on the data from the website (IDF, 2024). The data is obtained for 1995-2016 (no data for 2014, due to the annexation of Crimea and the east of the country. Since 2017, data for all regions have not been available, and their collection has not been continued.

Autocorrelation, also known as serial correlation, is the relationship between a signal (time series) and itself at different points in time. Informally, autocorrelation involves comparing two observations of the same signal at different times to assess how similar they are. Let's construct an autocorrelation function for the time series of the percentage of children with diabetes. To do this, our dataset must be fully known, so we impute the missing value for 2014 using the average of the nearest neighbours. Mathematically, the autocorrelation function provides a measure of the signal's characteristics in

the time domain, which is used to describe the relationship between the signal's value at one point in time and its value at another point in time (Fig. 1).

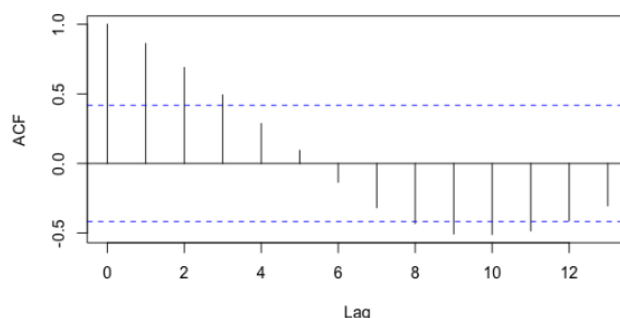


Fig. 1. Autocorrelation

To analyse trends in morbidity, time series analysis was used, with time series classified as stationary or non-stationary. Non-stationary data, which is typical of medical indicators, is presented as the sum of the trend $f(t)$ and the stationary process

$$X_t : Y_t = f(t) + X_t, t \in \{0, 1, \dots, n-1\}.$$

The ARIMA (Autoregressive Integrated Moving Average) model has been selected as the primary forecasting method; it is characterised by the following parameters (p, d, q) . Mathematically, the model for a non-stationary time series takes the form():

$$\Delta^d X_t = c + \sum_{i=1}^p a_i \Delta^d X_{t-i} + \sum_{j=1}^q b_j \epsilon_{t-j} + \epsilon_t,$$

where Δ_d is the difference operator of order d , and ϵ_t is a "white noise".

The study is based on morbidity data for the city of Kyiv for the period 1995–2016. The key indicators identified are:

- p is the percentage of children among all patients with diabetes mellitus (DM).

- fp is the proportion of children among those newly diagnosed with diabetes. To analyze the dataset, R programming language for statistical computing and data visualization was used.

Statistical analysis revealed significant variability in the fp value (coefficient of variation 36.0%) compared with p (19.6%).

Modelling and forecasting results. Two forecasting models have been developed for the p indicator. The first model is $ARIMA(0, 2, 1)$ model: The coefficients were selected using the Akaike Information Criterion (AIC). The model indicates an upward trend in the proportion of sick children over the next five years.

The second model was a linear regression model; the least-squares method (LSM) was used to estimate the trend parameters. The analysis of the parameters revealed that the coefficients were statistically significant (p -value < 0.05); however, the model explains only 20% of the variation in the series ($R = 0.2091$).

A comparison of the results reveals a discrepancy: the ARIMA model predicts a further rise in incidence, whilst linear regression indicates a slight downward trend, highlighting the advantage of dynamic models for short-term forecasting of health indicators.

2. Diabet prediction based on medical records

This section is devoted to the analysis of the diabetes prevalence dataset in women with 9 other variables: number of pregnancies, blood glucose, blood pressure, body mass index (BMI), genetic predisposition to diabetes, insulin, age, and presence of diabetes from (Smith et al., 1988). The dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the dataset is to diagnose whether a patient has diabetes or not based on certain diagnostic measurements included in the dataset. All patients in this dataset are women aged at least 21 years.

The contrast matrix (last column) in Fig. 2 visually demonstrates differences in diabetes incidence across several characteristics, namely the number of pregnancies, glucose level, body mass index (BMI), and age. However, these observed differences require further statistical verification to confirm their significance.

To analyze the relationships between variables in more detail, a matrix of pairwise correlation coefficients is constructed. This matrix reflects the degree of linear dependence between each pair of factors. The lower triangular part of the matrix contains the numerical values of the correlation coefficients, while the upper part provides a visual representation using circles of varying intensity: with blue shading indicating a positive outcome (with diabetes) and red shading indicating a negative outcome (without diabetes).

The Student's t-test is one of the basic statistical methods used to compare means between groups of data and to identify statistical differences. Comparing means allows you to determine whether the selected groups are significantly different in terms of the indicator or outcome under study. In the context of medical research, comparing mean values between groups, such as patients and healthy individuals, can help identify risk factors that contribute to the development of diseases. This approach allows you to determine which indicators or properties of the body differ between groups and may be associated with the onset and progression of specific diseases. This information can be important for developing

strategies to prevent and manage disease risks. Suppose we want to compare two sample means: X_1 , calculated from a sample with standard sd_1 deviation and sample size n_1 and X_2 calculated from a sample with standard deviation sd_2 and sample size n_2 . Then, we formulate the following hypotheses:

- H_0 : In the population, there is no difference between the mean values.
- H_1 The means in the population are not equal (the alternative hypothesis).



Fig. 2. Matrix of contrasts and pairwise coefficients of correlations depending on whether there is diabetes

First, we assume that the null hypothesis H_0 holds. This means that in the population, there is indeed no difference between the mean values X_1 and X_2 . If this is the case, then in repeated experiments (drawing samples and calculating the difference between the two sample means), the difference $X_1 - X_2$ should be close to zero. This is compared with the tabulated value of the Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom, which depends on the chosen significance level. In R, we will use the function `t.test` and compare the obtained p-value with the significance level 0.05 (Tab. 1). If the p-value is greater than 0.05, there are no grounds to reject the null hypothesis at the 0.95 confidence level (the means are statistically equal). Otherwise, the null hypothesis is rejected (the means differ with probability 0.95).

Table 1

Statistical significance of the difference between healthy and sick people			
Factor	t	p-value	Is the difference stat.significant?
Pregnancies	-5.907	3.411e-09	yes
Glucose 1	-13.752	2.2e-16	yes
BloodPressure	-1.7131	0.08735	no
Skin Thickness	-1.9706	0.05	no
Insulin	-3.3009	0.001047	yes
BMI	-8.6199	2.2e-06	yes
DPF	-4.5768	6.1e-06	yes
Age	-6.9207	1.202e-11	yes

For the classification of diabetes, three machine learning methods were applied: k-Nearest Neighbors, Random Forest, and Logistic Regression.

The k-Nearest Neighbors (k-NN) method is one of the simplest machine learning algorithms used for classification and regression. It is based on the assumption that similar objects have similar outcomes. In this method, there is no explicit training process. The algorithm simply stores all the training data along with their labels (in the case of classification) or values (for regression). To predict the class or value of a new object, the algorithm computes the distances between this object and all objects in the training set. For applying the k-NN method, all values were normalized (for each variable, the mean was subtracted and divided by the standard deviation). Next, the entire dataset was split into training and testing sets in an 80 – 20 % ratio. For the nearest neighbors method, we use the parameter $k = 5$. The `knn()` function is used to build the classification model, and a confusion matrix is constructed on the test set.

Random Forest is a commonly used machine learning algorithm that combines the outputs of multiple decision trees to achieve a single result. A random forest is an ensemble of decision trees, where each tree depends on a certain set of random variables. To apply the Random Forest method, all data are split into training and testing sets in an 80 – 20 % ratio. Using the `randomForest()` function, a classification model is created, and a confusion matrix is constructed.

Logistic regression is a statistical method used to model the relationship between a dependent variable (response) and one or more independent variables (predictors). Its primary purpose is to predict the probability of an event with a binary outcome, for example, success or failure, a positive or negative diagnosis, and so forth. The dependent variable (Y)

in logistic regression is categorical, most commonly binary ($Y \in \{0, 1\}$). In our case, 0 indicates no Diabetes, 1 indicates the presence of Diabetes.

The model estimates the probability p that the event will occur: $p = P(Y = 1|X)$.

To transform probabilities into a form suitable for linear modeling, logistic regression applies the logit function (the natural logarithm of the odds):

$$\text{logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k,$$

where $\beta_0, \beta_1, \dots, \beta_k$ are model parameters estimated from the data.

Using the `glm()` function, we obtained the following logistic regression model for our dataset:

$$\text{logit}(p) = -8.404 + 0.123X_1 + 0.025X_2 - 1.33X_3 + 0.001X_4 + 0.089X_6 + 0.945X_7 + 0.014X_8.$$

Based on this model, we classified individuals as either healthy or diabetic. The classification was performed using the following R command, which assigns an observation to the diabetic group if the predicted probability exceeds 0.5.

To evaluate the model's performance, a confusion matrix was generated using a table. This matrix provides a summary of correctly and incorrectly classified cases, allowing for an assessment of the model's accuracy.

3. Classification performance of quality metrics

Accuracy is one of the most widely applied quality metrics, defined as the proportion of correctly classified instances.

Sensitivity is used to assess the ability of a test or model to correctly identify positive cases. In other words, sensitivity indicates the proportion of truly positive cases that were correctly classified as positive. From the table, it can be seen that the method providing the highest sensitivity is Random Forest.

Specificity is used to evaluate the quality of a test or model. While sensitivity measures how well the test detects positive cases, specificity reflects how well it identifies negative cases. In other words, it is the ability of the test to correctly recognize the absence of a certain feature or disease. According to the table, Logistic Regression demonstrates the highest specificity.

Precision characterizes the degree of consistency of results when measurements or experiments are repeated. In other words, precision shows how closely the results of repeated measurements of the same quantity agree with each other. The highest level of precision is achieved by Random Forest.

Negative Predictive Value (NPV) indicates how reliable a negative test result is. In other words, NPV answers the question: "If the test result is negative, how likely is it that the individual truly does not have the condition?" The highest NPV is achieved by Logistic Regression (see Tab. 2).

Statistical significance of the difference between healthy and sick people

Table 2

	k-NN	Random Forest	Logistic Reg
Accuracy	0.7208	0.7662	0.78255
Sensitivity	0.8617	0.8842	0.58209
Specificity	0.5000	0.5763	0.89
Precision	0.7297	0.7706	0.73933
NPV	0.6977	0.7556	0.79892

Discussion and conclusions

This paper examined and analyzed type 2 diabetes mellitus. First, the prevalence of diabetes in Kyiv from 1995 to 2015 was considered. Using ARIMA processes and a regression model, the short-term incidence of diabetes in Kyiv was forecasted. Then we investigated and described another dataset consisting of one categorical variable (presence of diabetes) and eight numerical variables. To explore the data in more detail, graphical visualization methods were used. This allowed us to assess the distribution of each variable and detect potential outliers. To identify relationships between variables and their dependence on the presence of diabetes, a correlation matrix was built. Then, to test whether the mean values of indicators differ between diabetic and healthy individuals, a statistical test – Student's t-test – was applied. For patient classification into groups, three machine learning methods were used: k-nearest neighbors, logistic regression, and Random Forest. Logistic regression achieved the highest classification accuracy (0.78). The method with the highest sensitivity was Random Forest (0.88). Logistic regression achieved the highest specificity (0.89).

The conducted analysis shows that diabetes incidence is influenced by multiple factors, with glucose level, BMI, age, and number of pregnancies being statistically significant predictors. The time series modeling for Kyiv revealed that the ARIMA model is more suitable for short-term forecasting of non-stationary medical data, while linear regression provides less reliable results.

The comparison of classification methods demonstrated that Random Forest achieves the highest sensitivity and precision, making it more effective for detecting diabetes cases. Logistic regression showed higher specificity and negative predictive value, indicating better performance in identifying healthy individuals. The k-NN method yielded comparatively lower accuracy.

Overall, the results confirm that combining statistical methods with machine learning improves both forecasting and early diagnosis of diabetes, which is important for preventive healthcare and decision-making.

Authors' contribution: Mykola Pychta – conceptualization, methodology; Myroslava Lytvyn – software, empirical data collection and validation, empirical research, analysis of sources, preparation of literature review.

Acknowledgments. The authors are grateful to the Armed Forces of Ukraine. Without their heroic contributions, this and many other studies would not be possible.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Aguilera-Venegas, G., López-Molina, A., Rojo-Martínez, G., & Galán-García, J. L. (2023). Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *Journal of Computational and Applied Mathematics*, 427, 115115. <https://doi.org/10.1016/j.cam.2023.115115>
- IDF. (2024). *The the diabetes atlas*. <https://diabetesatlas.com.ua/ua/kyiv>.
- Iftikhar, K., Javaid, N., Ahmed, I., & Alrajeh, N. (2025). A novel explainable deep learning framework for accurate diabetes mellitus prediction. *Applied Sciences*, 15(16), 9162. <https://doi.org/10.3390/app15169162>
- Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3), 1728–1737. <https://doi.org/10.11591/eei.v12i3.4412>
- Livinska, H., Skrypnyk, D., & Galán-García, J. L. (2024). Comparative study of machine learning algorithms for diabetes detection using binary data. *Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics*, 78(1), 119–127. <https://doi.org/10.17721/1812-5409.2024/1.23>
- Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773. <https://doi.org/10.1016/j.cmpb.2022.106773>
- Orlova, N., Tkachenko, O., Herasymuk, K., Palamar, I., & Sereda, N. (2026). Epidemiological situation of type 2 diabetes mellitus in Ukraine: challenges and priorities for public health management. *Ukraine. Nation's Health*, 1, 28–37. <https://doi.org/10.32782/2077-6594/2026.1/03>
- Smith, J., Everhart, J., Dickinson, W., Knowler, W., & Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the symposium on computer applications and medical care* (pp. 261–265). IEEE Computer Society Press. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

Отримано редакцією журналу / Received: 23.10.25

Прорецензовано / Revised: 13.01.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Микола ПИХТАР, канд. пед. наук, доц.

ORCID ID: 0000-0003-4567-8901

e-mail: npihtar@gmail.com

Український державний університет імені Михайла Драгоманова, Київ, Україна

Мирослава ЛИТВИН, студ.

ORCID ID: 0009-0006-6792-6053

e-mail: myroslava.lytvyn.25@kse.org.ua

Київська школа економіки, Київ, Україна

СТАТИСТИЧНИЙ АНАЛІЗ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА ЗАХВОРЮВАНІСТЬ ЦУКРОВИМ ДІАБЕТОМ

Серйозність проблеми діабету фахівці пояснюють рівнем поширення хвороби. У 2021 р. приблизно 537 млн людей на всій Землі хворіли на цукровий діабет, що становить приблизно 6 % населення. З огляду на це дослідження та аналіз даних за динамікою захворювання цукрового діабету є актуальними як з позиції побудови прогнозу захворюваності, так і з позиції можливості державного регулювання та запобігання поширенню хвороби. Метою проєкту є статистичний аналіз захворюваності на цукровий діабет, побудова моделей динаміки захворюваності для міста Києва, а також аналіз впливу таких показників у жінок: кількість вагітностей, рівень глюкози в крові, кров'яний тиск, індекс маси тіла (ІМТ) тощо, на цукровий діабет. До цих змінних застосовано методи класифікації для побудови прогнозних значень за допомогою мови програмування R. Відповідно до поставленої мети в роботі розв'язано такі основні завдання: відшукування у відкритому доступі даних щодо цукрового діабету; опрацювання відомих підходів до аналізу таких даних; обчислення базових основних статистичних показників захворюваності та візуалізація даних за допомогою мови програмування R; знаходження тренду захворюваності; побудова моделі та прогнозних значень на основі процесів ARIMA та регресійного аналізу. За допомогою критерію Стьюдента здійснено перевірку на відмінності в середньому таких показників для здорових і хворих на цукровий діабет, як кількість вагітностей, рівень глюкози в крові, кров'яний тиск, ІМТ тощо. Прогнозування захворювання на діабет реалізовано за допомогою трьох алгоритмів машинного навчання, а саме: k-NN; Random forest; логістичної регресії. Виконано обчислення мір якості алгоритмів і порівняння точностей класифікації.

Ключові слова: часовий ряд, цукровий діабет, прогнозування та моделювання, k-NN, Random forest, логістична регресія, точність алгоритмів.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 004.94:004.89:378.1:681.5

DOI: <https://doi.org/10.17721/1812-5409.2026/1.32>

Tetiana SAVCHENKO¹, PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0002-8884-5360
e-mail: tsavchenko@ukma.edu.ua

Nataliia LUTSKA², DSc (Engin.), Prof.
ORCID ID: 0000-0001-8593-0431
e-mail: lutskanm2017@gmail.com

Lidiia VLASENKO³, PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0002-2003-6313
e-mail: vlasenko.lidia1@gmail.com

Ivan PARKHOMENKO⁴, PhD (Engin.), Assoc. Prof.
ORCID ID: 0000-0001-6889-9284
e-mail: ivan.parkhomenko@knu.ua

¹Kyiv National University "Kyiv-Mohyla Academy", Kyiv, Ukraine

²National University of Food Technologies, Kyiv, Ukraine

³State University of Trade and Economics, Kyiv, Ukraine

⁴Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

DIGITAL TWIN IN THE STRUCTURE OF THE MODERN EDUCATIONAL ENVIRONMENT

In today's world, the advantage of digital twins is the ability to create copies of various physical objects, systems or learning environments, etc. This is especially true in engineering disciplines, where models of digital twins of a given accuracy provide research and experimentation without physical risks and significant financial costs. In the context of the educational environment, this means ensuring the effective functioning of virtual educational laboratories equipped with real equipment. The article develops an educational environment using a digital twin of a SMART laboratory, which functions as a cyber-physical system with end-to-end integration of a sensor network, actuators and a virtual model of the room. The purpose of the work is to theoretically substantiate and experimentally verify the educational environment, that includes a digital twin, focused on improving energy efficiency, comfort and safety of educational spaces. The methodological basis of the study includes system analysis, mathematical modelling of microclimate dynamics, as well as imperative control logic using typical regulators.

Within the framework of the proposed architecture, a microclimate model has been developed that combines the subsystems of monitoring, forecasting and adaptive control. Actualization of the real and virtual state is carried out in real time, which provides the possibility of preliminary testing of control actions and assessment of their impact without risk to the material base. A series of simulations and experiments confirmed the ability of the developed system to reduce resource consumption and stabilize microclimate parameters under the influence of external disturbances. At the same time, the system maintains pedagogical flexibility, supporting various formats of organizing the educational process.

The obtained results testify to the scientific novelty of the proposed approach, which consists in combining models and control systems with educational scenarios in a single software environment. The practical significance of the work is determined by the possibility of scaling the developed solution to other types of laboratories and academic disciplines, as well as the promoting the formation of digital competencies of future specialists in the field of information technology and cyber-physical systems.

Key words: educational environment, SMART laboratory, digital twin, monitoring, KPIs, microclimate.

AMS 2020 classification: 68M99, 28-04, 93B45, 93C05.

Introduction

Digital twins (DTs) are one of the key technologies of the modern cyber-physical paradigm, enabling the integration of a physical object with its virtual representation in real time. They allow monitoring, forecasting, optimization, and management of objects and processes based on the constant interaction between physical and virtual environments (Thelen et al., 2022). This technology has become widespread in industry, energy, medicine, transport, as well as in Smart environments.

In Smart Home systems, DTs enable the effective solution of tasks related to automatic control of climatic conditions, energy consumption, safety, maintenance of engineering systems and ensuring a comfortable environment for residents. They function as an integration platform that integrates data from different sensors, analytical algorithms, and decision-making systems in real time. The use of digital twins in such systems allows you to optimize resource consumption, increase the level of comfort for users and ensure a prompt response to changes in the environment. At the same time, the use of DTs in the learning environment, including for building Smart labs, has a number of specific requirements. Specifically, there is a need to simulate typical educational processes, ensure the interaction of students with virtual and real equipment, the ability to scale experiments and provide a safe environment for learning. In this context, the digital twin acts not only as a means of automation, but also as a didactic tool for the development of digital, analytical and engineering competencies.

In the current conditions of digitalization of education, there is a need to introduce new forms of organization of laboratory practice, focused on accessibility, personalization and flexibility of the educational process. This problem becomes especially relevant in connection with the development of distance and hybrid learning, when direct access to physical laboratory equipment is limited. DTs enable the educational process to be implemented in the format of an interactive simulation that reproduces real conditions and provides real-time feedback. In addition, the introduction of DTs in training laboratories contributes to increasing the efficiency of the use of educational, technical and energy resources, which is especially important in the context of sustainable development and rational consumption. In particular, due to the integration with the engineering infrastructure of the premises, it is possible to implement the control of microclimate, energy consumption, security and technical diagnostics systems in an automated mode.

© Savchenko Tetiana, Lutska Nataliia, Vlasenko Lidiia, Parkhomenko Ivan, 2026

This approach is consistent with strategic documents, including the European Union's Digital Education Action Plan (2021–2027) (European Commission, 2023), which aims to develop high-quality, inclusive and accessible digital education in Europe. It is also in line with the Strategy for Digital Transformation of Education and Science of Ukraine (Ministry of Education and Science of Ukraine, 2025a), which provides for the modernization of educational infrastructure and the introduction of digital technologies into the educational process. In addition, the initiative is in line with the objectives of the Horizon Europe program, which supports research and innovation in the field of digital learning and the development of digital competences (the general objectives of the Horizon Europe program are confirmed in the context of Ukraine's innovation activities, in particular, in the Digital Development Strategy for Innovation until 2030 (Ministry of Education and Science of Ukraine, 2025b; Diia, 2025). These documents emphasize the importance of integrating digital technologies into the field of education to ensure its quality, accessibility and compliance with the modern requirements of the digital society.

The article considers the conceptual foundations of creating a digital twin of the Smart Laboratory, proposes the architecture of the model, and identifies the key technical, software and educational aspects of its implementation in the educational environment of a higher education institution.

The purpose and objectives of the study are the development of an educational environment in the form of a cyber-physical SMART laboratory focused on ensuring the adaptive, safe and resource-efficient functioning of the learning space, as well as the formalization and implementation of DT as a key tool for modeling, monitoring and supporting real-time decision-making.

The object of research is an educational SMART laboratory with laboratory equipment and technical means of automation.

Building a digital twin begins with the choice of architecture (Sai et al., 2024). They are usually multi-level. The literature most often considers the following five levels: Reference Architecture Model for Industry 4.0 (RAMI 4.0) (Junior, 2025) (Tab. 1).

Table 1

Levels of CD architecture

Level	Function	Technologies/Facilities	Literary sources
Physical Layer	Real physical objects that will have a digital twin	Sensors, controllers, actuators, boards	(Herrero, & Dasari, 2025; Hu et al., 2021; Piromalis, & Kantaros, 2022)
Communication Layer	Data transfer between physical devices and the digital layer	Protocols: MQTT, OPC UA, Modbus, Wi-Fi, USB	(Kober et al., 2024; Raza, et al., 2024)
Digital Model Layer	Digitized attributes, metadata, semantics of objects	JSON/XML, Digital Twin Model (DTD), Device Description Language (DDL), RDF, OWL, SPARQL	(Schlenger et al., 2025; Ramonell, Chacón, & Posada, 2023; Xia et al., 2022)
Cyber Layer	Calculations, algorithms, big data, risk management, optimization and predictive analytics	Cloud/on-premises computing, uses scalable and distributed computing technologies such as AWS, Azure Databricks, Hadoop and Google Analytics, AI models, rule engines, stimulators	(Xie et al., 2025; Gautam et al., 2025; Delgado, & Oyedele, 2021).
Application Layer	User interfaces, analytics, visualization, interaction with the educational environment	LMS, Web UI, Analytics Panels, REST API	(Saif, RazaviAlavi, & Kassem, 2024; Tuhaise, Tah, & Abanda, 2023)

It should be noted that each of the considered layers can have its own implementations in the DT SMART laboratory. The Physical Layer may contain sensors specific to SMART buildings (Peldon et al., 2024; Huang, Zhang, & Zeng, 2022), power supplies, oscilloscopes, mockups of instructional circuits, etc. The Communication Layer provides data transfer to the backend. The server (physical, cloud, or edge – depending on the application) is responsible for the complete DT simulation, including data collection, processing, and analysis, as well as performing calculations that support decision-making processes and providing services to users (Weil et al., 2023). The Digital Model Layer typically focuses on the digital representation of objects and the establishment of semantic relationships. Ontologies are prevalent in scientific literature for working with large arrays of heterogeneous data. Their advantages include: describing complex objects, such as digital twin components, by formalizing their structure (classes, properties) and interrelationships; ensuring integration with AI; and the fact that most ontologies are based on open standards and languages (OWL, RDF), which significantly facilitates their integration into cross-platform systems (Vlasenko et al., 2022; Lutska, et al., 2022). The digital layer provides a virtual representation of physical objects by formalizing their geometric and other properties in a digital format that can be stored and managed on a variety of electronic media. In contrast to the Digital Model Layer, the Cyber Layer utilizes scalable and distributed computing technologies. The SMART lab's DT Application Layer can be responsible for the user interface, graph viewing, tasks, progress analytics, and more (Van der Valk et al., 2024).

Depending on the tasks set at different stages of construction and levels of DT, different models are used to describe approaches to its creation, structure and functioning. Among the most used models are the following:

- mathematical models that are built on the basis of physical laws, material, energy balances, etc. (Jia et al., 2024; Nezzi et al., 2025);
- data-driven model based on statistical data, ML algorithms that learn on historical/real data (these include regressions, neural networks, decision trees) (Beke et al., 2021);
- hybrid models that combine mathematical models and machine learning, which improves adaptability and increases accuracy (Lu et al., 2024; Yu, Liu, & Li, 2020; Luo et al., 2020);

- agent-based models, which consist of a set of autonomous agents with certain behaviors, rules, and interactions (Pretel et al., 2024);
- ontological models to describe the semantics of objects (Roper, 2022);
- Knowledge graph-based models for representing DT as nodes and edges (Meyers et al., 2022).

The ISO 23247 series of standards forms a common glossary, conceptual model and multi-layer digital twin architecture for production systems, describing domains, entities and data flows throughout the full lifecycle of assets, thereby ensuring unified data exchange between physical and virtual components of production. According to a systematic review (Semeraro et al., 2021), such standardization enhances interoperability, simplifies the integration of IoT devices and enables real-time predictive maintenance and process optimization, and generalized analysis (Sharma et al., 2022) demonstrates additional benefits, including improved simulation accuracy, reduced design time, and support for physics-data-analytics feedback-based decision-making. In addition, the review (Javaid, Haleem, & Suman, 2023) confirms that the application of ISO 23247 minimizes the risks of testing new process scenarios and contributes to the safe transfer of the developed models to production lines. At the same time, the authors of these works and practitioners note a number of shortcomings: the focus of ISO 23247 mainly on the discrete manufacturing sphere, incomplete coverage of cybersecurity and semantic interoperability issues, as well as significant implementation costs and high staffing requirements, which limits the scaling of the standard to cross-industry ecosystems and small and medium-sized businesses (Wallner et al., 2023).

Building digital twins for educational laboratories has its own characteristics related to the educational context of their use. In such laboratories, it is important not only to ensure technical integration with physical devices and engineering infrastructure, but also to create conditions for simulating various experiments, accumulating and analyzing data, and interactive interaction of students with the laboratory environment. The DT of the training laboratory acts as a platform for the development of competencies in the field of the Internet of Things (IoT), automation, analytics and cyber-physical systems, contributing to the approximation of the educational process to the real conditions of modern engineering activities.

In the investigated educational lab, which is integrated with IoT technology, it is intended to monitor and maintain a wide range of variables to ensure effective, safe and adaptive educational interaction. These variables can be collected in real time, simulated with the digital environment, compared with DT model data, output to the digital interface of the system, providing visualization, fault forecasting, process adaptation and remote access to the laboratory. The laboratory model as part of the digital twin also has a simulation interface to study its features. To provide feedback from the model, elements of recommendations and forecasting are used, which are aimed at improving the key performance indicators (KPIs) of the laboratory.

1. General system architecture

To build a conceptual model (Fig. 1) of the educational digital environment, it is necessary to consider and analyze the physical and virtual components. The Physical World includes the object of study: an educational SMART laboratory equipped with sensors (measuring parameters affecting room comfort and safety), along with equipment and actuators that influence the efficiency of its functioning. External factors affecting the facility's operation include equipment usage, the presence of students and teachers in the lab, and the weather. The components of the Physical World interact with the virtual component.

The Virtual World comprises External Systems that ensure enhanced functionality and interaction between the DT and the Physical World. These External Systems include network exchange protocols between sensors and the digital environment, such as MQTT (running on TCP/IP) and Modbus, as well as databases. The DT itself consists of three main components: a Data Manager (which receives heterogeneous data from sensors or the Monitoring system), a Models Manager (which includes a mathematical model of the object), and a Services Manager (responsible for managing DT functions). Within the DT structure, in addition to the virtualization module, a component responsible for coordinating the physical and digital environments is implemented.

The collection and transmission of data from real sensors to the processing system is provided through the Monitoring Manager. Based on the collected information, the Decision-Making Manager generates appropriate actions that are transmitted to the level of executive devices. The Simulation Manager, in turn, is responsible for creating and maintaining simulation scenarios in a virtual environment. In this work, the implementation of the monitoring and DT subsystems is performed by the Node-RED tool for visual programming of data streams (Savchenko et al., 2025).

In terms of the RAMI 4.0 reference model presented in Tab. 1, the Physical World corresponds to the Physical Layer, including sensors, actuators, and laboratory equipment. The communication between physical devices and the DT infrastructure is implemented at the Communication Layer. The Models Manager represents the Digital Model Layer, while the computational services for simulation, forecasting, and KPI evaluation correspond to the Cyber Layer. Finally, visualization dashboards, user interaction tools, and learning interfaces belong to the Application Layer.

Thus, the developed architecture reflects the bidirectional integration of Physical World (laboratory equipment, sensors, actuators) and Virtual World (Data, Models, Services Manager), complemented by external factors and components of the educational digital platform (Fig. 1).

2. External data sources

The paper examines environmental and safety variables in the educational Smart Laboratory, which affect both the comfort and safety of staying in the room, and the efficiency of laboratory equipment. Data is collected by sensors and transmitted to a single digital environment in real time for visualization, analysis and adaptation of working conditions. Safety indicators are critical for protection of users and equipment. They allow you to quickly respond to emergencies, prevent accidents and keep an event log for audit. For each variable, the terms of reference are given: task, devices, control influences and operating range (Tab. 2, 3).

3. Methodology for Environment Assessment Based on KPI Criteria

To ensure the safe, comfortable, and energy-efficient operation of the digital environment, a multi-criteria system for assessing the condition by key performance indicators (KPIs) is proposed. A reasonable choice of criteria is based on modern regulatory documents [DSanPiN, ISO, WHO] and practical requirements for the microclimate and operational safety of technical equipment.

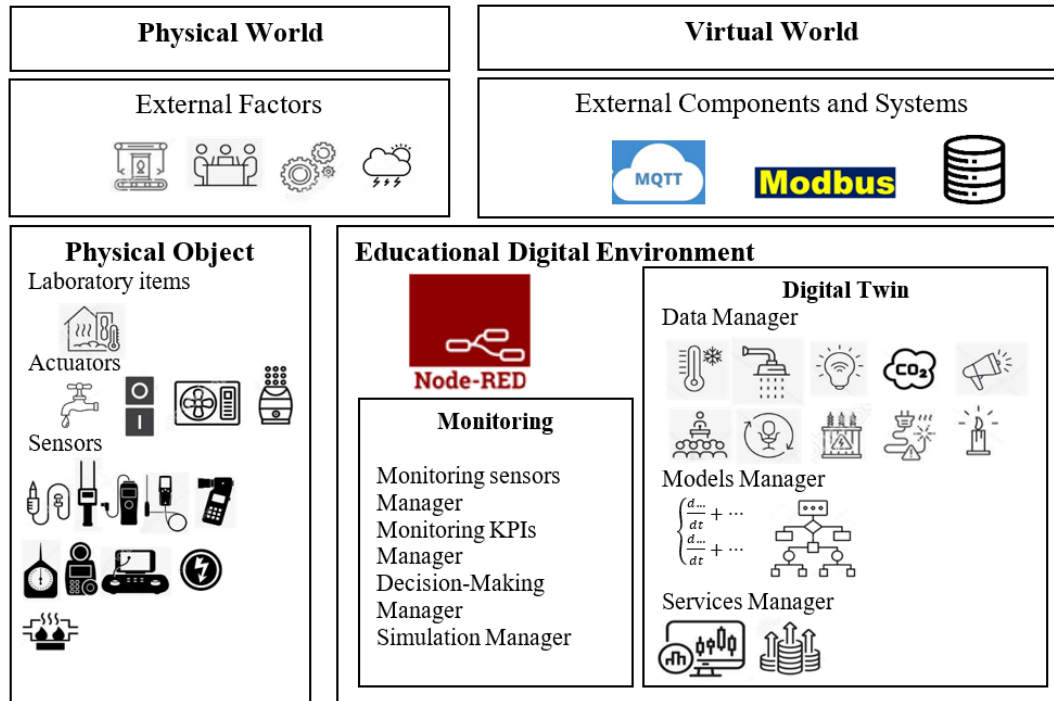


Fig. 1. A library's DT architecture

Table 2

The main variables affecting the comfort and safety of staying in the room

Variable	Problem statement	Sensor type	Control actions	Regulatory limits
Air temperature	Ensuring a comfortable microclimate and preventing overheating of equipment	Temperature sensors	Ventilation, air conditioning, notification control	+20...+24 °C (winter), +23...+26 °C (summer)
Air humidity	Humidity control for electronics accuracy and comfort	Humidity sensors	Switching on the humidifier / ventilation, logging	Optimal: 40-60%, acceptable: 30-70%
Illumination	Providing the appropriate level of lighting for visual work	Light sensors	Lighting adjustment, automatic on/off	300–500 lux (general works), > 500 lux (precision works)
Air quality	Monitoring CO ₂ , dust, and volatile organic compounds (VOCs) to provide fresh air	Air quality sensors	Ventilation control, alarms, logging	CO ₂ : < 1000 ppm; TVOC: < 500 ppb; PM2.5: < 25 µg/m ³
Noise (sound pressure)	Noise control for a comfortable learning environment	Acoustic sensors	Noise fixation, alarm, logging	Up to 50 dB (normal for classrooms)
Vibration	Detection of vibrations of equipment or structures	Vibration sensors	Emergency shutdown, alarm, logging	<0.005–0.012 m/s ² (depending on equipment sensitivity)

Table 3

Main variables affecting the efficiency of equipment

Variable	Problem statement	Sensor Type (Examples)	Control actions	Regulatory limits
User presence	Tracking the entrance to the laboratory within the allowed time. Ensuring the safety of equipment operation	Motion sensors	Presence logging, signaling in case of violation of access mode	User access only at allowed times
Overheating of equipment	Overheating detection to prevent equipment failure	Temperature sensors	Notifications, Auto Shutdown, Logging	The temperature should not exceed 70–80 °C (depending on the equipment)
Short circuit (overload)	Monitoring of the power grid to avoid emergencies	Current/Voltage Sensors	Emergency shutdown, alarm, logging	The current and voltage should not exceed the permissible values for the equipment (usually 10–16 A and 180–250 V)
Smoke and flame control	Prompt detection of fire hazards	Smoke and flame detectors	Emergency Power Off, Alert, Alarm	The presence of smoke or flames is a critical situation

The following KPIs were identified in the work:

✓ The comfort index (*Comfort Index*) determines how the environmental parameters correspond to those recommended for normal human activity. The set of controlled variables includes:

- air temperature (T , °C): recommended range of 20–26 °C;
- relative humidity (H , %): recommended range of 30–60 %;
- light level (L , lx): at least 300 lux;
- air quality (CO_2 level, ppm): CO_2 : <1000 ppm;
- noise level (N , dB): up to 50 dB;
- vibration level (V , m/s^2): up to 0.005 m/s^2 .

Assessment is carried out according to the formula:

$$\text{Comfort Index} = \frac{1}{6} \sum_{i=1}^6 \delta_i \cdot 100,$$

where

$$\delta_i = \begin{cases} 1, & \text{if the variable } i \text{ is within the standard} \\ 0, & \text{in other cases.} \end{cases}$$

Thus, the maximum value *Comfort Index* is 100 (perfect comfort), and the minimum is 0 (complete non-compliance of the conditions with the norms).

✓ The safety index (*Safety Index*) reflects the compliance of the technical condition of the environment with the criteria of electrical safety and fire safety. Contains controlled variables:

- temperature of equipment surfaces (T_{dev} , °C): up to 70 °C;
- consumption current (I , A): up to 16 A;
- voltage (U , V): in the range of 180–250 V;
- flame detection sign (F): Boolean signal.

The safety index is defined as:

$$\text{Safety Index} = \frac{1}{4} \sum_{j=1}^4 \sigma_j \cdot 100,$$

where

$$\sigma_j = \begin{cases} 1, & \text{if the variable } j \text{ is in safe range} \\ 0, & \text{in other cases.} \end{cases}$$

This index is also normalized in the range from 0 to 100, where 100 means full compliance with all safety requirements.

✓ The Energy Efficiency Index (*Efficiency Score*) reflects the ability of the system to consume electricity economically, in particular, during non-working periods. It takes into account:

- presence of personnel in the room (*presence*, Boolean signal);
- current energy consumption by lighting, temperature and equipment load.

The efficiency index is calculated as:

$$\text{Efficiency Score} = \left(1 - \frac{E_{\text{waste}}}{P + \varepsilon}\right) \cdot 100,$$

where ε is a small number to avoid division by zero. $\varepsilon = 0.001$.

Power consumption P is estimated as:

$$P = \frac{U \cdot I}{1000},$$

where U is the voltage (V), I is the current (A) and P is the power in kW.

Energy losses are defined as the weighted sum of consumption components when staff are absent:

$$E_{\text{waste}} = \sum_k \omega_k \cdot (1 - \text{presence}),$$

where $\omega_k \in [0,1]$ are the weighting factors for temperature, lighting, and consumption.

The Digital Twin dynamically adjusts the importance of specific criteria by modifying the weighting factors ω_k . For the practical implementation within the platform, three fundamental distribution profiles are defined:

- a profile with safety priority ($\omega_{\text{temp}} = 0.6$, $\omega_{\text{light}} = 0.2$, $\omega_{\text{cons}} = 0.2$);
- a profile with resource-saving priority ($\omega_{\text{temp}} = 0.2$, $\omega_{\text{light}} = 0.2$, $\omega_{\text{cons}} = 0.6$);
- 0.33, $\omega_{\text{light}} = 0.33$, $\omega_{\text{cons}} = 0.34$).

These values ensure the adaptability of the assessment model to different operational modes of the SMART laboratory.

The weighting coefficients satisfy the normalization condition $\sum_k \omega_k = 1$ and can be modified to reflect alternative pedagogical or operational strategies.

The general condition of the laboratory is determined by the number of critical deviations:

$$Overall\ Status = \begin{cases} stable, & \text{if } criticalCount = 0 \\ warning, & \text{if } 1 \leq criticalCount \leq 3 \\ critical, & \text{if } criticalCount > 3, \end{cases}$$

where $criticalCount$ is the number of indicators that have gone beyond the standard.

The proposed methodology for evaluating KPIs is distinguished by its versatility and scalability, which allows it to be applied in different types of environments. It combines technical and human comfort indicators, providing a comprehensive assessment. In addition, the approach is open to integration with modern monitoring and optimization systems, in particular based on IoT, artificial intelligence and energy audits.

4. DT Model

For the DT model, a discrete dynamic structure in the space of state variables with constant coefficients is chosen. Such a model provides flexibility in analysis, simplifies calculations and increases system stability, allowing for its effective adaptation the model to different scenarios. This approach also improves forecasting accuracy and process optimization due to its clear mathematical framework, making it a convenient tool for exploring and managing complex dynamical systems.

The model of the microclimate of the laboratory is presented in discrete time in the form of a system of difference equations of the type:

$$\begin{aligned} x[k+1] &= Ax[k] + Bu[k] + Ed[k], \\ y[k] &= Cx[k], \end{aligned} \quad (1)$$

where:

$x[k]$ is the vector of states at the time k ($x_0[k]$ is the air temperature in the room, °C; $x_1[k]$ – relative humidity, %; $x_2[k]$ – illumination, lx; $x_3[k]$ – CO₂ level, ppm; $x_4[k]$ – noise level, dB; $x_5[k]$ – vibration, 10⁻³ m/s²;

$u[k]$ – vector of control actions ($u_0[k]$ – air conditioning (cooling) (0 – 10), $u_1[k]$ – ventilation (0 – 10), $u_2[k]$ – humidifier (0 – 10), $u_3[k]$ – lighting (0 – 10);

$d[k]$ – perturbation vector (external conditions): $d_0[k]$ – external temperature, °C; $d_1[k]$ – external humidity, %; $d_2[k]$ – natural light (sun), Lux; $d_3[k]$ – number of people in the room; $d_4[k]$ – street noise, dB;

$y[k]$ – system output (measurement);

A, B, E, C – coefficient matrices: A – internal dynamics; B – influence of control actions; E – the influence of external factors; C – matrix of outputs/measurements.

The following values are used as initial values for these matrices (I – unit matrix):

$$A = \begin{bmatrix} 0.92 & 0.03 & 0 & 0.01 & 0 & 0 \\ 0.02 & 0.90 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.95 & 0 & 0 & 0 \\ 0.01 & 0.02 & 0 & 0.91 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.96 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.90 \end{bmatrix}, \quad B = \begin{bmatrix} -0.4 & -0.2 & 0 & 0 & 0 \\ 0 & -0.1 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & -0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$E = \begin{bmatrix} 0.25 & 0 & 0 & 0.3 & 0 \\ 0 & 0.2 & 0 & 0.2 & 0 \\ 0 & 0 & 0.3 & 0 & 0 \\ 0.15 & 0.1 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0.2 \end{bmatrix}, \quad C = I_{6 \times 6}.$$

Initial state:

$$x[0] = \begin{bmatrix} 22.0\ ^\circ C \\ 45.0\ \% \\ 400\ lux \\ 800\ ppm \\ 45.0\ dB \\ 0.006\ m/sec^2 \end{bmatrix}.$$

The structure of the model was determined on the basis of a preliminary analysis of the physical relationships between parameters (the effect of temperature on humidity, or the number of people on CO₂). The coefficients of the dynamics model matrices (A, B, E, C) were taken from the scientific literature and a priori assumptions based on the typical behavior of similar systems. Their values are consistent with theoretical principles and can be adapted to specific conditions (Yao et al., 2013). The initial conditions of the model were set on the basis of typical values characteristic of the laboratory environment, which allows obtaining plausible results already at the initial stages of the simulation. However, given the theoretical nature of the parameters, they can be clarified in the process of calibrating the model for specific experimental data collected in the process of work. Thus, such a DT model is a rapid implementation model that will be further identified/calibrated through operation or student training.

5. Algorithm for optimizing resources and comfort

The algorithm for improving resource efficiency and safety in the SMART laboratory is based on the integrated application of a digital twin (DT), which acts as a virtual model of the environment with the ability to predict, simulate scenarios and assess the impact of control actions before their implementation in physical space.

The main stages of the algorithm (Fig. 2):

1. Synchronized monitoring of the state of a physical object. All sensors send data to *Monitoring Manager sensors*, where pre-processing (filtering, normalization, time alignment) is performed. Data is duplicated in *DT Data Manager*, maintaining its up-to-date virtual state in real time.

2. Evaluation of quality indicators (KPIs). DT analyzes current variables and calculates three main indicators: Comfort Index – a generalized characteristic of microclimate favorability; Safety Index – an indicator of approaching dangerous modes (exceeding temperature, overloading power lines, etc.); Efficiency Score is the ratio between energy consumption and the achieved levels of comfort and safety.

3. Forecasting based on the DT model. DT simulates the development of events for the next few steps, taking into account the current state space ($x_1[k], \dots, x_5[k]$) and perturbations ($d_1[k], \dots, d_4[k]$) and current control actions ($u_1[k+1], \dots, u_3[k+1]$), behavioral patterns of laboratory use (presence of people, equipment activity). This forecast allows you to assess whether KPIs are expected to fall below critical thresholds and take action in advance.

4. Formation of recommended control actions. In case of detecting trends towards deterioration of conditions, the system generates recommendations through the DT model to optimize operation: changing the intensity of ventilation, heating or lighting; activation/deactivation of certain areas of the laboratory; switching the system to energy-saving mode in the absence of people. These actions are pre-simulated in DT, and only those that show an improvement in KPIs without deteriorating other indicators are suggested to the operator or applied automatically.

5. Implementation and monitoring of results. After implementing the recommended actions, DT conducts a performance analysis (before/after), updates the scenario database and forms new templates for successful configurations, this allows for improving the quality of recommendations over time based on accumulated experience.

It should be noted that the DT model takes into account which work scenario is activated (real, simulated, mixed) and automatically changes the weight of KPIs in the priority system. In particular, for a real scenario, safety is prioritized; for a simulated scenario, energy efficiency is prioritized; and for a mixed scenario, a balance among all indicators is maintained. Thus, there is an adaptation to the educational scenario.

Therefore, DT acts not only as a visualizer or a virtual copy of the laboratory, but as an active element of the decision support system. It provides behavior prediction, virtual testing of actions, assessment of efficiency without risk to the physical object, which significantly increases the safety, reliability and energy efficiency of the educational environment of the laboratory.

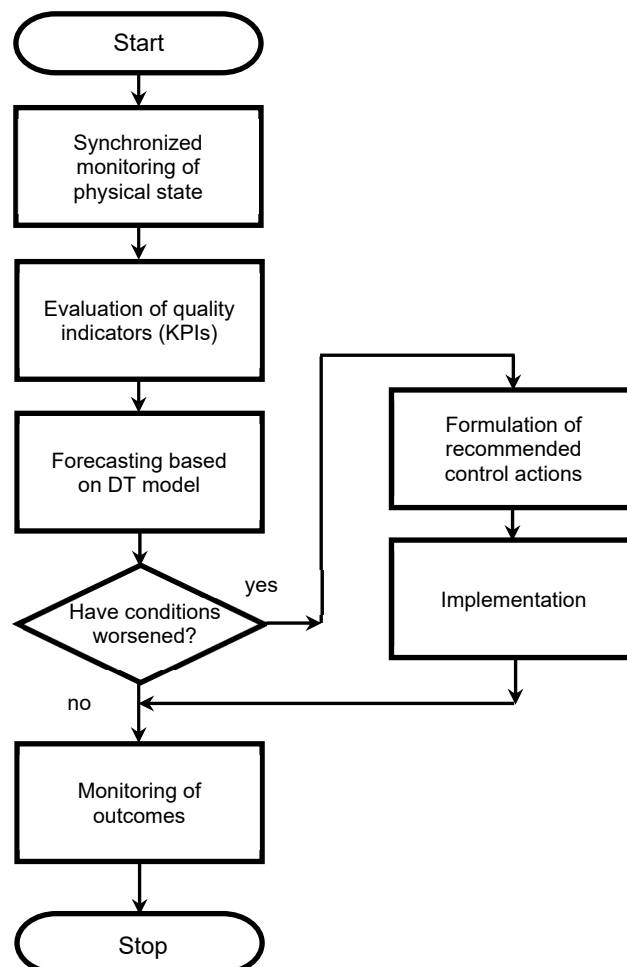


Fig. 2. The main stages of the control algorithm

6. Research results

Within the framework of this work, the educational environment is considered as a cyber-physical system. This system integrates the integration of physical objects (equipment, laboratory installations, sensors) and digital components (models, interfaces, analytical services) aimed at supporting the educational process. Such an environment not only reproduces the real conditions for the functioning of the laboratory space, but also serves as a tool for the development of digital competencies in students, in particular, in the field of information technology, IoT, data analytics, systems management and engineering thinking. The proposed architecture of the digital twin architecture for the laboratory covers all the key components of the educational environment: physical infrastructure, communication channels, virtual models, analytical algorithms, and user interfaces. In this paradigm, the educational environment becomes adaptive, context-sensitive, and focused on optimizing the learning process, thereby providing flexibility, accessibility, and personalization of educational interaction.

The developed DT SMART laboratory system was tested in simulation conditions with the involvement of real microclimate and safety parameters, as well as in the context of the implementation of the educational process. The results of the study showed the effectiveness of the selected architecture, control algorithms and approaches to resource monitoring.

Figure 3 demonstrates an online visualization of raw measurements with time graphs, deviation indicators, and color-coded risk levels, which facilitates the quick identification critical trends.

The panel in Fig. 4 combines the calculated Comfort, Safety and Efficiency indices, warning messages. The drill-down tools enable users to detail the contribution of individual parameters to the reduction of the generalized indicator.

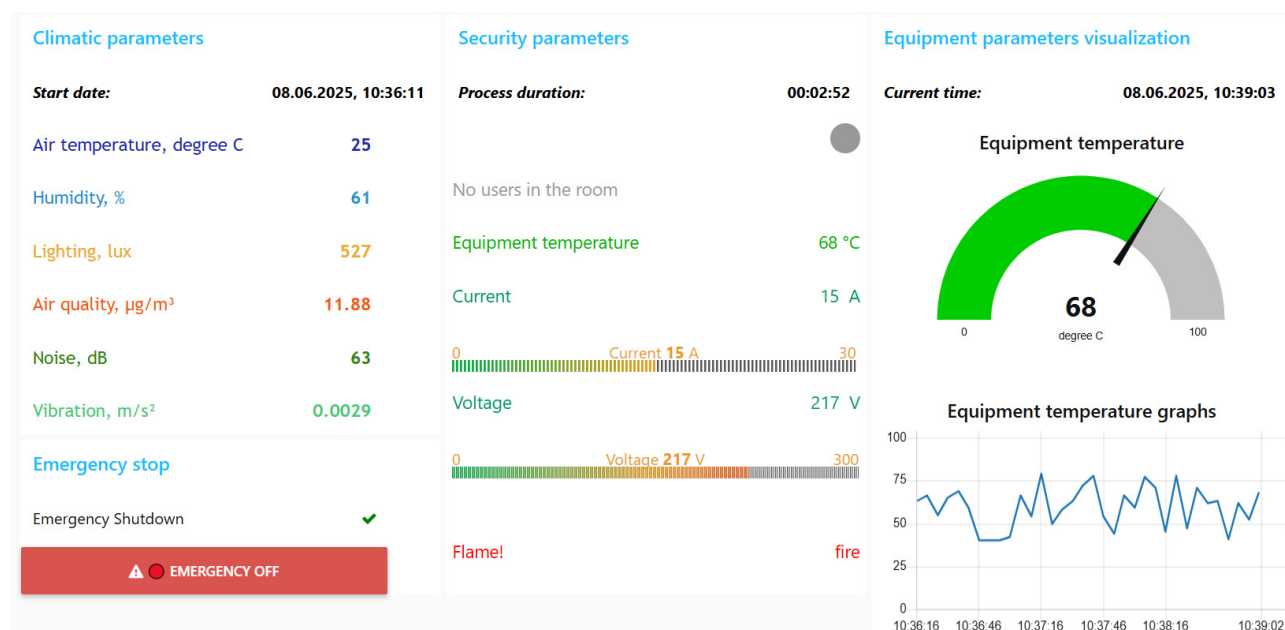


Fig. 3. Visualization Monitoring sensors Manager

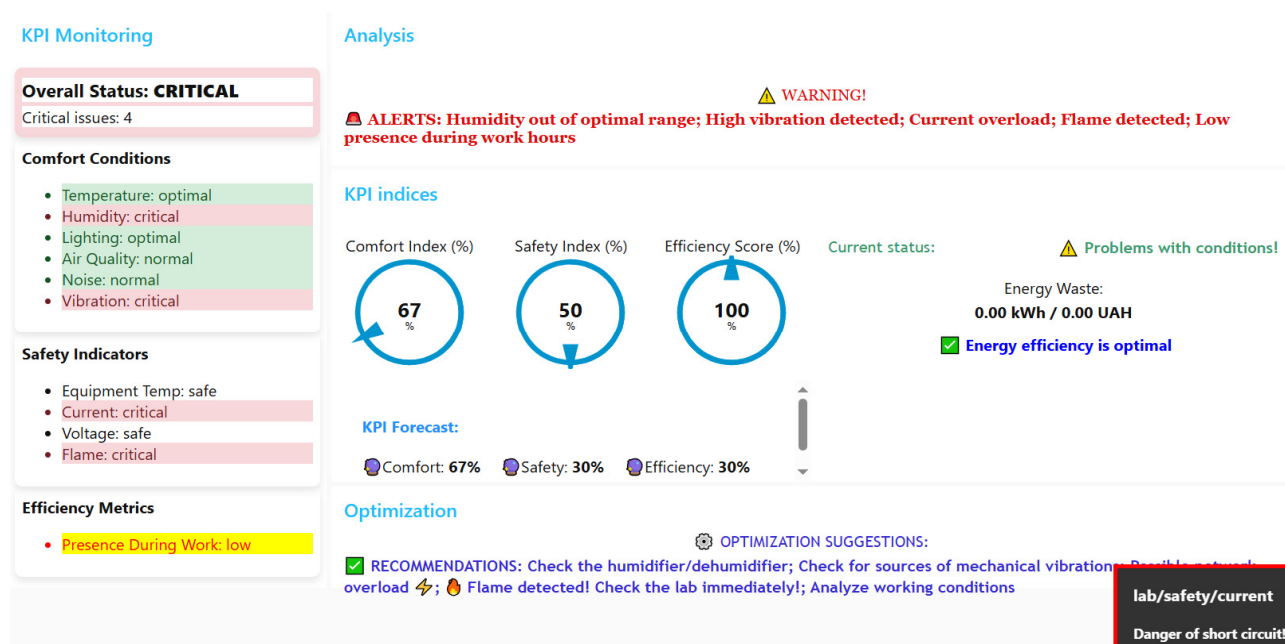


Fig. 4. Visualization Monitoring KPIs Manager

Figure 5 presents a fragment of the functional implementation of the DT SMART laboratory, developed in the Node-RED environment. The model integrates three main logical components: the microclimate model, perturbation actions and control actions, which are combined into a single computational flow with automatic regulation of environmental variables. Specifically, the matrix calculations for internal dynamics (A), control actions (B), and external disturbances (E) defined in (1) are handled by the 'Model: $x = Ax + Bu + Ed$ ' function node. This node performs discrete-time state-space integration at a 5-second frequency, serving as the core computational engine that links the theoretical mathematical model with the practical Node-RED data flow. This fragment simulates the dynamic changes in key microclimate variables such as temperature, humidity, lighting, CO₂ concentration, noise and vibration levels based on the input variables, with subsequent transmission of results to visualization and control nodes.

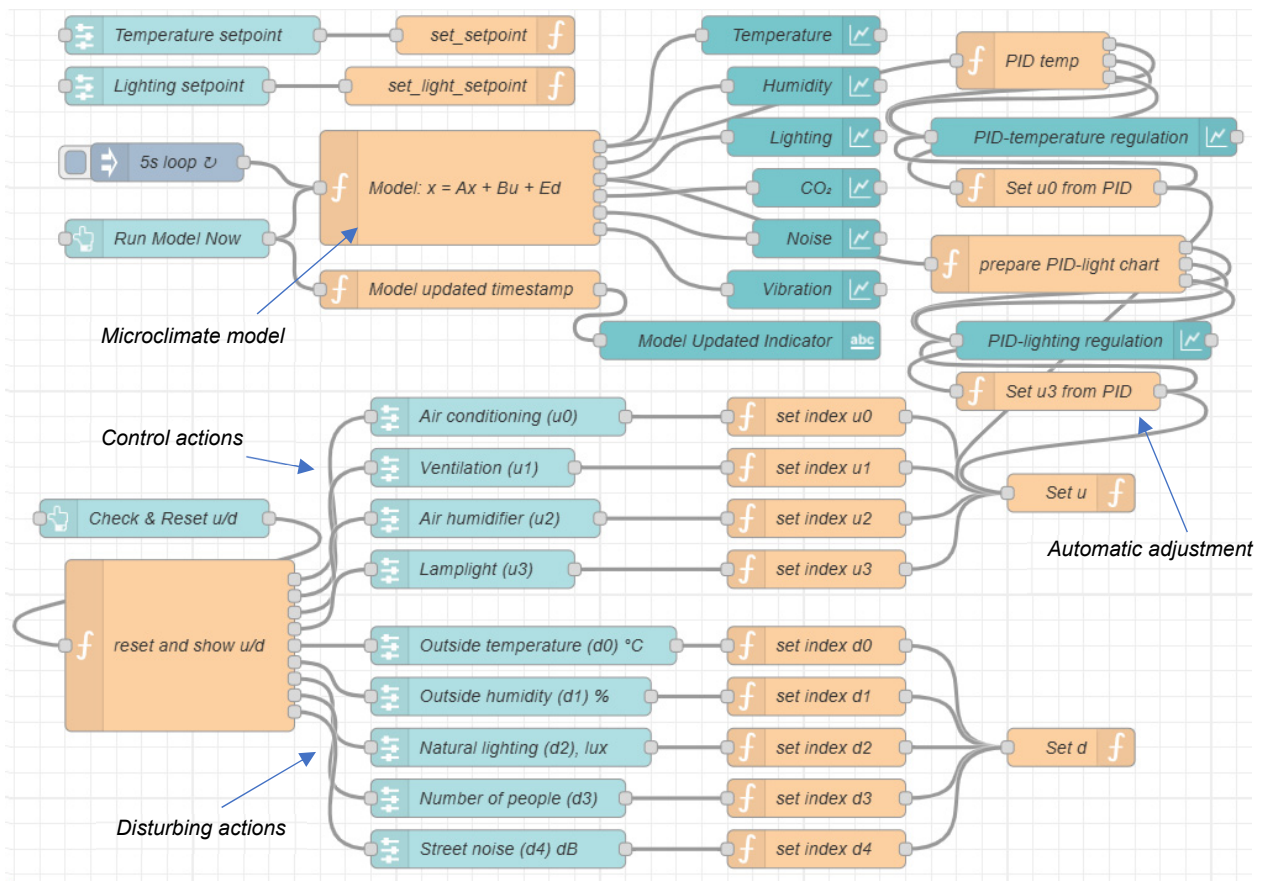


Fig. 5. Digital Twin Implementation

Disturbance inputs ($d_1, \dots, d_4$) are represented by signals which include outdoor temperature, humidity, natural light, the number of people in the room, and the level of street noise. These parameters can be entered manually or come from sensors, acting as external disturbances. Control actions ($u_1, \dots, u_3$) are responsible for operating actuators, specifically for air conditioning, ventilation, humidification, and lighting. Control can be carried out manually through the "set index" functions or automatically using PID controllers ("PID temperature controllers", "PID lighting controllers"), which analyze the simulated variables and generate appropriate control signals. The interconnection between the components demonstrates how DT integrates simulation, data collection, control, and regulation into a single, interactive, real-time environment.

Figure 6 displays a visualization of the DT model's simulation screen. In particular, users can set external perturbations and controls, then observe the predicted dynamics of state variables $x_1[k+1], \dots, x_5[k+1]$ calculated according to equations (1).

Figure 7 demonstrates how automatic temperature and lighting controls compensate for perturbations and return variables to standards, comparing system operation to a "no control" scenario.

Thus, DT integration with the laboratory's physical environment enables experiments, simulations, and real-time automatic control.

Discussion and conclusions

In the context of implementing a DT for an educational laboratory, three primary scenarios for conducting lab work are proposed:

1. Scenario of a real physics experiment with digital duplication: students work directly with physical equipment, with DT carrying out parallel monitoring and analysis of indicators. This allows you to compare real and virtual data, get an advanced view of the system's operation, detect obsolescence of DT model parameters, and automatically visualize and evaluate KPIs in real time.

2. Simulated experiment scenario without physical access: Under distance or hybrid learning conditions, DT fully simulates the behavior of the laboratory environment, allowing virtual experiments to be conducted, input variables can be modified, simulation results can be observed, and analyzed according to training tasks.



Fig. 6. Room climate simulation mode on the Digital Twin model

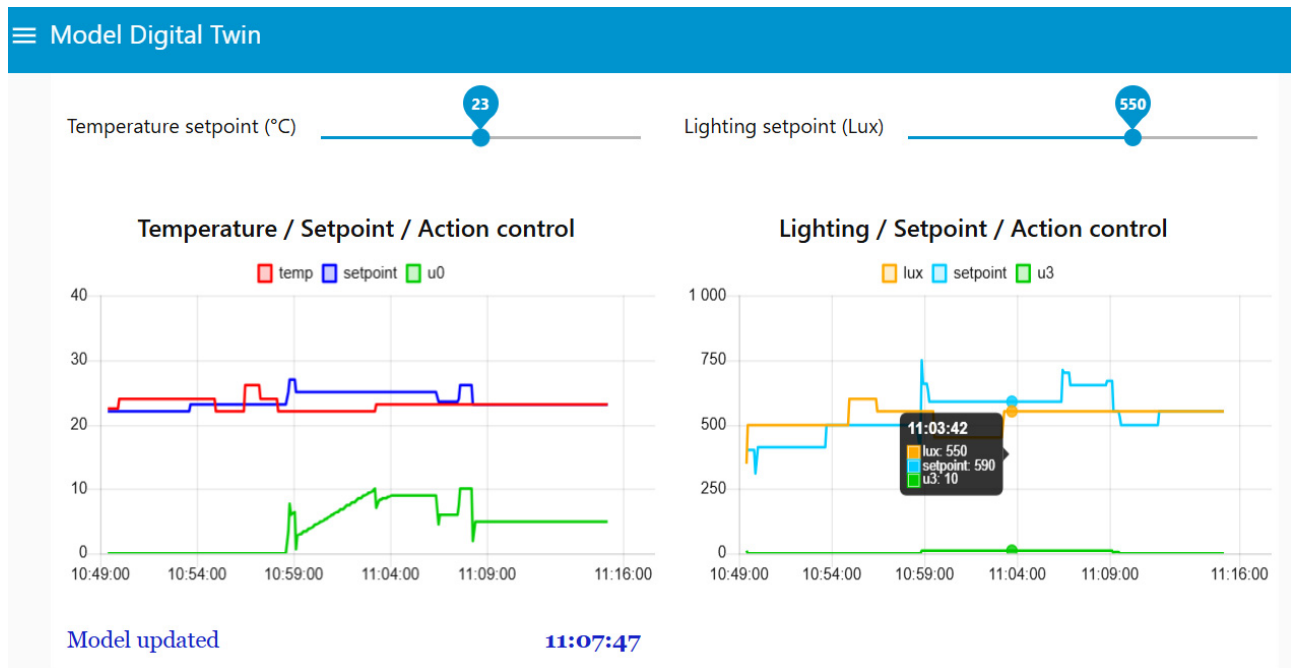


Fig. 7. Simulation mode of the automatic adjustment system on the Digital Twin model

3. Mixed-type scenario with adaptive environmental control: combines elements of real interaction and virtual model with the ability to control the microclimate, energy consumption and safety. Thanks to the use of DT, dynamic adjustment of the working conditions is implemented in accordance with the context (presence of students, type of experiment, resource constraints, etc.).

The proposed scenarios ensure the adaptability of the laboratory course to different learning formats, increase the efficiency of resource use, and create a safe and stimulating educational environment. In addition, each scenario is accompanied by an analytical block, which allows for continuous improvement of the methodology of conducting classes based on the analysis of the collected data.

The developed system of the educational environment of the SMART laboratory was tested in simulation conditions with the involvement of real microclimate and safety parameters, as well as in the context of the implementation of the educational process. The pedagogical evaluation of the proposed digital twin-based laboratory environment was conducted during regular laboratory classes in control systems and automation courses. The study involved 24 undergraduate students divided into small working groups. The experimental tasks included microclimate parameter analysis, PID controller tuning for temperature and lighting regulation (Fig. 7), and comparative analysis of system behavior under different disturbance scenarios. Data collection was carried out over a two-week period, covering six laboratory sessions with repeated task execution in real, simulated, and mixed scenarios.

The results of the study showed the effectiveness of the chosen architecture, management algorithms and approaches to resource monitoring. In this case, it is necessary to note:

1. Simulations using a discrete dynamic microclimate model and an automatic regulator showed that a system with a DT is capable of:

- stabilize the temperature and humidity within the standard values within 2–4 minutes after a disturbance;
- react to anomalies in the safety system (exceeding the current or temperature of the equipment) within 1-2 minutes;
- to reduce the total electricity consumption by approximately 5.4–11.2 % with the same quality of control compared to manual mode.

2. Average values of KPIs for the entire simulation period (7 days of continuous simulation):

- Comfort Index: 91 %
- Safety Index: 96 %
- Efficiency Score: 87 %

These values demonstrate the high quality of adaptive control and the consistency of all subsystem operations.

3. Within the framework of conducting classes according to three scenarios (real, simulated, mixed), the effectiveness assessment was carried out according to the following criteria:

- Number, variability, and speed of task completion: In the virtual scenario, an average of 23 % less time was spent on each task due to adjustable simulation time and the ability to work with saved data.
- Accuracy of measurements and repeatability of experiments. The use of DT reduced the variation of measurement results to 5.2 %, which is half that of traditional approaches.
- Interactivity and adaptability. A mixed scenario that combines physical access to equipment with remote simulation is the most effective for building analysis competencies and optimizing processes.

Thus, the DT integration into the educational space not only improves teaching quality but also ensures the flexibility of the educational process, which is critically important in today's environment. Virtualization allows you to scale the laboratory into different academic disciplines while maintaining the conceptual integrity of the environment. At the same time, further DT improvement requires model expansion and the development of an ontological knowledge base to ensure semantic interoperability between different types of laboratories.

Therefore, the study confirmed the feasibility and effectiveness of the developed system as both a technical and pedagogical tool for higher education. The developed educational environment in the form of a SMART laboratory with integrated DT is an example of the practical implementation of modern approaches to the digitalization of education. It combines physical measurements, mathematical modeling, automatic regulation and visualization in a single software environment. Such integration allows not only to effectively manage microclimate and safety indicators, but also creates conditions for the active involvement of students in the educational process.

The DT acts as a key element of adaptive management, providing not only the current display of the environment's state but also its forecasting, assessing the quality of the learning space using key indicators, forming recommendations for improvement, and storing successful configurations for future use.

In contrast to industrial standards like ISO 23247, which often entail high implementation costs and require specialized staffing, the proposed Node-RED-based environment offers a significantly more accessible alternative. Due to its open-source nature and low entry threshold provided by visual programming, Node-RED allows students to interact with complex digital twin structures without extensive coding knowledge. This cost-effective approach facilitates the scaling of SMART laboratories in academic settings where budget and time for specialized technical training are limited.

The system's implementation in the open Node-RED environment has demonstrated the platform's technical accessibility and flexibility, enabling the approach to be scaled to other educational laboratories. The implementation of DT also supports flexible forms of learning, in particular, distance and mixed formats, which is extremely important in the context of modern educational transformation. Overall, the proposed approach demonstrates a new level of interaction among the student, the digital model, and the physical object, not only enhancing learning quality but also fostering the digital competencies necessary for future specialists.

The results obtained prove the effectiveness of the concept and create the basis for scaling the approach to other types of classrooms and for further research towards the ontological integration of DTs from different laboratories into a unified educational ecosystem.

Authors' contribution: Natalia Lutska – conceptualization; methodology; Tetiana Savchenko – software; Tetiana Savchenko, Natalia Lutska – collection and verification of empirical data; Tetiana Savchenko, Natalia Lutska, Lidiya Vlasenko, Ivan Parkhomenko – empirical research; Lydia Vlasenko, Ivan Parkhomenko – analysis of sources, preparation of a literature review or theoretical foundations of research.

Sources of funding. This work was supported by a grant from the Simons Foundation International (SFI-PD-Ukraine-00014577; O.G.).

References

- Beke, Á. K., Gyürkés, M., Nagy, Z. K., Marosi, G., & Farkas, A. (2021). Digital twin of low dosage continuous powder blending–Artificial neural networks and residence time distribution models. *European journal of pharmaceuticals and biopharmaceutics*, 169, 64–77. <https://doi.org/10.1016/j.ejpb.2021.09.006>
- Delgado, J. M. D., & Oyedele, L. (2021). Digital Twins for the built environment: learning from conceptual and process models in manufacturing. *Advanced Engineering Informatics*, 49, 101332. <https://doi.org/10.1016/j.aei.2021.101332>
- Diia. (2025). *The Government approved the Strategy for Digital Development of Innovations until 2030*. <https://tinyurl.com/3y4zy6kz>
- European Commission. (2023). *Digital Education Action Plan (2021-2027)*. European Education Area. <https://education.ec.europa.eu/focus-topics/digital-education/action-plan>
- Gautam, A., Aryal, M. R., Deshpande, S., Padalkar, S., Nikolaenko, M., Tang, M., & Anand, S. (2025). IIoT-enabled digital twin for legacy and smart factory machines with LLM integration. *Journal of Manufacturing Systems*, 80, 511–523. <https://doi.org/10.1016/j.jmsy.2025.03.022>

- Herrero, R., & Dasari, M. (2025). Edge-based physical asset and digital twin virtualization framework to support cognitive digital twins. *Internet of Things*, 31, 101601. <https://doi.org/10.1016/j.iot.2025.101601>
- Hu, C., Fan, W., Zeng, E., Hang, Z., Wang, F., Qi, L., & Bhuiyan, M. Z. A. (2021). Digital twin-assisted real-time traffic data prediction method for 5G-enabled internet of vehicles. *IEEE Transactions on Industrial Informatics*, 18(4), 2811–2819. <https://doi.org/10.1109/TII.2021.3083596>
- Huang, W., Zhang, Y., & Zeng, W. (2022). Development and application of digital twin technology for integrated regional energy systems in smart cities. *Sustainable Computing: Informatics and Systems*, 36, 100781. <https://doi.org/10.1016/j.suscom.2022.100781>
- Javaid, M., Haleem, A., & Suman, R. (2023). Digital twin applications toward industry 4.0: A review. *Cognitive Robotics*, 3, 71–92. <https://doi.org/10.1016/j.cogr.2023.04.003>
- Jia, K., Wang, H., Ren, D., Liu, B., Zhao, Q., & Hong, J. (2024). A general mathematic model framework for assembly process driven digital twin of assembly precision. *Journal of Manufacturing Systems*, 77, 196–211. <https://doi.org/10.1016/j.jmsy.2024.09.007>
- Junior, O. C. (2025). Digital Twin in Industry 4.0: Systematic review and content analysis and an architectural reference model. *Procedia Computer Science*, 253, 2919–2928. <https://doi.org/10.1016/j.procs.2025.02.016>
- Kober, C., Medina, F. G., Benfer, M., Wulfsberg, J. P., Martinez, V., & Lanza, G. (2024). Digital twin stakeholder communication: characteristics, challenges, and best practices. *Computers in Industry*, 161, 104135. <https://doi.org/10.1016/j.compind.2024.104135>
- Lu, Y., Li, Y., Fu, G., Jiang, Y., Huang, Y., Zhu, J., & Sheng, B. (2024). The physical information LSTM surrogate model for establishing a digital twin model of reciprocating air compressors. *Applied Soft Computing*, 167, 112309. <https://doi.org/10.1016/j.asoc.2024.112309>
- Lutska, N., Vlasenko, L., Ladanyuk, A., Zaiets, N., & Korobiichuk, I. (2022). Ontological support system of managerial decision-making of production tasks for a food enterprise. *Machinery & Energetics*, 13(3), 53–61. [https://doi.org/10.31548/machenergy.13\(3\).2022.53-61](https://doi.org/10.31548/machenergy.13(3).2022.53-61)
- Luo, W., Hu, T., Ye, Y., Zhang, C., & Wei, Y. (2020). A hybrid predictive maintenance approach for CNC machine tool driven by Digital Twin. *Robotics and Computer-Integrated Manufacturing*, 65, 101974. <https://doi.org/10.1016/j.rcim.2020.101974>
- Meyers, B., Van Noten, J., Liettaert, P., Tielemans, B., Hristov, H., Maes, D., & Gadeyne, K. (2022). Knowledge graphs in digital twins for manufacturing-lessons learned from an industrial case at Atlas Copco Airpower. *IFAC-PapersOnLine*, 55(10), 13–18. <https://doi.org/10.1016/j.ifacol.2022.09.361>
- Ministry of Education and Science of Ukraine. (2025a). *Digital transformation of education and science*. <https://tinyurl.com/2bx5he74>
- Ministry of Education and Science of Ukraine. (2025b). *WINWIN: Ukraine approved the Strategy for Digital Development of Innovations until 2030*. <https://tinyurl.com/2vjz8ztd>
- Nezzi, C., Gufler, V., Vidoni, R., & Rauch, E. (2025). Kinematic and dynamic modeling of mechanical systems towards Digital Twins. *Results in Engineering*, 104874. <https://doi.org/10.1016/j.rineng.2025.104874>
- Peldon, D., Banihashemi, S., LeNguyen, K., & Derrible, S. (2024). Navigating urban complexity: The transformative role of digital twins in smart city development. *Sustainable Cities and Society*, 111, 105583. <https://doi.org/10.1016/j.scs.2024.105583>
- Pretel, E., Moya, A., Navarro, E., López-Jaquero, V., & González, P. (2024). Analysing the synergies between Multi-agent Systems and Digital Twins: A systematic literature review. *Information and Software Technology*, 174, 107503. <https://doi.org/10.1016/j.infsof.2024.107503>
- Piromalis, D., & Kantaros, A. (2022). Digital twins in the automotive industry: The road toward physical-digital convergence. *Applied System Innovation*, 5(4), 65. <https://doi.org/10.3390/asi5040065>
- Ramonell, C., Chacón, R., & Posada, H. (2023). Knowledge graph-based data integration system for digital twins of built assets. *Automation in Construction*, 156, 105109. <https://doi.org/10.1016/j.autcon.2023.105109>
- Raza, M., Sartori, A., Bilberg, A., & Schlette, C. (2024). Communication between two distinct digital counterparts of a robotic cell digital twin via MQTT. *Procedia CIRP*, 130, 959–963. <https://doi.org/10.1016/j.procir.2024.10.191>
- Ropero, J. (2022). Peer review of "Toward human digital twins for cybersecurity simulations on the metaverse: Ontological and network science approach". *JMIRx Med*, 3(2), e38583. <https://doi.org/10.2196/38583>
- Sai, A. M. V. V., Wang, C., Cai, Z., & Li, Y. (2024). Navigating the digital twin network landscape: A survey on architecture, applications, privacy and security. *High-Confidence Computing*, 4(4), 100269. <https://doi.org/10.1016/j.hcc.2024.100269>
- Saif, W., RazaviAlavi, S., & Kassem, M. (2024). Construction digital twin: a taxonomy and analysis of the application-technology-data triad. *Automation in Construction*, 167, 105715. <https://doi.org/10.1016/j.autcon.2024.105715>
- Savchenko, T., Lutska, N., Vlasenko, L., Sashnova, M., Zahorulko, A., Minenko, S., Ibaiev, E. & Tytarenko, N. (2025). Risk analysis and cybersecurity enhancement of Digital Twins in dairy production. *Technology audit and production reserves*, 2(2(82)), 37–49. <https://doi.org/10.15587/2706-5448.2025.325422>
- Semeraro, C., Lezoche, M., Panetto, H., & Dassisti, M. (2021). Digital twin paradigm: A systematic literature review. *Computers in Industry*, 130, 103469. <https://doi.org/10.1016/j.compind.2021.103469>
- Schlenger, J., Pluta, K., Mathew, A., Yeung, T., Sacks, R., & Borrmann, A. (2025). Reference architecture and ontology framework for digital twin construction. *Automation in Construction*, 174, 106111. <https://doi.org/10.1016/j.autcon.2025.106111>
- Sharma, A., Kosasih, E., Zhang, J., Brintrup, A., & Calinescu, A. (2022). Digital Twins: State of the art theory and practice, challenges, and open research questions. *Journal of Industrial Information Integration*, 30, 100383. <https://doi.org/10.1016/j.jii.2022.100383>
- Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B. D., Todd, M. D., Mahadevan, S., Hu, Ch., & Hu, Z. (2022). A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies. *Structural and Multidisciplinary Optimization*, 65(12), 354. <http://dx.doi.org/10.48550/arXiv.2208.14197>
- Tuhaise, V. V., Tah, J. H. M., & Abanda, F. H. (2023). Technologies for digital twin applications in construction. *Automation in Construction*, 152, 104931. <https://doi.org/10.1016/j.autcon.2023.104931>
- Van der Valk, H., Schmelzer, R., Rose, D., & Dinter, B. (2024). Towards a Process Model for Digital Twin Implementation: The Implementation Canvas. *Procedia CIRP*, 122, 903–908. <https://doi.org/10.1016/j.procir.2024.01.123>
- Vlasenko, L., Lutska, N., Zaiets, N., Korobiichuk, I., & Hrybkov, S. (2022). Core ontology for describing production equipment according to intelligent production. *Applied System Innovation*, 5(5), 98. <https://doi.org/10.3390/asi5050098>
- Wallner, B., Zwölfer, B., Trautner, T., & Bleicher, F. (2023). Digital twin development and operation of a flexible manufacturing cell using iso 23247. *Procedia CIRP*, 120, 1149–1154. <https://doi.org/10.1016/j.procir.2023.09.140>
- Weil, C., Bibri, S. E., Longchamp, R., Golay, F., & Alahi, A. (2023). Urban digital twin challenges: A systematic review and perspectives for sustainable smart cities. *Sustainable Cities and Society*, 99, 104862. <https://doi.org/10.1016/j.scs.2023.104862>
- Xia, H., Liu, Z., Efremochkina, M., Liu, X., & Lin, C. (2022). Study on city digital twin technologies for sustainable smart city design: A review and bibliometric analysis of geographic information system and building information modeling integration. *Sustainable Cities and Society*, 84, 104009. <https://doi.org/10.1016/j.scs.2022.104009>
- Xie, W., Zeng, Y., Zhang, X., Wong, H. Y., Zhang, T., Wang, Z., Wu, X., Shi, J., Huang, X., Xiao, F., & Usmani, A. (2025). AIoT-powered building digital twin for smart firefighting and super real-time fire forecast. *Advanced Engineering Informatics*, 65, 103117. <https://doi.org/10.1016/j.aei.2025.103117>

- Yao, Y., Yang, K., Huang, M., & Wang, L. (2013). A state-space model for dynamic response of indoor air temperature and humidity. *Building and environment*, 64, 26–37. <https://doi.org/10.1016/j.buildenv.2013.03.009>
- Yu, J., Liu, P., & Li, Z. (2020). Hybrid modelling and digital twin development of a steam turbine control stage for online performance monitoring. *Renewable and Sustainable Energy Reviews*, 133, 110077. <https://doi.org/10.1016/j.rser.2020.110077>

Отримано редакцією журналу / Received: 02.08.25
Прорецензовано / Revised: 29.01.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Тетяна САВЧЕНКО¹, канд. техн. наук, доц.
ORCID ID: 0000-0002-8884-5360
e-mail: tsavchenko@ukma.edu.ua

Наталія ЛУЦЬКА², д-р техн. наук, проф.
ORCID ID: 0000-0001-8593-0431
e-mail: lutskanm2017@gmail.com

Лідія ВЛАСЕНКО³, канд. техн. наук, доц.
ORCID ID: 0000-0002-2003-6313
e-mail: vlasenko.lidia1@gmail.com

Іван ПАРХОМЕНКО⁴, канд. техн. наук, доц.
ORCID ID: 0000-0001-6889-9284
e-mail: ivan.parkhomenko@knu.ua

¹Київський національний університет "Києво-Могилянська академія", Київ, Україна

²Національний університет харчових технологій, Київ, Україна

³Державний торговельно-економічний університет, Київ, Україна

⁴Київський національний університет імені Тараса Шевченка, Київ, Україна

ЦИФРОВИЙ ДВІЙНИК У СТРУКТУРІ СУЧАСНОГО ОСВІТНЬОГО СЕРЕДОВИЩА

У сучасному світі перевагою цифрових двійників є можливість створення копій різних фізичних об'єктів, систем або навчальних середовищ тощо. Це особливо актуально для інженерних дисциплін, де моделі цифрових двійників заданої точності забезпечують проведення досліджень та експериментів без фізичних ризиків і суттєвих фінансових витрат. У контексті освітнього середовища це означає забезпечення ефективного функціонування віртуальних навчальних лабораторій, оснащених реальним обладнанням.

У представленій статті розроблено освітнє середовище з використанням цифрового двійника SMART-лабораторії, що функціонує як кіберфізична система з наскрізною інтеграцією сенсорної мережі, виконавчих механізмів і віртуальної моделі приміщення. Мета роботи – теоретичне обґрунтування й експериментальна верифікація освітнього середовища, що містить цифровий двійник, орієнтований на підвищення енергоефективності, комфортності й безпеки навчальних просторів. Методологічною основою дослідження є системний аналіз, математичне моделювання динаміки мікроклімату, а також імперативна логіка регулювання з використанням типових регуляторів.

У межах запропонованої архітектури розроблено модель мікроклімату, що поєднує підсистеми моніторингу, прогнозування й адаптивного керування. Актуалізація реального та віртуального стану здійснюється в реальному часі, що забезпечує можливість попереднього тестування керуючих дій та оцінювання їхнього впливу без ризику для матеріальної бази. Проведена серія симуляцій та експериментів підтвердила здатність розробленої системи зменшувати споживання ресурсів і стабілізувати параметри мікроклімату за умов дії зовнішніх збурень. Водночас система зберігає педагогічну гнучкість, підтримуючи різні формати організації навчального процесу.

Одержані результати засвідчують наукову новизну запропонованого підходу, що полягає в поєднанні моделей і систем керування з освітніми сценаріями в єдиному програмному середовищі. Практичне значення роботи визначається можливістю масштабування розробленого рішення на інші типи лабораторій і навчальні дисципліни, а також сприянням формуванню цифрових компетентностей майбутніх фахівців у галузі інформаційних технологій і кіберфізичних систем.

Ключові слова: освітнє середовище, SMART-лабораторія, цифровий двійник, моніторинг, KPI, мікроклімат.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; у рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 004.932.2:004.85:681.5.033

DOI: <https://doi.org/10.17721/1812-5409.2026/1.33>

Yaroslav SMOLIN, PhD Student
ORCID ID: 0009-0007-9744-6496
e-mail: smolin.yaroslav@iill.kpi.ua
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

Viacheslav LISKIN, PhD, Assoc. Prof.
ORCID ID: 0000-0002-9418-0633
e-mail: liskin.viacheslav@iill.kpi.ua
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

META-ANALYSIS OF PARETO-OPTIMAL OBJECT DETECTION MODELS FOR UNMANNED AERIAL VEHICLES ON THE VISDRONE2019-DET DATASET

The subject of this research is modern deep learning models for object detection from unmanned aerial vehicles (UAVs), trained and evaluated on the benchmark VisDrone2019-DET dataset. The problem addressed was the lack of a systematic analysis of existing models based on the criterion of optimal balance between detection accuracy, computational speed, and model size, all of which are critical for real-time operation on UAV onboard computers with limited resources. The core result lies in a comprehensive meta-analysis of over 60 scientific publications and the identification of two groups of Pareto-optimal models: one focused on minimizing model parameters and the other on minimizing computational complexity. Through multi-criteria Pareto analysis and the creation of the most complete performance summary table to date, the study identified architectures that offer the best trade-offs between accuracy, speed, and size. These turned out to be optimized versions of YOLO and transformer-based detectors like DETR, a result explained by the efficiency of their core architectures and further tuning for the specifics of aerial imagery. Using the Hailo 8L accelerator as a case study, the research also demonstrated that a model's computational complexity is directly proportional (with a coefficient of determination up to 0.98) to its execution speed on the typical neural accelerator that is used in a UAV companion computer. This allows engineers to estimate relative runtime based on the number of operations, avoiding time-consuming benchmarking. Overall, the findings can be used by engineers to make informed choices about detector architectures depending on specific UAV hardware constraints, and by researchers as reference points for developing new models for autonomous UAV operation in aerial monitoring, infrastructure inspection, and search-and-rescue missions.

Key words: *unmanned aerial vehicles, VisDrone2019, object detection, deep learning models, Pareto analysis.*

AMS 2020 classification: 68T07, 68T10, 90C29.

Introduction

Unmanned aerial vehicles (UAVs) are increasingly used in surveillance, infrastructure inspection, precision farming, rescue operations, traffic monitoring, security operations, and reconnaissance (Cao et al., 2023; Du et al., 2024; Gutierrez et al., 2025; Jiao, & Abdullah, 2024; Li et al., 2023; Zhu et al., 2019). The unique aerial view, mobility, low cost, and ease of use of UAVs provide advantages over both ground-based systems and manned aviation (An et al., 2024; Cao et al., 2023; Hua, & Chen, 2025; Li et al., 2023; Mirzaei et al., 2023; Wan et al., 2021).

Autonomous or semi-autonomous operation of UAVs is becoming critically important due to radio horizon limitations, possible communication interference in difficult terrain or urban environments, delays in data transmission over long distances, and the need for instant response to dynamic changes in the environment without operator intervention, especially in the presence of electromagnetic interference or electronic countermeasures. In addition, full autonomy eliminates the human factor and allows the operator's work to be automated (Elmokadem, & Savkin, 2021).

A key technology for solving the autonomy problem is object detection, which allows UAVs to identify and locate objects in their environment, understand the scene, and provide operational information (Cao et al., 2023; Jiao, & Abdullah, 2024; Nikouei et al., 2025). Modern detectors, such as those based on transformers, are often too resource-intensive, making them unsuitable for use on board UAVs in real time, as they have significant limitations in terms of battery consumption, payload, computing power, and memory capacity (Chen et al., 2021; Suo et al., 2024). At the same time, faster models, such as those based on EfficientNet (Tan et al., 2020), are not accurate enough for stable detection of small objects in dense scenes. Despite significant research in this area, there is no systematic comparative analysis of existing models based on key parameters, which complicates the selection of the optimal solution for specific UAV hardware constraints. Researchers and engineers are faced with the problem of navigating a vast "zoo" of models, each of which demonstrates improvements in certain metrics, but without a clear understanding of the overall picture of trade-offs between accuracy, speed, and model size.

The practical value of this study lies in creating the most comprehensive comparison of object detection models for UAVs to date, based on the VISDRONE2019-DET dataset (Du et al., 2019) and evaluated according to three key parameters: average accuracy (mAP), number of model parameters (M), and computational complexity (GFLOPs). Using Pareto analysis (Nia et al., 2022), the study allows us to objectively identify models that demonstrate the best balance between these parameters. Although GFLOPs is not traditionally used as a direct measure of detection speed, the study shows that for models with similar architectures, this indicator allows approximating the relative execution time, providing a theoretical basis for the initial selection of models without the need for testing on specific hardware.

The relevance of the study lies in establishing a starting point for both researchers working on new architectures and practitioners interested in directly implementing models on board UAVs. A systematic meta-analysis allows for quick navigation among the multitude of existing solutions and select models that demonstrate the best balance in their "weight class" for specific hardware constraints.

1. Literature review and problem statement

Standardized datasets play an important role in the objective evaluation and comparison of object detection models in aerial photography. One such comprehensive reference dataset for object detection from UAVs is VisDrone2019 (Du et al., 2019), developed by the AISKEYE team at Tianjin University. It consists of 10209 images containing a variety of scenes under different weather and lighting conditions. The key challenges of this dataset are significant variations in viewpoints and scales, a large number of small objects, scenes with high density and overlap, and a complex background (Wan et al., 2021). The complexity of the VisDrone2019 dataset has stimulated the active adaptation and development of various object detection paradigms. However, analysis of existing solutions reveals a number of unresolved issues related to the fundamental trade-off between accuracy and computational efficiency on board UAVs. Detection models can be divided into three categories: two-stage, single-stage, and transformer detectors.

Traditional two-stage detectors, such as Faster R-CNN (Ren et al., 2016) and Cascade R-CNN (Cai, & Vasconcelos, 2018), although demonstrating high accuracy on standard datasets, remain computationally expensive and slow. This significantly complicates their use on UAV edge devices for real-time computing (Song et al., 2024).

Single-stage detectors, in particular the YOLO family (Redmon et al., 2016), SSD (W. Liu et al., 2016), and RetinaNet (Lin et al., 2020), offer significantly better speed, making them a popular choice for real-time systems (Vaddi et al., 2021). However, the standard configurations of these models do not always effectively handle the specifics of aerial images. Anchor-free approaches such as CornerNet (Law, & Deng, 2020), CenterNet (Duan et al., 2019), FCOS (Keles et al., 2022), and YOLOX (Camarlinghi et al., 2024) partially solve the problem of detecting objects with variable aspect ratios, but as noted in (Du et al., 2019; Xiao et al., 2024), they can have difficulty localizing very small objects, which is a critical drawback for VisDrone. Thus, the question of reliable detection of small targets using fast single-stage models remains open.

A new paradigm of transformer detectors (DETR) (Kong et al., 2024; Liu et al., 2025; Zhang et al., 2025; Zhao et al., 2024) offers an end-to-end approach with a global attention mechanism that also eliminates the need for additional Non-Maximum Suppression (NMS) post-processing. Despite this, early DETR models had significant drawbacks: high computational cost and low performance in detecting small objects (Rekavandi et al., 2023). Although recent models such as RT-DETR (Zhao et al., 2024) and specialized versions UAV-DETR (Zhang et al., 2025) and VRF-DETR (Kong et al., 2024) have become more competitive, they do not solve the problem completely. For the most part, achieving high accuracy still requires significant computational resources, and their lighter versions are inferior in accuracy to specialized convolutional architectures. Thus, the question of creating a transformer architecture that simultaneously retains its advantages and achieves parity with the best CNN models in terms of accuracy/efficiency ratio on small objects remains unresolved. To overcome these challenges, researchers have proposed a number of strategies, but each has its limitations.

For example, multiscale feature fusion techniques - Bidirectional and Recursive Feature Pyramid Networks (Xiao et al., 2024), Hierarchical Extended Path Aggregation Networks (Chen, & Zhang, 2025), self-attention mechanism and specialized "heads" (Chen, & Zhang, 2025) improve the detection of small objects, but often at the cost of increasing the number of parameters and computational complexity, which contradicts the requirements for peripheral devices.

The use of lightweight architectures such as YOLO-tiny/n/s (Redmon et al., 2016), EfficientDet (Tan et al., 2020), EdgeYOLO (Liu et al., 2023), RT-DETR-R18 (Zhao et al., 2024), VRF-DETR (Wenbin, 2025) leads to a reduction in computational complexity. The same applies to the use of efficient convolution operations: depthwise separable (X. Tang et al., 2024), partial (Wang et al., 2025) and gated (Wenbin, 2025) convolutions, and spatial-channel separable downsampling (He, & Wang, 2023; Peng et al., 2018) decrease computation complexity and parameter count. However, simplifying the architecture often leads to the loss of important features necessary for identifying the smallest and most complex objects, which in turn affects the accuracy of the model.

To improve detection accuracy and training stability for small objects, it is recommended to use specialized loss functions such as Normalized Wasserstein Distance (Xu et al., 2022), Shape-IoU (Zhang, H., & Zhang, S., 2024), or Wise-IoU (Tong et al., 2023). They are less sensitive to localization errors of small objects, as these functions do not "penalize" minor shifts of a few pixels, which are critical for standard IoU metrics. The danger of this approach lies in the potential increase in computational complexity during training and the risk of model over-specialization. If the loss function is too focused on small objects, it can potentially impair the model's ability to generalize to medium and large objects.

Several approaches are used to address detection on high-resolution images. One of them is to slice images (Akyon, Altınuç, & Temizel, 2022) into smaller, often overlapping blocks, allowing standard models to find small objects. Another approach is super-resolution (Xiao et al., 2024), which attempts to artificially increase the quality of areas with small objects before detection. However, the disadvantage of these methods is a significant increase in processing time, as the detector is run multiple times, and a loss of global context, which can lead to errors at block boundaries or misclassification. QueryDet (Yang, Huang, & Wang, 2022) is an alternative that avoids dense computations on high-resolution feature maps. Although more efficient, this approach complicates the architecture, and its success strongly depends on the quality of the initial, coarse detection stage, where errors cannot be corrected later.

Analysis shows that achieving state-of-the-art results on VisDrone is an evolutionary process of adapting existing models (YOLO, DETR) by integrating specific modules. However, most publications focus on demonstrating the accuracy gains from their proposed modifications, often without conducting an in-depth analysis of the trade-off with computational costs in the context of real UAV hardware platforms. Although there is dynamic competition between CNN and transformer approaches in the industry, there is no single methodology that would allow objective identification of models that offer the optimal balance for different hardware budgets. Researchers and practitioners face the problem of choosing among dozens of models without a clear understanding of their practical suitability.

Thus, an important unresolved problem is the lack of a systematic analysis and identification of models that form the Pareto front in the multidimensional "accuracy-speed-size" space for the task of object detection in aerial images. Existing studies offer local improvements but do not provide a global picture that would allow engineers to make informed choices about the architecture that best suits the hardware constraints of a particular UAV and mission requirements.

Problem statement: It is necessary to identify which of the existing object detection models not only achieve high accuracy on complex aerial images, but are also computationally efficient enough for real-time use on resource-constrained UAVs. This requires a careful analysis of the "accuracy-speed-model size" trade-off and a search for optimal solutions on the Pareto front.

2. Research objectives

The aim of the study is to conduct a critical comparative meta-analysis of modern deep learning models for object detection tested on the VisDrone2019-DET dataset to identify architectures that form the Pareto frontier in the "accuracy-speed-size" space. This will enable us to formulate informed recommendations for engineers and researchers on the selection of optimal solutions for practical applications on UAV onboard computers, taking into account their hardware limitations. To achieve this goal, the following tasks were set:

1. Collect and compare data on accuracy (mAP), number of parameters (M), and computational complexity (FLOPs) for a wide range of models tested on VisDrone2019-DET.
2. Identify models that demonstrate the best balance between accuracy and computational cost (Pareto front) for different levels of resource requirements.
3. Show how GFLOPs correlate with execution speed in ms for different types of models using the example of Hailo 8L NPU

3. Materials and research methods

This section begins with a definition of the key elements of the research.

Research object: modern deep learning models for object detection, optimized for aerial monitoring tasks using unmanned aerial vehicles (UAVs) and tested on the VisDrone2019-DET benchmark dataset.

The main research hypothesis: among the variety of existing architectures, it is possible to identify a distinct set of Pareto-optimal models that offer the best compromise between accuracy, speed, and size. It is assumed that this frontier is predominantly formed by specialized rather than standard detectors, thanks to their targeted adaptation to the challenges of aerial imagery.

Assumptions made in this work: model performance data collected from various scientific publications are comparable because they were obtained on the standardized VisDrone2019-DET benchmark. It is assumed that this dataset adequately represents real-world aerial monitoring scenarios.

Simplifications made in this work: the multifactorial task of optimizing a model for UAVs (accuracy, execution speed, energy consumption, size) is simplified to the analysis of three key metrics: average accuracy (mAP), number of parameters, and computational complexity (FLOPs). It is important to note that although FLOPs is a common metric for computational complexity, the actual speed of the model, measured in frames per second (FPS) or milliseconds of execution (inference time), depends on many other factors (Fernandez et al., 2023). These include the specifics of the hardware platform (e.g., GPU, NPU), software optimization, memory bandwidth, and other implementation aspects. FLOPs is a theoretical, indirect metric, while the actual speed depends on the cost of memory access and platform characteristics (Jayanth, Gupta, & Prasanna, 2024; Ma et al., 2018). To investigate the relationship between theoretical complexity (GFLOPs) and practical runtime on representative edge hardware, we analyzed the correlation using performance data from the Hailo-8L NPU (Hailo-8L Model Zoo, 2025). The Hailo-8L is a specialized AI accelerator designed for edge devices, representative of hardware used in modern UAV companion computers.

The performance data (latency in milliseconds) was sourced directly from the official Hailo Model Zoo benchmarks. These benchmarks are conducted on the Hailo-8L M.2 PCIe module using the HailoRT (Hailo Runtime) environment. The analysis includes approximately 40 object detection models from the YOLO and DETR families, providing a statistically robust dataset for this investigation.

Crucially, the models benchmarked on the Hailo-8L utilize Post-Training Quantization (PTQ) to convert the model weights and activations from 32-bit floating-point (FP32) to 8-bit integer (INT8) format. This quantization is necessary for efficient execution on the NPU. While the meta-analysis (Tab. 1) reports accuracy in FP32 precision—as this is the standard metric reported in the original research papers and allows for fair architectural comparison—it is important to acknowledge that INT8 quantization typically results in a minor accuracy degradation (usually 1–2 % mAP) in practical deployment. The Pareto analysis in this study is based on FP32 accuracy and theoretical GFLOPs to identify the most efficient architectures. The Hailo-8L analysis (Fig. 3) then serves to validate whether this theoretical efficiency (GFLOPs) translates to practical execution speed (latency) on quantized models running on target hardware.

The research methodology was based on a comparative meta-analysis of the collected data. The following key metrics were selected to evaluate the "speed-accuracy" trade-off:

- **Mean Average Precision (mAP):** The mAP@0.50 indicator was mainly used (detection is considered correct if the intersection over union (IOU) between the predicted and reference frames exceeds 0.50). Notably, mAP@0.50 is preferable to the more stringent mAP@0.5:0.95 (which averages precision across IoU thresholds from 0.50 to 0.95 in increments of 0.05) when evaluating small objects, as IoU-based assessments tend to underperform for diminutive targets due to their sensitivity to minor pixel-level discrepancies.
- **Number of parameters (Parameters, M):** This metric is measured in millions (M) and indicates the size of the model, its complexity, and, accordingly, the memory requirements for storing and running the model.
- **Number of floating point operations (FLOPs, G):** Measured in billions and quantifies the computational cost of a single forward pass of the model (inference). This metric is often used to estimate theoretical computational efficiency.

Performance data (mAP, Parameters, FLOPs) were primarily extracted from the original research papers. In cases where the number of parameters or FLOPs was not explicitly reported (marked with "*" in Tab. 1), they were estimated using a standardized methodology. The estimation relied on architectural similarity and theoretical calculation/profiling. If the model is a direct modification of a known baseline (e.g., YOLOv5s with an added module), the baseline's values were used as a starting point, and the cost of the added modules was estimated based on their description. We also reconstructed the model architecture based on the description provided in the paper, utilizing the official source code if available. The theoretical

FLOPs and parameter counts were then calculated using standard PyTorch profiling tools, such as torchinfo or the thop (PyTorch-OpCounter) library. For example, the estimations for the EDNet series (Nos. 5, 7, 9, 11, 30, 47, 48) were performed by profiling the architecture based on the details provided in (Song et al., 2024). All calculations assumed the input image resolution specified in the respective paper; if the resolution was not specified, a standard benchmark resolution (e.g., 640 x 640) was used for the estimation.

Pareto front analysis was used to visualize and analyze the trade-off between these metrics. The model A is considered Pareto-optimal in the set of models S if there is no other model $B \in S$ that is simultaneously inferior to A in all criteria and superior to it in at least one criterion. That is, A is Pareto optimal if $\nexists B \in S : [\forall i, f_i(B) \geq f_i(A)] \wedge [\exists j, f_j(B) > f_j(A)]$, where f_j are criterion functions (mAP, number of parameters, GFLOPs) that need to be maximized or minimized depending on the task. The models that form the Pareto front represent a set of optimal solutions for which it is impossible to improve one indicator (e.g., mAP) without worsening another (e.g., increasing the number of parameters or FLOPs).

4. Research results

We start by benchmarking the surveyed models on VisDrone2019 across three axes – mAP@0.5, parameter count, and GFLOPs-and we summarize the results in Table 1. The literature analysis revealed a number of architectures that were specifically designed or adapted with a focus on real-time performance and computational efficiency for deployment on UAVs. To quantitatively assess the accuracy-efficiency trade-off, the performance data of various models on the VisDrone2019 dataset was analyzed. Table 1 provides summary data for representative models, including those at the Pareto frontier and well-known baseline models.

Table 1

Comparison of the performance of selected object detection models on VisDrone2019-DET.

No	Object detection model	Family	mAP@.50 (%)	Params (M)	FLOPs (G)	Pareto optimal categories
1	Drone-DETR (Kong et al., 2024)	DETR	53.9	28.7	128.3	M, FLOPs
2	VRF-DETR (Wenbin, 2025)	DETR	51.4	13.5	44.3	M, FLOPs
3	UAV-DETR-R50 (Zhang et al., 2025)	DETR	51.1	42.0	170	-
4	RT-DETR-R50 (Wenbin, 2025; Zhao et al., 2024)	DETR	50.7	41.8	133.2	-
5	EDNet-X (Song et al., 2024)	YOLOv10	50.2	48.7	270.4*	-
6	DBYOLOv8s (Tan et al., 2025)	YOLOv8s	49.3	11	119.8	M
7	EDNet-L (Song et al., 2024)	YOLOv10l	49.0	31.7	179.2*	-
8	Modified YOLOv8l (Wang et al., 2025)	YOLOv8l	48.3	31.2	490.9	-
9	EDNet-B (Song et al., 2024)	YOLOv10	48.3	25.5	141.1*	-
10	RT-DETR-R18 (Wenbin, 2025; Zhao et al., 2024)	DETR	47.2	20.2	57.0	-
11	EDNet-M (Song et al., 2024)	YOLOv10	47.1	19.1	88.4*	-
12	SL-YOLO (Chen, & Zhang, 2025)	YOLOv8s	46.9	9.6	36.7*	M, FLOPs
13	YOLOv9e (Song et al., 2024; Wang et al., 2025)	YOLOv9	46.5	57.4	189.0	-
14	LSOD-YOLOv8s (Li, & Wu, 2024)	YOLOv8n	45.2	2.7	23.5	-
15	MSGD-YOLO (Xiao et al., 2024)	YOLOv8n	45.2	2.6	21.5	M, FLOPs
16	YOLOv10x (Song et al., 2024; Hussain, 2024)	YOLOv10	44.8	31.6	160.4	-
17	YOLOv5x (Song et al., 2024; Hussain, 2024)	YOLOv5	44.7	97.2	205.7	-
18	YOLOv8x (Song et al., 2024; Hussain, 2024)	YOLOv8	44.7	68.1	257.8	-
19	YOLOv5s VIOU+PFFN+ADhead (Lu et al., 2023)	YOLOv5	44.6	19.26	112.6*	-
20	HIC-YOLOv5 (Tang et al., 2024; Wenbin, 2025)	YOLOv5	44.3	10.5	31.2	-
21	YOLOv10l (Song et al., 2024; Hussain, 2024)	YOLOv10	44.3	25.7	120.3	-
22	YOLOv8l (Song et al., 2024; Hussain, 2024)	YOLOv8	43.6	43.6	165.2	-
23	YOLOv10-B (Hussain, 2024)	YOLOv10	43.4	20.4	92	-
24	YOLOv5s VIOU+PFFN (Lu et al., 2023)	YOLOv5	43.2	7.4	39.6*	-
25	YOLOv9m (Song et al., 2024; Wang et al., 2025)	YOLOv9	43.1	20.0	76.3	-
26	YOLOv9-C (Wang et al., 2025)	YOLOv9	43.1	25.3	102.1	-
27	YOLOv12-M (Wenbin, 2025)	YOLOv12	43.1	20.2	67.5	-
28	CPAM-YOLO (Wu, Zhu, & Cao, 2024)	YOLOv5l	43.0	27.5	106.4	-
29	YOLOv5l (Song et al., 2024; Hussain, 2024)	YOLOv5	42.9	53.1	109.1	-
30	EDNet-S (Song et al., 2024)	YOLOv10	42.5	9.3	37.1*	-
31	Modified YOLOv8s (Li et al., 2023)	YOLOv8	42.2	9.6	22*	-
32	YOLOv8m (Song et al., 2024; Hussain, 2024)	YOLOv8	41.9	25.9	78.9	-
33	DM-YOLOX (Li et al., 2024)	YOLOX	41.9	9.6	27.9	-
34	LV-YOLOv5 (Wang et al., 2022; Wang et al., 2025)	YOLOv5	41.7	36.6	38.8	-
35	YOLOv10m (Song et al., 2024; Hussain, 2024)	YOLOv10	41.5	16.5	59.1	-
36	PDWT-YOLO (Zhang et al., 2023)	YOLOv7-tiny	41.2	6.4	24.2	-
37	YOLOv5m (Song et al., 2024; Hussain, 2024)	YOLOv5	41.1	21.2*	49	-
38	UN-YOLOv5s (Guo et al., 2023)	YOLOv5	40.5	9.1*	37.4	-
39	LAI-YOLOv5s (Deng et al., 2023)	YOLOv5	40.4	6.3	29.0	-
40	YOLOv9s (Song et al., 2024; Wang et al., 2025)	YOLOv9	39.0	7.2	26.4	-
41	YOLOv8s (Jiao, & Abdullah, 2024)	YOLOv8	38.9	11.1	28.5	-
42	YOLOv5s (Wang et al., 2025)	YOLOv5	38.5	9.1	23.8	-
43	YOLOv10s (Song et al., 2024; Hussain, 2024)	YOLOv10	38.5	8.0	21.6	-
44	C3TB-YOLOv5 (Wu et al., 2024)	YOLOv5	38.3	8.0	19.7	FLOPs
45	MPE-YOLO (Su et al., 2024)	YOLOv8	37.0	4.4	11.9*	FLOPs
46	YOLOv7-tiny (Wang et al., 2025)	YOLOv7	33.9	6.0	13.3	-

47	EDNet-N (Song et al., 2024)	YOLOv10	33.8	2.9	15.2*	-
48	EDNet-Tiny (Song et al., 2024)	YOLOv10	33.3	1.8	14.2*	M
49	YOLOv8n (Wang et al., 2025)	YOLOv8	33.2	3.0	8.1	FLOPs
50	Faster R-CNN (ResNet50-FPN) (Kong et al., 2024)	R-CNN	33.2	41.2	207.1	-
51	DCE-YOLOv8n (An et al., 2024)	YOLOv8	32.7	4.82*	13*	-
52	YOLOv10n (Song et al., 2024; Hussain, 2024)	YOLOv10	32.3	2.7	6.7	FLOPs
53	YOLOv5n (Wang et al., 2025)	YOLOv5	32.2	2.5	7.1	-
54	YOLOv6s (Wang et al., 2025)	YOLOv6	29.1	16.3	44.2	-
55	CenterNet-ResNet-50 (Lu et al., 2023)	CenterNet	29.0	32*	120*	-
56	RetinaNet-ResNet-50-FPN (Kong et al., 2024)	RetinaNet	27.3	19.8	93.7	-
57	SSD-ResNet-50 (Kong et al., 2024)	SSD	23.9	24.5	87.9	-
58	YOLOv3-tiny (Wang et al., 2025)	YOLOv3	23.6	12.1	18.9	-
59	EfficientDet-D1 (Lu et al., 2023)	EfficientDet	21.2	6.6*	6.1*	FLOPs

Values marked with "*" were not directly reported in the original works, but were approximated by the authors using the methodology detailed in Section 3. In the "Pareto optimal" column, "M" denotes Pareto optimal models by the number of parameters, and FLOPs by the number of operations.

Building on these benchmarks, we apply Pareto analysis to reveal the best accuracy-efficiency trade-offs. The scatter plots (Fig. 1: parameters vs. mAP@0.5; Fig. 2: GFLOPs vs. mAP@0.5) make the Pareto front clear along the upper-left edge. The application of Pareto analysis in this task allows not only to systematize existing architectures, but also to objectively evaluate their effectiveness in a multidimensional space of criteria. The models that form the Pareto front are benchmarks in terms of the balance between accuracy, speed, and size, and they should be considered a priority for practical implementation. This approach eliminates subjective model selection based on a single criterion and promotes transparency and reproducibility of research in the field. This approach ensures that all models on the Pareto frontier are the best possible compromises for the given constraints and requirements. The results of the meta-analysis are presented in summary tables and graphs, which are a summary visualization demonstrating the best models for practical application on UAVs, where each of these models represents the optimal trade-off between accuracy and efficiency for different hardware constraints.

It is noteworthy that models such as MSGD-YOLO, SL-YOLO, VRF-DETR, and Drone-DETR appear on both sides of the Pareto front, indicating their exceptional balance in terms of both size and computational complexity relative to the achieved accuracy.

To connect these trade-offs with real device behavior, we relate GFLOPs to latency on a representative UAV companion computer – the Hailo-8L NPU, using YOLO/DETR models as a case study (Fig. 3). FLOPs are a useful indicator for comparing architectures, especially within a single model family or on a single hardware platform, but they are not an absolute measure of real-world performance, and testing on target hardware remains necessary. In Fig. 3, a bubble chart is shown demonstrating models for object detection of the YOLO and DETR families on Hailo 8l for another dataset popular in object detection – COCO (Hailo-8l Model Zoo, 2025). The horizontal axis shows the computational cost in gigaflops per image (GFLOPs, log₁₀ scale), and the vertical axis shows the processing delay of a single image in milliseconds (log₁₀ scale). The area of the bubble is proportional to the number of trained parameters (Params M), which forms the third dimension and reflects the memory requirement. The purple subset, which corresponds to YOLOv11 n/s/m/l/x, is also a Pareto optimal model according to (Hailo-8l Model Zoo, 2025).

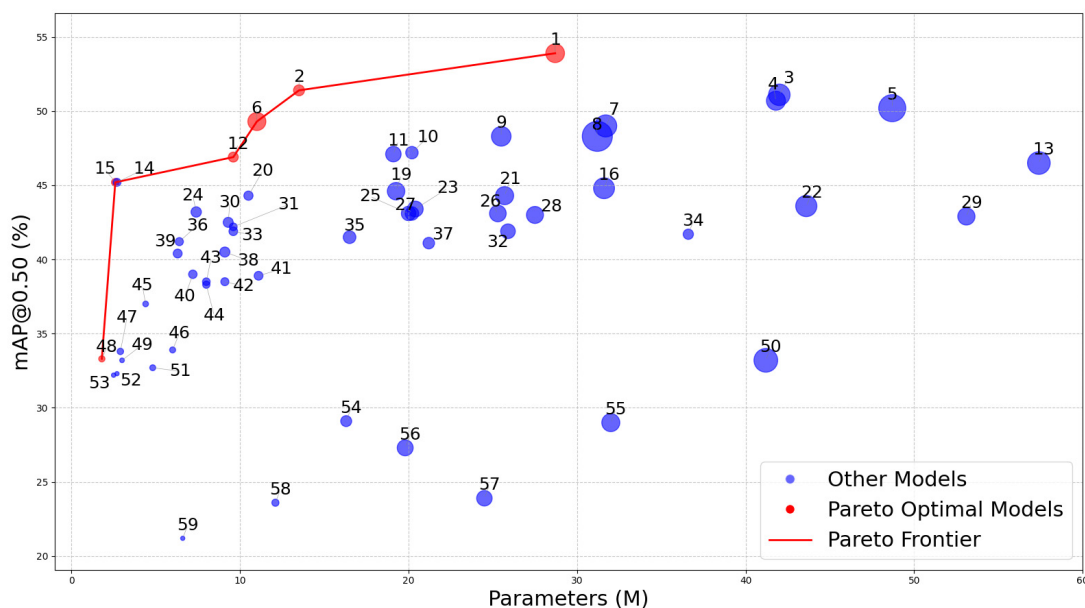


Fig. 1. Number of parameters (M) vsmAP@0.5. Pareto models are shown in red. The size of the bubble is proportional to GFLOPs, the number corresponds to the column "No." in Tab. 1

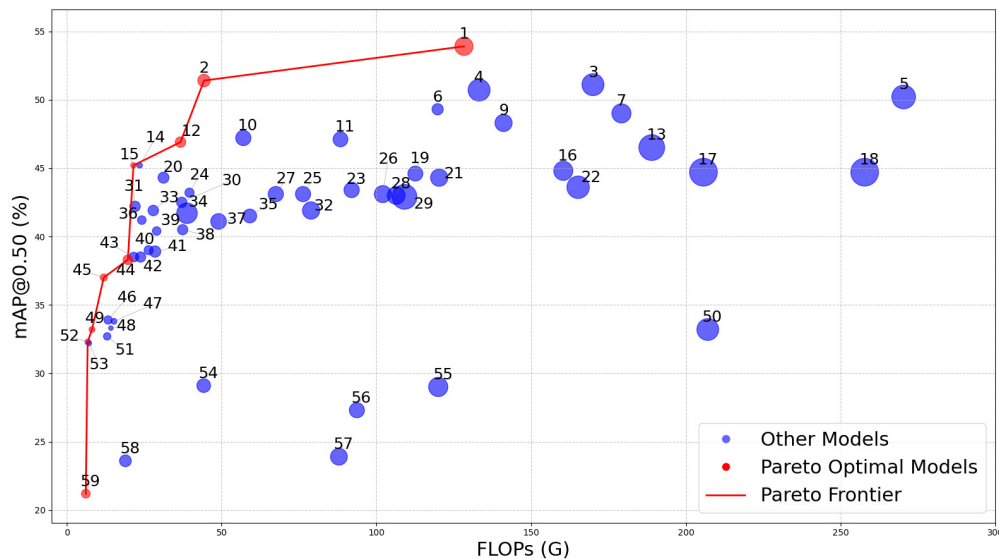


Fig. 2. Flops (G) vs mAP@0.5. Pareto models are shown in red.

The size of the bubble is proportional to the number of parameters, the number corresponds to the column "№" in Tab. 1

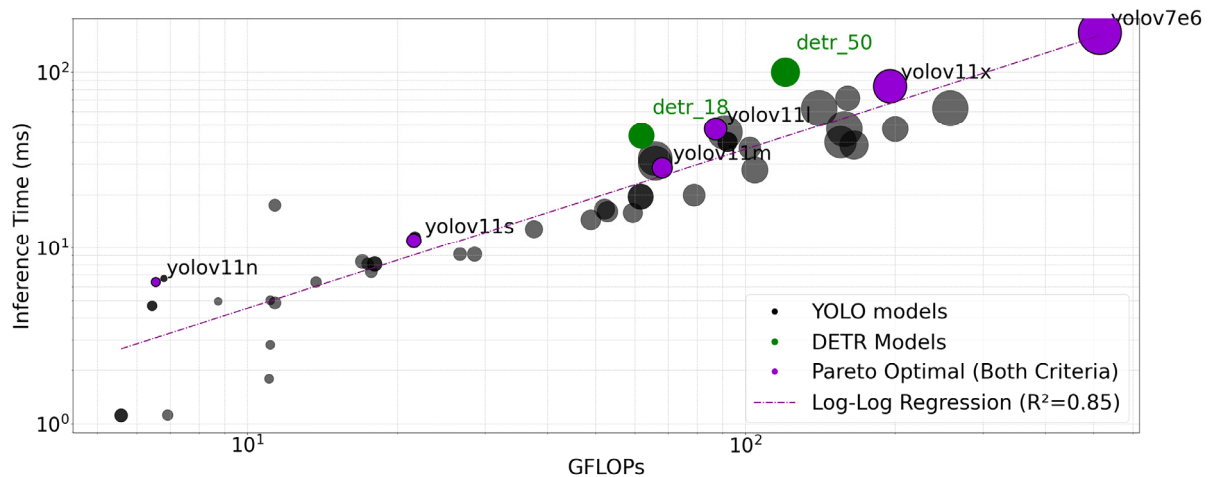


Fig. 3. Dependence of model speed on GFLOPs using the example of Hailo 8L NPU

Log-log orthogonal regression (purple dashed line) summarizes the global trend of the dependence of "GFLOPs on execution time". The analysis is based on approximately 40 data points spanning various YOLO and DETR architectures. The coefficient of determination $R^2 \approx 0.85$ globally, $R^2 \approx 0.92$ for the YOLO family, and an exceptionally high $R^2 \approx 0.98$ for the YOLOv11 subfamily. This strong correlation indicates a clear power-law relationship between inference time and GFLOPs on the target hardware, suggesting the data is statistically robust for this specific accelerator.

This high correlation confirms that GFLOPs can serve as a powerful proxy for runtime on the Hailo-8L, likely due to the high degree of optimization achieved by the Hailo Dataflow Compiler, which minimizes software overheads. However, it is crucial to discuss the limitations of this correlation. While GFLOPs measure the number of arithmetic operations, actual latency is also significantly influenced by Memory Access Cost (MAC), data movement, operator parallelism, and the efficiency of specific kernel implementations on the NPU (Ma et al., 2018).

The slightly lower global correlation ($R^2 \approx 0.85$) compared to intra-family correlation (up to 0.98) highlights architectural differences. For instance, DETR models might exhibit different latency characteristics compared to GFLOPs-equivalent YOLO models due to the differing computational nature and memory access patterns of attention mechanisms versus convolutions. Therefore, while GFLOPs provide a robust heuristic for pre-selecting models, the correlation might weaken when comparing architectures with vastly different operational profiles or when applied to different hardware platforms.

Discussion and conclusions

The results showed that among modern detectors for VisDrone2019, there is a wide range of trade-offs between accuracy and computational complexity. Table 1 shows how the average accuracy of mAP@0.50 varies from 33% for the simplest models to 54% for the best ones, while the number of parameters ranges from 2 million to 97 million and the complexity ranges from 12 to 490 GFLOPs. Thus, higher accuracy is usually achieved by models with higher resource costs, but numerous exceptions indicate significant progress in architectures. For example, the specialized transformer model Drone-DETR achieved the highest mAP = 53.9 % at 128 GFLOPs, while a typical two-stage detector achieved only 48 % at nearly 491 GFLOPs. Similarly, the improved single-stage SL-YOLO model achieved 46.9 % mAP at 36.7 GFLOPs, exceeding the accuracy of the standard large YOLOv5x (44.7 % mAP) with approximately 6 times fewer computations. Thus, through

targeted optimization for UAV applications, some models provide significantly better accuracy per unit of computation than classical detectors. This is particularly important for UAV onboard computers, where computational resources are limited.

To formalize the "accuracy-efficiency" balance, Pareto fronts were constructed based on parameters (Fig. 1) and GFLOPs (Fig. 2).

Pareto analysis objectively identified a limited set of the best solutions. In particular, the Pareto frontier mainly consists of improved versions of YOLO and DETR, specially adapted for small objects. These include MSGD-YOLO, SL-YOLO, VRF-DETR, and Drone-DETR, which are present in both diagrams (by parameters and FLOPs). This indicates their exceptional balance between model size and computational complexity while maintaining high accuracy. Other models on these fronts, such as DB-YOLOv8s, EfficientDet-D1, MPE-YOLO, and C3TB-YOLOv5, provide optimal compromises in their complexity ranges. Importantly, none of the traditional approaches, except for YOLOv8n and YOLOv10n, made it into the Pareto frontier, and none did so on both criteria simultaneously. They either lose in accuracy at equal cost or require more resources to achieve similar accuracy compared to the selected Pareto models. For example, the baseline YOLOv8x achieved 44.7 % mAP at 257.8 GFLOPs, while the much lighter MSGD-YOLO achieved 45.2 % at only 21.5 GFLOPs.

The dominance of YOLO and DETR variants on the Pareto front is consistent with general trends in object detection, but their success in the UAV domain (specifically VisDrone) warrants architectural discussion. The efficiency of YOLO architectures stems primarily from their single-stage design, which avoids the computational overhead of region proposal networks found in two-stage detectors. Furthermore, modern YOLO variants utilize efficient backbones and optimized neck/head structures that effectively balance feature fusion across scales – critical for detecting the small, densely packed objects common in aerial imagery. In contrast, DETR-based models leverage the global context provided by self-attention mechanisms. This allows them to better model relationships between objects and the complex backgrounds typical of VisDrone, potentially reducing false positives. While early DETR models struggled with small objects and high computational costs, specialized variants like Drone-DETR and VRF-DETR incorporate mechanisms (e.g., enhanced high-resolution features or specialized queries) that specifically address these limitations, making them competitive and often superior in this specific domain.

Thus, the problem of selecting the optimal model for a hardware-constrained UAV is actually solved by determining this Pareto set. The identified models achieve mAP@0.50 ~50 % with parameter sizes ≤ 15 M and costs ≤ 60 GFLOPs, which is consistent with the typical capabilities of onboard systems. This means that UAVs with different resources should use different models to achieve the best performance.

Results show how well the theoretical computational complexity of a model (GFLOPs) describes its actual performance on the target accelerator. Research using the Hailo 8L NPU confirmed a strong correlation between GFLOPs and inference time: models that perform fewer operations tend to run faster. In Fig. 3 for a set of YOLO and DETR family models, there is a near-exponential dependence between GFLOPs and execution speed, with a coefficient of determination $R^2 \approx 0.85$ globally (for all models) and up to 0.92–0.98 within a single architecture. In other words, in practice, reducing GFLOPs results in an almost proportional reduction in processing time, especially when comparing related models. This is an important result because it confirms the validity of using GFLOPs as a proxy for model speed in the early stages of selection. However, it should be noted that the correlation is not absolute: for example, DETR models may have slightly higher latency than GFLOPs-equivalent YOLO models due to differences in architecture, memory access, and optimization. Therefore, GFLOPs should only be interpreted as a rough guide, and final performance verification should be performed on the target hardware. Despite this caveat, the results obtained demonstrate that for models from the same family (e.g., different versions of YOLO), GFLOPs are almost directly proportional to inference time. Thus, computer vision system engineers for UAVs can confidently use GFLOPs as a preliminary selection metric for models capable of meeting real-time requirements, narrowing down the pool of candidates before detailed testing on hardware.

Data on the performance of over 60 object detection models on the VisDrone2019-DET dataset has been collected and systematized, creating the most comprehensive consolidated database for this task to date. The analysis revealed a wide range of characteristics: from lightweight models with 1.8M parameters (EDNet-Tiny) to heavy ones with 97.2M (YOLOv5x), with an mAP@0.50 accuracy ranging from 21.2 % to 53.9 %. This systematization allows you to quickly navigate the "zoo" of models and make objective comparisons.

Two groups of Pareto-optimal models have been identified, forming the front of the best compromises. Based on the parameter minimization criterion, the optimal models are EDNet-Tiny, MSGD-YOLO, SL-YOLO, DBYOLOv8s, VRF-DETR, and Drone-DETR, which achieve mAP@0.50 ranging from 33.3 % to 53.9 % with 1.8–28.7M parameters. Based on the criterion of minimizing GFLOPs, the optimal models are EfficientDet-D1, YOLOv10n, YOLOv8n, MPE-YOLO, C3TB-YOLOv5, MSGD-YOLO, SL-YOLO, VRF-DETR, and Drone-DETR with a range of 6.1–128.3 GFLOPs. It is noteworthy that four models (MSGD-YOLO, SL-YOLO, VRF-DETR, Drone-DETR) are present on both fronts, which indicates their exceptional balance.

The high correlation between the theoretical GFLOPs metric and the practical execution speed has been experimentally confirmed using the Hailo 8L NPU as an example. The coefficient of determination $R^2 \approx 0.85$ for all models and $R^2 \approx 0.92$ –0.98 for models of the same family indicates an almost exponential dependence, which validates the use of GFLOPs as the primary model selection criterion. This simplifies the process of selecting an architecture for specific UAV hardware constraints without the need for full-scale testing on the target platform.

Data availability: For a more detailed and interactive study of the data, allowing dynamic filtering and comparison of models, a web page has been created, available at: https://smolinyaroslav.github.io/visdrone_literature_review

Authors' contribution: Yaroslav Smolin – conceptualization, methodology, formal analysis, software, validation, writing – original draft. Viacheslav Liskin – supervision, guidance on research design and interpretation of results, writing – review & editing.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

- Akyon, F. C., Altınuç, S. O., & Temizel, A. (2022). Slicing aided hyper inference and fine-tuning for small object detection. In Y. Berthoumieu & P. Frossard (Eds.), *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 966–970). IEEE. <https://doi.org/10.1109/ICIP46576.2022.9897990>
- An, J., Lee, D., Putro, M. D., & Kim, B. (2024). DCE-YOLOv8: Lightweight and Accurate Object Detection for Drone Vision. *IEEE Access*, 12, 170898–170912. <https://doi.org/10.1109/ACCESS.2024.3481410>
- Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving Into High Quality Object Detection. In D. Forsyth, I. Laptev, D. Ramanan, & A. Oliva (Eds.), *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6154–6162). IEEE. <https://doi.org/10.1109/CVPR.2018.00644>
- Camarlinghi, N., Michelozzi, B., Martino, G., Di Tommaso, A., Fontanelli, G., & Masini, A. (2024). Drone-based monitoring on the edge using a high-resolution payload. In J. Dijk & J. L. Sanchez-Lopez (Eds.), *Autonomous Systems for Security and Defence* (p. 1–7). SPIE. <https://doi.org/10.1117/12.3031453>
- Cao, Z., Kooistra, L., Wang, W., Guo, L., & Valente, J. (2023). Real-Time Object Detection Based on UAV Remote Sensing: A Systematic Literature Review. *Drones*, 7(10), 620. <https://doi.org/10.3390/drones7100620>
- Chen, D., & Zhang, L. (2025). SL-YOLO: A Stronger and Lighter Drone Target Detection Model. arXiv. <https://doi.org/10.48550/arXiv.2411.11477>
- Chen, L., Hu, J., Li, X., Quan, F., & Chen, H. (2021). Onboard Real-time Object Detection for UAV with Embedded NPU. In X. Ning & Y. Haibin (Eds.), *2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 192–197). IEEE. <https://doi.org/10.1109/CYBER53097.2021.9588193>
- Deng, L., Bi, L., Li, H., Chen, H., Duan, X., Lou, H., Zhang, H., Bi, J., & Liu, H. (2023). Lightweight aerial image object detection algorithm based on improved YOLOv5s. *Scientific Reports*, 13(1), 7817. <https://doi.org/10.1038/s41598-023-34892-4>
- Du, B., Liao, Z., Zhang, Y., Cai, Z., Chen, J., & Huang, D. (2024). Centerness-based Instance-aware Knowledge Distillation with Task-wise Mutual Lifting for Object Detection on Drone Imagery. arXiv. <https://doi.org/10.48550/arXiv.2411.02861>
- Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., Bo, L., Shi, H., Zhu, R., Kumar, A., Li, A., Zinollayev, A., Askergaliyev, A., Schumann, A., Mao, B., Liu, Z. (2019). VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 213–226. <https://doi.org/10.1109/ICCVW.2019.00030>
- Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint Triplets for Object Detection. In I. S. Kweon, N. Paragios, M.-H. Yang, & S. Lazebnik (Eds.), *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6568–6577). <https://doi.org/10.1109/ICCV.2019.00667>
- Elmokadem, T., & Savkin, A. V. (2021). Towards Fully Autonomous UAVs: A Survey. *Sensors*, 21(18), 6223. <https://doi.org/10.3390/s21186223>
- Fernandez, J., Kahn, J., Na, C., Bisk, Y., & Strubell, E. (2023). The Framework Tax: Disparities Between Inference Efficiency in NLP Research and Deployment. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1588–1600). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.98>
- Guo, J., Liu, X., Bi, L., Liu, H., & Lou, H. (2023). UN-YOLOv5s: A UAV-Based Aerial Photography Detection Algorithm. *Sensors*, 23(13), 5907. <https://doi.org/10.3390/s23135907>
- Gutierrez, G., Llerena, J. P., Usero, L., & Patricio, M. A. (2025). A Comparative Study of Convolutional Neural Network and Transformer Architectures for Drone Detection in Thermal Images. *Applied Sciences*, 15(1), 109. <https://doi.org/10.3390/app15010109>
- Hailo-8l model zoo. (2025). https://github.com/hailo-ai/hailo_model_zoo/blob/master/docs/public_models/HAILO8L_object_detection.rst
- He, L., & Wang, M. (2023). SliceSamp: A Promising Downsampling Alternative for Retaining Information in a Neural Network. *Applied Sciences*, 13(21), 11657. <https://doi.org/10.3390/app132111657>
- Hua, W., & Chen, Q. (2025). A survey of small object detection based on deep learning in aerial images. *Artificial Intelligence Review*, 58(6), 162. <https://doi.org/10.1007/s10462-025-11150-9>
- Hussain, M. (2024). YOLOv5, YOLOv8, and YOLOv10: The Go-To Detectors for Real-time Vision. arXiv. <https://doi.org/10.48550/arXiv.2407.02988>
- Jayanth, R., Gupta, N., & Prasanna, V. (2024). Benchmarking Edge AI Platforms for High-Performance ML Inference. In E. Mortensen (Ed.), *2024 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1–7). <https://doi.org/10.1109/HPEC62836.2024.10938499>
- Jiao, L., & Abdullah, M. I. (2024). YOLO series algorithms in object detection of unmanned aerial vehicles: A survey. *Service Oriented Computing and Applications*, 18(3), 269–298. <https://doi.org/10.1007/s11761-024-00388-w>
- Keles, M. C., Salmanoğlu, B., Guzel, M. S., Gursoy, B., & Bostanci, G. E. (2022). Evaluation of YOLO Models with Sliced Inference for Small Object Detection. arXiv. <https://doi.org/10.48550/arXiv.2203.04799>
- Kong, Y., Shang, X., & Jia, S. (2024). Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model. *Sensors*, 24(17), 5496. <https://doi.org/10.3390/s24175496>
- Law, H., & Deng, J. (2020). CornerNet: Detecting Objects as Paired Keypoints. *International Journal of Computer Vision*, 128(3), 642–656. <https://doi.org/10.1007/s11263-019-01204-1>
- Li, H., & Wu, J. (2024). LSOD-YOLOv8s: A Lightweight Small Object Detection Model Based on YOLOv8 for UAV Aerial Images. *Engineering Letters*, 32(11), 2073–2082.
- Li, X., Wang, F., Wang, W., Han, Y., & Zhang, J. (2024). DM-YOLOX aerial object detection method with intensive attention mechanism. *The Journal of Supercomputing*, 80(9), 1–23. <https://doi.org/10.1007/s11227-024-05944-x>
- Li, Y., Fan, Q., Huang, H., Han, Z., & Gu, Q. (2023). A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones*, 7(5), 304. <https://doi.org/10.3390/drones7050304>
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
- Liu, S., Zha, J., Sun, J., Li, Z., & Wang, G. (2023). EdgeYOLO: An Edge-Real-Time Object Detector. In J. Han & J. Sun (Eds.), *2023 42nd Chinese Control Conference (CCC)* (pp. 7507–7512). IEEE. <https://doi.org/10.23919/CCC58697.2023.10239786>
- Liu, W., Angelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-0_2
- Liu, Y., He, M., & Hui, B. (2025). ESO-DETR: An Improved Real-Time Detection Transformer Model for Enhanced Small Object Detection in UAV Imagery. *Drones*, 9(2), 143. <https://doi.org/10.3390/drones9020143>
- Lu, S., Lu, H., Dong, J., & Wu, S. (2023). Object Detection for UAV Aerial Scenarios Based on Vectorized IOU. *Sensors*, 23(6), 3061. <https://doi.org/10.3390/s23063061>

- Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 122–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-01264-9_8
- Mirzaei, B., Nezamabadi-pour, H., Raoof, A., & Derakhshani, R. (2023). Small Object Detection and Tracking: A Comprehensive Review. *Sensors*, 23(15), 6887. <https://doi.org/10.3390/s23156887>
- Nia, V. P., Ghaffari, A., Zolnouri, M., & Savaria, Y. (2022). *Rethinking Pareto Frontier for Performance Evaluation of Deep Neural Networks*. arXiv. <https://doi.org/10.48550/arXiv.2202.09275>
- Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., & Moghaddam, M. E. (2025). *Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications*. arXiv. <https://doi.org/10.48550/arXiv.2503.20516>
- Peng, J., Xie, L., Zhang, Z., Tan, T., & Wang, J. (2018). *Accelerating Deep Neural Networks with Spatial Bottleneck Modules*. arXiv. <https://doi.org/10.48550/arXiv.1809.02601>
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In J. A. Schnabel & K. Mori (Eds.), *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788). IEEE. <https://doi.org/10.1109/CVPR.2016.91>
- Rekavandi, A. M., Rashidi, S., Boussaid, F., Hoefs, S., Akbas, E., & Bennamoun, M. (2023). *Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art*. arXiv. <https://doi.org/10.48550/arXiv.2309.04902>
- Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. Lawrence, & M. Sugiyama (Eds.), *Advances in Neural Information Processing Systems* (pp. 91–99). Curran Associates, Inc. <https://doi.org/10.48550/arXiv.1506.01497>
- Song, Z., Zhang, Y., & Ebayyeh, A. A. R. M. A. (2024). EDNet: Edge-Optimized Small Target Detection in UAV Imagery – Faster Context Attention, Better Feature Fusion, and Hardware Acceleration. In H. Liu & K. Qian (Eds.), *2024 IEEE Smart World Congress (SWC)* (pp. 829–838). <https://doi.org/10.1109/SWC62898.2024.00141>
- Su, J., Qin, Y., Jia, Z., & Liang, B. (2024). MPE-YOLO: Enhanced small target detection in aerial imaging. *Scientific Reports*, 14(1), 17799. <https://doi.org/10.1038/s41598-024-68934-2>
- Suo, J., Zhang, X., Shi, W., & Zhou, W. (2024). E\$^3\$-UAV: An Edge-based Energy-Efficient Object Detection System for Unmanned Aerial Vehicles. *IEEE Internet of Things Journal*, 11(3), 4398–4413. <https://doi.org/10.1109/IJOT.2023.3301623>
- Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and Efficient Object Detection. In G. Mori, K. Saenko, & S. Savarese (Eds.), *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10778–10787). <https://doi.org/10.1109/CVPR42600.2020.01079>
- Tan, Y., Xu, B., Sun, J., Xu, C., Pan, W., Dai, S., & Liu, H. (2025). DBYOLOv8: Dual-Branch YOLOv8 Network for Small Object Detection on Drone Image. *International Journal of Advanced Computer Science and Applications*, 16(1), 1301–1309. <https://doi.org/DOI10.14569/IJACSA.2025.01601124>
- Tang, S., Zhang, S., & Fang, Y. (2024). HIC-YOLOv5: Improved YOLOv5 For Small Object Detection. In Y. Hasegawa (Ed.), *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6614–6619). <https://doi.org/10.1109/ICRA57147.2024.10610273>
- Tang, X., Chen, X., Cheng, J., Wu, J., Fan, R., Zhang, C., & Zhou, Z. (2024). YOLO-Ant: A Lightweight Detector via Depthwise Separable Convolutional and Large Kernel Design for Antenna Interference Source Detection. *IEEE Transactions on Instrumentation and Measurement*, 73, 3503817. <https://doi.org/10.1109/TIM.2024.3379397>
- Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. arXiv. <https://doi.org/10.48550/arXiv.2301.10051>
- Vaddi, S., Kim, D., Kumar, C., Shad, S., & Jannesari, A. (2021). Efficient Object Detection Model for Real-time UAV Application. *Computer and Information Science*, 14(1), 45–55. <https://doi.org/10.5539/cis.v14n1p45>
- Wan, J., Zhang, B., Zhao, Y., Du, Y., & Tong, Z. (2021). VistrongerDet: Stronger Visual Information for Object Detection in VisDrone Images. In L. O’Conner (Ed.), *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2820–2829). IEEE. <https://doi.org/10.1109/ICCVW54120.2021.00316>
- Wang, C.-Y., Yeh, I.-H., & Mark Liao, H.-Y. (2025). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (pp. 1–21). Springer. https://doi.org/10.1007/978-3-031-72751-1_1
- Wang, H., Liu, J., Zhao, J., Zhang, J., & Zhao, D. (2025). Precision and speed: LSOD-YOLO for lightweight small object detection. *Expert Systems with Applications*, 269, 126440. <https://doi.org/10.1016/j.eswa.2025.126440>
- Wang, J., Liu, W., Zhang, W., & Liu, B. (2022). LV-YOLOv5: A light-weight object detector of Vit on Drone-captured Scenarios. In Z. Li (Ed.), *2022 16th IEEE International Conference on Signal Processing (ICSP)* (pp. 178–183). IEEE. <https://doi.org/10.1109/ICSP56322.2022.9965217>
- Wang, X., Peng, Y., & Shen, C. (2025). *Efficient Feature Fusion for UAV Object Detection*. arXiv. <https://doi.org/10.48550/arXiv.2501.17983>
- Wenbin, L. (2025). *An Efficient Aerial Image Detection with Variable Receptive Fields*. arXiv. <https://doi.org/10.48550/arXiv.2504.15165>
- Wu, H., Zhu, Y., & Cao, M. (2024). An algorithm for detecting dense small objects in aerial photography based on coordinate position attention module. *IET Image Processing*, 18(7), 1759–1767. <https://doi.org/10.1049/ipr2.13061>
- Wu, Q., Li, Y., Huang, W., Chen, Q., & Wu, Y. (2024). C3TB-YOLOv5: Integrated YOLOv5 with transformer for object detection in high-resolution remote sensing images. *International Journal of Remote Sensing*, 45, 2622–2650. <https://doi.org/10.1080/01431161.2024.2329528>
- Xiao, L., Li, W., Yao, S., Liu, H., & Ren, D. (2024). High-precision and lightweight small-target detection algorithm for low-cost edge intelligence. *Scientific Reports*, 14(1), 23542. <https://doi.org/10.1038/s41598-024-75243-1>
- Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., & Xia, G.-S. (2022). Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 79–93. <https://doi.org/10.1016/j.isprsjprs.2022.06.002>
- Yang, C., Huang, Z., & Wang, N. (2022). QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In K. Dana, G. Hua, S. Roth, S. Dimitris, & R. Singh (Eds.), *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13658–13667). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01330>
- Zhang, H., Liu, K., Gan, Z., & Zhu, G.-N. (2025). UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery. arXiv. <https://doi.org/10.48550/arXiv.2501.01855>
- Zhang, H., & Zhang, S. (2024). *Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale*. arXiv. <https://doi.org/10.48550/arXiv.2312.17663>
- Zhang, L., Xiong, N., Pan, X., Yue, X., Wu, P., & Guo, C. (2023). Improved Object Detection Method Utilizing YOLOv7-Tiny for Unmanned Aerial Vehicle Photographic Imagery. *Algorithms*, 16(11), 520. <https://doi.org/10.3390/a16110520>

- Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs Beat YOLOs on Real-time Object Detection. In E. Mortensen (Ed.), *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16965–16974). IEEE. <https://doi.org/10.1109/cvpr52733.2024.01605>
- Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., Nie, Q., Cheng, H., Liu, C., Liu, X., Ma, W., Wu, H., Wang, L., Schumann, A., Brown, C., Qian, C., Li, C., Li, D., Michail, E., & Song, Z. (2019). VisDrone-DET2018: The Vision Meets Drone Object Detection in Image Challenge Results. In L. Leal-Taixé & S. Roth (Eds.), *Computer Vision – ECCV 2018 Workshops* (Vol. 11134, pp. 437–468). https://doi.org/10.1007/978-3-030-11021-5_27

Отримано редакцією журналу / Received: 08.08.25

Прорецензовано / Revised: 15.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Ярослав СМОЛІН, асп.

ORCID ID: 0009-0007-9744-6496

e-mail: smolin.yaroslav@iit.kpi.ua

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

Вячеслав ЛІСКІН, канд. техн. наук, доц.

ORCID ID: 0000-0002-9418-0633

e-mail: liskin.viacheslav@iit.kpi.ua

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

МЕТААНАЛІЗ ПАРЕТО-ОПТИМАЛЬНИХ МОДЕЛЕЙ ВИЯВЛЕННЯ ОБ'ЄКТІВ ДЛЯ БЕЗПІЛОТНИХ ЛІТАЛЬНИХ АПАРАТІВ НА НАБОРІ ДАНИХ VISDRONE2019-DET

Предметом пропонованого дослідження є сучасні моделі глибокого навчання для виявлення об'єктів із безпілотних літальних апаратів (БПЛА), навчені й оцінені на еталонному наборі даних VisDrone2019-DET. Проблема, яку розглянуто в роботі, полягала у відсутності систематичного аналізу наявних моделей на основі критерію оптимального балансу між точністю виявлення, швидкістю обчислення та розміром моделі, що є критично важливими для роботи в режимі реального часу на бортових комп'ютерах БПЛА з обмеженими ресурсами. Основний результат полягає у всебічному метааналізі понад 60 наукових публікацій і визначенні двох груп Парето-оптимальних моделей: одна зосереджена на мінімізації параметрів моделі, а інша – на мінімізації обчислювальної складності. За допомогою багатокритеріального аналізу Парето та створення найповнішої на даний момент таблиці підсумкових показників ефективності, у ході дослідження було визначено архітектури, що забезпечують найкращий компроміс між точністю, швидкістю та розміром. Ними виявилися оптимізовані версії YOLO та детекторів на основі трансформерів, таких як DETR, що пояснюється ефективністю їхньої основної архітектури та додатковим налаштуванням під особливості аерофотознімок. Використовуючи прискорювач Nvivo 8L як приклад, дослідження також продемонструвало, що обчислювальна складність моделі прямо пропорційна (з коефіцієнтом детермінації до 0,98) її швидкості виконання на типовому нейронному прискорювачі, який використовується в комп'ютерах-супутниках БПЛА. Це дає змогу інженерам оцінювати відносний час виконання на основі кількості операцій, уникаючи трудомісткого тестування. Загалом отримані результати можуть бути використані інженерами для ухвалення обґрунтованих рішень щодо архітектури детекторів залежно від конкретних обмежень апаратного забезпечення БПЛА, а дослідниками – як орієнтири для розроблення нових моделей для автономної експлуатації БПЛА в повітряному моніторингу, інспекції інфраструктури та пошуково-рятувальних операціях.

Ключові слова: безпілотні літальні апарати, VisDrone2019, детекція об'єктів, моделі глибокого навчання, аналіз Парето.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; у рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 517.98

DOI: <https://doi.org/10.17721/1812-5409.2026/1.34>

Maryna CHIGIDINA¹, Student
ORCID ID: 0009-0005-2470-033X
e-mail: mchigidina@knu.ua

Iryna LAZARENKO², PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-3384-1186
e-mail: i_lazar@i.kpi.ua

Dmytro NOMIROVSKII¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0001-9482-776X
e-mail: kashpir74@gmail.com

Bogdan RUBLYOV¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0001-6618-797X
e-mail: rublyovbv@gmail.com

Vladimir SEMENOV¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-3280-8245
e-mail: semenov.volodya@knu.ua

Kateryna SVITOVENKO¹, Student
ORCID ID: 0009-0002-9815-5889
e-mail: svitovenko33@knu.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

BASIC PROPERTIES OF THE METRIC OF TANGENTS

The article studies a metric related to the boundaries of figures of a certain class. To properly define it, the class of figures, referred to as figures with convex-concave boundaries, is first introduced. For this, there must exist a finite set of points on the boundary of the figure, such that the arc between any two adjacent points is convex. This allows the boundary of the figure to be considered as a union of a finite number of convex arcs. Since the metric is defined using generalized tangents to the boundary of the figure, and for convex figures, a generalized tangent becomes a supporting line, this approach opens promising prospects for further study of figures of this class. A dense subset of this class is the collection of all polynomials. For practical applications, this serves as an efficient simplification, as figures are usually defined by a set of points on their boundary, and any approximation begins with approximating the figure by a polynomial.

For the given class of figures, a tangent metric is defined, which is determined through generalized tangents to the boundary of the figure. Previously, this metric was studied together with an initial term formed by the Hausdorff metric. This made it possible to avoid verifying the first axiom of the metric, as it followed from the Hausdorff metric. However, it was found that the Hausdorff metric is not a necessary component when defining the tangent metric, as demonstrated by the conducted research. In general, the presence of the Hausdorff metric stemmed from a certain similarity to the metric of the space of continuously differentiable functions, which has the metric of continuous functions as its first term. Therefore, this advancement, which enabled the elimination of the Hausdorff metric as a redundant component, proved to be pivotal. This will also streamline future algorithms for computing the tangent metric.

For further research, firstly, it will be necessary to narrow the class of functions to the set of all polynomials. As a result, studying the properties of the canonical decomposition of a figure within the class of all polynomials has become important. Additionally, for polynomials, constructing generalized tangents becomes a simpler problem.

Keywords: metric of tangents, generalized tangents, smooth metric, canonical decomposition, figure.

AMS 2020 classification: 52A10, 54E35.

Introduction

For various plane figures, the Hausdorff metric is highly popular (Conci, & Kubrusly, 2017; Ortigosa et al., 2021; Kumar et al., 2020; Huttenlocher, Klanderman, & Rucklidge, 1993). Under certain restrictions on the class of figures being considered, a sequence of figures converging in the Hausdorff metric will also have estimates for the convergence of their areas to the limiting figure. Finding analogous metrics for perimeter convergence is significantly more challenging. While some metrics evidently possess this property, they are all primitive, such as:

$$\rho(\Phi_1, \Phi_2) = \rho^{(0)}(\Phi_1, \Phi_2) + |P(\Phi_1) - P(\Phi_2)|,$$

where $\rho^{(0)}$ is the Hausdorff metric, and $P(\Phi)$ is the perimeter of the figure Φ .

The aim of this paper is to propose a fundamentally different approach to constructing such a metric, one that leverages the properties of the boundary of the figure without directly computing its perimeter. This study continues the research initiated in works (Rublev, & Petunin, 1999a, 1999b; Rublev, & Molodtsov, 2000).

1. Class of Figures with Convex-Concave Boundary

Let $S^{(1)}$ denote the class of planar simply connected regions bounded by closed rectifiable Jordan curves. Consider a figure $\Phi \in S^{(1)}$ and define an orientation ω on its boundary $\Gamma(\Phi)$ such that, when moving along the curve $\Gamma(\Phi)$ according to the orientation ω , the figure Φ remains on the right. Let $\overline{AB}$ denote the arc of the curve $\Gamma(\Phi)$, connecting points A and B , that lie on $\Gamma(\Phi)$ in the direction of the orientation ω .

Assume there exists a finite set of points $M_0, M_1, \dots, M_{n-1}$, ordered according to the orientation ω that satisfies the following conditions (for convenience, we assume $M_n = M_0$ and $M_{n+1} = M_1$):

- 1) The arc $\overline{M_i M_{i+1}}$ is distinct from the segment $M_i M_{i+1}$, $i = \overline{0, n-1}$;
- 2) The figure F_i , bounded by the arc $\overline{M_i M_{i+1}}$ and the segment $M_i M_{i+1}$, is convex, $i = \overline{0, n-1}$;
- 3) The figure G_i , bounded by the arc $\overline{M_i M_{i+2}}$ and the segment $M_i M_{i+2}$, is not convex, $i = \overline{0, n-1}$.

Such a set of points $P = \{M_i, i = \overline{0, n-1}\}$ is called the canonical partition of the boundary of the figure, and the figure $\Phi \in S^{(1)}$ is referred to as a figure with a convex-concave boundary (FCCB). The collection of all such figures will be denoted by S . An arc $\overline{M_i M_{i+1}} \subset \Gamma(\Phi)$ is called convex if there exists a segment PQ with endpoints on the arc $\overline{M_i M_{i+1}}$, that lies entirely within the figure Φ , otherwise, it is called concave (Conci, & Kubrusly, 2017; Huttenlocher, Klanderman, & Rucklidge, 1993).

Thus, if $\Phi \in S$, the boundary $\Gamma(\Phi)$ is divided into convex and concave sections by the points of the canonical partition. If the arcs $\overline{M_{i-1} M_i}$ and $\overline{M_i M_{i+1}}$ have different convexity characteristics, the point M_i is called an inflection point of the boundary of the figure Φ .

The properties of the canonical partition of a figure $\Phi \in S$ significantly depend on the structure of the boundary of Φ . We divide all such figures into three subsets. The subset S_1 includes all figures whose boundaries do not contain two points $A, B \in \Gamma(\Phi)$, such that $\overline{AB}$ coincides with the segment AB . The subset S_3 includes all polynomials, and the remaining figures form the subset S_2 , meaning the boundaries of these figures contain at least one straight segment. We define some interesting properties of the canonical partition, some of which significantly depend on the membership of a figure in a particular subset, while others hold for all figures of the class S .

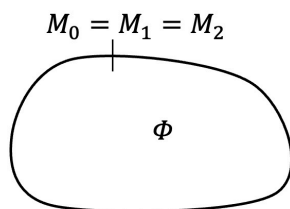


Fig. 1. Convexity and decomposition of the figure Φ

2. Properties (Canonical Decomposition)

Property 1. The canonical decomposition of a figure $\Phi \in S$ consists of a single point, i.e., $P = \{M_0\}$, if and only if the figure Φ is convex (see Fig. 1). In this case, the point M_0 can be any point on the boundary of the figure.

This property is obvious, but it is important to understand condition 3) in the definition of canonical decomposition. It should be noted that the arc $\overline{M_0 M_2}$ differs from the arc $\overline{M_0 M_1}$ in that the former must traverse the entire boundary of the figure twice.

Property 2. If the canonical decomposition of a figure $\Phi \in S^{(1)}$ consists of two points, i.e., $P = \{M_0, M_1\}$, then the arcs $\overline{M_0 M_1}$ and $\overline{M_1 M_2}$ can either have different convexities or both be convex.

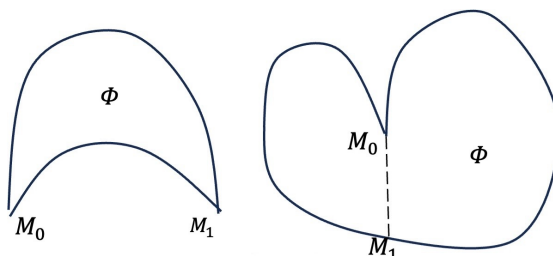


Fig. 2. Convexity and decomposition of the figure Φ

Neither arc can be concave, as it is always possible to draw a segment inside the figure. The other cases are illustrated in Fig. 2. Note that for a figure from the class S_3 this property can hold for the same figure $\Phi \in S_3$. Consider the quadrilateral $ABCD$ (see Fig. 3). If we choose the decomposition $P_1 = \{A, C\}$, the arc $\overline{ABC}$ is convex, while the arc $\overline{CDA}$ is concave. If we choose the decomposition $P_2 = \{B, D\}$, both arcs $\overline{BCD}$ and $\overline{DAB}$ are convex.

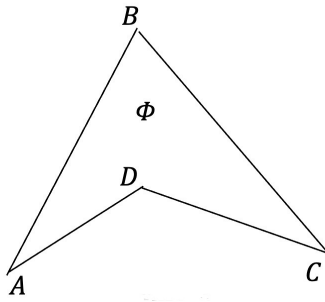


Fig. 3. The quadrilateral $ABCD$

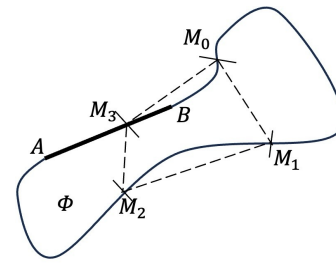


Fig. 4. Figure Φ with segment AB as part of its boundary

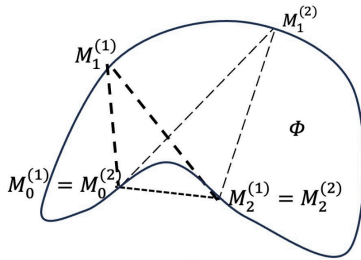


Fig. 5. Figure Φ with two canonical decompositions P_1 and P_2

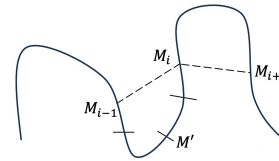


Fig. 6. Figure Φ with a canonical decomposition

Property 3. For certain figures $\Phi \in S$ there exist multiple canonical decompositions.

This property is obvious if $\Phi \notin S_1$, i.e., if the boundary of the figure contains a straight-line segment. In Fig. 4, a figure $\Phi \in S_2$ is shown, where part of the boundary is formed by the segment AB . The canonical decomposition P is defined by the points $\{M_0, M_1, M_2, M_3\}$. At the same time, any point of the segment AB could be chosen as the point M_3 .

It is more interesting to come up with a similar example for a figure from the subset S_1 . This is illustrated in Fig. 5. Here, we have two canonical decompositions: $P_1 = \{M_0^{(1)}, M_1^{(1)}, M_2^{(1)}\}$ and $P_2 = \{M_0^{(2)}, M_1^{(2)}, M_2^{(2)}\}$.

Note that any point on the arc $\widehat{M_1^{(1)}M_1^{(2)}}$ can be chosen as a point of the canonical decomposition of the figure Φ .

Property 4. If a figure $\Phi \in S_1$ has a canonical decomposition where all the decomposition points are inflection points, then this canonical decomposition is unique for such a figure.

Consider a decomposition $P = \{M_i, i = \overline{0, n-1}\}$ of the figure $\Phi \in S_1$, where all the points are inflection points. Suppose there exists another canonical decomposition of the boundary of this figure. Let M' be a point from the other decomposition P' of the figure Φ , and let it lie on the arc $\widehat{M_{i-1}M_i}$ (see Fig. 6). Where can the next point, M'' , in the decomposition P' be located? It cannot be located after the point M_i , because the figure between the segment $M'M''$ and the arc $\widehat{M'M''}$ is not convex. Thus, it must lie on the arc $\widehat{M'M_i}$. Similarly, the point of the canonical decomposition P' , that precedes the point M' , must lie on the arc $\widehat{M_{i-1}M'}$. Therefore, the point M' is redundant according to the definition of a canonical decomposition. The contradiction obtained completes the proof of the property.

In particular, it follows that if for a certain figure $\Phi \in S$ the arcs between any two adjacent points $A, B \in \Gamma(\Phi)$ have different convexities and $\widehat{AB}$ does not contain straight line segments, then the points of any canonical decomposition on this arc are uniquely determined.

Property 5. For figures $\Phi \in S_1$, i.e., for polynomials, there exist canonical decompositions that do not have any common points. Moreover, for one of the decompositions, all the arcs will be convex, while for the other, they will be concave.

One example of such a figure is shown in Fig. 7. Here, for the polynomial $ABCDEF$, there are two canonical decompositions: $P_1 = \{A, C, E\}$ and $P_2 = \{B, D, F\}$. In the first case, all the arcs are concave, while in the second case, they are convex.

Property 6. If a figure $P_1 = \{M_0, M_1, \dots, M_{m-1}\}$ and $P_2 = \{N_0, N_1, \dots, N_{n-1}\}$, that do not have any common points, then $\frac{m}{n} \leq 2$.

Proof. In Fig. 8, a figure is shown for which $\frac{m}{n} = 2$. Here, the case $m = 6$ and $n = 3$ is depicted, which can be easily generalized for any natural values. We see that the figures bounded by $\widehat{M_iM_{i+1}}$ and the segment $[M_iM_{i+1}]$ are either triangles or curvilinear triangles. The figures bounded by $\widehat{N_jN_{j+1}}$ and the segment $[N_jN_{j+1}]$ are curvilinear quadrilaterals.

We will show that this is the maximum possible value of such a ratio. Consider the decomposition $P_2 = \{N_0, N_1, \dots, N_n\}$, and for each point N_j associate the arc $\widehat{N_jN_{j+1}}$. On this arc, no more than two points from another canonical decomposition

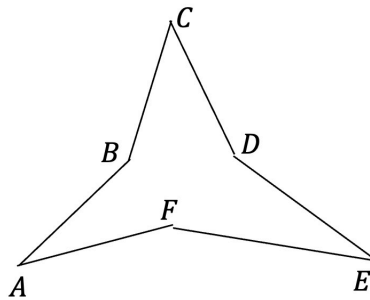


Fig. 7. The polynomial $ABCDEF$

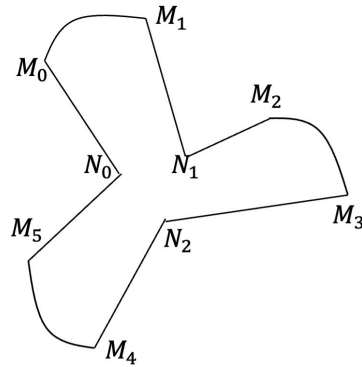


Fig. 8. A figure illustrating the case $\frac{m}{n} = 2$ with $m = 6$ and $n = 3$

can be located. Indeed, if we assume that there are at least three such points, consider three consecutive points in the direction of traversal: A, B, C . Then, the figure bounded by the segment AC and the arc $\widehat{AC}$ is convex, since it lies inside a convex figure bounded by $\widehat{N_j N_{j+1}}$ and the segment $[N_j N_{j+1}]$ (see Fig. 9). This contradicts the definition of a decomposition, so it is not possible. Thus, for each decomposition point N_j , no more than two points from another decomposition can be associated with it, which leads to the inequality $\frac{m}{n} \leq 2$, completing the proof. $\square$

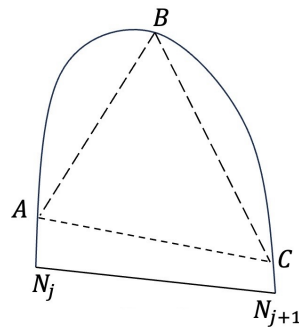


Fig. 9. The figure bounded by the segment AC and the arc $\widehat{AC}$

It is clear why we are focusing so much on polynomials. For practical applications of the metric, all images will have boundaries in the form of polynomials, possibly with a large number of vertices. Therefore, the properties of polynomials themselves will be particularly important.

3. Generalized tangent to figures of class S

In this regard, the concept of FCCB (Figure with Convex-Concave Boundary) without specifying the canonical decomposition will not be entirely correct. Therefore, we will further use the term figure Φ from class S (or $\Phi \in S$, or FCCB) to denote not only the figure Φ , but also the canonical decomposition $P = \{M_0, M_1, \dots, M_{n-1}\}$ of its boundary $\Gamma(\Phi)$. If there are multiple decompositions, we will take any of them and continue the investigation with that decomposition.

For an arbitrary point $M \in \Gamma(\Phi)$, we define the concept of a generalized tangent to the figure $\Phi \in S$, passing through the point M . If $M \notin P$, then $M \in \widehat{M_i M_{i+1}}$. We choose two arcs $\widehat{DM}$ and $\widehat{MC}$, such that $\widehat{DC} \subset \widehat{M_i M_{i+1}}$. Due to the finiteness of the set of points in the canonical decomposition, this choice is always possible. If $M = M_i$ for some $i = \overline{0, n-1}$, then we select arcs $\widehat{DM}$ and $\widehat{MC}$, such that $\widehat{DM} \subset \widehat{M_{i-1} M_i}$ and $\widehat{MC} \subset \widehat{M_i M_{i+1}}$. Next, we construct a sequence of points $(D_k) \subset \widehat{DM}$, such that it converges monotonically to M along $\widehat{DM}$ as $k \rightarrow \infty$. By the convexity of $\widehat{DM}$, the sequence of lines $D_k M$ converges to a line, which we denote as l_1 . Similarly, we define the line l_2 as the limit of lines $C_k M$ when the sequence of points $(C_k) \subset \widehat{MC}$ approaches M . Clearly, the lines l_1 and l_2 do not depend on the choice of sequences (D_k) and (C_k) .

A line l , passing through a point $M \in \Gamma(\Phi)$, is called a generalized tangent (GT) to the figure Φ at point M (or to the curve $\Gamma(\Phi)$), if one of the following three conditions holds (see Fig. 10):

1. $\exists \delta > 0$: The segment of line l , of length δ centered at M , is entirely contained within the figure Φ ;
2. $\exists \delta > 0$: The segment of line l , of length δ centered at M , has no interior points in common with the figure Φ ;
3. The line l coincides with one of the two lines l_1 or l_2 .

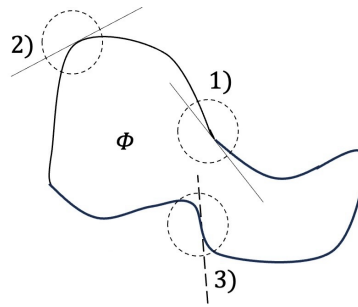


Fig. 10. The generalized tangent (GT) to the figure Φ

Lemma 1 (Set of Generalized Tangents). *The generalized tangent (GT) does not depend on the specific form of the canonical partition of $\Gamma(\Phi)$.*

Proof. The proof follows from the fact that the convexity of an arc does not change when the canonical partition is modified. For a fixed point $M \in \Gamma(\Phi)$, the extreme generalized tangents l_1 and l_2 remain unchanged. Consequently, the set of all generalized tangents at M also remains unchanged. Thus, the lemma is proven. $\square$

Lemma 2 (Connection Between Generalized Tangents and Supporting Lines). *Every generalized tangent to the figure $\Phi \in S$ at a point $M \in \widehat{M_i M_{i+1}} \subset \Gamma(\Phi)$, $M \notin \{M_i, M_{i+1}\}$, is a supporting line of the convex figure F_i , which is bounded by the arc $\widehat{M_i M_{i+1}}$ and the segment $M_i M_{i+1}$.*

Proof. The proof follows from the fact that for the part of the arc $\widehat{M_i M_{i+1}}$, containing the point M , the generalized tangent is defined analogously to a supporting line for the convex figure F_i . The lemma is proven. $\square$

Lemma 3 (Number of Generalized Tangents in a Given Direction). *For each line m , there exists a finite number of generalized tangents to the figure $\Phi \in S$, that are parallel to the given line m .*

Proof. Each generalized tangent that is parallel to the line m either passes through a point of the canonical partition (and thus their number is finite) or is a supporting line to the figure F_i , which is bounded by the arc $\widehat{M_i M_{i+1}}$ and the segment $M_i M_{i+1}$, $i = \overline{(0, n-1)}$. The number of such figures is finite, and each of them has exactly two supporting lines parallel to m . This conclusively proves the lemma on their finiteness. The lemma is proven. $\square$

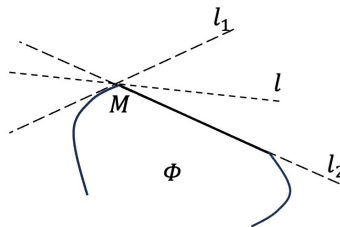


Fig. 11. Illustration of more than one generalized tangent at point M

A point M on the boundary of a figure Φ , where more than one supporting line to Φ exists, is called a **corner point** (or angular point) of the curve Γ (Fig. 11). The remaining points of the curve are referred to as **simple points** of the curve Γ .

A corner point M , for which $l_1 = l_2$, is called a **cusp** (Fig. 12). If M is a cusp, then any line passing through M will also be a GT to Φ at this point; whereas at a simple point of the curve $\Gamma(\Phi)$, the GT is unique. It should also be noted that if M is a simple point and an inflection point of $\Gamma(\Phi)$, it satisfies only condition 3) of the GT definition.

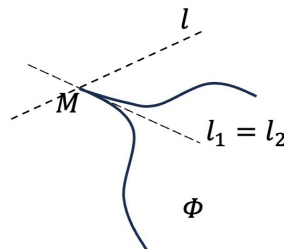


Fig. 12. A return point M

4. Definition of the tangent metric for figures of class S

Let $\Phi_1, \Phi_2 \in S$ be given figures. Choose an arbitrary point $M_1 \in \Gamma(\Phi_1)$, let l_1 be a generalized tangent (GT) to Φ_1 at point M_1 , and $l_2^{(1)}, l_2^{(2)}, \dots, l_2^{(m)}$ be all the GTs to figure Φ_2 , that are parallel to line l_1 . Choose the line $l_1^{(j)}$, that is closest to l_1 , i.e.,

$$\rho(l_2^{(j)}, l_1) = \min_{1 \leq i \leq m} \rho(l_2^{(i)}, l_1).$$

Next, we sequentially define:

$$r_1(M_1, l_1) = \rho(l_2^{(j)}, l_1), \quad r_1(M_1) = \max_{l_1 \notin M_1} r_1(M_1, l_1), \quad r_1 = \max_{M_1 \in \Gamma(\Phi_1)} r_1(M_1).$$

Similarly, by swapping the figures Φ_1 and Φ_2 , we define the quantity r_2 . The tangent metric in the class of figures S is defined as the function:

$$\rho_1(\Phi_1, \Phi_2) = \max\{r_1, r_2\}.$$

In works (Rublev, & Petunin, 1999a, 1999b), a function called the smooth metric was studied:

$$\rho^{(1)}(\Phi_1, \Phi_2) = \rho^{(0)}(\Phi_1, \Phi_2) + \rho_1(\Phi_1, \Phi_2),$$

where $\rho^{(0)}(\Phi_1, \Phi_2)$ is the Hausdorff metric between the figures Φ_1 and Φ_2 . It was proven that the smooth metric is indeed a metric.

5. The tangent metric is a metric.

To prove that this functional is a metric, it requires verification of three axioms. The second axiom is obvious, and the third axiom has been proven in works (Rublev, & Petunin, 1999a, 1999b). However, the first axiom for the smooth metric followed from the fact that it holds for the Hausdorff metric, one of the terms. The main goal of this work is to prove that the first axiom holds for the tangent metric.

Theorem 1 (First Axiom for the Tangent Metric). *If for any figures $\Phi_1, \Phi_2 \in S$ the tangent metric $\rho_1(\Phi_1, \Phi_2) = 0$, then $\Phi_1 = \Phi_2$.*

Proof. Let us start with the case where $\Phi_1, \Phi_2 \in S$ are polynomials. We will prove the statement by contradiction. Suppose that $\Phi_1 \neq \Phi_2$. Without loss of generality, this means there exists a vertex A of the polynomial Φ_1 , that is not a vertex of the polynomial Φ_2 . Since this is a vertex, it is a corner point, and therefore, there is an infinite number of tangent lines to Φ_1 at this point. Let us consider five of them: $m_1, \dots, m_5$. Each of these must be a tangent line to Φ_2 , as follows from the theorem's condition. Thus, on each of these lines, there is at least one point $\Gamma(\Phi_2)$, for which the corresponding line is a tangent line. Let these points be denoted as $B_1, \dots, B_5$. Three of them are located on the same side of point A . Let us denote them as B_1, B_2, B_3 (see Fig. 13). Then, the corresponding convex figure F_i , which has part of the line $\Gamma(\Phi_2)$, as its boundary, cannot contain all points B_1, B_2, B_3 . Thus, between m_1 and m_2 there is at least one convex segment. Since such sets of lines can be constructed infinitely, this would imply that the canonical partition of the boundary of such a figure must consist of an infinite number of points, which contradicts the definition of figures in class S . The resulting contradiction completes the proof.

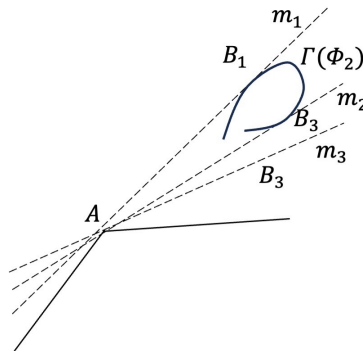


Fig. 13. Tangent lines

For arbitrary figures in class S consider a vertex $A \in \Gamma(\Phi_1)$, which does not belong to $\Gamma(\Phi_2)$. Then, there exists a sufficiently small arc $\widehat{A_1 A_2}$, for which $A \in \widehat{A_1 A_2} \subset \Gamma(\Phi_1)$, and which does not have any points in common with $\Gamma(\Phi_2)$. Next, using similar reasoning, we examine tangent lines to the points of this arc and obtain that Φ_2 does not belong to class S . The theorem is proven. $\square$

Discussion and conclusions

The article studies a metric related to the boundaries of figures of a certain class. To properly define it, the class of figures, referred to as figures with convex-concave boundaries, is first introduced. Since the metric is defined using generalized tangents to the boundary of the figure, and for convex figures, a generalized tangent becomes a supporting line, this approach opens promising prospects for further study of figures of this class. A dense subset of this class is

the collection of all polynomials. For practical applications, this serves as an efficient simplification, as figures are usually defined by a set of points on their boundary, and any approximation begins with approximating the figure by a polynomial.

Further research on the tangent metric should focus on two directions. First, an attempt should be made to generalize it to a broader class of figures and investigate the completeness of the obtained spaces. Second, algorithms for computing the tangent metric should be developed, along with its application for image processing.

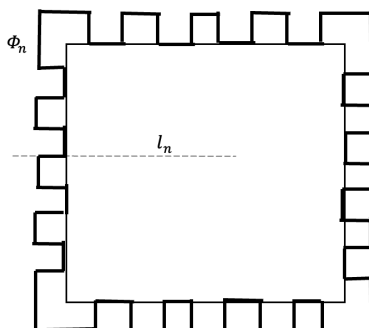


Fig. 14. Example of a sequence of figures (Φ_n) that does not converge to Φ in the tangent metric

In the future, it is planned to compare the tangent metric with other metrics. Regarding the Hausdorff metric, the following example is very illustrative. Consider a figure Φ — a square with a side of 1, and a sequence of figures (Φ_n) , as shown in Fig. 14. If the small squares have a side $\varepsilon = \frac{1}{2n}$, then for the Hausdorff metric we have:

$$\rho^{(0)}(\Phi, \Phi_n) = \varepsilon\sqrt{2} = \frac{\sqrt{2}}{2n} \rightarrow 0, n \rightarrow \infty.$$

For the tangent metric, if we consider the GT l_n drawn through the middle of the side of the square Φ , we have:

$$\rho^{(1)}(\Phi, \Phi_n) = 2 \not\rightarrow 0, n \rightarrow \infty.$$

As can be easily seen, the convergence of perimeters $P(\Phi_n)$ to the perimeter of the figure Φ also does not occur.

Authors' contribution: Maryna Chigidina – study of the properties of the class of figures, Iryna Lazarenko – study of the properties of the class of figures, Dmytro Nomirovskii – proof of the axiom of the metric, Bogdan Rublyov – definition of the tangent metric, Vladimir Semenov – definition of the tangent metric, Kateryna Svitovenko – proof of the axiom of the metric.

Acknowledgments. The authors would like to thank the anonymous referee who provided useful and detailed comments on an earlier version of the manuscript.

Sources of funding. Vladimir Semenov was partially supported by the NASU (0124U002162).

References

- Conci, A., & Kubrusly, C. S. (2017). Distance between sets: A survey. *Advances in Mathematical Sciences and Applications*, 26, 1–18.
- Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 850–863. <https://doi.org/10.1109/34.232073>
- Kumar, K. S., Manigandan, T., Chitra, D., & Murali, L. (2020). Object recognition using Hausdorff distance for multimedia applications. *Multimedia Tools and Applications*, 79, 4099–4114. <https://doi.org/10.1007/s11042-019-07774-z>
- Ortigosa, R., Martínez-Frutos, J., Mora-Corral, C., Pedregal, P., & Períago, F. (2021). Optimal control of soft materials using a Hausdorff distance functional. *SIAM Journal on Control and Optimization*, 59, 393–416. <https://doi.org/10.1137/19M1307299>
- Rublev, B. V., & Molodtsov, O. I. (2000). Tangent metric as a smooth metric for studying plane figures. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics*, 1, 276–281.
- Rublev, B. V., & Petunin, Y. I. (1999a). The smooth metric in the class of plane figures with piecewise-convex and concave boundary. Part I. *Cybernetics and Systems Analysis*, 35, 118–127. <https://doi.org/10.1007/BF02667921>
- Rublev, B. V., & Petunin, Y. I. (1999b). The smooth metric in the class of plane figures with piecewise-convex and concave boundary. Part II. *Cybernetics and Systems Analysis*, 35, 271–278. <https://doi.org/10.1007/BF02733474>

Отримано редакцією журналу / Received: 04.04.25

Прорецензовано / Revised: 26.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Марина ЧІГІДІНА¹, студ.
ORCID ID: 0009-0005-2470-033X
e-mail: mchigidina@knu.ua

Ірина ЛАЗАРЕНКО², канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-3384-1186
e-mail: i_lazar@iitl.kpi.ua

Дмитро НОМІРОВСЬКИЙ¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-9482-776X
e-mail: kashpir74@gmail.com

Богдан РУБЛЬОВ¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-6618-797X
e-mail: rublyovbv@gmail.com

Володимир СЕМЕНОВ¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-3280-8245
e-mail: semenov.volodya@knu.ua

Катерина СВИТОВЕНКО¹, студ.
ORCID ID: 0009-0002-9815-5889
e-mail: svitovenko33@knu.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна

ОСНОВНІ ВЛАСТИВОСТІ МЕТРИКИ ДОТИЧНИХ

Досліджено метрику, що пов'язана з межами фігур певного класу. Для її коректного визначення спочатку визначено клас фігур, який названий фігурами з опукло-угнутими межами. Задля цього на межі фігури має існувати скінченний набір точок, для яких дуга між кожними двома сусідніми точками є опуклою. Це дає змогу розглядати межу фігури як об'єднання скінченної кількості опуклих дуг, оскільки метрику визначено за допомогою узагальнених дотичних для межі фігури, а для опуклих фігур узагальнена дотична стає опорною прямою. Зазначений підхід відкриває гарні перспективи для подальшого вивчення фігур цього класу. Щільною підмножиною в цьому класі є сукупність усіх многокутників. Для практичних застосувань це є достатньо ефективним спрощенням, бо зазвичай фігури задаються набором точок на їхній межі і будь-яка апроксимація починається з наближення фігури многокутником. Для наведеного класу фігур визначено метрику дотичних, яку задано через узагальнені дотичні до межі фігури. Раніше цю метрику вивчали разом з першим доданком, який утворював метрику Хаусдорфа. Це давало змогу не перевіряти першу аксіому метрики, що впливала з метрики Хаусдорфа. Проте метрика Хаусдорфа виявилася не обов'язковим доданком при визначенні метрики дотичних, що прямо впливає з проведених досліджень. Загалом метрика Хаусдорфа виникла з певної подібності до метрики простору неперервно диференційованих функцій, яка має першим доданком метрику неперервних функцій. Тому дуже важливим став цей крок, який дозволив позбутися метрики Хаусдорфа як зайвого доданку. Це так само надалі має спростити алгоритми з обчислення метрики дотичних. Для подальших досліджень першим кроком стане звуження класу функцій до сукупності усіх багатокутників, тож важливим є вивчення властивостей канонічного розбиття фігури саме в класі усіх поліномів. Також для поліномів стає простішим питання побудови узагальнених дотичних.

Ключові слова: метрика дотичних, узагальнена дотична, гладка метрика, канонічна декомпозиція, фігура.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.35>

Nataliya SHCHESTYUK, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-7652-8157
e-mail: n.shchestyuk@ukma.edu.ua
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Dmytro SLUCHYNSKYI, PhD Student
ORCID ID: 0009-0003-4178-1608
e-mail: d.sluchynskyi@ukma.edu.ua
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Fedir SERDUK, Master of Arts
ORCID ID: 0009-0001-5785-3480
e-mail: f.serdiuk@ukma.edu.ua
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

SUBDIFFUSIVE MODELS WITH LOCAL VOLATILITY FOR ILLIQUID MARKETS

The paper focuses on subdiffusive models with local volatility that extend classical approaches to capture illiquidity effects and trapping events through stochastic time changes.

The illiquidity effects mean lower trading volume, wider bid-ask spreads, greater price volatility, and long periods without any trading for risky assets. These effects occur during crisis periods that negatively affect financial activity, or in emerging markets where the number of participants and, thus, the number of transactions is relatively low. All models based on Brownian motion are perpetually moving. They are not adapted for modeling periods with motionless stock returns. Nevertheless, in statistical physics, there are subdiffusion processes that are well suited to describe the dynamics of underlying returns with trapping events. However, for modeling the dynamics of the financial market with illiquid effects, we need not only to describe stock prices, but also to propose a method for option evaluation. The main option pricing tool in this paper is a fractional extension of the Dupire equation, which gives the fair price for call and put options for subdiffusive underlying assets and takes into account local volatility. It is a forward partial differential equation in which the derivative with respect to time is replaced by a Dzerbayshan – Caputo ($D - C$) derivative. The form of the $D - C$ derivative and fractal Dupire equation depends upon the chosen Lévy subordinator. We apply the equation for subdiffusive in a Brownian setting in the case of the inverted tempered stable subordinator and demonstrate numerical results for real financial data.

Due to the proposed approaches and numerical algorithm, the investor gets tools that allow him to take into account both the market's illiquidity and local volatility.

Key words: option pricing, subdiffusion, local volatility, tempered stable subordinator.

AMS 2020 classification: 91G20, 91G60, 60G40.

Introduction

An illiquid market refers to the state of a stock, bond, or other asset that cannot easily and readily be sold or exchanged for cash without a substantial loss in value. It happens for crisis periods that negatively affect financial activity or for emerging markets in which the number of participants, and thus the number of transactions, is rather low. As a result, illiquid assets tend to have lower trading volume, wider bid-ask spreads, greater price volatility, and long periods without any trading. To account for the market's ability to accelerate or slow down during periods of varying trading intensity, fractal models of activity over time have been proposed, see for example (Heyde, 1999; Casteli, Leonenko, & Shchestyuk, 2017; Leonenko, Liu, & Shchestyuk, 2025), etc. In the context of this model, the "activity time" process can be interpreted as time over which market prices evolve, and it is often associated with trading volume or the flow of new price-sensitive information. However, these models and any other models based on Brownian motion are perpetually moving; they are not adapted for modeling periods with motionless stock returns.

Nevertheless, in statistical physics, there are subdiffusion processes that are well suited to describe the dynamics of underlying returns with trapping events (Metzler, & Klafter, 2000). Magdyarz was the first to propose a subdiffusion extension of the Bachelier and Black-Scholes models (see (Magdziarz, Orzeł, & Weron, 2011; Magdziarz, 2009)). He considered the standard diffusive process $S_t = S_{H_t}$, which is time-changed by some stochastic process H_t called the inverse subordinator ("hitting time"). In general, for a given subordinator G_t , the inverse subordinator H_t is defined as

$$H_t = \inf \{ \tau > 0 : G_\tau \geq t \}, \quad (1)$$

and interpreted as the first time at which G_t hits the barrier t . The subordinator G_t in (1) is generally Lévy process with Laplace transform in the following form:

$$E \left(e^{-\theta G_t} \right) = e^{-t \Psi(\theta)},$$

where $\Psi(\theta)$ is called Lévy exponent, which can be written in the following driftless form

$$\Psi(\theta) = \int_0^{+\infty} \left(1 - e^{-\theta x} \right) \tilde{\nu}(dx),$$

$\tilde{\nu}(dx)$ is an appropriate Lévy measure. By construction, the inverted process may be constant, and any process subordinated by H_t exhibits motionless periods. Magdziarz constructed the corresponding martingale measure, showed that

the model is incomplete, and derived the formulas for European put and call option prices using the martingale method (Magdziarz, Orzeł, & Weron, 2011; Magdziarz, 2009). A general result that implies that subdiffusive markets are generally incomplete is given in (Torricelli, 2020), Theorem 2. In (Scalas, & Toaldo, 2021), Theorem 5, a pseudo-differential equation that generalizes the Black and Scholes equation for the underlying time-changed geometric Brownian motion was derived. However, Magdziarz's approach has two drawbacks. Firstly, following the martingale method, the option price can be evaluated by computing an integral that includes the hitting time density. Thus, the density should admit a closed form. An alternative is to calculate the price using a Monte Carlo simulation. These solutions are explored in (Magdziarz, 2009) in a Brownian setting for α -stable subordinator, in (Shchestyuk, & Tyshchenko, 2025) for inverse Gaussian subordinator, and in (Shchestyuk, Drin, & Tyshchenko, 2024) for Gamma subordinator. Unfortunately, in the general case, the closed form of the hitting time density is unknown, and we need to look for other ways. Secondly, Magdziarz considered subdiffusive models in a Brownian setting with constant volatility. In fact, the realized volatility of an underlying asset actually varies with time and with the underlying itself, and it is worth taking into account.

The object of the research is a subdiffusive model with local volatility that extend classical approaches to capture illiquidity effects and trapping events through stochastic time changes. To price options, we use the Dupire fractal equation, which does not require a closed form for hitting time and takes into account local volatility. It is a forward partial differential equation in which the derivative with respect to time is replaced by a Dzerbayshan – Caputo (D – C) derivative. In (Hainaut, & Leonenko, 2021) was proved that option prices for subdiffusion can be found as solutions of the fractal Dupire equation. The form of the D – C derivative and fractal Dupire equation depends upon the chosen Lévy subordinator. In (Hainaut, & Leonenko, 2021), it was detailed for α -stable and Poisson subordinators in the jump-diffusion setting.

The aim of the paper is to propose the subdiffusion model in case of a Brownian setting for a *Tempered Stable Subordinator* $TS(\alpha, \lambda, c)$, develop a numerical method for the corresponding fractional Dupire equation, and apply the model to real market data for demonstration of its practical relevance.

The paper is organized as follows. In the second section, we consider the subdiffusive model for the dynamics of the illiquid market and formulate the proposition about the form of the fractal Dupire equation for option pricing in the case of the tempered stable subordinator. After that, we shortly discuss the issues of the existence of arbitrage opportunities and the completeness of the TSS subdiffusion market, the interpretation of the TSS parameters in light of the market price of risk. Beside this we mention that inverse Gaussian $IG(\delta, \gamma)$ subordinator can be considered as tempered stable Lévy subordinator $TS(1/2, \gamma^2, \delta)$ and demonstrate the form of the corresponding fractal Dupire equation in this case. The third section is devoted to the numerical approach and pricing of options for real financial data.

1. Subdiffusive option pricing via fractal Dupire equation

We assume that the subdiffusive market consists of at least three components: a risk-free bond, a risky asset, and an option. The idea of subdiffusion models is to replace the calendar time t in the risk-free bond motion and the classical geometrical Brownian motion (GBM) by some stochastic process H_t that represents the hitting time, defined as (1). Thus, the bond B_t and stock S_t are the solutions of the time-changed stochastic differential equations:

$$\begin{aligned} \frac{dB_{H_t}}{B_{H_t}} &= r dH_t, B_0 = 1, \\ \frac{dS_{H_t}}{S_{H_t}} &= \left(\mu + \frac{\sigma_{H_t}^2}{2} \right) dH_t + \sigma_{H_t} dB_{H_t}, \quad t > 0, \end{aligned}$$

where $\sigma_t = \sigma_{H_t}$ is local volatility.

To model option pricing, we use the fractional approach that was introduced in (Hainaut, & Leonenko, 2021) in a jump-diffusion setting, where the derivative for the time in the Dupire equation is replaced by a Dzerbayshan – Caputo derivative for jump-diffusion. The applications of anomalous diffusions for option pricing and the volatility term structure were considered in (Jacquier, & Torricelli, 2020; Torricelli, 2020). Option pricing in time-changed Lévy models with compound Poisson jumps was explored in (Ivanov, & Ano, 2018). The theory in these papers was detailed for inverse α -stable (Magdziarz, 2009; Magdziarz, Orzeł, & Weron, 2011; Hainaut, & Leonenko, 2021), Mittag – Leffler (Scalas, & Toaldo, 2021) and inverted Poisson processes (Hainaut, & Leonenko, 2021). We denote the option price as $C_H(T, k)$, where T is the time to maturity and the logarithmic strike price as $k = \log K$. We then replace the standard time derivative in the classical Dupire equation with a convolution-type derivative. Thus, the Fractal Dupire equation can be written in the form

$${}^{\Psi}D C_H(T, k) = -r \frac{\partial}{\partial k} C_H(T, k) + \frac{1}{2} \sigma^2(T, S_{H_t}) \frac{\partial^2}{\partial k^2} C_H(T, k), \quad (2)$$

where ${}^{\Psi}D u(t)$ is the convolution-type derivative for an absolutely continuous on $(0, \infty)$ function $u(t)$, called the Dzerbayshan – Caputo (D – C) derivative and defined as

$${}^{\Psi}D u(t) = \int_0^t \frac{\partial}{\partial t} u(t-s) \nu(s) ds. \quad (3)$$

In the (3) the tail of the Lévy measure, or kernel $\nu(s)$, is obtained as

$$\nu(s) ds = \int_s^\infty \tilde{\nu}(dx) ds,$$

where the Lévy measure $\tilde{\nu}(dx)$ for the subordinator G_t with Lévy exponent Ψ can be expressed through the inverse Laplace transform

$$\tilde{\nu}(dx) = \mathcal{L}^{-1} \left[\frac{d}{dx} \Psi(x) \right] (x).$$

We apply the equation (2) to diffusion in a Brownian setting with the inverse tempered stable subordinator $TS(\alpha, \lambda, c)$. The $TS(\alpha, \lambda, c)$ subordinator is a type of Lévy subordinator with Lévy exponent

$$\Psi(x) = c((\lambda + x)^\alpha - \lambda^\alpha), \quad (4)$$

where $0 < \alpha < 1$ is stability parameter, $\lambda > 0$ is tempering parameter and $c > 0$. The corresponding hitting time process H_t is called an inverse tempered stable subordinator (ITS). Its analytical properties are fully detailed in (Kumar, & Vellaisamy, 2015) and (Alrawashdeh et al., 2017).

Proposition 1. *The European call option price $C_H(T, k) = C_H(S, K, T, \sigma_H)$ for a time-changed version of the Brownian setting with local volatility, in case of an inverse tempered stable subordinator H_t is a solution of the fractional Dupire equation:*

$$c \frac{\alpha \lambda^\alpha}{\Gamma(1-\alpha)} \int_0^T \frac{\partial}{\partial T} C_H(T-s, k) \Gamma(\alpha, \lambda s) ds = -r \frac{\partial}{\partial k} C_H(T, k) + \frac{1}{2} \sigma^2(T, S_{H_t}) \frac{\partial^2}{\partial k^2} C_H(T, k), \quad (5)$$

with the initial condition $C_H(0, K) = (S_{H_0} - K)^+$. The European put option price is a solution of the same fractional Dupire equation but with initial conditions $P_H(0, K) = (K - P_{H_0})^+$.

Proof. Using (4) we compute

$$\frac{d\Psi(x)}{dx} = c \cdot \alpha (\lambda + x)^{\alpha-1},$$

and obtain Lévy measure:

$$\tilde{\nu}(dx) = c \frac{\alpha}{\Gamma(1-\alpha)} x^{-\alpha-1} e^{-\lambda x} I_{x>0} dx.$$

To compute the kernel $\nu(s)$, we evaluate the tail integral:

$$\nu(s) ds = \int_s^\infty \tilde{\nu}(dx) ds = c \frac{\alpha}{\Gamma(1-\alpha)} \int_s^\infty x^{-\alpha-1} e^{-\lambda x} dx ds.$$

Using the change of variables $x = \frac{t}{\lambda} \Rightarrow dx = \frac{dt}{\lambda}$ we have

$$\nu(s) ds = c \frac{\alpha \lambda^\alpha}{\Gamma(1-\alpha)} \int_{\lambda s}^\infty t^{-\alpha-1} e^{-t} dt ds = c \frac{\alpha \lambda^\alpha}{\Gamma(1-\alpha)} \Gamma(-\alpha, \lambda s) ds,$$

where $\Gamma(-\alpha, \lambda s)$ denotes the upper incomplete gamma function. Then D-C derivative (3) for $TS(\alpha, \lambda, c)$ subordinator can be written as

$${}^\Psi D C_H(T, k) = c \frac{\alpha \lambda^\alpha}{\Gamma(1-\alpha)} \int_0^T \frac{\partial}{\partial T} C_H(T-s, k) \Gamma(\alpha, \lambda s) ds. \quad (6)$$

Thus, if we substitute (6) in (2), the proposition follows. $\square$

We can notice that the derivative we obtain is a tempered D-C derivative (see (Sabzikar, Meerschaert, & Chen, 2015; Scalas, & Toaldo, 2021)). In general, these kinds of convolution operators have been introduced in (Kochubei, 2011) and then further explored in (Toaldo, 2015) and (Chen, 2017).

Remark 1. The issues of the existence of arbitrage opportunities and the completeness of the TSS subdiffusion market were investigated in (Torricelli, 2020; Toaldo, 2015), where was proved that subdiffusive TSS market is arbitrage-free and incomplete.

Remark 2. The presented option pricing subdiffusive model with $TS(\alpha, \lambda, c)$ subordinator has an interesting interpretation of its parameters in light of the market price of risk. Thanks to the theorem about the incompleteness of the subdiffusive market, the existence of a market price of duration risk naturally arises (see (Torricelli, 2020)). Once S_t is calibrated to liquid market prices, the market price of duration risk will be reflected in the martingale density parameters, and these same features will thus be reflected in the (risk-adjusted) frequency and duration of trade pauses incorporated in hitting time H_t . In (Torricelli, 2020) was shown the parameter λ expresses belief in the speed at which the granular price evolution of the asset will revert to a fully "liquid" state, which is well approximated by a standard Lévy-driven diffusion. So, λ captures the risk of latency to liquidity, while absolute trade duration risk is defined by α . Therefore, $TS(\alpha, \lambda, c)$ models offer a unified framework in which ultra short-term option pricing accounts for the microstructural duration effect, whereas options at typical maturities are priced according to the usual Lévy paradigm: the cut-off between the two states is encoded in λ .

Remark 3. Since the inverse Gaussian $IG(\delta, \gamma)$ subordinator can be considered as tempered stable Lévy subordinator $TS(1/2, \gamma^2, \delta)$, we can refer to the paper (Shchestsyuk, & Tyshchenko, 2025), where the fair option price for IG market was

explored as a solution of the PIDE:

$$2\delta \int_0^T \frac{\partial}{\partial t} C_H(T-s, k) \left(\gamma(\Phi(\gamma\sqrt{s}) - 1) + \frac{e^{-\frac{s\gamma^2}{2}}}{\sqrt{2\pi s}} \right) ds = -r \frac{\partial}{\partial k} C_H(T, k) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial k^2} C_H(T, k),$$

with the initial condition $C_H(0, K) = (S_{H_0} - K)^+$, and $k = \ln K$. Here, parameter γ captures the risk of latency to liquidity, while absolute trade duration risk remains constant and is defined by $1/2$ (see (Shchestyuk, & Tyshchenko, 2025; Torricelli, 2020)).

Remark 4. If $\gamma = 0$ for $TS(1/2, \gamma^2, \delta)$, we get a regime where prices exhibit maximum staleness at all temporal scales. In fact,

$$\Psi_D C_H(T, k) = 2\delta \int_0^T \frac{\partial}{\partial T} C_H(T-s, k) \cdot \frac{1}{\sqrt{2\pi s}} ds.$$

Using the substitution $z = T - s \Rightarrow ds = -dz$, we get

$$\Psi_D C_H(T, k) = \frac{2\delta}{\sqrt{2\pi}} \Gamma\left(\frac{1}{2}\right) \frac{1}{\Gamma\left(1 - \frac{1}{2}\right)} \int_0^T \frac{\partial}{\partial z} C_H(z, k) (T-z)^{1/2} dz. \quad (7)$$

We remind the definition of the Caputo derivative

$$D_\alpha u(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial}{\partial z} u(z) (t-z)^\alpha dz, \quad (8)$$

and use $\Gamma(1/2) = \sqrt{\pi}$, then we can rewrite (7) in the form

$$\Psi_D C_H(T, k) = \frac{2\delta}{\sqrt{2}} D_{1/2} C_H(T, k).$$

This leads to the fractional Dupire-type equation:

$$\frac{\delta}{\sqrt{2}} D_{1/2} C_H(T) = -\frac{r}{2} \frac{\partial}{\partial k} C_H(T, k) + \frac{\sigma^2(T, S_{H_t})}{4} \frac{\partial^2}{\partial k^2} C_H(T, k). \quad (9)$$

2. Numerical approach and illustration

In this section, a numerical approach involving the discretization of the integro-differential equation (9) is applied. The Caputo derivative on the left-hand side of (9) admits a finite difference representation (see, for example, (Deng, 2008)). In what follows, we use the discretization formula given in Eq. (37) of (Hainaut, & Leonenko, 2021):

$$\begin{aligned} \mathcal{D}_\alpha C_H(k, j) &\approx \frac{(\Delta t)^{-\alpha}}{\Gamma(1-\alpha)} (C_H(k, j) - C_H(k-1, j)) + \\ &+ \frac{(\Delta t)^{-\alpha}}{\Gamma(1-\alpha)} \sum_{m=0}^{k-2} (k-m)^{-\alpha} (C_H(m+1, j) - C_H(m, j)). \end{aligned}$$

The matrix form of Caputo derivative for the case $\alpha = \frac{1}{2}$ can be written as

$$\begin{aligned} \frac{1}{\sqrt{2}} \mathcal{D}_{1/2} C_H(k, \cdot)^T &\approx \frac{1}{\sqrt{2}} \frac{(\Delta t)^{-\frac{1}{2}}}{\sqrt{\pi}} (C_H(k, \cdot) - C_H(k-1, \cdot))^T + \\ &+ \frac{1}{\sqrt{2}} (D(k) C_H(0 : k-1, \cdot))^T \mathbf{1}_{k-1}, \end{aligned}$$

where $\mathbf{1}_{k-1}$ is a $(k-1)$ -vector of ones, $D(k)$ is a special matrix of dimension $(k-1) \times k$ with appropriate coefficients:

$$D(k) = \frac{(\Delta t)^{-\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} \begin{pmatrix} 0 & \dots & \dots & 0 & -(2)^{-\frac{1}{2}} & (2)^{-\frac{1}{2}} \\ \vdots & \ddots & 0 & -(3)^{-\frac{1}{2}} & (3)^{-\frac{1}{2}} & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ -(k)^{-\frac{1}{2}} & (k)^{-\frac{1}{2}} & 0 & \dots & \dots & 0 \end{pmatrix}.$$

The first and second order derivatives with respect to k on the right side of (9) can be calculated by a central finite difference approximation:

$$\begin{aligned} \frac{\partial C_H(k, j)}{\partial k} &\approx \frac{C_H(k, j+1) - C_H(k, j-1)}{2\Delta_K}, \quad \frac{\partial C_H(k, 0)}{\partial k} \approx \frac{C_H(k, 1) - C_H(k, 0)}{\Delta_K}, \\ \frac{\partial C_H(k, n_K+1)}{\partial k} &\approx \frac{C_H(k, n_K+1) - C_H(k, n_K)}{\Delta_K}, \end{aligned}$$

$$\frac{\partial^2 C_H(k, j)}{\partial k^2} \approx \frac{C_H(k, j+1) - 2C_H(k, j) + C_H(k, j-1)}{\Delta_K^2},$$

$$\frac{\partial^2 C_H(k, 0)}{\partial k^2} \approx \frac{\partial^2 C_H(k, 1)}{\partial k^2}, \frac{\partial^2 C_H(k, n_K+1)}{\partial k^2} \approx \frac{\partial^2 C_H(k, n_K)}{\partial k^2},$$

where $n_K = \left\lceil \frac{K_{max} - K_0}{\Delta_K} \right\rceil$. These derivatives are represented in matrix form:

$$\frac{\partial C_H(k, j)}{\partial k} \approx \bar{K} R_1 C_H(k, \cdot)^T, \quad \frac{\partial^2 C_H(k, j)}{\partial k^2} \approx \bar{K}^2 R_2 C_H(k, \cdot)^T,$$

where

$$R_1 = \frac{1}{\Delta_K} \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}, \quad R_2 = \frac{1}{\Delta_K^2} \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix},$$

$\bar{K} = \text{diag}(K_j)_{j \in 0, 1, \dots, n_K}$, the diagonal matrix of strikes and $C_H(k, \cdot)$ is the k -th line of the matrix of call prices, denoted by C_H (see (Hainaut, & Leonenko, 2021)).

Therefore, the final numerical iterative scheme for solving (9) in the case $\delta = 1$ can be written in the form:

$$C_H(k, \cdot)^T \approx \left(\bar{K} R_1 \frac{r}{2} - \bar{K}^2 R_2 \frac{\sigma_H^2}{4} + I_{n_K+1} \frac{1}{\sqrt{2}} \frac{(\Delta t)^{-\frac{1}{2}}}{\sqrt{\pi}} \right)^{-1} \times$$

$$\times \left(\frac{1}{\sqrt{2}} \frac{(\Delta t)^{-\frac{1}{2}}}{\sqrt{\pi}} (D(k) C_H(k-1, \cdot))^T \mathbf{1}_{k-1} \right),$$

where I_{n_K+1} is the identity matrix.

Numerical experiments were performed to compute European call option prices on AMD stock for the date 13/02/2025 with the following parameters: strike price $K = 111$, spot price $S = 111$, and interest rate $r = 5\%$. The annual constant volatility from historical data equals $\sigma = 47.37\%$. Local volatility σ_{local} was constructed iteratively using the recurrence relation:

$$\sigma_{\text{local}}(i) = \sigma_{\text{local}}(i-1) \cdot (\beta_0 + \beta_1 i^\gamma + \beta_2 i^{2\gamma}), \quad \text{for } i = 1, \dots, n_t,$$

with initial value $\sigma_{\text{local}}(0) = \sigma$ and parameters $\beta_0 = 0.99$, $\beta_1 = -1.10 \times 10^{-11}$, $\gamma = 1.63$, which was found by optimization procedure.

Three models were evaluated:

- Classical Black-Scholes model;
- Subdiffusive model with constant volatility;
- Subdiffusive model with local volatility.

Figure 1 presents the comparison of call option model prices and actual market prices (premiums) for varying times of maturity.

As we can see from the graphics in Fig. 1, the Black-Scholes option pricing model shows better results in the short-term period, while the subdiffusive model is more effective in the long-term perspective. To assess and compare the predictive performance of the models, for two time periods, the following accuracy metrics were calculated: **MAE** – Mean Absolute Error, **MSE** – Mean Squared Error, **RMSE** – Root Mean Squared Error, **ARPE** – Average Relative Pricing Error.

The results are presented in the Tab. 1.

Table 1

Accuracy metrics by time interval and model					
Interval	Model	MAE	MSE	RMSE	ARPE
2025-03-21 to 2025-09-19	Black-Scholes	0.753	0.652	0.807	0.061
	Subdiffusive (const)	3.888	15.533	3.941	0.369
	Subdiffusive (local)	3.087	10.218	3.197	0.302
2025-10-17 to 2026-01-16	Black-Scholes	1.066	1.140	1.068	0.054
	Subdiffusive (const)	2.128	4.656	2.158	0.109
	Subdiffusive (local)	0.888	0.974	0.987	0.046

Discussion and conclusions

The paper develops a subdiffusive model with local volatility time-changed by inverse TSS subordinator. The main pricing tool for European call and put options in this paper is a fractional extension of Dupire's equation in a Brownian setting with local volatility. We apply the equation for TSS -subordinator and demonstrate numerical results for real historical financial data.

The results of comparing the studied model with the classical one and real option prices show that the classical Black-Scholes model demonstrates better results in the short-term period, while the subdiffusive models are more effective in the long-term perspective. These results, based on the fractal Dupire equation technique, are similar to those obtained with the

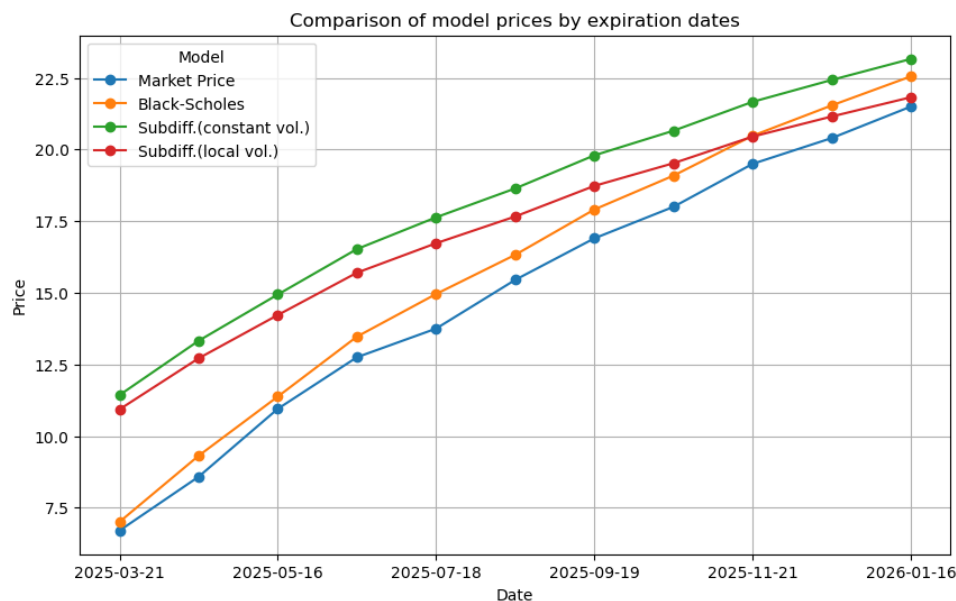


Fig. 1. Comparison of model call option prices with actual market premiums

$IG(\delta, \gamma)$ subordinator using the Monte Carlo method in the already published article (Shchestyuk, & Tyshchenko, 2025). Moreover, the accuracy of the subdiffusive model with local volatility prevails over the subdiffusive model with constant volatility.

It may be noted that in the articles of other authors dealing with option pricing based on subdiffusion (Magdziarz, Orzel, & Weron, 2011; Magdziarz, 2009; Hainaut, & Leonenko, 2021), the fair prices of options were not compared with real market prices (premiums).

Due to the proposed approaches and numerical algorithm, the investor gets tools that allow him to take into account both the market's illiquidity and local volatility.

Authors' contribution: Nataliya Shchestyuk – conceptualization, methodology, analysis of sources, preparation of literature review or theoretical foundations of research; Dmytro Sluchynskiy – methodology, software, empirical data collection and validation, empirical research, analysis of sources, preparation of literature review or theoretical foundations of research; Fedir Serduk – software, empirical research.

Acknowledgments. Nataliya Shchestyuk acknowledges N. Leonenko for some ideas developed in this article.

Sources of funding. Dmytro Sluchynskiy acknowledges financial support from the project "Platform for model based solutions to systemic complexity challenges: bridging leading research schools of Finland and Ukraine by Double Doctor Degree" funded by the TFK program of the Finnish National Agency for Education EDUF.

References

- Alrawashdeh, M., Kelly, J., Meerschaert, M., & Scheffler, H. (2017). Applications of inverse tempered stable subordinators. *Computers and Mathematics with Applications*, 73, 892–905. <https://doi.org/10.1016/j.camwa.2016.07.026>
- Casteli, F., Leonenko, N., & Shchestyuk, N. (2017). Student-like models for risky asset with dependence. *Stochastic Analysis and Applications*, 35, 452–464. <https://doi.org/10.1080/07362994.2016.1266945>
- Chen, Z.-Q. (2017). Time fractional equations and probabilistic representation. *Chaos, Solitons and Fractals*, 102, 168–174.
- Deng, W. (2008). Finite element method for the space and time fractional Fokker-Planck equation. *SIAM Journal on Numerical Analysis*, 47, 204–226. <https://doi.org/10.1137/080714130>
- Hainaut, D., & Leonenko, N. (2021). Option pricing in illiquid markets: A fractional jump–diffusion approach. *Journal of Computational and Applied Mathematics*, 381, . <https://doi.org/10.1016/j.cam.2020.112995>
- Heyde, C. (1999). A risky asset model with strong dependence through fractal activity time. *Journal of Applied Probability*, 36, 1234–1239.
- Ivanov, R., & An, K. (2018). Option pricing in time-changed Lévy models with compound poisson jumps. *Modern Stochastics: Theory and Applications*, 6(1), 81–107.
- Jacquier, A., & Torricelli, L. (2020). Anomalous diffusions in option prices: connecting trade duration and the volatility term structure. *SIAM Journal on Financial Mathematics*, 11(4), 1137–1167.
- Kochubei, A. N. (2011). General fractional calculus, evolution equations, and renewal processes. *Integral Equations and Operator Theory*, 71(4), 583–600.
- Kumar, A., & Vellaisamy, P. (2015). Inverse tempered stable subordinators. *Statistics and Probability Letters*, 103, 134–141. <https://doi.org/10.1016/j.spl.2015.04.010>
- Leonenko, N., Liu, A., & Shchestyuk, N. (2025). Student models for a risky asset with dependence: Option pricing and greeks. *Austrian Journal of Statistics*, 54(1), 138–165. <https://doi.org/10.17713/ajs.v54i1.1952>
- Magdziarz, M. (2009). Black–Scholes formula in subdiffusive regime. *Journal of Statistical Physics*, 136, 553–564. <https://doi.org/10.1007/s10955-009-9791-4>

- Magdziarz, M., Orzel, S., & Weron, A. (2011). Option pricing in subdiffusive Bachelier model. *Journal of Statistical Physics*, 145, 187–202. <https://doi.org/10.1007/s10955-011-0310-z>
- Metzler, R., & Klafter, J. (2000). The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1), 1–77. [https://doi.org/10.1016/S0370-1573\(00\)00070-3](https://doi.org/10.1016/S0370-1573(00)00070-3)
- Sabzikar, F., Meerschaert, M., & Chen, J. (2015). Tempered fractional calculus. *Journal of Computational Physics*, 293, 14–28.
- Scalas, E., & Toaldo, B. (2021). Limit theorems for prices of options written on semi-Markov processes. *Theory of Probability and Mathematical Statistics*, 105, 3–33.
- Shchestsyuk, N., Drin, S., & Tyshchenko, S. (2024). Risk evaluating for subdiffusive option price model with gamma subordinator. In *Mathematical and statistical methods for actuarial sciences and finance: Maf 2024, Le Havre, France* (pp. 286–291). Springer, Cham. https://doi.org/10.1007/978-3-031-64273-9_47
- Shchestsyuk, N., & Tyshchenko, S. (2025). Subdiffusive option price model with inverse Gaussian subordinator. *Modern Stochastics: Theory and Applications*, 12(2), 136–152. <https://doi.org/10.15559/24-VMSTA265>
- Toaldo, B. (2015). Convolution-type derivatives, hitting-times of subordinators and time-changed C_0 -semigroups. *Potential Analysis*, 42, 115–140. <https://doi.org/10.1007/s11118-014-9426-5>
- Torricelli, L. (2020). Trade duration risk in subdiffusive financial models. *Physica A: Statistical Mechanics and its Applications*, 541, 123694. <https://doi.org/10.1016/j.physa.2019.123694>

Отримано редакцією журналу / Received: 04.08.25

Прорецензовано / Revised: 18.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Наталія ЩЕСТИЮК, канд. фіз.-мат. наук, доц.

ORCID ID: 0000-0002-7652-8157

e-mail: n.shchestsyuk@ukma.edu.ua

Національний університет «Києво-Могилянська академія», Київ, Україна

Дмитро СЛУЧИНСЬКИЙ, асп.

ORCID ID: 0009-0003-4178-1608

e-mail: d.sluchynskiy@ukma.edu.ua

Національний університет «Києво-Могилянська академія», Київ, Україна

Федір СЕРДЮК, магістр

ORCID ID: 0009-0001-5785-3480

e-mail: f.serdiuk@ukma.edu.ua

Національний університет «Києво-Могилянська академія», Київ, Україна

СУБДИФУЗІЙНІ МОДЕЛІ З ЛОКАЛЬНОЮ ВОЛАТИЛЬНІСТЮ ДЛЯ НЕЛІКВІДНИХ РИНКІВ

Розглянуто субдифузійні моделі з локальною волатильністю, що розширюють класичні підходи з метою врахування ефектів неліквідності ринку та затримок (пасток) цін шляхом стохастичних змін часу.

Ефекти неліквідності означають менший обсяг торгів, ширші спреди між цінами попиту та продажу, більшу волатильність цін і тривалі періоди без будь-якої торгівлі ризикованими активами. Ці ефекти виникають у кризові періоди, які негативно впливають на фінансову активність, або на ринках, що розвиваються, на яких кількість учасників, а отже, і кількість транзакцій, досить низька. Оскільки усі моделі, засновані на броунівському процесі, описують постійний неперервний хаотичний рух, то вони не пристосовані для моделювання цін акцій зі сталими періодами, коли ціни можуть не змінюватись протягом певного часу. Одним із варіантів виходу з такої ситуації є те, що у статистичній фізиці існують процеси субдифузії, які добре підходять для опису динаміки базових активів з ефектами потрапляння частинки в пастку, де вона може перебувати певний час без руху. Однак для моделювання динаміки фінансового ринку з ефектами неліквідності нам потрібно не лише описати ціни акцій, а й запропонувати метод оцінки опціонів. Основним інструментом ціноутворення опціонів у цій статті є дробове (фрактальне) розширення рівняння Дюпіра, що дає справедливу ціну для опціонів типу кол та пут у випадку субдифузійних базових активів і враховує локальну волатильність. Це диференціальне рівняння з частинними похідними, в якому похідна відносно часу замінюється фрактальною похідною Дзербайшана – Капуто ($D - C$). Вид похідної $D - C$ та фрактального рівняння Дюпіра залежить від обраного субординатора Леві. У цій статті ми застосовуємо зазначене рівняння для геометричної броунівської субдифузії у випадку оберненого TS (tempered stable) субординатора та демонструємо числові результати для реальних фінансових даних.

Завдяки запропонованим підходам і числовому алгоритму інвестор отримує інструменти, що дають йому можливість враховувати як неліквідність ринку, так і локальну волатильність.

Ключові слова: оцінка опціонів, субдифузія, локальна волатильність, темперований стабільний субординатор.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

СУЧАСНА ФІЗИКА

UDC 621 #(621.37/.39; 621.382.2/.3; 621.387)

DOI: <https://doi.org/10.17721/1812-5409.2026/1.36>

Ihor BEKH, PhD (Phys. & Math.), Assoc. Prof.

ORCID ID: 0000-0002-7359-8192

e-mail: igor_ivanovich_bekh@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Roman BOHDANOV, PhD (Phys. & Math.), Assist.

ORCID ID: 0009-0007-8449-1466

e-mail: roman.bohdanov@knu.ua

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

INFLUENCE OF THE POSITIVE FEEDBACK RESISTOR RESISTANCE ON THE AMPLITUDE-FREQUENCY RESPONSE OF THE GYRATOR BASED ON THE OPERATIONAL AMPLIFIER

You can get rid of the use of inductors in active RLC filters of the second and higher orders by using a gyrator – an RC cell included in the positive feedback circuit of an operational amplifier. This "active inductor" circuit allows you to implement active filters even using microelectronics devices. The purpose of our work is to develop practical recommendations for choosing the value of the feedback resistor resistance in gyrator circuits and to investigate its influence on the amplitude-frequency characteristic of the device.

As a research method, computer modeling was used in the LTSpice software application, during which the main characteristics of the oscillatory LC circuit and the circuit with a gyrator on an LT1006 operational amplifier were compared, and the corresponding amplitude- and phase-frequency characteristics were obtained.

It is shown that the position of a maximum of amplitude-frequency response of the circuit with a gyrator is closer to the resonant frequency of the RLC oscillating circuit, when the resistance value of the coupling resistor is closest to the output resistance of the operational amplifier, but at the same time it must be at least 100 Ohm. To ensure a stable transmission coefficient of the circuit with a gyrator when the resonant frequency changes tenfold, the resistance must be about half the operational amplifier's output resistance, 300 Ω in this case.

Key words: active filter, gyrator, feedback, operational amplifier, amplitude-frequency response.

Classification according to AMS 2020: 94C60.

Introduction

In radio engineering and radio electronics, there is a need to pass and amplify signals only in a certain frequency range, the so-called passband, or to suppress them outside the selected range – in the suppression band. This task is assigned to four-poles (filters), the amplitude-frequency characteristics (AFC) of which are non-zero in the passband and close to zero at its outside (Levytskyi, 2007; Bekh, 2024; Carter, & Mancini, 2018).

Passive second-order RLC filters are more efficient than first-order RC filters, but the latter, in turn, do not contain inductors, the dimensions and other parasitic parameters (resistance, interturn capacitance, etc.) of which increase significantly at low frequencies close to the frequencies of the audio range (Bekh, 2024; Carter, & Mancini, 2018). Another disadvantage of passive RC filters is the attenuation of the useful signal in the passband. This drawback is eliminated by combining passive RC filters with amplifier stages covered by various types of feedback. Such combined filters are called active (Bekh, 2024; Carter, & Mancini, 2018).

The object of the study is the gyrator circuit – an RC cell connected in the positive feedback loop of an operational amplifier (Bekh, 2024; "Old Hack EE", 2022; National Semiconductor Corp, 1980) for a change from the use of inductors in active filters with order higher than two. The voltage across capacitor C in the gyrator circuit causes the voltage and current at the output of the circuit to behave similarly to those values across the inductor. Similar "active inductor" circuits allow implementing active filters in a wide frequency range (up to units of GHz (Kilinc et al., 2025)) even by means of microelectronics, where, however, they are more often implemented on inverting and non-inverting amplifiers included in the feedback loop on separate field-effect transistors (Patel, & Oza, 2018), which complicates the calculation of such circuits due to the need to take into account the steepness and nonlinearity of their transmission characteristics (Patel, & Oza, 2018; Saad, & Garrab, 2025; Kilinc et al., 2025).

The literature (Bekh, 2024; Carter, & Mancini, 2018) provides practical recommendations for choosing the value of the resistance of the feedback resistor R_2 in the gyrator circuit on the operational amplifier (Fig. 1 b) – it should be equal to the minimum load resistance of the operational amplifier. However, the minimum load resistance of the operational amplifier itself is not often included in its passport data. In addition, there is no information about the influence of the value of the resistance R_2 on the frequency properties of the resulting device.

The goal of the study is to develop practical recommendations for choosing the value of the feedback resistor R_2 in the gyrator circuit and to study its influence on the amplitude-frequency characteristic of the device.

Computer simulation in the LTSpice software environment using the LT1006 operational amplifier available in it (Linear Technology, 2007) was chosen as the research method.

1. Modeling Details

Figure 1 shows a schematic of an oscillating circuit with an inductor (Fig. 1 a) and a circuit with a gyrator (Fig. 1 b) based on the LT1006 operational amplifier, whose "inductance" is determined by the circuit elements, together with the operational amplifier, and is equal to the inductance of the circuit's coil.

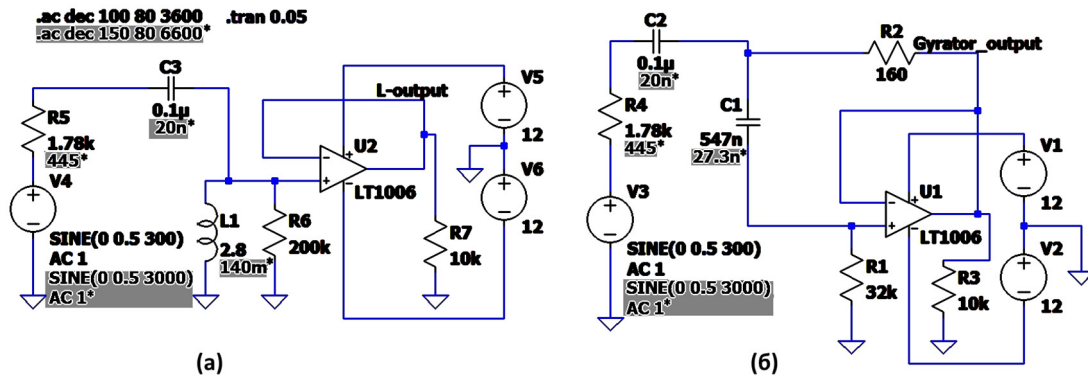


Fig. 1. Schematic of an RLC oscillating circuit with a repeater based on an operational amplifier (output – a "L-output") (a) and a circuit with a gyrator (output – "Gyrator_output") (b). The values of the components without a background correspond to a resonant frequency of approximately 300 Hz, while those with a gray background and an asterisk (*) correspond to – 3 kHz

The resistance of the load resistors (Fig. 1 a and Fig. 1 b) for the LT1006 operational amplifier, as specified in its documentation (Linear Technology, 2007), ranged from 600 Ω to 10 kΩ. A final value of 10 kΩ was selected to eliminate the influence of this parameter.

The circuit with an inductor is loaded by a repeater on the operational amplifier and a resistor connected in parallel to it (Fig. 1 a), which acts as a load resistance even when the operational amplifier's input impedance is very high. Reducing the resistance leads to a decrease in the quality factor of the oscillating circuit, but this changes the results only quantitatively.

The parameters of the gyrator components (Fig. 1 b) were calculated according to the algorithm (Bekh, 2024):

- the resistance of the feedback resistor was selected R_2 ;
- The resistance of the resistor was selected R_1 , such that it satisfies the conditions $R_1 \gg R_2$ (but not more than $200 \cdot R_2$) and $R_1 < 0,1 \cdot R_{BX}$ (where R_{BX} – is the input resistance of the operational amplifier);
- For this resonant frequency f_0 the values of the circuit capacitor C_2 and the capacitor C_1 were selected based on the inductance of the coil L_1 (Fig. 1 a), де – $L_1 = R_1 \cdot R_2 \cdot C_1$, (Fig. 1 b).

The Q-factor of the gyroscope, obtained from the expression (Bekh, 2024):

$$Q = \frac{1}{2} \cdot \sqrt{\frac{R_1}{R_2}},$$

has always been 7,07.

The ranges of parameter variation for the circuit (Fig. 1) elements are listed in Tab. 1. The step size for changing the resistance of the feedback resistor R_2 was 1,5 to 2 times the previous value. This change also determined the values of the parameters R_1 and C_1 . The resonant frequency of the circuit f_0 was 300 Hz in the first stage of the study (the "300 Hz" case in Tab. 1) and 3 kHz in the second stage (the "3 kHz" case in Tab. 1).

During computer simulation using the LTSpice software environment, several frequency response and phase response curves were obtained. They are:

- amplitude-frequency dependences (characteristics) of the transmission coefficient K (in dB) as a function of signal amplitudes (hereinafter referred to as AFCs).
- phase-frequency characteristics, which hereinafter are referred to as phase responses or PFCs or *Phase* (in degrees).

In the frequency response graphs obtained using cursor measurements in LTSpice at the maximum transmission coefficients $K_{0-Gyrator-1}$ and $K_{0-Gyrator-2}$ the resonance frequencies $f_{0-Гіратор-1}$ and $f_{0-Гіратор-2}$, as well as the circuit passbands (at the -3 dB level) were determined; based on these, the Q-factor of the circuit with coils Q_{RLC-1} and Q_{RLC-2} and with gyrators $Q_{Gyrator-1}$ and $Q_{Gyrator-2}$ was calculated. The resistance value of the resistor R_2 was normalized to 1 kΩ to evaluate the relative change in resonant frequencies, the resonant frequency of the circuit with the gyrator was normalized to the frequency of the R5C3L1 oscillating circuit with coils $f_{0-RLC-1}$ and $f_{0-RLC-2}$.

2. Experimental results

Figures 2 and 3 show the frequency response and phase response curves of the circuits under study for the "300 Hz" and "3 kHz" cases, respectively. Noticed that in both cases, the circuit with real inductance (dark solid line) has higher values compared to the same parameter of the circuit with a gyrator (light solid line), while the frequency response behavior of both circuits in the vicinity of the resonant frequency (dotted lines) is similar.

Table 1

Changing the parameters of the RLC oscillating circuit and the gyrator-based circuit for two frequencies

Parameter	Circuit 1 – "300 Hz"		Circuit 2 – "3 kHz"	
Resonant frequency f_0 , Hz (theoretical)	301		3041	
	Minimum	Maximum	Minimum	Maximum
R_1 , kΩ	10	320	8	256
R_2 , Ω	50	1600	40	1280
R_3 , kΩ	10		10	
R_4, R_5 , kΩ	1,78		0,445	
C_1	5,6 μF	5,47 nF	438 nF	427 pF
C_2, C_3	0,1 μF		20 nF	
L_1	2,8 H		140 mH	

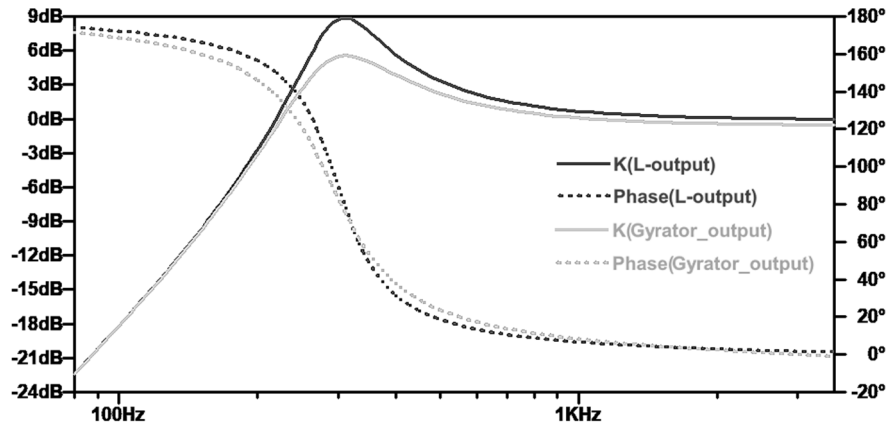


Fig. 2. Amplitude-frequency response (solid lines) and phase responses (dotted lines) plots of an oscillating circuit with an inductor (dark lines) and a circuit with a gyrator (light lines), obtained in LTSpice with a feedback resistor $R_2 = 160 \Omega$ and a resonance frequency of 300 Hz

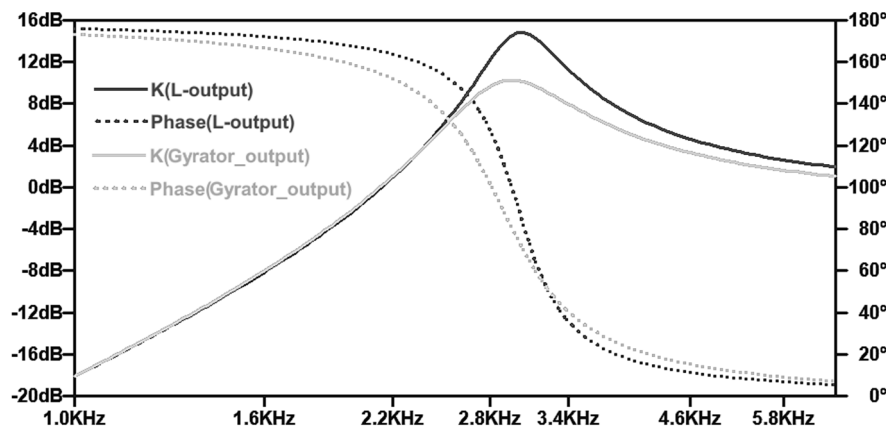


Fig. 3. Amplitude-frequency response (solid lines) and phase responses (dotted lines) plots of an oscillating circuit with an inductor (dark lines) and a circuit with a gyrator (light lines), obtained in LTSpice with a feedback resistor $R_2 = 160 \Omega$ and a resonance frequency of 3 kHz

According to cursor measurements, the resonant frequency of the circuit $R5C3L1$ is $f_{0-RLC-1} = 311,236$ Hz for the "300 Hz" case and $f_{0-RLC-2} = 3041,515$ Hz for the "3 kHz" case.

Figure 4 shows graphs of the relative change in the resonant frequency of a circuit with a gyrator $f_{0-Gyrator-1}$ and $f_{0-Gyrator-2}$ compared to the resonant frequency of a circuit with inductance $f_{0-RLC-1}$ and $f_{0-RLC-2}$ as a function of a resistance of the resistor R_2 for both simulation stages. Measurements showed that the resonant frequency of the circuit with a gyrator $f_{0-Gyrator-1}$ and $f_{0-Gyrator-2}$ can differ from the frequency of the circuit with real inductance in the upward direction up to 12% at resistance of the resistor $R_2 > 300 \Omega$, and the closest values are observed at a resistor resistance of $R_2 = \{200..400\} \Omega$. At a resistance of $R_2 < 100 \Omega$ in the "300 Hz" case, a sharp increase in the resonant frequency of the circuit with the gyrator is observed; at the same time, in the "3 kHz" case at $R_2 < 300 \Omega$, a further decrease in the resonant frequency of the circuit with the gyrator relative to the circuit $R5C3L1$ frequency is observed.

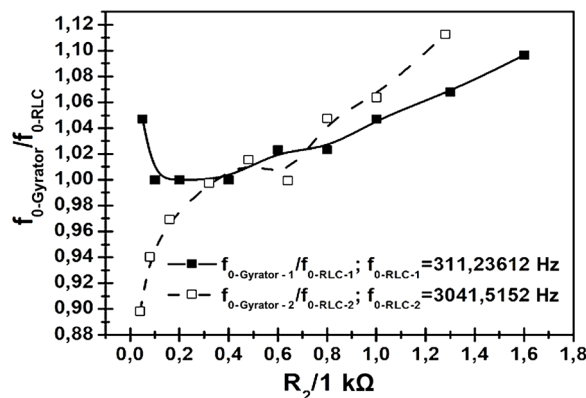


Fig. 4. Graphs showing the relative change in the resonant frequency of a circuit with a gyrator $f_{0-Gyrator-1}$ and $f_{0-Gyrator-2}$ compared to the resonant frequency of a circuit with inductance $f_{0-RLC-1}$ and $f_{0-RLC-2}$ as a function of the resistance of the feedback resistor R_2

Figure 5 shows graphs of the transmission coefficient of the circuit $K_{0-Gyrator-1}$ and $K_{0-Gyrator-2}$ with a gyrator at the resonant frequency as a function of the relative change in the resistance of the feedback resistor R_2 , normalized to the transmission coefficient of the circuit with the coil, which was approximately $K_{0-RLC-1} = 8,908$ dB for the "300 Hz" case and $K_{0-RLC-2} = 14,844$ dB for the "3 kHz" case. Analysis of the graphs showed that for the "300 Hz" case, the maximum transmission coefficient of the circuit with a gyrator is achieved at $R_2 = 400 \Omega$ and does not exceed 0,74 of the transmission coefficient of the circuit with inductance, while for the "3 kHz" case, at $R_2 = 200 \Omega$ it does not exceed 0,7. With a coupling resistance of $R_2 < 100 \Omega$, the transmission coefficient of the circuit with a gyrator decreases rapidly (Fig. 5).

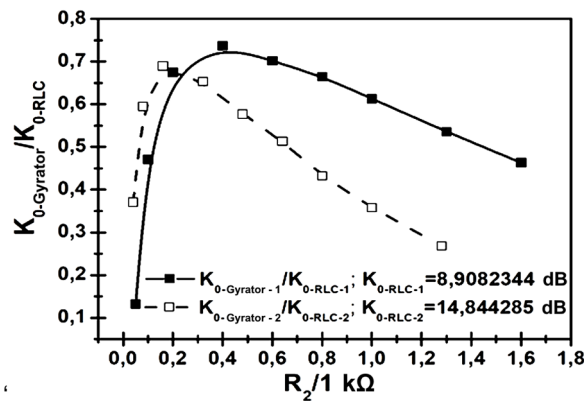


Fig. 5. Graphs showing the dependence of the transmission coefficient $K_{0-Gyrator-1}$ and $K_{0-Gyrator-2}$ on the resonant frequency of the gyrator normalized to the transmission coefficient of the coil-based oscillating circuit relative to the feedback resistor R_2

As is well known, the transmission coefficient of the circuits under study is related to the Q-factor of both the oscillating circuit with a coil and the circuit with a gyrator. Figure 6 shows graphs of the Q-factor of the circuit with a gyrator $Q_{Gyrator-1}$ and $Q_{Gyrator-2}$ normalized to the Q-factor of the RLC circuit with a coil $Q_{RLC-1} = 2,439$ and $Q_{RLC-2} = 5,36$ respectively, as a function of the relative change in the resistance of the feedback resistor R_2 . By comparing the graphs shown in Fig. 5 and Fig. 6, one can see that the dependence of the relative Q-factor of studied circuits with a gyrator mirrors the behavior of the dependence of their transmission coefficients, albeit with slightly lower values.

In conclusion, it can be stated that the optimal resistance of the feedback resistor R_2 should not exceed the output impedance of the operational amplifier. For the circuits studied in this work, at both values of the resonant frequency, a resistance of 300Ω can be considered optimal, i.e., half of the recommended load resistance for the operational amplifier (Linear Technology, 2007).

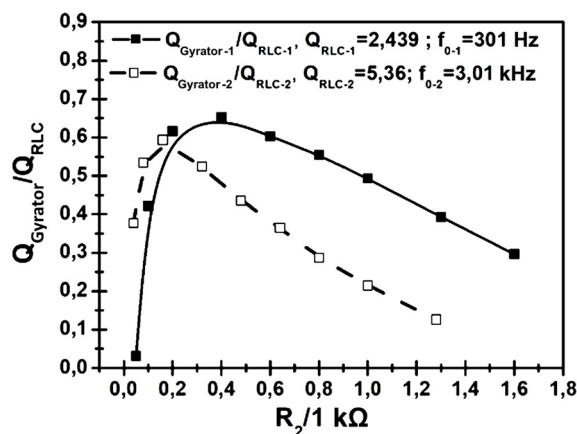


Fig. 6. Graphs showing the dependence of the Q-factor of the circuit with a gyrator ($Q_{Gyrator-1}$ and $Q_{Gyrator-2}$), normalized to the quality factor of the RLC circuit with a coil (Q_{RLC-1} and Q_{RLC-2}), on the relative change in the resistance of the feedback resistor R_2

Discussion and conclusions

To obtain a resonance frequency for a circuit with a gyrator based on the LT1006 operational amplifier that is as close as possible to the resonance frequency of a circuit with an inductor, the resistance of the feedback resistor R_2 , which is part of the gyrator, must be no less than 100Ω , and to ensure a stable gain of the circuit with the oscillator when the resonant frequency changes tenfold, it (R_2) should be approximately half the output impedance of the operational amplifier used (for the LT1006, the resistor R_2 value will be 300Ω). Note that these values depend on the type of operational amplifier, specifically its output resistance, which must be accounted when selecting values of R_2 in practice.

Author's contributions: Ihor Bekh – conceptualization; methodology; development of the theoretical framework; analysis of the results; Roman Bohdanov – analysis of sources; preparation of the literature review; software; collection and verification of empirical data; empirical research and analysis of the results.

Sources of funding. The study was conducted mostly at the authors' own expense. Some financial support was partially provided by Taras Shevchenko National University of Kyiv.

References

- Bekh, I. (2024). *Analog electronic devices*. Publishing Laboratory of the Faculty of Radiophysics, Electronics and Computer Systems of Taras Shevchenko National University of Kyiv [in Ukrainian]. [Бех, І. (2024). *Аналогові електронні пристрої*. Видавнича лабораторія факультету радіофізики, електроніки та комп'ютерних систем КНУ імені Тараса Шевченка].
- Carter, B., Mancini, R. (2018). *Op Amps for Everyone* (5th ed.). Elsevier.
- Kilinc, M., Ormanci, M.A., Kilinc, S., & Kacar, F. (2025). CMOS Active Inductor Using Gm-Boosting Technique with Resistive Feedback and Its Broadband RF Application. *Electronics* 2025, 14, 4776. <https://doi.org/10.3390/electronics14234776>
- Levytskyi, S. (2007). *Basics of radio electronics*. PPC "Kyiv university" [in Ukrainian]. [Левитський, С. (2007). *Основи радіоелектроніки*. ВПЦ "Київський університет"].
- Linear Technology (2007). *LT1006: Precision, Single Supply Op Amp Data Sheet*. Analog Devices. <https://www.analog.com/media/en/technical-documentation/data-sheets/LT1006.pdf>
- National Semiconductor Corp. (1980). *Audio/Radio Handbook*. Internet Archive. https://archive.org/details/bitsavers_nationaldaAudioRadioHandbook_17034677/page/n1/mode/2up
- "Old Hack EE". (2022). (2022). *The Op Amp Gyrator Demystified*. YouTube. <https://www.youtube.com/watch?v=zpGm4R9eGJk>
- Patel, D. P., & Oza S. (2018). *CMOS Active Inductor: A Technical Review*. International Journal of Applied Engineering Research, 13(11) 9680–9685, Research India Publications. https://www.researchgate.net/publication/362909404_CMOS_Active_Inductor_A_Technical_Review
- Saad, S., & Garrab, H. (2025). Chapter 3: Recent Advances in Single-Ended Gyrator–C Active Inductor Designs (Innovations from 2008). Deep Science Publishing, 2025. https://doi.org/10.70593/978-93-7185-269-2_3

Отримано редакцією журналу / Received: 27.07.25

Прорецензовано / Revised: 30.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Ігор БЕХ, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0002-7359-8192
e-mail: igor_ivanovich_bekh@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Роман БОГДАНОВ, канд. фіз.-мат. наук, асист.
ORCID ID: 0009-0007-8449-1466
e-mail: roman.bohdanov@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

ВПЛИВ ОПОРУ РЕЗИСТОРА ПОЗИТИВНОГО ЗВОРОТНОГО ЗВ'ЯЗКУ НА АМПЛІТУДНО-ЧАСТОТНУ ХАРАКТЕРИСТИКУ ГІРАТОРА НА ОПЕРАЦІЙНОМУ ПІДСИЛЮВАЧІ

Позбутися від використання котушок індуктивності в активних RLC-фільтрах другого та більше порядків можна за допомогою гіратора – RC-комірки, увімкненої в коло позитивного зворотного зв'язку операційного підсилювача. Така схема «активної котушки індуктивності» дає змогу реалізовувати активні фільтри навіть засобами мікроелектроніки.

Метою пропонованого дослідження є розроблення практичних рекомендацій щодо вибору величини опору резистора зворотного зв'язку у схемі гіратора та його впливу на амплітудно-частотну характеристику (АЧХ).

Як метод дослідження було використано комп'ютерне моделювання у програмному застосунку LTSpice, під час якого було порівняно основні характеристики коливального LC-контуру та контуру з гіратором на операційному підсилювачі LT1006 й отримано відповідні амплітудно- та фазово-частотні характеристики.

Показано, що положення максимуму АЧХ контуру з гіратором тим ближче до резонансної частоти RLC контуру, чим опір резистора зв'язку ближчий до вихідного опору операційного підсилювача. Проте він має бути не меншим за 100 Ом, а для забезпечення сталого значення коефіцієнта передачі контуру з гіратором, у разі зміни резонансної частоти у десять разів, має становити близько половини вихідного опору операційного підсилювача, у цьому випадку – 300 Ом.

Ключові слова: активний фільтр, гіратор, зворотний зв'язок, операційний підсилювач, амплітудно-частотна характеристика.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

UDC 544.35

DOI: <https://doi.org/10.17721/1812-5409.2026/1.37>

Yuri ZABASHTA¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0003-2835-4057
e-mail: yu.zabashta@gmail.com

Yaroslav LAZARENKO⁴, student of grade 10
ORCID ID: 0009-0007-8716-2684
e-mail: yaroslav.lazer@gmail.com

Oleksandr ALEKSEEV¹, PhD (Phys. & Math.), Senior Research
ORCID ID: 0000-0002-6230-2093
e-mail: alan@univ.kiev.ua

Lena VERGUN¹, PhD (Phys. & Math.)
ORCID ID: 0009-0005-6086-6548
e-mail: liena.vergun@knu.ua

Kateryna HNATIUK¹, PhD (Phys. & Math.)
ORCID ID: 0000-0001-8605-5253
e-mail: 20gnatk@gmail.com

Dmytro ANDRUSENKO¹, PhD (Phys. & Math.)
ORCID ID: 0000-0002-3061-4498
e-mail: and_@univ.kiev.ua

Kateryna YABLOCHKOVA¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-8719-5776
e-mail: yablochkova@knu.ua

Roman DINZHOS^{2,3}, DSc (Technical Sciences), Prof.
ORCID ID: 0000-0003-1105-2642
e-mail: dinzhosrv@gmail.com

Neli KOSEVA³, PhD, Prof.
ORCID ID: 0000-0002-8445-8953
e-mail: koseva@polymer.bas.bg

Maksym LAZARENKO¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0003-0719-3522
e-mail: lazmaks@knu.ua

¹Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

²Petro Mohyla Black Sea National University, Mykolayiv, Ukraine

³Institute of Polymers, Bulgarian Academy of Sciences, Sofia, Bulgaria

⁴Buchanan Lyceum No. 4, Bucha, Kyiv region, Ukraine

EFFECT OF THE POLYMER NETWORK ON THE DIFFUSION OF FOREIGN PARTICLES IN HYDROGELS

In the present work, we introduce a physical model of foreign particle diffusion in polymer hydrogels that accounts for the network architecture of these systems. The model employs a classical diffusion theory framework in condensed media to derive a theoretical dependence of the diffusion coefficient on the volume occupied by the hydrogel's network. We demonstrate that the changes in the diffusion of foreign particles can be attributed to the changes in the free activation energy due to the presence of the polymer network. Using optical image analysis, we obtained the dependence of the diffusion coefficient in agarose and gelatine hydrogels on the concentration of the gelling agent. The experimental data obtained are in good agreement with the theoretical dependences predicted by the proposed physical model. Furthermore, when the size of the diffusing particles is much smaller than the hydrogel pore size, the diffusion coefficient appears to be independent of the nature of the hydrogel and determined solely by the gelling agent concentration.

Keywords: diffusion, hydrogels, agarose, gelatine, rhodamine, polymer network.

PACS classification: 66.30.-h, 82.70.Gg.

Introduction

A polymer hydrogel is a water-polymer system in which polymer chains form a network (Yuan et al., 2025; Zhu et al., 2023; Alekseev et al., 2019). These hydrogels are widely used in the pharmaceutical industry (Clegg et al., 2024; Yang, 2022; Mehta et al., 2023; Ho et al., 2022) for drug delivery systems. When a hydrogel 'loaded' with an active pharmaceutical ingredient is brought into contact with tissue, the diffusion flow of the API molecules is formed (Richtering et al., 2016). The rate of this flow is given by the diffusion equation (Birkhoff, 2015):

$$J = -D\nabla c, \quad (1)$$

where c is the concentration of API, and D is a diffusion coefficient.

This implies that the duration of drug action is, in part, determined by the rate of diffusion. Engineering the rate of diffusion thus offers an opportunity to modify a drug's half-life, as described, for instance, in (Smeets, & Hoare, 2013; Wang et al., 2020; Armstrong et al., 2014; Velasco et al., 2011), where the authors' work is empirical and practical. The knowledge gap, however, exists in the absence of a physical model of the diffusive motion of particles of enriching agents in hydrogels. The

© Zabashta Yuri, Lazarenko Yaroslav, Alekseev Oleksandr, Vergun Lena, Hnatiuk Kateryna, Andrusenko Dmytro, Yablochkova Kateryna, Dinzhos Roman, Koseva Neli, Lazarenko Maksym, 2026

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

difference in the diffusion of such particles in hydrogels, compared to water, is primarily due to the network architecture (Roca-Arroyo et al., 2025; Zhu et al., 2025). Thus, in this work, we present the model of diffusion of foreign particles that accounts for the presence of a polymer network.

1. Theoretical model

1.1. Diffusion mechanism in condensed matter: general characteristics

The current understanding of the diffusion mechanism in condensed matter is based on the classical theory described in (Frenkel, 1955). This theory can be summarized as follows.

The particle that undergoes diffusion has position vectors given by $r_1, r_2, \dots, r_n$ at instances $t_1, t_2, \dots, t_n$. Each of the positions of the particle corresponds to a certain locally equilibrium state of the system, which contains this particle (Hnatiuk et al., 2023; Lazarenko et al., 2024; 2025). Let's denote the free energy of each of these states as $F_1, F_2, \dots, F_n$. The volumes in which the local equilibria are established, and which contain the particle in question, will be denoted $V_1, V_2, \dots, V_n$.

The change in the position vectors of the particle in two consecutive states is

$$l_i = r_{i+1} - r_i, \quad (2)$$

an elementary displacement of the particle.

The time interval θ_i during which the elementary displacement takes place is thus

$$\theta_i = t_{i+1} - t_i. \quad (3)$$

As the particle undergoes diffusive motion, an intermediate locally equilibrium state of free energy F'_i is formed between the pair of states with free energies F_i and F_{i+1} . The volume where this locally equilibrium state is formed is denoted V'_i .

The difference in energies

$$E_i = F'_i - F_i, \quad (4)$$

is usually referred to as a free activation energy.

In the diffusion motion the following condition is met

$$F'_i > F_i, F_{i+1}. \quad (5)$$

As seen from the inequality (5), for the particle to undergo an elementary displacement, the system must overcome the energy barrier E_i . This occurs due to thermal fluctuations, and thus the transition from state with free energy F_i into the state with free energy F_{i+1} is referred to as the thermal fluctuation event. The state with free energy F'_i is referred to as an activated state and the states with free energies F_i and F_{i+1} as non-activated. Finally, the diffusive motion itself is a thermally activated motion.

Let's denote the lifetimes of states with free energies F_i and F'_i as τ_i and τ'_i , respectively. The lifetime of the elementary thermal fluctuation act, and thus the duration of the elementary displacement of the particle is determined by the following sum:

$$\theta_i = \tau_i + \tau'_i. \quad (6)$$

According to Boltzmann's principle (Bagchi 2023), the following equation stands

$$\frac{\tau_i}{\tau'_i} = \exp\left(-\frac{E_i}{k_B T}\right), \quad (7)$$

where k_B is the Boltzmann constant and T is the absolute temperature.

The following inequality must also stand:

$$\frac{E_i}{k_B T} \gg 1. \quad (8)$$

With this in mind, equation (6) becomes

$$\theta_i \approx \tau_i. \quad (9)$$

In other words, the duration of the elementary displacement is the lifetime of the state with free energy F_i , or the lifetime of the equilibrium state between the displacements.

1.2. Diffusion of particles in hydrogel: a physical model

The general features of the diffusion mechanism in condensed matter, as mentioned above, should be inherent in the diffusion of foreign particles in a hydrogel. The free activation energy plays the main role in this mechanism (Wolde-Kidan et al., 2021). Within this mechanism, the fluctuation associated with the formation of the activated state, occurring in the region with the initial volume V' , tends to increase with increasing volume. This tendency meets resistance from the environment. The amount of the resistance experienced must then determine the numerical value of the free activation energy. The polymer network of the hydrogel must significantly increase this resistance (compared to water) and thus must substantially increase the free activation energy.

For each position of the particle, the value of E_i is determined by many factors: the configuration of the polymer network in the vicinity of the particle, the network's rigidity, the particle's location relative to the network, etc. In other words, the value of E_i is determined by the local supramolecular structure of the hydrogel. The strict calculation of the free activation energy could be performed using methods of statistical physics. Instead, in the present work, we introduce a simplified thermodynamic model that accounts for the presence of the polymer network when describing the diffusion of foreign particles in a hydrogel. This model is based on approximations used in the classical diffusion theory (Frenkel, 1955).

The first approximation we use in the model is the following: we assume that the elementary displacements are equal in magnitude, and so are the durations of each elementary displacement. We denote the elementary displacements as l and the durations of these displacements as τ . Within such an approximation, the values of the diffusion coefficient must be a function of both l and τ :

$$D = D(l, \tau). \quad (10)$$

The unit of the diffusion coefficient, as indicated by (1), is $\text{m}^2 \cdot \text{s}^{-1}$. Based on this, the function in (10) is (up to a factor of order one) equal to:

$$D = l^2 / \tau. \quad (11)$$

After substitution of (7) into (11) we get:

$$D = \frac{l^2}{\tau} \exp\left(-\frac{E}{k_B T}\right). \quad (12)$$

For the foreign particles in water (denoted by index "0") diffusion coefficient can be written as:

$$D_0 = \frac{l_0^2}{\tau_0} \exp\left(-\frac{E_0}{k_B T}\right). \quad (13)$$

The second assumption made in the model is that elementary displacements l and l_0 must be of the order of magnitude of the size of a particle, while τ and τ_0 must be of the same order of magnitude (Frenkel, 1955). This implies that:

$$\frac{l^2}{\tau} = \frac{l_0^2}{\tau_0}. \quad (14)$$

Consequently, we can arrive at the following expression for the diffusion coefficient:

$$D = D_0 \exp\left(-\frac{\Delta E}{k_B T}\right), \quad (15)$$

where

$$\Delta E = E - E_0. \quad (16)$$

According to (14), within the classical theory formalism, the presence of the polymer network changes the value of D as it changes the value of the free activation energy: ΔE .

Within the assumptions made, as indicated by expressions (10) and (14), ΔE must be the quantity averaged over the system. Therefore, it must be the function of macroscopic parameters, holistically characterizing the system. One of such parameters is the concentration q , a relative volume occupied by the polymer network.

Consequently, we assume the existence of a function of q which gives the values of ΔE :

$$\Delta E = f(q). \quad (17)$$

As the polymer network slows down the diffusion of the foreign particles, this function can be safely assumed to be monotonously increasing. In this case, we can expand this function into a power series and limit this expansion to the first term:

$$\Delta E = \left. \frac{df}{dq} \right|_0 q. \quad (18)$$

Then expression (15) becomes:

$$D = D_0 \exp\left[-\frac{q}{Q(T)}\right], \quad (19)$$

where

$$Q(T) = \frac{k_B T}{\left. \frac{df}{dq} \right|_0}. \quad (20)$$

Expression (19) summarizes the essence of the model proposed. The presence of a polymer network changes the diffusion of the foreign particles via the change imparted on $Q(T)$. This quantity can be thought of as the value of concentration q that prevents the diffusion of foreign particles in the volume occupied by a hydrogel.

2. Materials and methods

In the present work, we used the following two gelling agents: agarose (purchased from Sigma-Aldrich) and gelatine (purchased from Bloom 200, France). Agarose is the main component (70–80%) of agar-agar natural extract obtained from *Florideophyceae* algae (Jiang et al., 2023). Gelatine is an animal-sourced protein, predominantly consisting of collagen. It can be found in animal skin, bones, and ligaments and is 90% protein (with such amino acids as glycine, proline, and hydroxyproline) (Milano et al., 2023). Both agarose- and gelatine- based hydrogels are bio-compatible. They have a porous structure (Lazarenko et al., 2025), that resembles a honeycomb structure, as revealed by cryo-scanning electron microscopy (Rahbani et al., 2013). Agarose pores are spherical in shape, with radii between 370 and 700 nm. Gelatine pores are smaller, more uniform, with radii between 320 and 650 nm. The foreign particles introduced into the hydrogel were rhodamine 6G (R6G) molecules. The hydrodynamic radius of the R6G molecule is only 0.65 nm (Majer, & Melchior, 2014), several orders of magnitude smaller than the size of the pores.

Unstained hydrogels were prepared by mixing gelling agents with distilled water and keeping them for 10 minutes at 90 °C. Hydrogels stained with a dye were prepared by mixing gelling agents with 0.1 % aqueous solution of R6G. The following samples were prepared: 2, 4, 6, 8 % agarose gels; 16, 20, 25 % gelatine gels; 10 % (2 % agarose + 8 % gelatine) and 14 % (4 % agarose + 10 % gelatine) mixtures.

The systems in which diffusion was observed were prepared as follows: first, 1 mL of pure hydrogel at 90 °C was placed in a 2 mL syringe. The hydrogel was allowed to cool to room temperature. Then 1 more milliliter of stained hydrogel at 60 °C was added to the syringe. The resulting sample with a boundary between the stained and unstained gel is shown in Fig. 1. Subsequent diffusion was observed in the syringe at 24 °C.

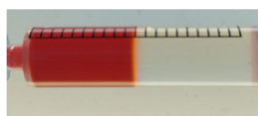


Fig. 1. Syringe with hydrogel at the beginning of the experiment: the left side contains rhodamine dye

To observe the diffusion process, the image processing analysis was utilized (Davydiuk et al., 2024). The syringe was uniformly illuminated with white, diffused light from two 40 cm – long rectangular white LED lamps, which were placed horizontally on a stand in front of a digital camera. The images were taken at equal time intervals, starting from the moment the stained and unstained hydrogels were brought into contact. The time intervals for hydrogels with higher concentrations were longer than for those with lower concentrations. The RGB images obtained were processed in ImageJ software. As R6G

absorbs blue light well, only the blue channel in the RGB image was used for processing and analysis. The changes in the blue channel intensity were normalized for each sample.

3. Results and their discussion

Figure 2 shows the snapshots of the syringes with hydrogels taken at 0, 2, 4, 6, and 14 hours after the start of the experiment. The rows of the figure represent the hydrogels with, respectively, 4 % (agarose), 8 % (agarose), and 16 % (gelatine) concentrations.

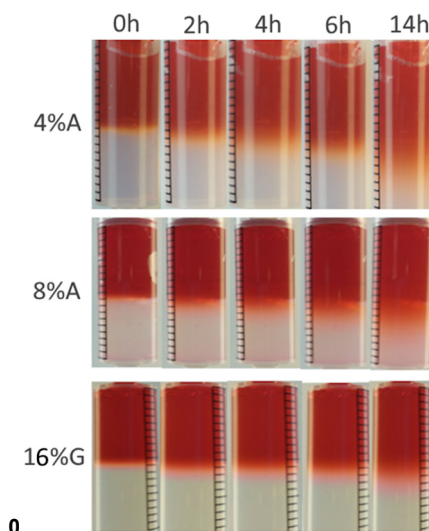


Fig. 2. Syringes with gels of different concentrations: 4 % (agarose), 8 % (agarose), and 16 % (gelatine). Images of the syringes were taken at the start of the experiment and after 2, 4, 6, and 14 hours, respectively

The initial analysis of Fig. 2 shows that the stained/unstained boundary moves more slowly for gels with higher R6G concentration. To quantify the diffusion rate, ImageJ was used. Firstly, the Image\Color\Split Channels function was used to split the RGB image into separate channels. Then, the blue channel image was converted to greyscale (Fig. 3 b).

The central section of the image was then selected for analysis using the Analyze\Plot Profile function (Fig. 3 c). This function was used to plot the intensity vs. coordinate function (Figure 3d), which correlates with the spatial distribution of the R6G relative concentration in the gel.

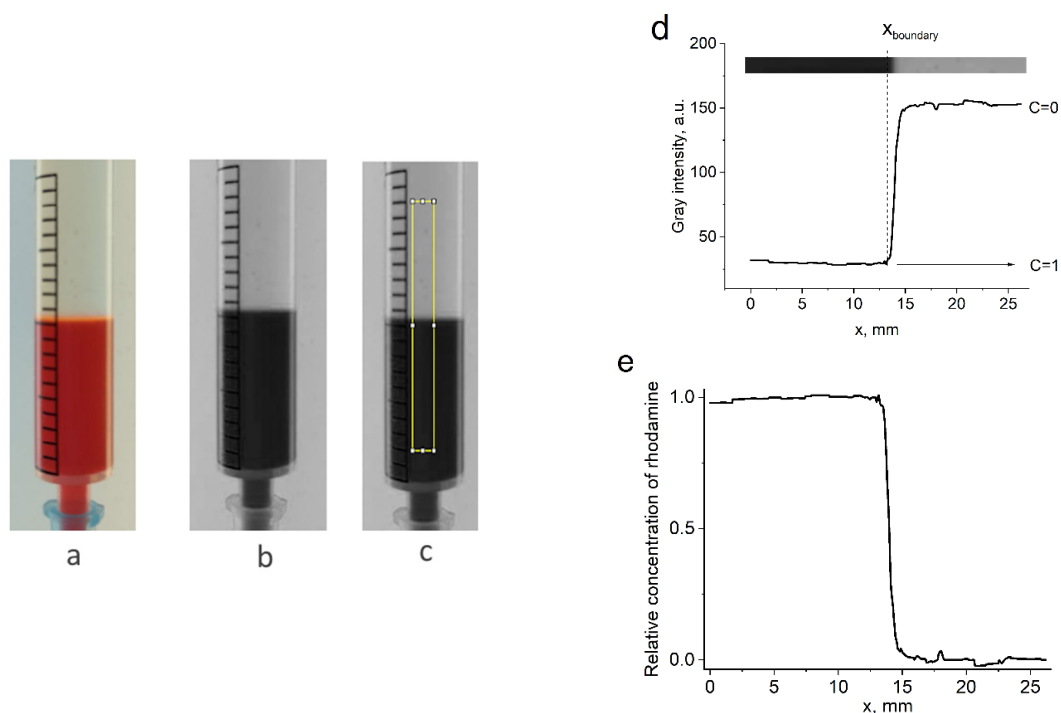


Fig. 3. Digitization of the images representing the dye concentration boundary (front) position in space. a – original image, b – the same image after converting into greyscale by blue color intensity, c – the area of the image used for the intensity distribution analysis, d – intensity of the greyscale image vs. coordinate, e – relative concentration of R6G vs. coordinate

In Fig. 3 e, the highest relative concentration (C) of R6G corresponds to $C = 1$, and the absence of R6G corresponds to $C = 0$. To determine the numerical value of the diffusion coefficient of the R6G in the hydrogel, a multistep algorithm was utilized. We use the data from Fig. 4, representing relative concentration vs. coordinate for the 25 % gelatine hydrogel, to illustrate this algorithm. The data in the graph shows the concentration of the dye at different points in the sample at different times.

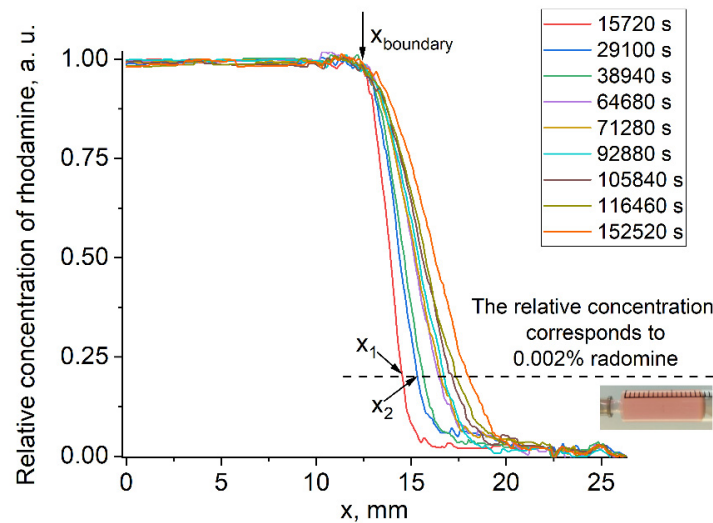


Fig. 4. Relative concentration of R6G vs. coordinate recorded at different times for the 25 % gelatine gel

The change in the concentration of R6G must obey Fick's second law (Crank, 1979):

$$\frac{\partial C(x,t)}{\partial t} = D \frac{\partial^2 C(x,t)}{\partial x^2}, \quad (21)$$

where $C(x,t)$ is the concentration of the substance that undergoes diffusion, D is a diffusion coefficient, $\frac{\partial C}{\partial t}$ is the partial first derivative of the concentration function with respect to time, whereas $\frac{\partial^2 C}{\partial x^2}$ is the second partial derivative of the concentration with respect to coordinate x .

The solution of (21) for a half-space with a constant concentration C_0 of the diffusant at the boundary is a Gauss error function (Crank, 1979):

$$C(x,t) = C_0 \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right), \quad (22)$$

where x is the position of the boundary, and t is time.

Which gives the following expression:

$$D = \frac{x^2/t}{4(\operatorname{erfc}^{-1}(C(x,t)/C_0))^2}. \quad (23)$$

To calculate the value of $\operatorname{erfc}^{-1}(C(x,t)/C_0)$ let's examine a gel with R6G concentration of 0.002 % (Fig. 4). The initial concentration of the rhodamine in the stained gel was $C_s = 0.1$ %. The boundary corresponds to the half of this value: $C_0 = C_s/2 = 0.05$ %. The 0.002 % concentration is thus 25 times smaller than that at the boundary. Such a value allows us to neglect the non-linear effects of a camera at high concentrations of the dye. The numerical value of the ratio $C(x,t)/C_0 = 0.02$. The value of $\operatorname{erfc}^{-1}(0.02) \approx 1.453$.

Then the coefficient of diffusion is given by

$$D = 0.118 x^2/t. \quad (24)$$

Expression (24) gives the value of the diffusion coefficient, given the position of the boundary at a given time instant.

In Fig. 4, a horizontal line corresponding to the given (0.002 %) concentration of R6G intersects the graphs plotted for different time intervals at different values of coordinate x . The intersections provide the data point pairs x_1 and t_1 , x_2 and t_2 , and so on, corresponding to the same dye concentration. This data, x^2 vs. t is plotted in Fig. 5. As shown in this figure, the squared position of the given R6G concentration in a hydrogel appears to vary linearly with time, with a slope of $1.627 \cdot 10^{-4} \text{ mm}^2 \cdot \text{s}^{-1}$.

Substituting this numerical value into the equation (24) we get the following for the diffusion coefficient: $D = 0.118 \cdot 1.627 \cdot 10^{-4} = 19.2 \cdot 10^{-6} \text{ mm}^2 \cdot \text{s}^{-1}$. The same algorithm is applied to determine the diffusion coefficient for gels with different gelling agent concentrations. This plot of diffusion coefficient vs. the concentration of a gelling agent in a hydrogel is shown in Fig. 6.

The variation of the diffusion coefficient with the concentration of the gelling agent appears to be exponentially decaying in agreement with (19) This behavior persists despite the different nature of the gelling agents – all points seem to fit the same curve, even though some of them correspond to agarose hydrogels, other to gelatine hydrogel, and the rest to hydrogels formed by the mixture of agarose and gelatine. The latter fact can be interpreted as follows: when the size of the diffusant molecule (R6G in this case) is much smaller than the pore size in the gel, the nature of the gel does not affect the diffusion coefficient.

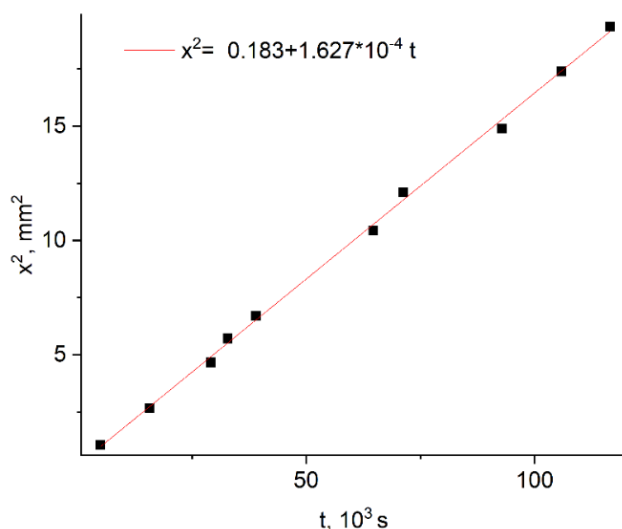


Fig. 5. Squared position of the given (0.002 %) concentration of R6G in a sample of 25 % gelatine hydrogel vs. time

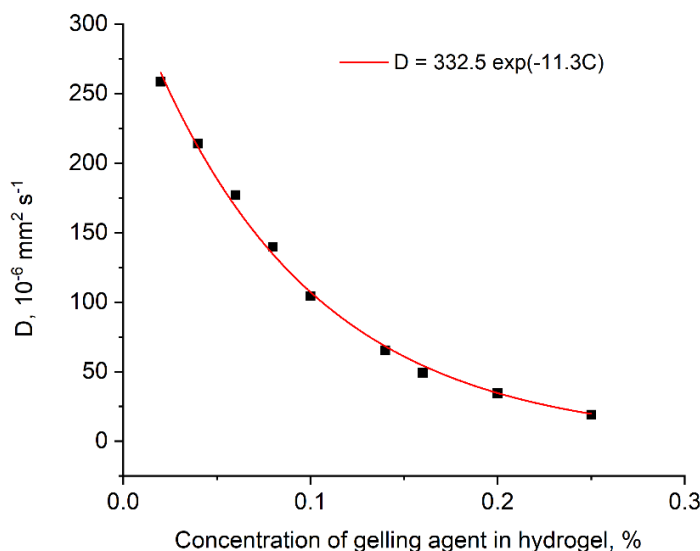


Fig. 6. Diffusion coefficient of R6G vs. gelling agent concentration in the hydrogel

Discussion and conclusions

The diffusive motion of foreign particles in hydrogels has features inherent to the diffusion of particles in any condensed matter. It can be thought of as a sequence of elementary displacements, in each of which the system, due to thermal fluctuations, overcomes an energy barrier called free activation energy.

Diffusion of foreign particles in the hydrogel is altered by the presence of a polymer network, which creates obstacles to the diffusion. Such interference forces the diffusion coefficient of the foreign particles to become a function of the relative volume occupied by the polymer network. This function has the form of an exponential decay. The constant that determines the rate of decay serves as a parameter characterizing the effect of the polymer network on the diffusion of foreign particles in the hydrogel.

The experiment shows that (within experimental uncertainty) this constant is identical for hydrogels of various natures. Thus, the exponential decay observed can be considered a universal process in polymer hydrogels. And latter fact may be of interest when engineering hydrogels intended for controlled drug delivery.

Authors' contribution: Yuri Zabashta – conceptualization; methodology; Yaroslav Lazarenko – investigation, formal analysis; Oleksandr Alekseev – investigation, conceptualization; Lena Vergun – investigation, writing – original draft; Kateryna Hnatuk – investigation; Dmytro Andrusenko – investigation, validation, formal analysis; Kateryna Yablochkova – Writing – review & editing; Roman Dinzhos – data curation, formal analysis; Neli Koseva – methodology, investigation; Maksym Lazarenko – investigation, validation, visualization, writing – review & editing, supervision.

Sources of funding. The study was financed by the National Research Foundation of Ukraine [project № 2025.07/0341].

References

- Alekseev, O. M., Zabashta, Y. F., Kovalchuk, V. I., Lazarenko, M. M., & Bulavin, L. A. (2019). The structure of polymer clusters in aqueous solutions of hydroxypropyl Cellulose. *Ukrainian journal of physics*, 64(3), 238–238. 10.15407/ujpe64.3.238
- Armstrong, M., Walker, S., Mains, J., & Wilson, C. G. (2014). Drug delivery across the nasal mucosa. *Mucoadhesive Materials and Drug Delivery Systems*, 61–82. <https://doi.org/10.1002/9781118794203.ch03>
- Bagchi, B. (2023). *Nonequilibrium statistical mechanics: an introduction with applications*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003157601>
- Birkhoff, G. (2015). *Hydrodynamics*. Princeton University Press.
- Clegg, J. R., Adebawale, K., Zhao, Z., & Mitragotri, S. (2024). Hydrogels in the clinic: An update. *Bioengineering & translational medicine*, 9(6), e10680. <https://doi.org/10.1002/btm2.10680>
- Crank, J. (1979). *The mathematics of diffusion*. Oxford university press.
- Davydiuk, A. L., Andrusenko, D. A., Lazarenko, M. M., Burbelo, R. M., Kuzmich, A. G., Yablochkova, K. S., Dinzhos R. V., & Alekseev, O. M. (2024). A new photoacoustic method to study diffusion in soft matter. *Molecular Crystals and Liquid Crystals*, 768(16), 858–867. <https://doi.org/10.1080/15421406.2024.2372126>
- Frenkel, J. (1955). *Kinetic Theory of Liquids*. Dover Publications. Inc., New York.
- Hnatiuk, K. I., Zabashta, Y. F., Lazarenko, M. M., Alekseev, S. A., Yablochkova, K. S., Ushcats, M. V., Dinzhos, R.V., Vergun, L. Yu., Andrusenko, D. A., Alekseev, A. N., & Bulavin, L. A. (2023). Dielectric Relaxation in Nanocrystals: A Scale Effect. *The Journal of Physical Chemistry C*, 127(25), 12107–12117. <https://doi.org/10.1021/acs.jpcc.3c01523>
- Ho, T. C., Chang, C. C., Chan, H. P., Chung, T. W., Shu, C. W., Chuang, K. P., Duh, T. H., Yang, M. H., & Tyan, Y. C. (2022). Hydrogels: properties and applications in biomedicine. *Molecules*, 27(9), 2902. <https://doi.org/10.3390/molecules27092902>
- Jiang, F., Xu, X. W., Chen, F. Q., Weng, H. F., Chen, J., Ru, Y., Xiao, Q., & Xiao, A. F. (2023). Extraction, modification and biomedical application of agarose hydrogels: a review. *Marine Drugs*, 21(5), 299. <https://doi.org/10.3390/md21050299>
- Lazarenko, M. M., Zabashta, Y. F., Alekseev, O. M., Alekseev, S. A., Yablochkova, K. S., Vergun, L. Y., Andrusenko, D. A., Cherevko, K. V., Shevchenko, V. B., Dinzhos, R. V., & Bulavin, L. A. (2024). Thermally activated processes: the underlying mechanism of activated state formation. *RSC advances*, 14(53), 39242–39252. 1 <https://doi.org/10.1039/D4RA06983H>
- Lazarenko, M. M., Zabashta, Y. F., Honcharuk, D. K., Alekseev, O. M., Yablochkova, K. S., Vergun, L. Y., Andrusenko, D. A., Lazarenko, M. V., & Bulavin, L. A. (2025). Determining hydrogel porosity through dielectric relaxation intensity ratios between water and hydrogel. *Soft Matter*, 21(21), 4298–4305. <https://doi.org/10.1039/D5SM00077G>
- Majer, G., & Melchior, J. P. (2014). Characterization of the fluorescence correlation spectroscopy (FCS) standard Rhodamine 6G and calibration of its diffusion coefficient in aqueous solutions. *The Journal of Chemical Physics*, 140(9). <https://doi.org/10.1063/1.4867096>
- Mehta, P., Sharma, M., & Devi, M. (2023). Hydrogels: An overview of its classifications, properties, and applications. *Journal of the Mechanical Behavior of Biomedical Materials*, 147, 106145. <https://doi.org/10.1016/j.jmbbm.2023.106145>
- Milano, F., Masi, A., Madaghiele, M., Sannino, A., Salvatore, L., & Gallo, N. (2023). Current trends in gelatin-based drug delivery systems. *Pharmaceutics*, 15(5), 1499. <https://doi.org/10.3390/pharmaceutics15051499>
- Rahbani, J., Behzad, A. R., Khashab, N. M., & Al-Ghoul, M. (2013). Characterization of internal structure of hydrated agar and gelatin matrices by cryo-SEM. *Electrophoresis*, 34(3), 405–408. <https://doi.org/10.1002/elps.201200434>
- Richtering, W., Potemkin, I. I., Rudov, A. A., Sellge, G., & Trautwein, C. (2016). Could multiresponsive hollow shell-shell nanocontainers offer an improved strategy for drug delivery? *Nanomedicine*, 11(22), 2879–2883. <https://doi.org/10.2217/nnm-2016-0327>
- Roca-Arroyo, A. F., Gutierrez-Rivera, J. A., Morton, L. D., & Castilla-Casadiago, D. A. (2025). Hydrogel network architecture design space: impact on mechanical and viscoelastic properties. *Gels*, 11(8), 588. <https://doi.org/10.3390/gels11080588>
- Smeets, N. M., & Hoare, T. (2013). Designing responsive microgels for drug delivery applications. *Journal of Polymer Science Part A: Polymer Chemistry*, 51(14), 3027–3043. <https://doi.org/10.1002/pola.26707>
- Velasco, D., Danoux, C. B., Redondo, J. A., Elvira, C., San Roman, J., Wray, P. S., & Kazarian, S. G. (2011). pH-sensitive polymer hydrogels derived from morpholine to prevent the crystallization of ibuprofen. *Journal of controlled release*, 149(2), 140–145. <https://doi.org/10.1016/j.jconrel.2010.10.015>
- Wang, S., Liu, R., Fu, Y., & Kao, W. J. (2020). Release mechanisms and applications of drug delivery systems for extended-release. *Expert Opinion on Drug Delivery*, 17(9), 1289–1304. <https://doi.org/10.1080/17425247.2020.1788541>
- Wolde-Kidan, A., Herrmann, A., Prause, A., Gradzielski, M., Haag, R., Block, S., & Netz, R. R. (2021). Particle diffusivity and free-energy profiles in hydrogels from time-resolved penetration data. *Biophysical Journal*, 120(3), 463–475. <https://doi.org/10.1016/j.bpj.2020.12.020>
- Yang, D. (2022). Recent advances in hydrogels. *Chemistry of Materials*, 34(5), 1987–1989. <https://doi.org/10.1021/acs.chemmater.2c00188>
- Yuan, Y., Zhang, Q., Lin, S., & Li, J. (2025). Water: The soul of hydrogels. *Progress in Materials Science*, 148, 101378. <https://doi.org/10.1016/j.pmatsci.2024.101378>
- Zhu, T., Ni, Y., Biesold, G. M., Cheng, Y., Ge, M., Li, H., Huang, J., Lin, Z., & Lai, Y. (2023). Recent advances in conductive hydrogels: classifications, properties, and applications. *Chemical Society Reviews*, 52(2), 473–509. <https://doi.org/10.1039/D2CS00173J>
- Zhu, Y., Zhang, H., Tan, Q., Wu, Y., Du, M., Hadjichristidis, N., Jiang Yu., & Zhang, D. (2025). Strong and ultra-tough hydrogel with hierarchical cross-linking network architecture constructed by a hyperbranched topological structure. *Macromolecules*, 58(13), 6512–6523 <https://doi.org/10.1021/acs.macromol.5c00513>

Отримано редакцією журналу / Received: 09.01.26

Прорецензовано / Revised: 17.03.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Юрій ЗАБАШТА¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0003-2835-4057
e-mail: yu.zabashta@gmail.com

Ярослав ЛАЗАРЕНКО⁴, учень 10 кл.
ORCID ID: 0009-0007-8716-2684
e-mail: yaroslav.lazer@gmail.com

Олександр АЛЕКСЄЄВ¹, канд. фіз.-мат. наук, ст. наук. співроб.
ORCID ID: 0000-0002-6230-2093
e-mail: alan@univ.kiev.ua

Лена ВЕРГУН¹, канд. фіз.-мат. наук
ORCID ID: 0009-0005-6086-6548
e-mail: lena.vergun@knu.ua

Катерина ГНАТЮК¹, канд. фіз.-мат. наук
ORCID ID: 0000-0001-8605-5253
e-mail: 20gnatk@gmail.com

Дмитро АНДРУСЕНКО¹, канд. фіз.-мат. наук
ORCID ID: 0000-0002-3061-4498
e-mail: and_@univ.kiev.ua

Катерина ЯБЛОЧКОВА¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-8719-5776
e-mail: yablochkova@knu.ua

Роман ДІНЖОС^{2,3}, д-р техн. наук, проф.
ORCID ID: 0000-0003-1105-2642
e-mail: dinzhosrv@gmail.com

Нелі КОСЕВА³, д-р філософії, проф.
ORCID ID: 0000-0002-8445-8953
e-mail: koseva@polymer.bas.bg

Максим ЛАЗАРЕНКО¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0003-0719-3522
e-mail: lazmaxs@knu.ua

¹Київський національний університет імені Тараса Шевченка, Київ, Україна

²Чорноморський національний університет імені Петра Могили, Миколаїв, Україна

³Інститут полімерів Болгарської академії наук, Софія, Болгарія

⁴Бучанський ліцей № 4, м. Буча, Київська область, Україна

ВПЛИВ КАРКАСА НА ДИФУЗИЮ СТОРОННІХ ЧАСТИНОК У ПОЛІМЕРНИХ ГІДРОГЕЛЯХ

Запропоновано фізичну модель, що описує вплив каркаса на дифузію сторонніх частинок у полімерних гідрогелях. Під час побудови цієї моделі використано традиційний підхід, властивий класичній теорії дифузії в конденсованих середовищах. Отримано теоретичну залежність коефіцієнта дифузії від об'єму, зайнятого каркасом. Показано, що вплив каркаса на дифузію сторонніх частинок у гідрогелі визначається переважно змінами вільної енергії активації, викликаними наявністю каркаса.

За допомогою методу аналізу оптичних зображень отримано залежності коефіцієнта дифузії в гідрогелях агарози та желатини з різною концентрацією гелеутворювача. Встановлено, що отримані експериментальні залежності узгоджуються з теоретичною залежністю, передбаченою запропонованою фізичною моделлю. Показано, що коли розмір часток дифузанта набагато менший від розміру пори гідрогелю, то коефіцієнт дифузії залежить не від природи гідрогелю, а від концентрації гелеутворювача.

Ключові слова: дифузія, гідрогелі, агароза, желатин, родамін, полімерна сітка.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

Наукове видання



ВІСНИК
КИЇВСЬКОГО НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ ІМЕНІ ТАРАСА ШЕВЧЕНКА
ФІЗИКО-МАТЕМАТИЧНІ НАУКИ

Випуск 1(82)

Оригінал-макет виготовлено Видавничо-поліграфічним центром "Київський університет"

Автори опублікованих матеріалів несуть повну відповідальність за підбір, точність наведених фактів, цитат, економіко-статистичних даних, відповідної галузевої термінології, власних імен та інших відомостей. Редколегія залишає за собою право скорочувати та редагувати подані матеріали.

Opinions, statements, accuracy of the quotations, economic and statistical data, terminology, proper names and other information are made on the responsibility of the authors. The Editorial Board reserves the right to shorten and edit the submitted materials.



Формат 60x84^{1/8}. Обл. вид. арк. 27,02. Ум. друк. арк. 34,0. Наклад 300. Зам. № 226-11795.
Гарнітура Arial. Папір офсетний. Друк офсетний. Вид. №К1.
Підписано до друку 05.06.26

Видавець і виготовлювач
ВПЦ "Київський університет"
Б-р Тараса Шевченка 14, м. Київ, 01030
☎ (38044) 239 32 22; (38044) 239 31 58; (38044) 239 31 28
e-mail: vpc@knu.ua
<http://vpc.knu.ua>

Свідоцтво суб'єкта видавничої справи ДК № 1103 від 31.10.02