

ISSN 2686-9667 (Print)
ISSN 2782-3342 (Online)

ВЕСТНИК РОССИЙСКИХ УНИВЕРСИТЕТОВ МАТЕМАТИКА

Научно-теоретический журнал

RUSSIAN UNIVERSITIES REPORTS MATHEMATICS

Scientific-theoretical journal



Том 29
№ 148
2024

ВЕСТНИК РОССИЙСКИХ УНИВЕРСИТЕТОВ МАТЕМАТИКА

Научно-теоретический
журнал

Том 29, № 148,
2024

Издается с 14 июня 1996 года
Выходит 4 раза в год

Журнал включен в «Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук» ВАК при Минобрнауки России по научным специальностям:
1.1.1. Вещественный, комплексный и функциональный анализ (физико-математические науки);
1.1.2. Дифференциальные уравнения и математическая физика (физико-математические науки)

Индексируется в базе данных Scopus, Russian Science Citation Index (RSCI) на платформе Web of Science, РИНЦ (входит в ядро РИНЦ), Math-Net.Ru, ВИНТИ РАН, Zentralblatt MATH, SciLIT, Ulrich's Periodicals Directory, НЭБ «eLIBRARY.RU», ЭБ «КиберЛенинка», Норвежский реестр научных журналов, серий и издателей первого уровня (NSD)

СОДЕРЖАНИЕ

CONTENTS	379
НАУЧНЫЕ СТАТЬИ	
<i>Е.О. Бурлаков, В.А. Ярема</i>	О задачах импульсного управления, возникающих при разработке автоматизированных систем контроля численности вредителей в тепличных комплексах 381
<i>Н.Д. Золотарева</i>	О пересчете эллипсоидов при оценке погрешности неявного метода Штермера для линейного дифференциального уравнения второго порядка 391
<i>А.Ф. Измаилов, Е.И. Усков</i>	Ускорение сходимости ньютоновских методов к особым решениям нелинейных уравнений 401
<i>С.В. Людковский</i>	Разложение модулей над обобщенными алгебрами Диксона 425
<i>В.Г. Пименов, А.В. Лекомцев</i>	Компактная схема для решения супердиффузионного уравнения с несколькими переменными запаздываниями 440
<i>В.И. Сумин, М.И. Сумин</i>	Регуляризация классических условий оптимальности в задачах оптимизации линейных распределенных систем вольтеррова типа с поточечными фазовыми ограничениями 455
<i>Р. Чакар, С. Дехилис, В. Мерчела, Х. Геббай</i>	Теорема о неподвижной точке $\rho - F$ -сжатия 485
<i>Д. Эттайб</i>	О λ -коммутативности, левом (правом) псевдоспектре и левом (правом) условном псевдоспектре линейных непрерывных операторов на ультраметрических банаховых пространствах 494

Учредитель: Федеральное государственное бюджетное образовательное учреждение высшего образования «Тамбовский государственный университет имени Г.Р. Державина»
(ОГРН 1026801156689) (392000, Тамбовская обл., г. Тамбов, ул. Интернациональная, д. 33)

ГЛАВНЫЙ РЕДАКТОР д. ф.-м. н., проф. Е.С. Жуковский (г. Тамбов, Российская Федерация)

РЕДАКЦИОННАЯ КОЛЛЕГИЯ ЖУРНАЛА: к. ф.-м. н., доц. Е.А. Панасенко (научный редактор) (г. Тамбов, Российская Федерация), И.В. Ильина (отв. секретарь) (г. Тамбов, Российская Федерация), д. ф.-м. н., проф. А.В. Аругюнов (г. Москва, Российская Федерация), д. ф.-м. н., доц. М.В. Балашов (г. Москва, Российская Федерация), д. ф.-м. н., проф. Л.М. Березанский (г. Беэр-Шева, Израиль), д. ф.-м. н., проф. А.Г. Кушнер (г. Москва, Российская Федерация), д. ф.-м. н., проф. Е.Б. Лансеев (г. Москва, Российская Федерация), д. ф.-м. н., проф. В.Ф. Молчанов (г. Тамбов, Российская Федерация), доктор, проф. М. Певзнер (г. Реймс, Франция), доктор, проф. А.В. Поносов (г. Ос, Норвегия), д. ф.-м. н., проф. В.И. Сумин (г. Нижний Новгород, Российская Федерация), д. ф.-м. н., проф. М.И. Сумин (г. Нижний Новгород, Российская Федерация), доктор, проф. Г. Хельминк (г. Амстердам, Нидерланды), член-корр. РАН, д. ф.-м. н., проф. А.Г. Ченцов (г. Екатеринбург, Российская Федерация)

Адрес редакции: 392000, Тамбовская обл., г. Тамбов, ул. Интернациональная, д. 33

Телефон редакции: +7(4752)72-34-34 доб. 0440

Электронная почта: zukovskys@mail.ru; ilina@tsutmb.ru

Сайт: <http://journals.tsutmb.ru/mathematics/>; <http://journals.tsutmb.ru/mathematics-en/>

Издание зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор), выписка из реестра зарегистрированных средств массовой информации (реестровая запись) от 03.07.2019 ПИ № ФС77-76133

ISSN 2686-9667 (Print) ISSN 2782-3342 (Online)

Редакторы: М.А. Сенина, М.И. Филатова
Редакторы английских текстов: И.С. Голыбина, М.А. Сенина
Технический редактор Ю.А. Бирюкова
Технический секретарь М.В. Борзова
Администратор сайта М.И. Филатова

Для цитирования:

Вестник российских университетов. Математика. 2024. Т. 29. № 148. 140 с. <https://doi.org/10.20310/2686-9667-2024-29-148>

Подписано в печать 27.11.2024. Дата выхода в свет

Формат А4 (60×84 1/8). Гарнитура «Times New Roman». Печать на ризографе.

Печ. л. 17,5. Усл. печ. л. 16,3. Тираж 1000 экз. Заказ № 24216. Свободная цена.

Издатель: ФГБОУ ВО «Тамбовский государственный университет имени Г.Р. Державина»

Адрес издателя: 392000, Тамбовская обл., г. Тамбов, ул. Интернациональная, д. 33

Отпечатано с готового оригинал-макета в отделе оперативной печати Издательского дома «Державинский»

ФГБОУ ВО «Тамбовский государственный университет имени Г.Р. Державина».

392008, Тамбовская обл., г. Тамбов, ул. Советская, д. 190г. Эл. почта: izdat_tsu09@mail.ru



Контент доступен под лицензией [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)

© ФГБОУ ВО «Тамбовский государственный университет имени Г.Р. Державина», 2024

© Журнал «Вестник российских университетов. Математика», 2024

При перепечатке, при цитировании материалов, в том числе в электронных СМИ, ссылка на журнал обязательна.

Ответственность за содержание публикаций несет автор

RUSSIAN UNIVERSITIES REPORTS MATHEMATICS

Scientific-theoretical
journal

**Volume 29, no. 148,
2024**

Published since June 14, 1996
Issued 4 times a year

The journal is on the “List of peer-reviewed scientific periodicals recommended by Higher Attestation Commission at Ministry of Science and Higher Education for publication of scientific results of dissertations for academic degree of candidate of science, doctor of science on physical and mathematical sciences in the scientific specialties:
1.1.1. Real, complex and functional analysis (physical and mathematical sciences);
1.1.2. Differential equations and mathematical physics (physical and mathematical sciences)”

Indexed in the Scopus database, Russian Science Citation Index (RSCI) on Web of Science platform, RSCI (included in the RSCI core), Math-Net.Ru, VINITI RAS, Zentralblatt MATH, SciLIT, Ulrich’s Periodicals Directory, Scientific Electronic Library “eLIBRARY.RU”, Electronic Library “CyberLeninka”, Norwegian Register of Scientific Journals, Series and Publishers Level 1 (NSD)

CONTENTS

SCIENTIFIC ARTICLES

<i>E.O. Burlakov, V.A. Yarema</i>	On impulse control problems arising in automatization of pests control in greenhouses	381
<i>N.D. Zolotareva</i>	On the recalculation of ellipsoids in estimating the error of the implicit Stormer method for a second-order linear differential equation	391
<i>A.F. Izmailov, E.I. Uskov</i>	Accelerating convergence of Newton-type methods to singular solutions of nonlinear equations	401
<i>S.V. Ludkovsky</i>	Decomposition of modules over generalized Dickson algebras	425
<i>V.G. Pimenov, A.V. Lekomtsev</i>	A compact scheme for solving a superdiffusion equation with several variable delays	440
<i>V.I. Sumin, M.I. Sumin</i>	Regularization of classical optimality conditions optimization problems of linear distributed Volterra-type systems with pointwise state constraints	455
<i>R. Chakar, S. Dehilis, W. Merchela, H. Guebbai</i>	$\rho - F$ -contraction fixed point theorem	485
<i>J. Ettayb</i>	On λ -commuting and left (right) pseudospectrum and left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces	494

From June 14, 1996 to May 27, 2019, the journal was published under the name
“Tambov University Reports. Series: Natural and Technical Sciences”. ISSN 1810-0198

Founder: Federal State Budgetary Educational Institution of Higher Education “Derzhavin Tambov State University”
(ОГРН 1026801156689) (33 Internatsionalnaya St., Tambov 392000, Tambov Region)

EDITOR-IN-CHIEF: Dr., Prof. Zhukovskiy, Evgeny S. (Tambov, Russian Federation)

EDITORIAL BOARD OF THE JOURNAL: Cand., Assoc. Prof. Panasenko, Elena A. (Scientific Editor) (Tambov, Russian Federation), Ilyina, Irina V. (Executive Editor) (Tambov, Russian Federation), Dr., Prof. Arutyunov, Aram V. (Moscow, Russian Federation), Dr., Assoc. Prof. Balashov, Maxim V. (Moscow, Russian Federation), Dr., Prof. Berezansky, Leonid (Beersheba, State of Israel), Dr., Prof. Kushner, Alexei G. (Moscow, Russian Federation), Dr., Prof. Laneev, Evgenii B. (Moscow, Russian Federation), Dr., Prof. Molchanov, Vladimir F. (Tambov, Russian Federation), Dr., Prof. Pevzner, Michael (Reims, French Republic), Dr., Prof. Ponosov, Arcady V. (Ås, Kingdom of Norway), Dr., Prof. Sumin, Vladimir I. (Nizhnii Novgorod, Russian Federation), Dr., Prof. Sumin, Mikhail I. (Nizhnii Novgorod, Russian Federation), Dr., Prof. Helminck, Gerard (Amsterdam, Netherlands), Corresponding Member of RAS, Dr., Prof. Chentsov, Alexander G. (Yekaterinburg, Russian Federation)

Editors address: 33 Internatsionalnaya St., Tambov 392000, Tambov Region
Telephone number: +7(4752)-72-34-34 extension 0440
E-mail: zukovskys@mail.ru; ilina@tsutmb.ru
Web-site: <http://journals.tsutmb.ru/mathematics/>; <http://journals.tsutmb.ru/mathematics-en/>

The publication is registered by the Federal Service for Supervision of Communications, Information Technology and Mass Media (Roskommnadzor), extract from the register of registered mass media (register entry dated) 03.07.2019 ПИИ no. ФС77-76133

ISSN 2686-9667 (Print) ISSN 2782-3342 (Online)

Editors: M.A. Senina, M.I. Filatova
English texts editors: I.S. Golybina, M.A. Senina
Technical editor Y.A. Biryukova
Technical secretary M.V. Borzova
Web-site administrator M.I. Filatova

For citation:

Vestnik Rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics, **29**:148 (2024), 140 p.
<https://doi.org/10.20310/2686-9667-2024-29-148>

Signed for printing 27.11.2024. Release date
Format A4 (60×84 1/8). Typeface “Times New Roman”. Printed on risograph.
Pr. sheet 17,5. Conv. pr. sheet 16,3. Copies printed 1000. Order no. 24216. Free price

Publisher: FSBEI of HE “Derzhavin Tambov State University”
Publisher’s address: 33 Internatsionalnaya St., Tambov 392000, Tambov Region

Published basing on ready-to-print file in Instant Print Department of Publishing House “Derzhavinskiy”
of FSBEI of HE “Derzhavin Tambov State University”.
190g Sovetskaya St., Tambov 392008, Tambov Region. E-mail: izdat_tsu09@mail.ru



This work is licensed under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)

© FSBEI of HE “Derzhavin Tambov State University”, 2024
© The journal “Russian Universities Reports. Mathematics”, 2024
While reprinting, citing materials, including in electronic media,
a reference to the journal is required.
The author is responsible for the contents of publications

НАУЧНАЯ СТАТЬЯ

© Бурлаков Е.О., Ярема В.А., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-381-390>

УДК 51-76, 519.62



О задачах импульсного управления, возникающих при разработке автоматизированных систем контроля численности вредителей в тепличных комплексах

Евгений Олегович БУРЛАКОВ^{1,2,3}, Виталий Алексеевич ЯРЕМА³

¹ ФГБУН «Институт проблем управления им. В. А. Трапезникова РАН»
117997, Российская Федерация, г. Москва, ул. Профсоюзная, 65

² ФГБОУ ВО «Тамбовский государственный университет им. Г.Р. Державина»
392000, Российская Федерация, г. Тамбов, ул. Интернациональная, 33

³ ФГАОУ ВО «Тюменский государственный университет»
625003, Российская Федерация, г. Тюмень, ул. Володарского, 6

Аннотация. В работе предлагается математическая модель на основе непрерывной динамической системы, формализующая взаимодействие популяций насекомого-вредителя тепличной агрокультуры и энтомофага. В рамках модели энтомофаг, согласно концепции комплексной борьбы с вредителями, выступает агентом экологического контроля популяции насекомого-вредителя. Проводится идентификация параметров модели на основе экспериментальных данных, полученных в лабораторных условиях для тепличной системы «огурец обыкновенный — паутинный клещ — клещ-акарифаг». Для предложенной модели формулируется задача импульсного управления, отвечающая реализации контроля популяции вредителя при помощи энтомофага, исследуются вопросы существования решений и их непрерывной зависимости от управления. На основе исследованной управляемой системы обсуждаются вопросы построения экономически эффективных управляющих стратегий, позволяющих реализовывать востребованное при проектировании «умных» теплиц математическое обеспечение программно-аппаратных средств автоматизированных систем контроля численности вредителей.

Ключевые слова: математическое моделирование экологических систем, идентификация параметров математических моделей, импульсное управление, корректность

Благодарности: Исследование выполнено за счет гранта Российского научного фонда (проект № 20-11-20131, <https://rscf.ru/project/20-11-20131/>).

Для цитирования: Бурлаков Е.О., Ярема В.А. О задачах импульсного управления, возникающих при разработке автоматизированных систем контроля численности вредителей в тепличных комплексах // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 381–390. <https://doi.org/10.20310/2686-9667-2024-29-148-381-390>

SCIENTIFIC ARTICLE

© E. O. Burlakov, V. A. Yarema, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-381-390>

On impulse control problems arising in automatization of pests control in greenhouses

Evgenii O. BURLAKOV^{1,2,3}, Vitaly A. YAREMA³¹ V. A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
65 Profsoyuznaya St., Moscow 117997, Russian Federation² Derzhavin Tambov State University
33 Internatsionalnaya St., Tambov 392000, Russian Federation³ Tyumen State University
6 Volodarskogo St., Tyumen 625003, Russian Federation

Abstract. The paper proposes a mathematical model based on a continuous dynamic system that formalizes the interaction of populations of a greenhouse agricultural insect pest and an entomophage acting as an agent of environmental control of the pest population in the framework of the concept of integrated pest management paradigm. The model parameters are identified based on experimental data obtained in laboratory conditions for the greenhouse system “cucumber — spider mite — acariphage mite”. For the proposed system, an impulse control problem that corresponds to the implementation of the pest population control using an entomophage is formulated, and the issues of the existence of solutions and their continuous dependence on control are investigated. Based on the studied control system, the issues of constructing cost-effective control strategies that allow implementing the mathematical software for automated pest control systems in greenhouses are discussed.

Keywords: mathematical models in ecology, parameters identification, impulse control problems, well-posedness

Acknowledgements: The research was supported by the Russian Science Foundation (project no. 20-11-20131, <https://rscf.ru/project/20-11-20131/>).

Mathematics Subject Classification: 37N25, 34H05.

For citation: Burlakov E.O., Yarema V.A. On impulse control problems arising in automatization of pests control in greenhouses. *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 381–390.

<https://doi.org/10.20310/2686-9667-2024-29-148-381-390> (In Russian, Abstr. in Engl.)

Введение

Современные инновационные решения в сельском хозяйстве направлены на то, чтобы потребители получали качественную агропродукцию, выращенную с минимальным применением химических средств. Одним из подходов к решению этой задачи является использование концепции комплексной борьбы с вредителями (Integrated Pest Management). Данная концепция подразумевает защиту агрокультур, основанную на применении комплекса мероприятий, учитывающих циклы совместного развития представленных в теплице растений, насекомых и микроорганизмов, а также их взаимодействия с питательными веществами [1, 2]. Современные подходы, направленные на снижение зависимости агрокультур от пестицидов, требуют от тепличных экосистем способности самостоятельно восстанавливаться [3]. Для создания подобных устойчивых экосистем в институте X-BIO Тюменского государственного университета ведется разработка программно-аппаратного комплекса для выработки и реализации мер профилактики и борьбы с наиболее распространенными вредителями сельскохозяйственно значимых растений, выращиваемых в закрытом грунте. Одним из модулей данного комплекса должен стать программный комплекс автоматизации экологического контроля вредителей тепличных агрокультур, использующий биопрепараты на основе энтомофагов. В основу математического обеспечения программного комплекса будут положены дифференциальные системы типа «хищник-жертва», содержащие управляющие параметры. В настоящей работе рассматривается подобная дифференциальная модель, содержащая импульсное управление, формализующее серию управляющих внесений биопрепарата-энтомофага в тепличную систему.

В первом разделе данной работы решается задача восстановления параметров не поддерживающей управления модели «хищник-жертва» на основе экспериментальных данных, полученных для тепличной системы «огурец обыкновенный — паутинный клещ — клещ-акарифаг». Второй раздел работы посвящен исследованию корректности задачи импульсного управления для изученной в первом разделе системы, а также вопросу построения экономически эффективных управляющих стратегий на основе рассматриваемой модели.

1. Математическая модель

Рассмотрим систему

$$\begin{aligned} \frac{d}{dt}x(t) &= a\left(1 - \frac{x(t)}{K}\right)x(t) - f(x(t))y(t), \\ \frac{d}{dt}y(t) &= cf(x(t), y(t))y(t) - my(t), \end{aligned} \quad (1.1)$$

где $t \in [0, T]$ ($T > 0$), $x(t)$ и $y(t)$ — значения плотности популяции вредителя и энтомофага, соответственно (в таблице 1 приведена интерпретация параметров модели). Функция f — трофическая функция Холлинга второго типа [4], имеющая вид $f(x) = \frac{bx}{d+x}$.

Восстановим значения параметров модели (1.1), опираясь на данные полевых экспериментов, проведенных сотрудниками Института исследования растений в Департаменте сельского хозяйства округа Виктория, Австралия (экспериментальные данные, а также подробное описание постановки экспериментов доступны в работе [5]), где роли тепличной агрокультуры, вредителя и энтомофага играли обыкновенный огурец *Cucumis sativus*, паутинный клещ *Tetranychus urticae* и клещ-акарифаг *Tetranychus persimilis*, соответственно.

Таблица 1

Описание параметров модели (1.1)

Параметр	Описание
$a > 0$	Скорость прироста популяции вредителя
$K > 0$	Плотность популяции вредителя, приводящая к полному поражению им агрокультуры
$b > 0$	Максимальная скорость истребления вредителя энтомофагом
$0 < c < 1$	Эффективность конверсии питания энтомофага вредителем в прирост популяции энтомофага
$0 < b/d < 1$	Эффективность поиска вредителем энтомофага
$m > 0$	Скорость естественной гибели энтомофага

Запишем систему (1.1) в сокращенном виде

$$\dot{\bar{X}}(t) = G(\bar{X}(t), \bar{P}), \quad (1.2)$$

где $\bar{X}(t) = (x(t), y(t))^T = (x(t, \bar{P}), y(t, \bar{P}))^T$, $G : \mathbb{R}^2 \times M \rightarrow \mathbb{R}^2$ — функция, соответствующая правой части (1.1), $M \subset \mathbb{R}^5$ — множество допустимых значений параметров. Для решения задачи идентификации параметров системы (1.1) используем метод максимального правдоподобия. Введем функционал $\mathcal{F}$, определяющий отклонение решения модели от экспериментальных данных:

$$\mathcal{F}(\bar{X}) = \sum_{t_i \in \mathbb{I}} \left| \bar{X}(t_i, \bar{P}) - \tilde{X}_i \right|.$$

Здесь t_i — i -ая неделя эксперимента, $\tilde{X}_i$ — вектор экспериментальных данных по численностям популяций на неделю t_i , $\bar{P}$ — вектор параметров модели (см. таблицу 1), $\bar{X}(t_i, \bar{P})$ — значение $\bar{X}$ в момент времени t_i при значениях параметров $\bar{P}$, полученное в результате решения системы (1.1), символом $|\cdot|$ обозначено евклидово расстояние в $\mathbb{R}^2$. Для идентификации параметров модели необходимо минимизировать функционал $\mathcal{F}$ относительно переменной $\bar{P}$, предварительно оценив области определения ее компонент (границы множества M).

Исходя из биологического смысла параметра K , примем его значение $K = 1$ (100%). Отметим, что действительная плотность популяции вредителя может достигать значений, превышающих 1, однако в настоящем исследовании нас интересует режим работы системы (1.1), при котором $x(t) < K$ при всех $t \in [0, T]$, что соответствует сохранению тепличной культуры на рассматриваемом интервале времени (вышеупомянутые экспериментальные данные также согласуются с условием $x(t) < K$).

Оценим скорость a прироста популяции вредителя исходя из данных, собранных на экспериментальных площадках до момента введения в эксперимент клеща-акарифага (см. таблицу 2). Проинтегрируем первое уравнение системы уравнение (1.1) в отсутствие энтомофага (т. е. при $y \equiv 0$). Получим

$$a = \ln \left(\frac{x(0)(K - x(1))}{x(1)(K - x(0))} \right).$$

Будем использовать в наших вычислениях среднее значение $a = 0.6159$ скорости прироста популяции вредителя.

Таблица 2

Экспериментальные данные по приросту численности паутинного клеща
в отсутствие клеща-акарифага

Номер эксперимента	$x(0)$	$x(1)$	a
I	0.65	0.76	0.5336
II	0.36	0.56	0.8165
III	0.94	0.96	0.4265
IV	0.64	0.70	0.6897

Для биологически релевантной оценки границ скорости b максимального потребления добычи энтомофагом, будем опираться на данные по потреблению клещом *P. persimilis* яиц (см. [6–9]), личинок (см. [8, 10]) и взрослых особей (см. [8]) паутинного клеща, а также на сведения о типичной пропорции представленности вышеуказанных стадий развития паутинного клеща в популяции (см. [8]). Получим оценку $b \in [3.49, 4.92]$.

Согласно [6], эффективность конверсии питания *P. persimilis* в прирост популяции имеет приблизительное значение $c \approx 0.7$. В наших расчетах выберем диапазон $c \in [0.6, 0.8]$.

Оценку значений параметра d в терминах биологически интерпретируемого параметра b/d , характеризующего в рассматриваемой тепличной экосистеме эффективность поиска добычи хищником, произвести довольно затруднительно ввиду существенной зависимости этого параметра от ряда условий эксперимента (например, от температуры в тепличной системе), не представленных в формализации (1.1). В связи с этим, при восстановлении параметра d будем использовать стандартную для *P. persimilis* оценку $b/d \in [0.2, 0.65]$ (см. [5]), откуда получим $d \in (1.53b, 5b)$.

Коэффициент смертности клеща-акарифага m полагается нами варьирующимся от 1.4 до 2.8 на основе сведений, содержащихся в работах [11, 12].

Для решения задачи поиска компонент вектора $\bar{P}$, соответствующих минимуму функционала $\mathcal{F}$, используем метод Нелдера–Мида [13]. Интегральные кривые системы (1.1) при найденном минимизирующем функционал значении $\bar{P}_0$ (см. таблицу 3) изображены на рисунке 1. Система (1.1) обладает единственным (неустойчивым) положением равновесия E , имеющим координаты (32.67, 4.99) на фазовой плоскости и порождающим предельный цикл.

Таблица 3

Восстановленные по экспериментальным данным значения параметров системы (1.1)

a	b	c	d	m
0.6159	4.31	0.68	20.52	1.8

2. Задача импульсного управления

Продемонстрируем применимость модели (1.1) для автоматизации экологического контроля вредителей тепличных агрокультур. Для всякого начального условия $(x(0), y(0)) = (x_0, y_0) \in \mathbb{R}^2$, уравнение (1.2) можно представить в виде

$$\bar{X}(t) = \bar{X}_0 + \int_a^t G_0(\bar{X}(s)) ds, \quad t \in [0, T], \quad (2.1)$$

где $\bar{X}(0) = (x_0, y_0)^T$, $G_0(\bar{X}) = G(\bar{X}, \bar{P}_0)$.

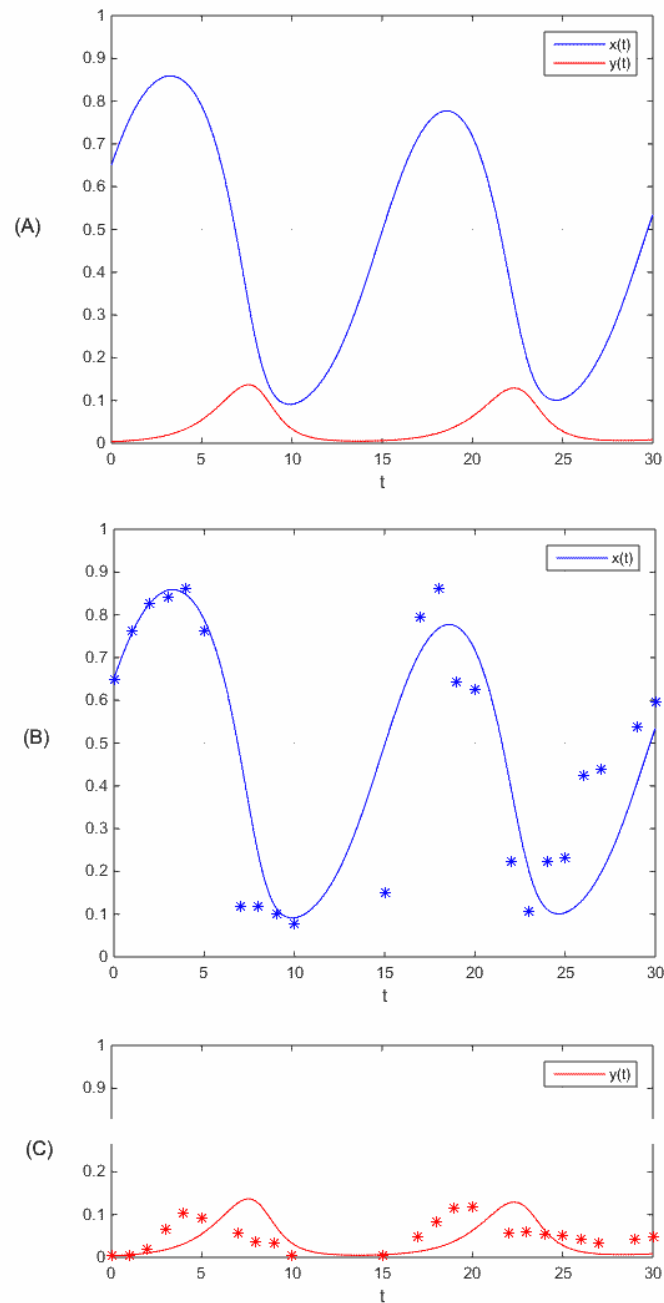


Рис. 1. Динамика компонент системы (1.1) при идентифицированных значениях параметров (А). Сравнение восстановленной динамики компонент $x(t)$ (В) и $y(t)$ (С) системы (1.1) с соответствующими экспериментальными данными (отмечены символами " * ")

Для системы (2.1) возможность контроля численности вредителей, основанная на внесении в тепличную систему биопрепаратов на основе энтомофагов, может быть формализована следующим образом:

$$\overline{X}(t) = \overline{X}_0 + \int_a^t G_0(\overline{X}(s))ds + \overline{U}(t), \quad (2.2)$$

где $\overline{U}$ — импульсное управление вида $(0, \sum_{\forall k} \chi_{[t_k, T]}(t)y_k)^T$; χ_A — характеристическая функция множества $A \subset [0, T]$, $\{t_k\} \subset [0, T]$ и $y_k > 0$ — моменты времени и величины импульсных воздействий, соответственно, причем $0 \leq k \leq \hat{K} \in \mathbb{N}$. Таким образом,

различные управления $\bar{U}$ могут состоять из разного количества импульсных воздействий, общее количество которых для каждого управления не превосходит $\hat{K}$.

Определим множество $\mathcal{Y}$ функций вида $\bar{Y} = \bar{X} + \bar{U}$, где $\bar{X}$ — функция из класса непрерывных двумерных вектор-функций. Множество $\mathcal{Y}$ образует полное метрическое пространство относительно метрики

$$\rho_{\mathcal{Y}}(\bar{Y}_1, \bar{Y}_2) = \|\bar{X}_1 - \bar{X}_2\|_C + \int_0^T \left| \sum_{k: t_k^1 \in [a, b]} \chi_{[t_k^1, b]}(t) y_k^1 - \sum_{k: t_k^2 \in [a, b]} \chi_{[t_k^2, b]}(t) y_k^2 \right| dt, \quad (2.3)$$

где символом $\|\cdot\|_C$ обозначена норма в соответствующем пространстве непрерывных функций (см. [14]). Подобная метрика была использована в работе [15] при исследовании отклика нейронных систем на зрительные стимулы.

Решением задачи управления (2.2) будем считать функцию $\bar{Y} \in \mathcal{Y}$, удовлетворяющую уравнению (2.2).

Утверждение 2.1. *При любом управлении $\bar{U}$ задача (2.2) имеет единственное решение. Кроме того, если некоторая последовательность управлений $\bar{U}^i$ сходится к управлению $\bar{U}^0$ в смысле сходимости $\int_0^T \left| \sum_{k: t_k^i \in [0, T]} \chi_{[t_k^i, T]}(t) V_k^i dt - \sum_{k: t_k^0 \in [0, T]} \chi_{[t_k^0, T]}(t) V_k^0 \right| dt \rightarrow 0$ при $i \rightarrow \infty$, то последовательность решений $\bar{Y}^i$ системы (2.2), соответствующих управлениям $\bar{U}^i$, сходится к решению $\bar{Y}^0$, соответствующему управлению $\bar{U}^0$, в метрике (2.3).*

Справедливость утверждения 2.1 следует из основного результата работы [14].

Наибольший практический интерес представляет использование результатов данного исследования для решения задач экономической оптимизации экологического контроля вредителей, значимость которых неоднократно подчеркивалась в профильной литературе по направлению комплексной борьбы с вредителями (см., например, работы [16, 17], а также обзор [18] и библиографию). Утверждение 2.1 обосновывает корректность задачи импульсного управления для рассматриваемой системы «хищник-жертва», открывая возможности отыскания наиболее эффективных с экономической точки зрения управляющих стратегий, в том числе, с использованием численных методов. Подобные задачи оптимизации могут быть формализованы, например, следующим образом:

$$\begin{aligned} \bar{X}(t) &= \bar{X}_0 + \int_a^t G(\bar{X}(s)) ds + \sum_{\forall k} \chi_{[t_k, T]}(t) y_k, \quad t \in [0, T], \\ \|(\bar{X}, (1, 0)^T)\|_C &< \Theta, \\ \sum_{\forall k} y_k &\rightarrow \min, \end{aligned}$$

где $\Theta > 0$ — критический уровень поражения тепличной агрокультуры вредителем, символом $(\cdot, \cdot)$ обозначено скалярное произведение векторов.

Настоящая работа не преследует цели отыскания оптимальных управляющих стратегий, однако мы приведем эвристическую схему построения стратегии импульсного управления системой «хищник-жертва», исследованной в предыдущем параграфе. Пусть $\Theta = 0.5$, $\bar{X}_0 = (0.01, 0)$. Фазовый портрет данной системы изображен на рисунке 2 (А). С учетом имеющегося знания о координатах положения равновесия $E = (32.67, 4.99)$ рассматриваемой системы, применим операцию внесения энтомофагов в тепличную систему в количестве $y_1 = 4.99$ в такой момент времени t_1 , что $x(t_1) \approx 32.67$. В силу того, что численность

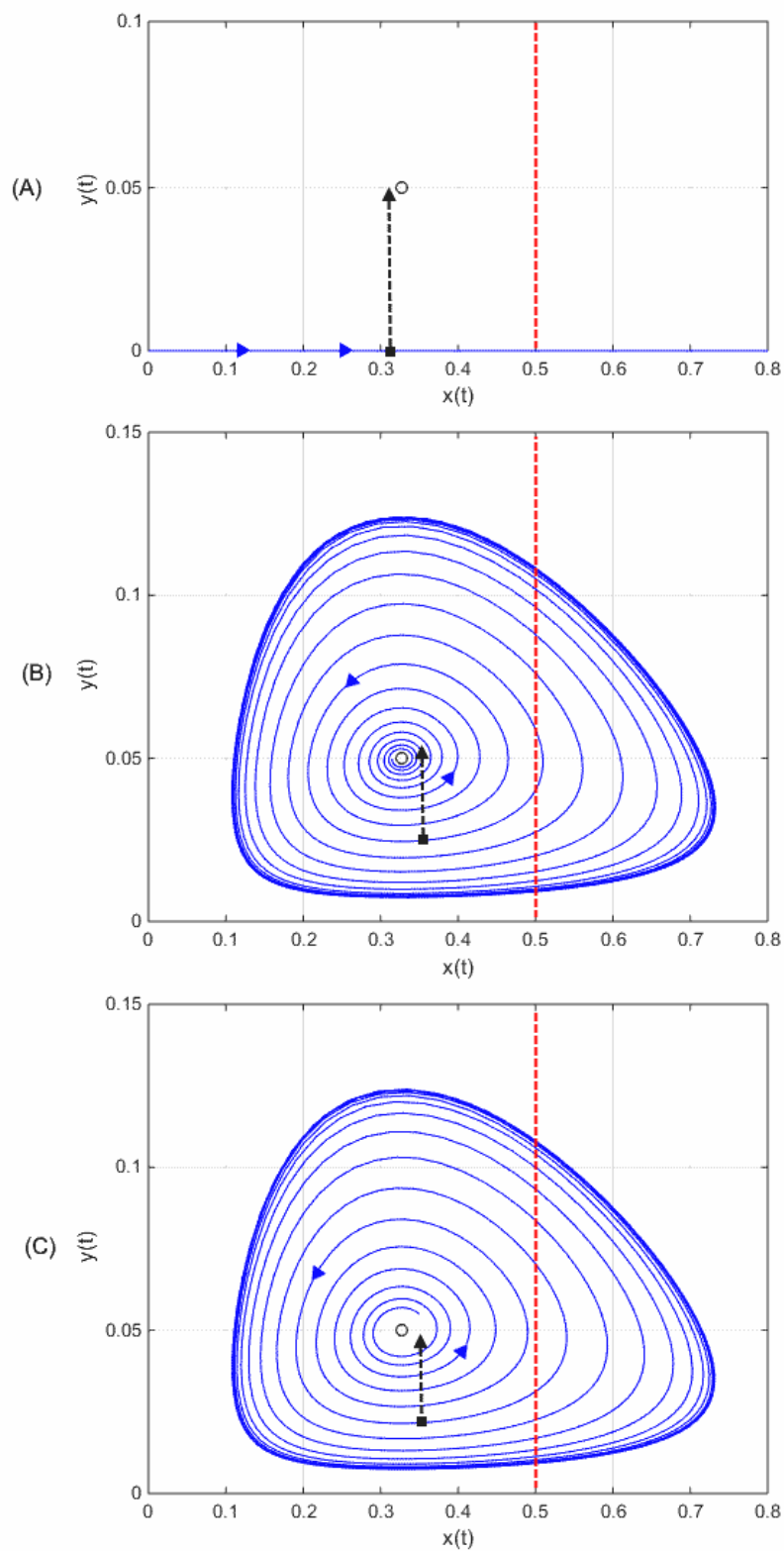


Рис. 2. Шаги построения допустимого управления в рамках системы (2.2). Синие стрелки показывают направление движения по кривым фазового состояния. Черной окружностью изображено положение равновесия E системы (1.1), красной пунктирной линией изображено пороговое значение Θ поражения агрокультуры, черные пунктирные векторы отображают импульсные воздействия на экосистему теплицы. Рисунки (А), (В) и (С) отвечают первому, второму и третьему импульсному воздействию, соответственно

популяций в тепличной экосистеме может быть оценена при помощи системы (1.1) лишь приближенно, будем полагать, что момент реализации импульсного воздействия определяется квадратом на конце черного пунктирного вектора (при этом длина вектора равна величине $y_1 = 4.99$, см. рисунок 2 (А)). Фазовое состояние системы, таким образом (с учетом погрешностей), переместилось в окрестность положения равновесия E (положение равновесия изображено черной окружностью, см. рисунок 2 (В)). Следующим моментом применения управления предлагается наибольшее значение t_2 , $x(t_2) \approx 32.67$, предшествующее возникновению общих точек кривой фазового состояния и линии порогового значения поражения агрокультуры (отображена красным пунктиром). С учетом неизбежных погрешностей, будем полагать, что момент импульсного воздействия определяется квадратом на конце черного пунктирного вектора (при этом длина вектора равна величине $y_2 = 4.99 - y(t_2)$, см. рисунок 2 (В)). Фазовое состояние системы переместилось в окрестность положения равновесия E (изображено черной окружностью, см. рисунок 2 (С)). Следующим моментом применения управления предлагается наибольшее значение t_3 , $x(t_3) \approx 32.67$ (см. рисунок 2 (С)), предшествующее возникновению общих точек кривой фазового состояния и линии порогового значения поражения агрокультуры. Модуль y_3 применяемого импульсного воздействия при этом равен $4.99 - y(t_2)$. И так далее ... Итоговая стоимость поддержания популяции вредителя в допороговом состоянии будет пропорциональна сумме длин черных пунктирных векторов. Естественным образом, ключевую роль в снижении стоимости реализации подобных управляющих стратегий играет высокая точность оценки численности популяций вредителя и энтомофага, позволяющая, путем внесения необходимых количеств особей энтомофага в тепличный комплекс, переводить состояние системы как можно ближе к ее положению равновесия.

References

- [1] S. K. Dara, "The new integrated pest management paradigm for the modern age", *Journal of Integrated Pest Management*, **10**:1 (2019), 1–9.
- [2] S. E. Naranjo, P. C. Ellsworth, G. B. Frisvold, "Economic value of biological control in integrated pest management of managed plant systems", *Annual Review of Entomology*, **60** (2015), 621–645.
- [3] S. Mouden, K. F. Sarmiento, P. G. L. Klinkhamer, K. A. Leiss, "Integrated pest management in western flower thrips: past, present and future", *Pest Management Science*, **73**:5 (2017), 813–822.
- [4] C. S. Holling, "Some characteristics of simple types of predation and parasitism", *Canadian Entomologist*, **91**:7 (1959), 385–398.
- [5] I. Kozlova, *A Numerical Study of Modelling in Mathematical Biology*, Swinburne University of Technology, Melbourne, Australia, 2002.
- [6] M. W. Sabelis, *Agricultural Research Report 910*, Pudoc, Wageningen, The Netherlands, 1981.
- [7] K. Fujita, T. Inoue, A. Takafuji, "Systems analysis of an acarine predator-prey system. I", *Researches on Population Ecology*, **21** (1979), 105–119.
- [8] D. Elliott, *Biological Technical Manual*, Applied Bio-nomics Ltd., Melbourne, Australia, 1993.
- [9] A. Takafuji, D. A. Chant, "Comparative studies of two species of predacious phytoseiid mites (Acarina: Phytoseiidae), with special reference to their responses to the density of their prey.", *Researches on Population Ecology*, **17** (1976), 255–310.
- [10] M. H. Malais, W. J. Ravensberg, *Knowing and Recognizing: The Biology of Glasshouse Pests and Their Natural Enemies*, Reed Business Information, Amsterdam, 2003.
- [11] G. Nachman, "Systems analysis of acarine predator-prey interactions. I. A stochastic simulation model of spatial processes", *Journal of Animal Ecology*, **56**:1 (1987), 247–265.
- [12] J. S. Bancroft, D. C. Margolie, "An individual-based model of an acarine tritrophic system: lima bean, *Phaseolus lunatus* L., twospotted spider mite, *Tetranychus urticae* (Acari: Tetranychidae), and *Phytoseiulus persimilis* (Acari: Phytoseiidae)", *Ecological Modelling*, **123**:2-3 (1999), 161–181.

- [13] J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright, “Convergence properties of the Nelder–Mead simplex method in low dimensions”, *SIAM Journal on Optimization*, **9**:1 (1998), 112–147.
- [14] Е. О. Бурлаков, Е. С. Жукковский, “On well-posedness of generalized neural field equations with impulsive control”, *Russian Mathematics (Iz VUZ)*, **60** (2016), 66–69.
- [15] Е. О. Бурлаков, В. М. Верхлютов, И. Н. Мальков, “О корректности математической модели вызванной активности первичной зрительной коры”, *Вестник российских университетов. Математика*, **29**:145 (2024), 43–50. [Е. О. Burlakov, V. M. Verkhlyutov, I. N. Malkov, “On well-posedness of a mathematical model of evoked activity in the primary visual cortex”, *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:145 (2024), 43–50 (In Russian)].
- [16] P. Tiftikci, S. K  k, I. Kasap, “Biological control of twospotted spider mites [*Tetranychus urticae* Koch (Acari: Tetranychidae)] using *Phytoseiulus persimilis* Athias-Henriot (Acari: Phytoseiidae) at different ratios of release on field-grown tomatos”, *Biological control*, **151** (2020), Article number: 104404.
- [17] G. P. Opit, J. R. Nechols, D. C. Margolies, “Biological control of twospotted spider mites, *Tetranychus urticae* Koch (Acari: Tetranychidae), using *Phytoseiulus persimilis* Athias-Henriot (Acari: Phytoseiidae) on ivy geranium: assessment of predator release ratios”, *Biological Control*, **29**:3 (2004), 445–452.
- [18] M. Jakubowska, R. Dobosz, D. Zawada, J. Kowalska, “A Review of Crop protection methods against the twospotted spider Mite-*Tetranychus urticae* Koch (Acari: Tetranychidae) – with special reference to alternative methods”, *Agriculture*, **12** (2022), Article number: 898.

Информация об авторах

Бурлаков Евгений Олегович, кандидат физико-математических наук, научный сотрудник Института проблем управления им. В. А. Трапезникова РАН, г. Москва; научный сотрудник НОЦ «Фундаментальные математические исследования», Тамбовский государственный университет им. Г. Р. Державина, г. Тамбов; научный сотрудник института X-BIO, Тюменский государственный университет, г. Тюмень, Российская Федерация. E-mail: eb_@bk.ru

ORCID: <https://orcid.org/0000-0002-7286-9456>

Ярема Виталий Алексеевич, аспирант, институт X-BIO, Тюменский государственный университет, г. Тюмень, Российская Федерация. E-mail: i.n.malkov@yandex.ru

ORCID: <http://orcid.org/0000-0001-5845-5591>

Конфликт интересов отсутствует.

Для контактов:

Бурлаков Евгений Олегович
E-mail: eb_@bk.ru

Поступила в редакцию 29.08.2024 г.

Поступила после рецензирования 01.11.2024 г.

Принята к публикации 22.11.2024 г.

Information about the authors

Evgenii O. Burlakov, PhD, Researcher at the V. A. Trapeznikov Institute of Control Sciences of RAS, Moscow; Researcher at the Scientific and Educational Center “Fundamental Mathematical Research”, Derzhavin Tambov State University, Tambov; Researcher at the Institute X-BIO, Tyumen State University, Tyumen, Russian Federation. E-mail: eb_@bk.ru

ORCID: <https://orcid.org/0000-0002-7286-9456>

Vitaly A. Yarema, Post-Graduate Student, Institute X-BIO, Tyumen State University, Tyumen, Russian Federation. E-mail: yaremavertical@mail.ru

ORCID: <http://orcid.org/0009-0001-7784-6720>

There is no conflict of interests.

Corresponding author:

Evgenii O. Burlakov
E-mail: eb_@bk.ru

Received 29.08.2024

Reviewed 01.11.2024

Accepted for press 22.11.2024

НАУЧНАЯ СТАТЬЯ

© Золотарева Н.Д., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-391-400>

УДК 519.622



О пересчете эллипсоидов при оценке погрешности неявного метода Штермера для линейного дифференциального уравнения второго порядка

Наталья Дмитриевна ЗОЛОТАРЕВА

ФГБОУ ВО «Московский государственный университет им. М. В. Ломоносова»

119991, ГСП-2, Российская Федерация, г. Москва, Ленинские горы, 1

Аннотация. В настоящей работе предложен новый способ построения оценки погрешности численного решения задачи Коши дифференциального уравнения второго порядка, полученного с помощью неявного метода Штермера. В отличие от ранее предложенных способов, он позволяет учитывать знаки малых слагаемых при пересчете эллипсоидов, содержащих точное решение, в случае неявного многошагового численного метода. Это приводит к более точной оценке погрешности численного решения и применимости метода эллипсоидов на больших интервалах. Приведен численный эксперимент, демонстрирующий эффективность предложенного метода получения гарантированной оценки погрешности неявного метода Штермера.

Ключевые слова: метод эллипсоидов, оценка погрешности, неявный метод Штермера, численное решение задачи Коши для ОДУ второго порядка

Для цитирования: Золотарева Н.Д. О пересчете эллипсоидов при оценке погрешности неявного метода Штермера для линейного дифференциального уравнения второго порядка // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 391–400. <https://doi.org/10.20310/2686-9667-2024-29-148-391-400>

SCIENTIFIC ARTICLE

© N. D. Zolotareva, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-391-400>

On the recalculation of ellipsoids in estimating the error of the implicit Stormer method for a second-order linear differential equation

Natalia D. ZOLOTAREVA

Lomonosov Moscow State University

1 Leninskiye Gory, Moscow 119991, Russian Federation

Abstract. In the paper, a new method is proposed for constructing an error estimate for the numerical solution of the Cauchy problem for a second-order differential equation obtained using the implicit Stormer method. Unlike the previously proposed methods, it allows one to take into account the signs of small terms when recalculating ellipsoids containing the exact solution in the case of implicit multistep numerical method. This leads to a more accurate estimation of the error of the numerical solution and the applicability of the ellipsoid method over large intervals. A numerical experiment is presented demonstrating the effectiveness of the proposed method for obtaining a guaranteed error estimate of the implicit Stormer method.

Keywords: ellipsoid method, error estimation, the Stormer method, numerical solution of the Cauchy problem for second-order ODEs

Mathematics Subject Classification: 65L70, 65L05.

For citation: Zolotareva N.D. On the recalculation of ellipsoids in estimating the error of the implicit Stormer method for a second-order linear differential equation. *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 391–400. <https://doi.org/10.20310/2686-9667-2024-29-148-391-400> (In Russian, Abstr. in Engl.)

Введение

Оценка погрешности приближенного решения дифференциального уравнения при по координатном оценивании имеет экспоненциальный рост, даже если сама погрешность растет не так быстро. Такой эффект называется эффектом раскрутки или эффектом Мура [1, с. 155].

Один из способов построения оценки погрешности, позволяющий избежать эффекта раскрутки, был предложен в [2]. Численное решение заключалось в эллипсоид, который содержал также и точное решение. С использованием разностного уравнения для погрешности этот эллипсоид пересчитывался на каждом шаге.

Для явного метода Штермера в [3], благодаря более точному оцениванию малых слагаемых, был предложен способ получения более точной оценки погрешности. В статье [4] этим способом был обобщен метод Нумерова (неявный двухшаговый метод Штермера).

В данной работе предлагается распространить результат работы [4] на неявный многошаговый метод любого порядка.

1. Постановка задачи

Для вычисления приближенного решения задачи Коши

$$y'' = f(x, y), \quad y(0) = y_0, \quad y'(0) = y'_0, \quad 0 \leq x \leq X_0, \quad y, f \in \mathbb{R} \quad (1.1)$$

будем использовать неявную k -шаговую формулу метода Штермера (см. [5, с. 143])

$$\nabla^2 y_m = h^2 f(x_{m-1}, y_{m-1}) + h^2 \sum_{i=2}^k \beta_i \nabla^i f(x_m, y_m) + w_m, \quad m \geq k.$$

Здесь y_m является приближенным значением для $y(x_m)$ в точке $x_m = hm$; h — постоянный шаг; β_i — коэффициенты метода Штермера; ∇ — конечная разность (назад), т. е. $\nabla a_m = a_m - a_{m-1}$; ∇^i — конечная разность i -го порядка, $\nabla^i a_m = \nabla(\nabla^{i-1} a_m)$, $i = 2, 3, \dots$, в частности конечная разность второго порядка ∇^2 определяется формулой $\nabla^2 a_m = \nabla(\nabla a_m) = a_m - 2a_{m-1} + a_{m-2}$; w_m — ошибка, получаемая из-за округлений и обрыва итераций.

Погрешность $z_m = y(x_m) - y_m$ численного решения удовлетворяет разностному уравнению

$$\begin{aligned} \nabla^2 z_m &= h^2 (f(x_{m-1}, y(x_{m-1})) - f(x_{m-1}, y_{m-1})) \\ &+ h^2 \sum_{i=2}^k \beta_i \nabla^i (f(x_m, y(x_m)) - f(x_m, y_m)) + N_m - w_m, \end{aligned} \quad (1.2)$$

где N_m — ошибка k -го метода Штермера на m -м шаге:

$$N_m = -\beta_{k+1} h^{k+3} y^{(k+3)}(\xi), \quad x_{m-k} \leq \xi \leq x_m.$$

Целью работы является получение гарантированной оценки погрешности z_m .

2. Оценивание погрешности с помощью эллипсоидов

В статье [2] был предложен способ построения оценки погрешности, не приводящий к эффекту раскрутки. Численное решение заключалось в эллипсоид, содержащий точное решение, который пересчитывался на каждом шаге.

Способ получения более точной оценки погрешности для явного метода Штермера был предложен в [3], для неявного двухшагового численного метода (метода Нумерова) — в [4].

В данной работе полученные ранее результаты будут обобщены на неявный многошаговый метод любого порядка.

В случае, когда правая часть дифференциального уравнения (1.1) является линейной по y , т. е. $f(x, y) = A(x)y$, уравнение (1.2) можно записать в виде разностного уравнения

$$\nabla^2 z_m = h^2 A_{m-1} z_{m-1} + h^2 \sum_{i=2}^k \beta_i \nabla^i (A_m z_m) + N_m - w_m, \quad (2.1)$$

где $A_m = A(x_m)$. Пусть при всех рассматриваемых значениях m справедливы оценки $N \geq |N_m|$ и $w \geq |w_m|$.

Так же как и в [2] введем вспомогательную переменную v_m и представим (2.1) в виде системы

$$\begin{cases} v_m = v_{m-1} + h A_{m-1} z_{m-1} + h^{-1} Q_m, & m \geq k, \\ z_m = z_{m-1} + h v_m + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m), & m \geq k-1, \end{cases} \quad (2.2)$$

где $Q_m = N_m - w_m$, с начальными значениями

$$v_0 = v_1 = \dots = v_{k-2} = 0, \quad v_{k-1} = \frac{z_{k-1} - z_{k-2}}{h} - h \sum_{i=2}^k \beta_i \nabla^{i-1} (A_{k-1} z_{k-1}).$$

Подставим v_m из первого уравнения (2.2) во второе и перейдем к матричной записи:

$$\begin{pmatrix} v_m \\ z_m \end{pmatrix} = \begin{pmatrix} 1 & h A_{m-1} \\ h & 1 + h^2 A_{m-1} \end{pmatrix} \begin{pmatrix} v_{m-1} \\ z_{m-1} \end{pmatrix} + h^2 \begin{pmatrix} 0 \\ \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m) \end{pmatrix} + \begin{pmatrix} h^{-1} Q_m \\ Q_m \end{pmatrix}. \quad (2.3)$$

В [2] оценка векторов погрешности $\vec{Z}_m = (v_m, z_m)^T$ (здесь и ниже верхний индекс T — знак транспонирования) сводилась к последовательному нахождению оценок из уравнения для вектора погрешности

$$\vec{Z}_m = \sum_{i=0}^{k-1} C_i \vec{Z}_{m-i} + \vec{Q}_m, \quad m \geq k.$$

Вектор погрешностей $\vec{Z}_m$ заключался в эллипсоид (см. [6]), и этот эллипсоид пересчитывался на каждом шаге. Для симметрической положительно-определенной матрицы B размера (2×2) эллипсоид $E(c, B)$ с центром в c определялся как $B^{\frac{1}{2}} S_1 + c$, где $S_1 = E(0, I)$ — единичный шар с центром начале координат.

Для неявного двухшагового метода (метода Нумерова) в [4] для последовательного нахождения оценок векторов погрешностей было предложено записать представление

$$\vec{Z}_m = C_0 \vec{Z}_m + C_1 \vec{Z}_{m-1} + \vec{Q}_m, \quad m \geq 2,$$

где

$$C_0 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{h^2}{12} A_m \end{pmatrix}, \quad C_1 = \begin{pmatrix} 1 & hA_{m-1} \\ h & 1 + \frac{11h^2}{12} A_{m-1} \end{pmatrix},$$

в виде

$$\vec{Z}_m = D_0 \vec{Z}_m + \frac{h^2}{12} \begin{pmatrix} 0 \\ \nabla(A_m z_m) \end{pmatrix} + \vec{Q}_m, \quad m \geq 2, \quad \text{где} \quad D_0 = \begin{pmatrix} 1 & hA_{m-1} \\ h & 1 + h^2 A_{m-1} \end{pmatrix}.$$

В этом случае точность оценки погрешности увеличивалась за счет того, что использовались не оценки величин $A_{m-i} z_{m-i}$, а оценки их первых разностей, то есть оценка происходила с учетом знаков малых слагаемых.

Проведем подобное улучшение для неявного многошагового метода. Для этого выразим в (2.3) i -е разности через $\nabla(A_{m-i} z_{m-i})$:

$$\begin{pmatrix} v_m \\ z_m \end{pmatrix} = \begin{pmatrix} 1 & hA_{m-1} \\ h & 1 + h^2 A_{m-1} \end{pmatrix} \begin{pmatrix} v_{m-1} \\ z_{m-1} \end{pmatrix} + h^2 \sum_{i=0}^{k-1} \gamma'_i \begin{pmatrix} 0 \\ \nabla(A_{m-i} z_{m-i}) \end{pmatrix} + \begin{pmatrix} h^{-1} Q_m \\ Q_m \end{pmatrix}, \quad (2.4)$$

где коэффициенты γ'_i выражаются через β_i .

Второе слагаемое правой части равенства содержит величину $\nabla(A_{m-i} z_{m-i})$. Выразим оценку первых разностей $P_l^1 \geq \max_{0 \leq i \leq l} |\nabla(A_i z_i)|$ через оценки

$$L \geq |A_i|, \quad L' \geq |\nabla A_i|/h, \quad z_l^B \geq \max_{0 \leq i \leq l} |z_i|, \quad v_l^B \geq \max_{0 \leq i \leq l} |v_i|.$$

Из второй строчки (2.2) следует, что

$$\nabla z_m = h v_m + h^2 \sum_{i=0}^{k-1} \gamma'_i \nabla(A_{m-i} z_{m-i}). \quad (2.5)$$

В силу неравенства $|\nabla(A_i z_i)| \leq |\nabla A_i| z_l^B + L |\nabla z_i|$, используя прямо получаемую из (2.5) оценку, имеем

$$|\nabla(A_i z_i)| \leq h L' z_l^B + L (h v_l^B + h^2 L P_l^1 \sum_{i=0}^{k-1} |\gamma'_i|).$$

Откуда следует, что

$$P_l^1 \leq \frac{h}{1 - h^2 L \sum_{i=0}^{k-1} |\gamma'_i|} (L' z_l^B + L v_l^B). \quad (2.6)$$

Заметим, что шаг h должен быть выбран таким малым, чтобы знаменатель дроби в (2.6) был положительным.

Пусть погрешность начальных значений не превосходит δ , тогда $\vec{Z}_i \in E(0, \mathbf{Z}_i)$ при $i = 0, \dots, k-1$, где $\mathbf{Z}_i = \begin{pmatrix} 0 & 0 \\ 0 & \delta^2 \end{pmatrix}$ при $i = 0, \dots, k-2$ и $\mathbf{Z}_{k-1} = 2 \begin{pmatrix} (v_{k-1}^B)^2 & 0 \\ 0 & \delta^2 \end{pmatrix}$. Здесь

$$|v_{k-1}| \leq v_{k-1}^B = \frac{2\delta}{h} + h\delta L \sum_{i=0}^k |\gamma_i|,$$

где γ_i — коэффициенты, которые возникнут в (2.5) после расписывания первых разностей.

Предлагается, начиная с этих значений, на каждом шаге пересчитывать эллипсоид $E(0, \mathbf{Z}_m)$, содержащий вектор погрешностей $(v_m, z_m)^T$.

При умножении матрицы D на эллипсоид $E(0, B)$ надо использовать (см., например, [6, с. 74]) равенство

$$D \cdot E(0, B) = E(0, DBD^T).$$

При сложении двух эллипсоидов надо использовать вложение

$$E(0, B_1) + E(0, B_2) \subset E(0, (1+p)B_1 + (1+1/p)B_2).$$

Параметр p можно выбрать (см. [7]) из условия минимальности суммы квадратов полуосей ($p = \sqrt{\text{Tr}B_2/\text{Tr}B_1}$) или минимальности объема ($p = \sqrt{n^{-1}\text{Tr}(B_1^{-1}B_2)}$) эллипсоида, содержащего сумму данных эллипсоидов.

Суммой второго и третьего слагаемых в правой части (2.4) является вектор

$$\begin{pmatrix} h^{-1}Q_m \\ h^2 \sum_{i=0}^{k-1} \gamma'_i \nabla(A_{m-i}z_{m-i}) + Q_m \end{pmatrix},$$

содержащийся в эллипсоиде $E(0, \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix})$, где $a = h^{-1}Q$, $b = h^2 P_m^1 \sum_{i=0}^{k-1} |\gamma'_i| + Q$, $Q = N + w$.

В результате, общая последовательность действий выглядит следующим образом.

1. Определяются начальные значения:

$$z_0^* = z_1^* = \dots = z_{k-1}^* = \delta, \quad v_{k-1}^B = \frac{2\delta}{h} + h\delta L \sum_{i=0}^k |\gamma_i|.$$

Устанавливается $m = k$.

2. Вычисляется предварительная оценка погрешности:

$$\begin{cases} r_m^B = r_{m-1}^B + \frac{h}{1 - |\alpha_0|h^2L} \left(z_{m-1}^* L(2|\alpha_0| + |\alpha_1|) + z_{m-2}^* L(|\alpha_0| + |\alpha_2|) \right. \\ \quad \left. + L \sum_{i=3}^k |\alpha_i| z_{m-i}^* \right) + (N + w)(h^{-1} + |\alpha_0|L), \\ z_m^B = z_{m-1}^* + hr_m^B, \quad m \geq k. \end{cases} \quad (2.7)$$

Здесь z_i^* — улучшенная оценка ошибки, полученная на предыдущих шагах, начальные значения $z_i^* = \delta$ при $i = 0, \dots, k-1$, $r_{k-1}^B = \frac{2\delta}{h}$; шаг h такой, что $|\alpha_0|h^2L < 1$.

3. По предварительной оценке z_m^B из (2.7) находится эллипсоид $E(0, \mathbf{Z}_m)$, содержащий вектор $\vec{Z}_m = (v_m, z_m)^T$ из (2.4). Для этого вычисляется эллипсоид, в котором лежит сумма эллипсоидов, содержащих слагаемые правой части (2.4). Далее вычисляется улучшенная оценка погрешности

$$z_m^* = \sqrt{\mathbf{Z}_m^{2,2}},$$

где эллипсоид $E(0, \mathbf{Z}_m) = \mathbf{Z}_m^{\frac{1}{2}} S_1$, $S_1 = E(0, I)$ — единичный шар.

Затем m увеличивается на 1 и происходит переход к пункту 2.

3. Получение оценки погрешности с большей точностью

Для того, чтобы получить оценку погрешности с еще большей точностью, надо второе слагаемое в (2.3) представить в виде суммы величин двух типов. К первому типу отнесем члены, содержащие z_i и v_i только с индексом $m-1$ и присоединим их к первому слагаемому (2.3). Ко второму типу отнесем «малые» члены — разности порядка не меньше чем $G \in \{0, 1, \dots, k-1\}$ и, возможно, слагаемые, содержащие неоднородность Q_m . Прием, предложенный в пункте 2., соответствует $G = 1$, а метод статьи [2] соответствует $G = 0$.

Для $G \geq 2$ представим

$$\sum_{i=2}^k \beta_i \nabla^{i-1}(A_m z_m) = \sum_{i=2}^G \beta_i \nabla^{i-1}(A_m z_m) + \sum_{i=G+1}^k \beta_i \nabla^{i-1}(A_m z_m).$$

Второе слагаемое можно оценить через оценку разностей G -го порядка, предварительно выразив все разности $(i-1)$ -го порядка через разности порядка G .

Первое слагаемое нужно преобразовать таким образом, чтобы в него входили члены только либо первого, либо второго типа.

Для этого выполним следующие преобразования:

$$\begin{aligned} S &= \sum_{i=2}^G \beta_i \nabla^{i-1}(A_m z_m) = \sum_{i=2}^G \beta_i \sum_{q=0}^{i-1} C_{i-1}^q (\nabla^{i-1-q} A_{m-q}) (\nabla^q z_m) \\ &= \sum_{i=2}^G \beta_i (\nabla^{i-1} A_m) \cdot z_m + \sum_{i=2}^G (i-1) \beta_i (\nabla^{i-2} A_{m-1}) \cdot \nabla z_m + S_3, \end{aligned}$$

где слагаемое $S_3 = \sum_{i=3}^G \beta_i \sum_{q=2}^{i-1} C_{i-1}^q (\nabla^{i-1-q} A_{m-q}) (\nabla^q z_m)$ появляется только в случае $G \geq 3$.

С учетом значения $\nabla^2 z_m$ из (2.1), имеем

$$\begin{aligned} S_3 &= \sum_{i=3}^G \beta_i \sum_{q=2}^{i-1} C_{i-1}^q (\nabla^{i-1-q} A_{m-q}) (h^2 \nabla^{q-2} (A_{m-1} z_{m-1}) \\ &\quad + h^2 \sum_{l=2}^k \beta_l \nabla^{l+q-2} (A_m z_m) + \nabla^{q-2} Q_m). \end{aligned}$$

Сначала преобразуем z_m и ∇z_m в S , используя вторую строку (2.2). Получим

$$\begin{aligned} S &= \sum_{i=2}^G \beta_i (\nabla^{i-1} A_m) \cdot (z_{m-1} + h v_m + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m)) \\ &\quad + \sum_{i=2}^G (i-1) \beta_i (\nabla^{i-2} A_{m-1}) \cdot (h v_m + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m)) + S_3. \end{aligned}$$

Теперь преобразуем v_m , используя первую строку (2.2). Получим

$$\begin{aligned} S &= \sum_{i=2}^G \beta_i (\nabla^{i-1} A_m) \cdot (z_{m-1} + h(v_{m-1} + h A_{m-1} z_{m-1} + h^{-1} Q_m) \\ &\quad + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m)) + \sum_{i=2}^G (i-1) \beta_i (\nabla^{i-2} A_{m-1}) \cdot (h(v_{m-1} + h A_{m-1} z_{m-1} + h^{-1} Q_m) \\ &\quad + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1} (A_m z_m)) + S_3. \end{aligned}$$

Члены полученной суммы, содержащие в качестве множителей z_{m-1} и v_{m-1} , надо присоединить к первому слагаемому (2.4). Для остальных членов, которые не принадлежат ко второму типу, нужно повторить всю процедуру разбиения на слагаемые первого и второго типов заново.

Посмотрим, как предложенный метод работает при $G = 2$. В этом случае $S_3 = 0$,

$$S = \frac{1}{12}(\nabla A_m + h^2 A_m A_{m-1})z_{m-1} + \frac{h}{12}A_m v_{m-1} + \frac{1}{12}A_m(Q_m + h^2 \sum_{i=2}^k \beta_i \nabla^{i-1}(A_m z_m)).$$

Следовательно, запись (2.3) равносильна записи

$$\begin{pmatrix} v_m \\ z_m \end{pmatrix} = \begin{pmatrix} I & hA_{m-1} \\ h(I + \frac{h^2}{12}A_m) & I + h^2(A_{m-1} + \frac{\nabla A_m}{12} + \frac{h^2}{12}A_m A_{m-1}) \end{pmatrix} \begin{pmatrix} v_{m-1} \\ z_{m-1} \end{pmatrix} + h^2 \begin{pmatrix} 0 \\ \frac{1}{12}A_m Q_m + \frac{h^2}{12}A_m \sum_{i=2}^k \beta_i \nabla^{i-1}(A_m z_m) + \sum_{i=3}^k \beta_i \nabla^{i-1}(A_m z_m) \end{pmatrix} + \begin{pmatrix} h^{-1}Q_m \\ Q_m \end{pmatrix}. \quad (3.1)$$

Найдем оценку погрешности неявного метода Штермера 5-го порядка.

В этом случае справедлива следующая оценка второй координаты второго слагаемого правой части (3.1):

$$\begin{aligned} \left| \frac{1}{12}A_m Q_m + \frac{h^2}{12}A_m \sum_{i=2}^4 \beta_i \nabla^{i-1}(A_m z_m) + \sum_{i=3}^4 \beta_i \nabla^{i-1}(A_m z_m) \right| \\ \leq \frac{L}{12}Q + \frac{h^2 L}{144}P_m^1 + \frac{1}{120}P_m^2(1 + \frac{h^2 L}{12}). \end{aligned}$$

Здесь в качестве оценки первых разностей $P_l^1 : \max_{0 \leq i \leq l} \|\nabla(A_i z_i)\| \leq P_l^1$ используется оценка (2.6). Третьи разности были выражены через вторые, для которых ниже будет получена оценка $P_l^2 : \max_{0 \leq i \leq l} \|\nabla^2(A_i z_i)\| \leq P_l^2$.

Пусть выполнено неравенство $L'' \geq h^{-2}\|\nabla^2 A_i\|$. Тогда, используя оценки ∇z_i и $\nabla^2 z_i$, полученные из (2.5) и (2.1) соответственно, для $k \geq 4$ имеем

$$\begin{aligned} \|\nabla^2(A_i z_i)\| &\leq \|(\nabla^2 A_i)z_i + 2\nabla A_{i-1}\nabla z_i + A_{i-2}\nabla^2 z_i\| \\ &\leq h^2 L'' z_i^B + 2hL'(h v_i^B + h^2 \sum_{j=0}^{k-1} |\gamma'_j| P_l^1) + L(h^2 L z_{i-1}^B + h^2 \sum_{j=0}^{k-1} |\gamma'_j| P_l^2 + Q). \end{aligned}$$

Откуда

$$P_l^2 \leq \frac{1}{1 - h^2 L \sum_{j=0}^{k-1} |\gamma'_j|} \left(h^2 z_i^B (L'' + L^2) + 2h^2 L' v_i^B + 2h^3 L' \sum_{j=0}^{k-1} |\gamma'_j| P_l^1 + Q \right).$$

4. Численный эксперимент

Продемонстрируем работу предложенного метода оценки погрешности на конкретном примере. Вычисления будем проводить с точностью до 56-го двоичного знака, т. е. с $\delta = \frac{1}{2}2^{-56} = 2^{-57}$. Рассмотрим линейное уравнение

$$y'' = -\frac{9 \cos^2 x}{2 + \cos^2 x} y$$

с точным решением $y(x) = \sin x + \frac{1}{9} \sin 3x$.

Численное решение будем также считать методом Штермера 5-го порядка с точностью $\delta = 2^{-57}$ с шагом $h = 2^{-8}$.

Оценка $N = \frac{1}{240} |y^{(7)}| h^7 \leq 1.5 \cdot 10^{-17}$, т. к. $|y^{(8)}| \leq 244$. Оценка $w = 7 \cdot 10^{-18}$.

В таблице приведены только оценки, не превосходящие 10^{-4} .

X_0	100π	200π	400π	600π
$G = 0$	$4 \cdot 10^{-6}$			
$G = 1$	$9 \cdot 10^{-8}$	$8 \cdot 10^{-7}$	$3 \cdot 10^{-5}$	
$G = 2$	$9 \cdot 10^{-8}$	$7 \cdot 10^{-7}$	$5 \cdot 10^{-6}$	$2 \cdot 10^{-5}$

Результаты численных экспериментов показывают, что при учете знаков малых слагаемых точность оценки погрешности неявного метода Штермера значительно увеличивается. Кроме того, происходит увеличение длины интервала применимости метода оценивания погрешности с помощью эллипсоидов.

References

- [1] R. E. Moore, R. B. Kearfott, M. J. Cloud, *Introduction to interval analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 2009, 234 pp.
- [2] Н. Д. Золотарева, “Метод эллипсоидов для оценки глобальной ошибки метода Штермера”, *Вестн. Моск. ун-та. Сер. 15. Вычисл. матем. и киберн.*, 2002, № 1, 18–23; англ. пер.: N. D. Zolotareva, “Ellipsoid method for estimating the global error of the Stormer method”, *Moscow University Computational Mathematics and Cybernetics*, 2002, № 1, 20–26.
- [3] Н. Д. Золотарева, “О новом способе получения гарантированной оценки ошибки метода Штермера с помощью эллипсоидов”, *Вестн. Моск. ун-та. Сер. 15. Вычисл. матем. и киберн.*, 2002, № 3, 3–9. [N. D. Zolotareva, “New approach to Shtermer’s method guaranteed error assessment using ellipsoids”, *Moscow University Computational Mathematics and Cybernetics*, 2002, № 3, 3–9 (In Russian)].
- [4] Н. Д. Золотарева, “О новом способе получения гарантированной оценки ошибки метода Нумерова с помощью эллипсоидов”, *Вестник российских университетов. Математика.*, **27**:139 (2022), 261–269. [N. D. Zolotareva, “On a new method for obtaining a guaranteed error estimate for Numerov’s method using ellipsoids”, *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **27**:139 (2022), 261–269 (In Russian)].
- [5] О. Б. Арушанян, С. Ф. Залеткин, *Численное решение обыкновенных дифференциальных уравнений на фортране*, Издательство Московского университета, М., 1990, 335 с. [O. B. Arushanyan, S. F. Zaletkin, *Numerical Solution of Ordinary Differential Equations in Fortran*, Moscow University Publishing House, Moscow, 1990 (In Russian), 335 pp.]
- [6] Ф. Л. Черноусько, *Оценивание фазового состояния динамических систем*, Наука, М., 1988. [F. L. Chernousko, *Estimation of the Phase State of Dynamic Systems*, Nauka Publ., Moscow, 1988 (In Russian)].
- [7] Ю. Н. Решетняк, “Суммирование эллипсоидов в задаче гарантированного оценивания”, *Прикладная математика и механика*, **53**:2 (1989), 249–254; англ. пер.: Yu. N. Reshetnyak, “Summation of ellipsoids in the guaranteed estimation problem”, *Journal of Applied Mathematics and Mechanics*, **53**:2 (1989), 193–197.

Информация об авторе

Золотарева Наталья Дмитриевна, кандидат физико-математических наук, доцент кафедры вычислительных методов, Московский государственный университет им. М. В. Ломоносова, г. Москва, Российская Федерация. E-mail: zolotareva-vmk@yandex.ru

ORCID: <https://orcid.org/0000-0003-1490-2091>

Поступила в редакцию 29.09.2024 г.

Поступила после рецензирования 18.11.2024 г.

Принята к публикации 22.11.2024 г.

Information about the author

Natalia D. Zolotareva, Candidate of Physical and Mathematical Sciences, Associate Professor of Computational Methods Department, Lomonosov Moscow State University, Moscow, Russian Federation. E-mail: zolotareva-vmk@yandex.ru

ORCID: <https://orcid.org/0000-0003-1490-2091>

Received 29.09.2024

Reviewed 18.11.2024

Accepted for press 22.11.2024

НАУЧНАЯ СТАТЬЯ

© Измаилов А.Ф., Усков Е.И., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-401-424>

УДК 519.6



Ускорение сходимости ньютоновских методов к особым решениям нелинейных уравнений

Алексей Феридович ИЗМАИЛОВ¹, Евгений Иванович УСКОВ²¹ ФГБОУ ВО «Московский государственный университет им. М. В. Ломоносова»
119991, ГСП-2, Российская Федерация, г. Москва, Ленинские горы, 1² ФГБОУ ВО «Тамбовский государственный университет им. Г. Р. Державина»
392000, Российская Федерация, г. Тамбов, ул. Интернациональная, 33

Аннотация. Рассматривается простейшая процедура экстраполяции, а именно, удвоение шага, для ускорения сходимости ньютоновских методов к особым решениям гладких нелинейных уравнений. Демонстрируется, что ускоряющий эффект этой процедуры для разных ньютоновских методов может быть разным. Для линейно-квадратичных уравнений приводятся теоретические результаты, дающие количественные оценки потенциального эффекта от экстраполяции для методов Ньютона, Левенберга–Марквардта, а также недавно предложенного ньютоновского метода с подзадачами линейного программирования, в определенном смысле объясняющие наблюдаемую разницу. Теоретически анализ основан на интерпретации этих методов как возмущенного метода Ньютона с соответствующими оценками на возмущения, а также на тонких результатах, количественно характеризующих шаг такого возмущенного метода и его локальную сходимость с линейной скоростью к особым решениям, в которых выполняется условие 2-регулярности по направлению из ядра первой производной. Также проводятся численные эксперименты для глобализаций указанных алгоритмов, снабженных выбором параметра длины шага, на двух тестовых наборах. Экспериментальные наблюдения подтверждают теоретические результаты, а также демонстрируют, что в случаях, когда уравнение содержит нелинейные и неквадратичные члены, эффект от экстраполяции выравнивается.

Ключевые слова: нелинейное уравнение, особое решение, метод Ньютона, метод Левенберга–Марквардта, ньютоновский метод с подзадачами линейного программирования, экстраполяция

Благодарности: Исследование выполнено за счет гранта Российского научного фонда (проект № 24-21-00015, <https://rscf.ru/project/24-21-00015/>).

Для цитирования: Измаилов А.Ф., Усков Е.И. Ускорение сходимости ньютоновских методов к особым решениям нелинейных уравнений // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 401–424. <https://doi.org/10.20310/2686-9667-2024-29-148-401-424>

SCIENTIFIC ARTICLE

© A. F. Izmailov, E. I. Uskov, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-401-424>

Accelerating convergence of Newton-type methods to singular solutions of nonlinear equations

Alexey F. IZMAILOV¹, Evgeniy I. USKOV²

¹ Lomonosov Moscow State University

1 Leninskiye Gory, Moscow 119991, Russian Federation

² Derzhavin Tambov State University

33 Internatsionalnaya St., Tambov 392000, Russian Federation

Abstract. We consider the simplest extrapolation procedure, specifically doubling the step, intended for acceleration of convergence of Newton-type methods to singular solutions of smooth nonlinear equations. We demonstrate that the acceleration effect of this procedure can be different for different Newton-type methods. For linear-quadratic equations we provide theoretical results yielding quantitative estimates of the potential effect of extrapolation for the Newton method, for the Levenberg–Marquardt method, and for the recently proposed LP-Newton method, in some sense explaining the observed difference. Theoretical analysis relies on interpretation of these methods as a perturbed Newton method with the appropriate estimates of perturbations, as well as on sharp results yielding a quantitative characterization of a step of such perturbed method, and its local convergence at a linear rate to singular solutions satisfying the 2-regularity condition in a direction from the null space of the first derivative. Furthermore, we perform numerical experiments with globalized versions of the algorithms in question, equipped with choosing the stepsize parameter, on two sets of test problems. Experimental observations confirm the theoretical results, and also demonstrate that in cases when the equation contains nonlinear and nonquadratic terms, the effect of extrapolation is evened out.

Keywords: nonlinear equation, singular solution, Newton method, Levenberg–Marquardt method, LP-Newton method, extrapolation

Acknowledgements: The research was supported by the Russian Science Foundation (project no. 24-21-00015, <https://rscf.ru/project/24-21-00015/>).

Mathematics Subject Classification: 47J05, 65H10.

For citation: Izmailov A.F., Uskov E.I. Accelerating convergence of Newton-type methods to singular solutions of nonlinear equations. *Vestnik Rossiyskikh Universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 401–424.

<https://doi.org/10.20310/2686-9667-2024-29-148-401-424> (In Russian, Abstr. in Engl.)

Введение

Известным средством ускорения сходимости ньютоновских методов к особым решениям нелинейных уравнений является экстраполяция [1, 2]. Однако, как демонстрируется в этой статье, ускоряющий эффект этой процедуры для разных ньютоновских методов может быть разным, и целью работы является теоретическое объяснение этой разницы для линейно-квадратичных уравнений, а также численный анализ эффекта от экстраполяции для глобализованных одномерным поиском методов Ньютона, Левенберга–Марквардта, а также ньютоновского метода с подзадачами линейного программирования (соответствующие ссылки на литературу приводятся ниже).

Используемые в статье обозначения вполне традиционны. Через $L^\perp$ обозначается ортогональное дополнение линейного подпространства L . Если не указано иное, будем считать, что $\|\cdot\|$ — евклидова норма; в некоторых случаях будет также использоваться ∞ -норма, которая для $u \in \mathbb{R}^p$ определяется как $\|u\|_\infty = \max\{|u_j| \mid j \in \{1, \dots, p\}\}$. Для заданных $\bar{u}, \bar{v} \in \mathbb{R}^p$, $\varepsilon > 0$ и $\delta > 0$ определим множество

$$K_{\varepsilon, \delta}(\bar{u}; \bar{v}) = \{u \in \mathbb{R}^p \mid \|u - \bar{u}\| \leq \varepsilon, \ ||\bar{v}\|(u - \bar{u}) - \|u - \bar{u}\|\bar{v}\| \leq \delta\|u - \bar{u}\|\|\bar{v}\|\}.$$

Через I обозначается единичная матрица, а через $\text{im } A$ и $\ker A$ — образ (множество значений) и ядро (множество нулей) матрицы (линейного оператора) A . Для симметричного билинейного отображения $B : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ и для заданного $v \in \mathbb{R}^p$ через $B[v]$ обозначается матрица линейного оператора $B[v, \cdot]$. Наконец, запись $O(t)u$ используется для произведения вида Au с матрицей $A = O(t)$ в случаях, когда другая специфика этой матрицы не имеет значения.

1. Постановка задачи и ньютоновские методы

Рассматриваются ньютоновские методы для решения нелинейного уравнения

$$\Phi(u) = 0 \tag{1.1}$$

с достаточно гладким отображением $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^p$. Такие методы по имеющемуся приближению $u \in \mathbb{R}^p$ генерируют следующее приближение как $u + v$, где $v \in \mathbb{R}^p$ есть решение подзадачи, характеризующей конкретный метод. Например:

- подзадача метода Ньютона (NM; см., например, [3, разд. 2.1.1]) — это линеаризованное уравнение (1.1), т. е.

$$\Phi(u) + \Phi'(u)v = 0; \tag{1.2}$$

- подзадача метода Левенберга–Марквардта (LM; происходит из работ [4, 5]; см. также, например, [6, разд. 10.2], а также обзор современных теорий сходимости в [7]) имеет вид

$$\frac{1}{2}\|\Phi(u) + \Phi'(u)v\|^2 + \frac{1}{2}\sigma(u)\|v\|^2 \rightarrow \min, \tag{1.3}$$

с функцией $\sigma : \mathbb{R}^p \rightarrow \mathbb{R}_+$, определяющей значения параметра регуляризации, что равносильно линейному уравнению

$$(\Phi'(u))^\top \Phi(u) + ((\Phi'(u))^\top \Phi'(u) + \sigma(u)I)v = 0; \tag{1.4}$$

- подзадача ньютоновского метода с подзадачами линейного программирования (LPN; [8]) имеет вид

$$\gamma \rightarrow \min, \quad \|\Phi(u) + \Phi'(u)v\| \leq \gamma \|\Phi(u)\|^2, \quad \|v\| \leq \gamma \|\Phi(u)\|, \quad (1.5)$$

относительно (v, γ) с $\gamma \in \mathbb{R}$, и если используется ∞ -норма, то это задача линейного программирования, что и дает название методу.

Как обсуждается ниже, все перечисленные методы могут интерпретироваться как возмущенный метод Ньютона (pNM) с подзадачей

$$\Phi(u) + (\Phi'(u) + \Omega(u))v = \omega(u), \quad (1.6)$$

где отображения $\Omega : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}$ и $\omega : \mathbb{R}^p \rightarrow \mathbb{R}^p$ характеризуют различные возмущения базового NM. Свойства базового NM и pNM вблизи неособых решений излагаются в [3, разд. 2.1.1]. Здесь решение $\bar{u}$ уравнения (1.1) называется неособым, если $\Phi'(\bar{u})$ — невырожденная матрица, и особым в противном случае.

Данная работа посвящена именно случаю особого решения $\bar{u}$. Ключевую роль при этом играет условие 2-регулярности отображения Φ в точке $\bar{u}$ по направлению $v \in \mathbb{R}^p$ [9, разд. 1.3.3], сводящееся в рассматриваемом здесь случае отображения из $\mathbb{R}^p$ в $\mathbb{R}^p$ к невырожденности линейного оператора $u \mapsto \Pi \Phi''(\bar{u})[v, u] : \ker \Phi'(\bar{u}) \rightarrow (\operatorname{im} \Phi'(\bar{u}))^\perp$, где Π — оператор ортогонального проектирования на $(\operatorname{im} \Phi'(\bar{u}))^\perp$. Важное свойство, которое легко проверить, состоит в том, что из 2-регулярности Φ в точке $\bar{u}$ по направлению v вытекает существование $\varepsilon > 0$ и $\delta > 0$ таких, что для $u \in K_{\varepsilon, \delta}(\bar{u}, v)$ матрица $\Phi'(u)$ невырождена, причем

$$(\Phi'(u))^{-1} = O(\|u - \bar{u}\|^{-1}) \quad (1.7)$$

при $u \rightarrow \bar{u}$. С использованием этого факта, следующий результат о локальной сходимости pNM с линейной скоростью был получен в [10, теорема 1] (см. также [11, теорема 3.1], где используются более слабые предположения гладкости), и является развитием результатов, полученных [1, 2] для базового NM. Для всякого $u \in \mathbb{R}^p$ введем однозначное разложение $u = u_1 + u_2$, $u_1 \in (\ker \Phi'(\bar{u}))^\perp$, $u_2 \in \ker \Phi'(\bar{u})$.

Теорема 1.1. Пусть $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^p$ дважды дифференцируемо вблизи $\bar{u} \in \mathbb{R}^p$, и его вторая производная липшицева относительно $\bar{u}$, т. е.

$$\Phi''(u) - \Phi''(\bar{u}) = O(\|u - \bar{u}\|)$$

при $u \rightarrow \bar{u}$. Пусть $\bar{u}$ является решением уравнения (1.1), причем Φ 2-регулярно в точке $\bar{u}$ по направлению $\bar{v} \in \ker \Phi'(\bar{u}) \setminus \{0\}$. Пусть для $\Omega : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}$ и $\omega : \mathbb{R}^p \rightarrow \mathbb{R}^p$ существует $\delta > 0$ такое, что оценки

$$\Omega(u) = O(\|u - \bar{u}\|), \quad \omega(u) = O(\|u - \bar{u}\|^2),$$

$$\Pi \Omega(u) = O(\|u_1 - \bar{u}_1\|) + O(\|u - \bar{u}\|^2)$$

$$\Pi \omega(u) = O(\|u - \bar{u}\| \|u_1 - \bar{u}_1\|) + O(\|u - \bar{u}\|^3)$$

выполняются для $u \in K_{\varepsilon, \delta}(\bar{u}; \bar{v})$ при $\varepsilon \rightarrow 0+$.

Тогда для любых $\hat{\varepsilon} > 0$ и $\hat{\delta} > 0$ найдутся $\varepsilon = \varepsilon(\bar{v}) > 0$ и $\delta = \delta(\bar{v}) > 0$ такие, что для всякой начальной точки $u^0 \in K_{\varepsilon, \delta}(\bar{u}; \bar{v})$ существует единственная последовательность

$\{u^k\} \subset \mathbb{R}^p$ такая, что для каждого k выполняется $u^{k+1} = u^k + v^k$, где v^k удовлетворяет (1.6), и для этой последовательности и для каждого k имеет место $u_2^k \neq \bar{u}_2$, $u^k \in K_{\varepsilon, \delta}(\bar{u}; \bar{v})$, $\{u^k\}$ сходится к $\bar{u}$, $\{\|u^k - \bar{u}\|\}$ сходится к нулю монотонно,

$$\frac{\|u_1^{k+1} - \bar{u}_1\|}{\|u_2^{k+1} - \bar{u}_2\|} = O(\|u^k - \bar{u}\|)$$

при $k \rightarrow \infty$, и

$$\lim_{k \rightarrow \infty} \frac{\|u_2^{k+1} - \bar{u}_2\|}{\|u_2^k - \bar{u}_2\|} = \frac{1}{2}.$$

Характер сходимости в теореме 1.1 служит основой для техник ускорения сходимости ньютоновских методов к особым решениям [1, 2], одной из которых является экстраполяция, которая в своем простейшем варианте состоит в генерировании, параллельно с основной последовательностью $\{u^k\}$, вспомогательной последовательности $\{\hat{u}^k\}$ посредством удвоения шага (p)NM:

$$\hat{u}^{k+1} = u^k + 2v^k. \quad (1.8)$$

Как показано в [2, Theorem 4.1], для базового NM можно ожидать $\{\hat{u}^k\}$ линейной сходимости с асимптотическим общим частным $1/4$, в то время как для основной итерационной последовательности $\{u^k\}$ оценки в теореме 1.1 гарантируют такое свойство только с $1/2$. Заметим, что (1.8) не связано ни с какими дополнительными вычислительными затратами, кроме одного лишнего вычисления Φ на каждой итерации, что требуется для оценки качества полученного приближения $\hat{u}^{k+1}$. Кроме того, экстраполяция никак не влияет на основную последовательность $\{u^k\}$, и поэтому легко комбинируется с различными реализациями pNM, и в частности, с глобализациями сходимости методов этого класса, в том числе с рассматриваемыми в разд. 4. алгоритмами, использующими одномерный поиск для адаптивного выбора параметра длины шага. Однако, эффект от экстраполяции можно ожидать только при асимптотическом принятии полного ньютоновского шага, поскольку данная процедура основана на характере сходимости именно полношагового pNM; см. [12, 13].

Возвращаясь к базовому pNM, для метода LM, использующего $\sigma(u) = O(\|\Phi(u)\|^\theta)$ с фиксированным $\theta > 0$, на основе (1.7) в [10, разд. 3.1] показано, что при 2-регулярности Φ в точке $\bar{u}$ по направлению $\bar{v} \in \ker \Phi'(\bar{u}) \setminus \{0\}$ найдутся $\varepsilon > 0$ и $\delta > 0$ такие, что для $u \in K_{\varepsilon, \delta}(\bar{u}, \bar{v})$ (1.4) записывается в виде (1.6) с $\omega(\cdot) \equiv 0$ и

$$\Omega(u) = \sigma(u)((\Phi'(u))^{-1})^\top = O(\|u - \bar{u}\|^{-1} \|\Phi(u)\|^\theta) \quad (1.9)$$

при $u \rightarrow \bar{u}$. В [14, разд. 3.2] предложена альтернативная интерпретация подзадачи (1.3) метода LM как подзадачи (1.6) pNM, с $\Omega(\cdot) \equiv 0$ и соответствующей оценкой на ω . В отличие от приведенной выше интерпретации, явно основанной на форме (1.4) итерационной подзадачи метода LM, альтернативная интерпретация применима и в случае уравнения с дополнительным ограничением, но получаемые при этом оценки несколько слабее и не подходят для использования в разд. 3 ниже.

Для метода LPN, снова с использованием (1.7), в [10, разд. 3.2] показано, что в тех же предположениях найдутся $\varepsilon > 0$ и $\delta > 0$ такие, что для $u \in K_{\varepsilon, \delta}(\bar{u}, \bar{v})$ любое решение подзадачи (1.5) удовлетворяет (1.6) с $\Omega(\cdot) \equiv 0$ и некоторым ω , для которого

$$\|\omega(u)\| \leq \gamma(u) \|\Phi(u)\|^2 = O(\|u - \bar{u}\| \|\Phi(u)\|) \quad (1.10)$$

при $u \rightarrow \bar{u}$, где $\gamma(u)$ — оптимальное значение задачи (1.5).

2. Сравнительный эффект от экстраполяции

Начнем с простого примера, хорошо демонстрирующего эффекты, о которых идет речь в этом разделе.

Пример 2.1. Пусть $p = 1$, $\Phi(u) = u^2$. Единственным решением уравнения (1.1) является $\bar{u} = 0$, оно является особым, и Φ 2-регулярно в точке 0 по любому ненулевому направлению.

Итерационное уравнение (1.2) NM имеет вид $u^2 + 2uv = 0$, и при $u \neq 0$ его единственным решением является $v = -u/2$. Поэтому $u + 2v = 0$, т.е. NM с экстраполяцией за один шаг попадает в точное решение из любого начального приближения.

Далее, итерационное уравнение (1.4) метода LM имеет вид $2u^3 + (4u^2 + \sigma(u))v = 0$, и при $u \neq 0$ его единственным решением является $v = -2u^3/(4u^2 + \sigma(u))$. Если $\sigma(u) = |u|^{2\theta}$, то

$$u + 2v = \frac{\sigma(u)u}{(4u^2 + \sigma(u))} = \frac{|u|^{2\theta+1}}{(4u^2 + |u|^{2\theta})},$$

что при $\theta > 1$ дает оценки

$$u + 2v = \frac{1}{4}|u|^{2\theta-1} + o(|u|^{2\theta-1}) = O(|u|^{2\theta-1}),$$

что при $\theta \geq 2$ превращается в оценку $u + 2v = O(|u|^3)$.

Наконец, итерационная подзадача (1.5) метода LPN имеет вид

$$\gamma \rightarrow \min, \quad -\gamma u^4 \leq u^2 + 2uv \leq \gamma u^4, \quad -\gamma u^2 \leq v \leq \gamma u^2,$$

и, скажем, при $u > 0$ единственным решением этой задачи линейного программирования является $(v, \gamma) = (-u/(2+u), 1/(2u+u^2))$. Тогда $u + 2v = u^2/(2+u)$, что дает оценки

$$u + 2v = \frac{1}{2}u^2 + o(u^2) = O(u^2).$$

Функция в примере 2.1 является квадратичной скалярной функцией одной переменной. Обратимся теперь к случаю общего линейно-квадратичного отображения Φ : пусть

$$\Phi(u) = Au + \frac{1}{2}B[u, u], \quad (2.1)$$

где A — $(n \times n)$ -матрица, а $B : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ — симметричное билинейное отображение. Тогда $\Phi'(u) = A + B[u]$, и уравнение (1.1) имеет решение $\bar{u} = 0$, в котором $\Phi'(0) = A$, и, соответственно, это решение является особым, если матрица A вырождена. Условие 2-регулярности Φ в точке 0 по направлению $v \in \mathbb{R}^p$ при этом состоит в невырожденности линейного оператора $\hat{B}(v) : \ker A \rightarrow (\operatorname{im} A)^\perp$, $\hat{B}(v)u = \Pi B[v, u]$, где Π — оператор ортогонального проектирования на $(\operatorname{im} A)^\perp$.

Итерационная подзадача (1.2) NM принимает вид

$$(A + B[u])v = -Au - \frac{1}{2}B[u, u]. \quad (2.2)$$

Если $u \in \ker A$, то правую часть (2.2) можно заменить на $-(A + B[u])u/2$, и тогда уравнение (2.2) имеет решение $v = -u/2$, причем это решение единственно в случае невырожденности $A + B[u]$. При этом $u + 2v = 0$, т.е. экстраполяция немедленно дает точное

решение. В чисто квадратичном случае, когда $A = 0$, условие $u \in \ker A$ выполняется автоматически, и значит, экстраполяция дает точное решение за один шаг из любого начального приближения u , в котором матрица $B[u]$ невырождена.

Далее, итерационная подзадача (1.6) pNM принимает вид

$$(A + B[u] + \Omega(u))v = -Au - \frac{1}{2}B[u, u] + \omega(u), \quad (2.3)$$

и при $u \in \ker A$ правую часть можно заменить на $-(A + B[u] + \Omega(u))u - \Omega(u)u/2 + \omega(u)$. Если матрица $A + B[u] + \Omega(u)$ обратима, то единственным решением (2.3) является

$$v = -\frac{1}{2}u + (A + B[u] + \Omega(u))^{-1} \left(\frac{1}{2}\Omega(u)u + \omega(u) \right). \quad (2.4)$$

При $u \in \ker A$ имеет место оценка $\Phi(u) = O(\|u\|^2)$, и, предполагая 2-регулярность Φ в точке 0 по направлению $\bar{v} \in \ker A \setminus \{0\}$, для метода LM с $\theta > 1$, в силу (1.9), для $u \in K_{\varepsilon, \delta}(0, \bar{v})$ при достаточно малых $\varepsilon > 0$ и $\delta > 0$ имеем

$$\Omega(u) = O(\|u\|^{2\theta-1}), \quad (2.5)$$

и матрица $A + B[u] + \Omega(u)$ обратима, причем, аналогично (1.7),

$$(A + B[u] + \Omega(u))^{-1} = O(\|u\|^{-1}). \quad (2.6)$$

Тогда из (2.4) (и того, что $\omega(\cdot) \equiv 0$) получаем

$$v = -\frac{1}{2}u + O(\|u\|^{2\theta-1}) \quad (2.7)$$

при $u \rightarrow 0$. Поэтому экстраполяция дает приближение

$$u + 2v = O(\|u\|^{2\theta-1}), \quad (2.8)$$

и, в частности, при $\theta \geq 2$ имеет место оценка

$$u + 2v = O(\|u\|^3)$$

при $u \rightarrow 0$.

Для метода LPN аналогичным образом при $u \in \ker A$ из (1.10) выводится оценка

$$\omega(u) = O(\|u\|^3), \quad (2.9)$$

и в силу (2.4), (2.6) (и того, что $\Omega(\cdot) \equiv 0$)

$$v = -\frac{1}{2}u + O(\|u\|^2),$$

и следовательно,

$$u + 2v = O(\|u\|^2) \quad (2.10)$$

при $u \rightarrow 0$.

Тем самым доказано

Предложение 2.1. Пусть A — $(n \times n)$ -матрица, а $B : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ — симметричное билинейное отображение, и пусть для некоторого $\bar{v} \in \ker A \setminus \{0\}$ линейный оператор $\hat{B}(\bar{v})$ невырожден.

Тогда найдутся $\varepsilon > 0$ и $\delta > 0$ такие, что для $u \in K_{\varepsilon, \delta}(0, \bar{v}) \cap \ker A$, для уравнения (1.1) с линейно-квадратичным отображением Φ из (2.1) справедливо следующее:

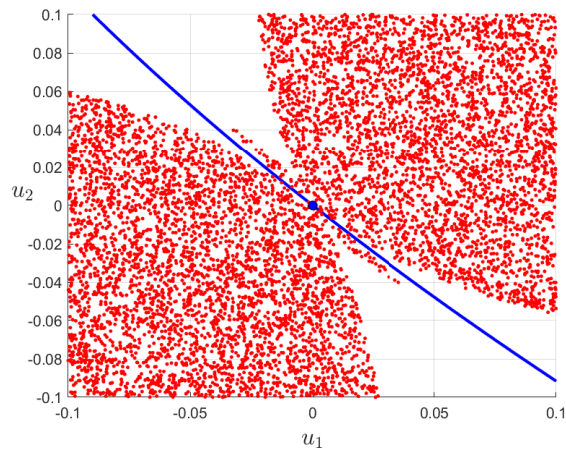
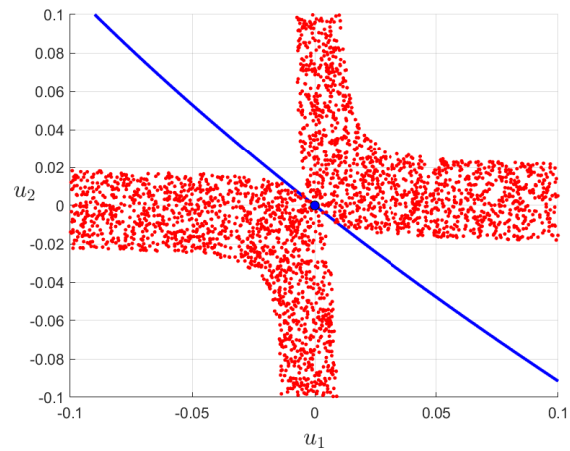
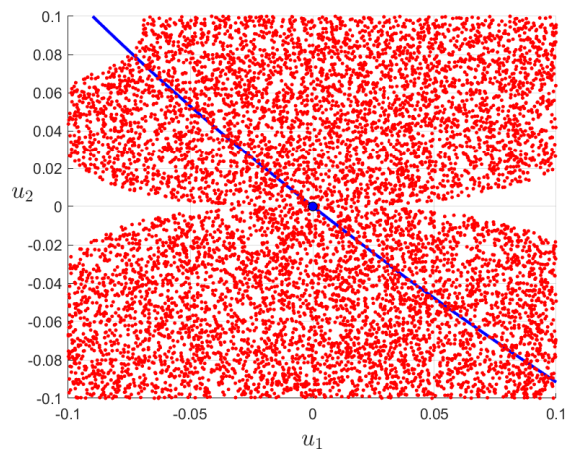
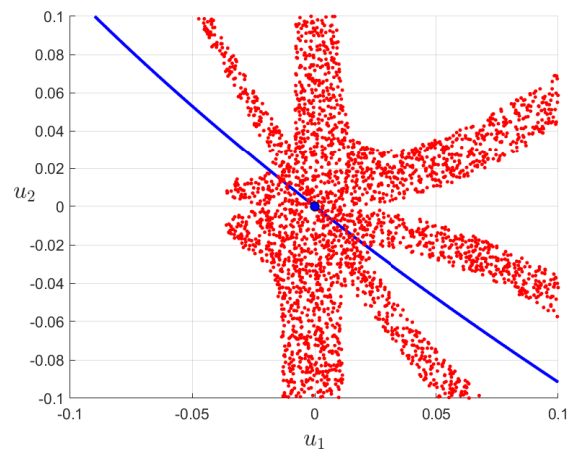
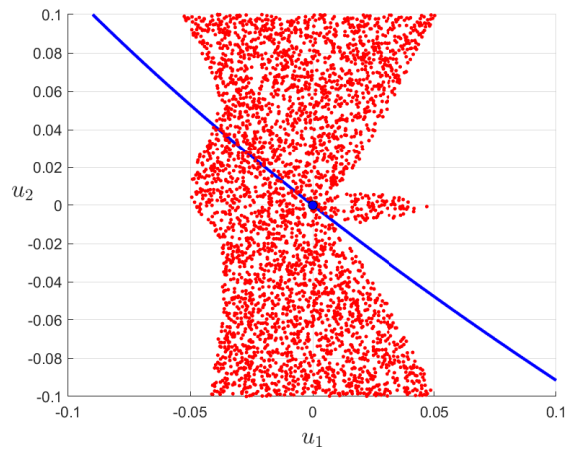
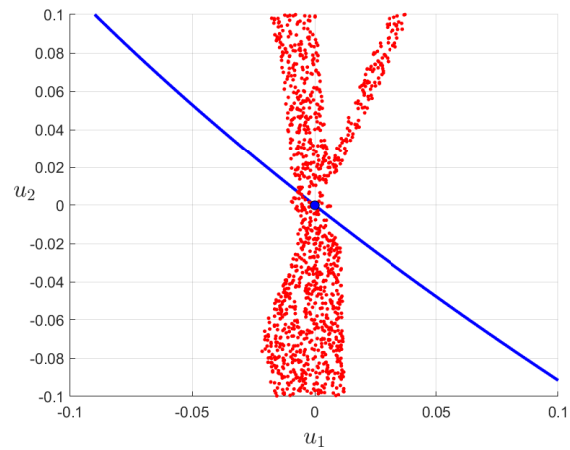
- NM с подзадачей (1.2) однозначно определяет направление v , и для него экстраполяция дает $u + 2v = 0$;
- метод LM с подзадачей (1.3), использующей $\sigma(u) = O(\|\Phi(u)\|^\theta)$ с фиксированным $\theta > 1$, однозначно определяет направление v , и для него экстраполяция дает приближение $u + 2v$, для которого выполняется оценка (2.8);
- метод LPN с подзадачей (1.5) определяет некоторое направление v , и для всякого такого направления экстраполяция дает приближение $u + 2v$, для которого выполняется оценка (2.10).

В частности, при $\theta > 1$ оценка (2.8) для метода LM лучше, чем оценка (2.10) для метода LPN, и это может служить объяснением большего эффекта экстраполяции для метода LM, чем для LPN.

В общем линейно-квадратичном случае (когда $A \neq 0$, и, соответственно, выполнение $u \in \ker A$ не является автоматическим), и даже просто в случае общего достаточно гладкого Φ , согласно [15, теорема 6.1] (см. также [13, лемма 7]), за один шаг NM «приближенно» попадает в $\ker A$ из асимптотически плотного звездного множества начальных точек $u^0 \in \mathbb{R}^p$, а именно, попадает в $K_{\varepsilon, \delta}(0, \bar{v})$ с соответствующим $\bar{v} = \bar{v}(u^0/\|u^0\|)$, при любых наперед заданных $\varepsilon > 0$ и $\delta > 0$. (Точного попадания в $\ker A$ может не быть даже для NM, даже в линейно-квадратичном случае, что демонстрируется примером 2.2 ниже, в котором $\dim \ker A = 1$; аналогичный эффект при $\dim \ker A = 2$ наблюдается в тесте Misc 10 в разд. 4.) Однако, как указано в комментариях в конце [10, разд. 2], гарантировать указанное свойство для общего pNM не удастся, если только не накладывать жесткие предположения на глобальное (не только вблизи ядра) поведение Ω и ω , а именно, $P\Omega(u) = O(\|u\|^2)$ и $P\omega(u) = O(\|u\|^3)$ для u из рассматриваемого звездного множества начальных приближений. Для метода LM эти предположения сводятся к оценке $\sigma(u) = O(\|u\|^3)$, что можно гарантировать только при $\theta \geq 3$. Для метода LPN гарантировать требуемое можно при выполнении оценки $\gamma(u) = O(\|u\|)$, а эта оценка ниоткуда не следует. Тем не менее, на практике попадание LM и LPN в $K_{\varepsilon, \delta}(0, \bar{v})$ с достаточно малыми $\varepsilon > 0$ и $\delta > 0$ обычно наблюдается, если не за одну, то за небольшое количество начальных итераций.

Пример 2.2. Пусть $p = 2$, $\Phi(u) = (u_1 + u_2^2/2, u_1 u_2 + u_2^2/2)$. Единственным решением уравнения (1.1) является $\bar{u} = 0$, оно является особым, $\ker \Phi'(0)$ натянуто на $\bar{v} = (0, 1)$, и Φ 2-регулярно в точке 0 по направлению $\bar{v}$.

На рис. 1 показана линия, задаваемая равенством $\det \Phi'(u) = 0$, а также области начальных точек, один шаг рассматриваемых методов из которых приводит в $K_{\varepsilon, \delta}(0, \bar{v})$ с $\varepsilon = 0.1$, при двух разных значениях δ . Эти области уменьшаются при уменьшении δ , но, по-видимому, остаются асимптотически плотными в точке 0.

(a) NM, $\delta = 0.05$ (b) NM, $\delta = 0.01$ (c) Метод LM, $\delta = 0.05$ (d) Метод LM, $\delta = 0.01$ (e) Метод LPN, $\delta = 0.05$ (f) Метод LPN, $\delta = 0.01$ Рис. 1. Пример 2.2: области попадания одного шага NM в $K_{\varepsilon, \delta}(0, \bar{v})$ с $\varepsilon = 0.1$

В соответствии со сказанным выше, далее будем рассматривать случай, когда $u^0 \in K_{\varepsilon, \delta}(0, \bar{v})$. По-прежнему будем использовать разложение всякого $u \in \mathbb{R}^p$ в сумму $u = u_1 + u_2$, где в линейно-квадратичном случае $u_1 \in (\ker A)^\perp$, $u_2 \in \ker A$.

Согласно [11, лемма 3.1] для общего (не обязательно линейно-квадратично) случая, и даже всего лишь при сильной полугладкости Φ' в решении $\bar{u} = 0$, для $u \in K_{\bar{\varepsilon}, \bar{\delta}}(0, \bar{v})$ при малых $\bar{\varepsilon} > 0$ и $\bar{\delta} > 0$ направление v NM однозначно определяется и удовлетворяет оценкам

$$u_1 + v_1 = O(\|u_1\|\|u\|) + O(\|u\|^3), \quad (2.11)$$

$$u_2 + v_2 = \frac{1}{2}u_2 + O(\|u_1\|) + O(\|u\|^2) \quad (2.12)$$

при $u \rightarrow 0$. При этом $\|u + v\| \geq \gamma\|u\|$ с некоторым фиксированным $\gamma > 0$, и, поскольку $\|u_1\| = O(\|u\|)$,

$$\frac{\|u_1 + v_1\|}{\|u + v\|^2} = O\left(\frac{\|u_1 + v_1\|}{\|u\|^2}\right) = O\left(\frac{\|u_1\|}{\|u\|}\right) + O(\|u\|) = O(1), \quad (2.13)$$

т. е. $u_1 + v_1 = O(\|u + v\|^2)$. Далее, согласно [11, теорема 3.1], если $\varepsilon > 0$ и $\delta > 0$ достаточно малы, то для любого $u^0 \in K_{\varepsilon, \delta}(0, \bar{v})$ последовательность $\{u^k\}$ с шагами v^k NM однозначно определяется, сходится к 0, причем $u^k \in K_{\bar{\varepsilon}, \bar{\delta}}(0, \bar{v})$ для любого k , откуда и из оценки (2.13) следует, что $u_1^k = O(\|u^k\|^2)$ при $k \rightarrow \infty$. Тогда, снова привлекая (2.11),

$$\frac{\|u_1^k + v_1^k\|}{\|u^k + v^k\|^3} = O\left(\frac{\|u_1^k + v_1^k\|}{\|u^k\|^3}\right) = O\left(\frac{\|u_1^k\|}{\|u^k\|^2}\right) + O(1) = O(1),$$

т. е. имеет место оценка

$$u_1^k = O(\|u^k\|^3) \quad (2.14)$$

при $k \rightarrow \infty$, которая согласуется с [2, теорема 2.1, (iii)], где эта оценка приведена без доказательства.

Для рNM, согласно [11, лемма 3.1], оценки (2.11), (2.12) заменяются на

$$u_1 + v_1 = O(\|u_1\|\|u\|) + O(\|u\|\|\Omega(u)\|) + O(\|\omega(u)\|) + O(\|u\|^3), \quad (2.15)$$

$$u_2 + v_2 = \frac{1}{2}u_2 + O(\|u_1\|) + O(\|\Pi\Omega(u)\|) + O(\|u\|^{-1}\|\Pi\omega(u)\|) + O(\|u\|^2) \quad (2.16)$$

при $u \rightarrow 0$. Для метода LM, согласно (2.5), поскольку в этом случае $\omega(\cdot) \equiv 0$, эти оценки принимают вид

$$u_1 + v_1 = O(\|u_1\|\|u\|) + O(\|u\|^{2\theta}) + O(\|u\|^3),$$

$$u_2 + v_2 = \frac{1}{2}u_2 + O(\|u_1\|) + O(\|u\|^{2\theta-1}) + O(\|u\|^2),$$

что при $\theta \geq 3/2$ дает (2.11), (2.12). Для метода LPN, согласно (2.9), поскольку в этом случае $\Omega(\cdot) \equiv 0$, оценки (2.15), (2.16) сразу принимают вид (2.11), (2.12). Значит, и для этих методов в указанных случаях проходят рассуждения, проведенные выше для NM, и приводящие к оценке (2.14).

Следующая лемма уточняет в линейно-квадратичном случае оценку (1.7).

Лемма 2.1. Пусть A — $(n \times n)$ -матрица, а $B : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ — симметричное билинейное отображение, и пусть для некоторого $\bar{v} \in \mathbb{R}^p$ линейный оператор $\widehat{B}(\bar{v})$ невырожден.

Тогда найдутся $\bar{\varepsilon} > 0$ и $\bar{\delta} > 0$ такие, что для $u \in K_{\bar{\varepsilon}, \bar{\delta}}(0, \bar{v})$ матрица $A + B[u]$ невырождена, причем для $w \in \operatorname{im} A$ справедлива оценка

$$(A + B[u])^{-1}w = O(\|w\|) \quad (2.17)$$

при $w \rightarrow 0$.

Доказательство. Для заданных $u \in \mathbb{R}^p$ и $w \in \operatorname{im} A$, умножая левую и правую части уравнения

$$(A + B[u])v = w \quad (2.18)$$

на $I - \Pi$ и Π , получаем уравнения

$$(A + (I - \Pi)B[u])v_1 = w - (I - \Pi)B[u]v_2, \quad \Pi B[u]v_2 = -\Pi B[u]v_1. \quad (2.19)$$

Обозначим через $\widehat{A} : (\ker A)^\perp \rightarrow \operatorname{im} A$ сужение линейного оператора A на $(\ker A)^\perp$, а через $\mathcal{A}(u) : (\ker A)^\perp \rightarrow \operatorname{im} A$ — сужение линейного оператора $A + (I - \Pi)B[u]$ на $(\ker A)^\perp$. Тогда

$$\mathcal{A}(u) = \widehat{A} + O(\|u\|),$$

причем оператор $\widehat{A}$ обратим, откуда и, например, из [3, лемма A.6] следует, что для u достаточно близких к 0 оператор $\mathcal{A}(u)$ тоже обратим, причем

$$(\mathcal{A}(u))^{-1} = \widehat{A}^{-1} + O(\|u\|)$$

при $u \rightarrow 0$. Тогда первое уравнение в (2.19) при любом $v_2 \in \mathbb{R}^p$ имеет единственное решение

$$v_1 = (\mathcal{A}(u))^{-1}(w - (I - \Pi)B[u]v_2) = O(\|w\|) + O(\|u\|)v_2 \quad (2.20)$$

при $u \rightarrow 0$ и $w \rightarrow 0$.

Подставляя выражение из (2.20) во второе уравнение в (2.19), получаем

$$\mathcal{B}(u)v_2 = \Pi O(\|u\|\|w\|), \quad (2.21)$$

где линейный оператор $\mathcal{B}(u) : \ker A \rightarrow (\operatorname{im} A)^\perp$ в левой части имеет вид

$$\mathcal{B}(u) = \widehat{B}(u) + O(\|u\|^2).$$

Из невырожденности $\widehat{B}(\bar{v})$, снова применяя [3, лемма A.6], получаем существование $\bar{\varepsilon} > 0$ и $\bar{\delta} > 0$ таких, что для $u \in K_{\bar{\varepsilon}, \bar{\delta}}(0, \bar{v})$ оператор $\mathcal{B}(u)$ невырожден, причем

$$(\mathcal{B}(u))^{-1} = (\widehat{B}(u))^{-1} + O(1) = O(\|u\|^{-1})$$

при $u \rightarrow 0$. Отсюда следует, что (2.21), а значит, и второе уравнение в (2.19) имеет единственное решение

$$v_2 = O(\|w\|) \quad (2.22)$$

при $u \rightarrow 0$ и $w \rightarrow 0$.

Тем самым показано, что для $u \in K_{\bar{\varepsilon}, \bar{\delta}}(0, \bar{v})$ уравнение (2.18) имеет единственное решение $v = v_1 + v_2$, т.е. матрица $A + B[u]$ невырождена, причем из (2.20) и (2.22) вытекает оценка (2.17). $\square$

Для NM, согласно (2.2),

$$(A + B[u])v = -\frac{1}{2}(A + B[u])u - \frac{1}{2}Au,$$

и в предположениях леммы 2.1 для $u \in K_{\varepsilon, \bar{\delta}}(0, \bar{v})$

$$v = -\frac{1}{2}u - \frac{1}{2}(A + B[u])^{-1}Au = -\frac{1}{2}u + O(\|Au\|) = -\frac{1}{2}u + O(\|u_1\|)$$

при $u \rightarrow 0$. Поэтому, с учетом (2.14), если $u^0 \in K_{\varepsilon, \delta}(0, \bar{v})$ с достаточно малыми $\varepsilon > 0$ и $\delta > 0$, то

$$u^k + 2v^k = O(\|u_1^k\|) = O(\|u^k\|^3)$$

при $k \rightarrow \infty$.

Для метода LM, аналогично построениям выше, для $u \in K_{\varepsilon, \bar{\delta}}(0, \bar{v})$ выводится следующее обобщение оценки (2.7):

$$v = -\frac{1}{2}u + O(\|u\|^{2\theta-1}) + O(\|Au\|) = -\frac{1}{2}u + O(\|u\|^{2\theta-1}) + O(\|u_1\|)$$

при $u \rightarrow 0$, откуда и из (2.14) вытекает оценка

$$u^k + 2v^k = O(\|u^k\|^{2\theta-1}) + O(\|u_1^k\|)$$

при $k \rightarrow \infty$. Тогда, согласно (2.14),

$$u^k + 2v^k = O(\|u^k\|^{\min\{2\theta-1, 3\}}) \quad (2.23)$$

и, в частности, при $\theta \geq 2$

$$u^k + 2v^k = O(\|u^k\|^3) \quad (2.24)$$

при $k \rightarrow \infty$, как и для NM.

Наконец, для метода LPN аналогичным образом для $u \in K_{\varepsilon, \bar{\delta}}(0, \bar{v})$ выводится оценка

$$v = -\frac{1}{2}u + O(\|u\|^2) + O(\|Au\|) = -\frac{1}{2}u + O(\|u\|^2) + O(\|u_1\|)$$

при $u \rightarrow 0$, а значит, согласно (2.14), и оценка

$$u^k + 2v^k = O(\|u^k\|^2) \quad (2.25)$$

при $k \rightarrow \infty$.

Предложение 2.2. В предположениях предложения 2.1 найдутся $\varepsilon > 0$ и $\delta > 0$ такие, что для $u^0 \in K_{\varepsilon, \delta}(0, \bar{v})$, для уравнения (1.1) с линейно-квадратичным отображением Φ из (2.1) справедливо следующее:

- NM с подзадачей (1.2) однозначно определяет последовательности $\{v^k\}$ шагов и $\{u^k\}$ приближений, и для соответствующих экстраполированных приближений выполняется оценка (2.24) при $k \rightarrow \infty$;

- метод LM с подзадачей (1.3), использующей $\sigma(u) = O(\|\Phi(u)\|^\theta)$ с фиксированным $\theta \geq 3/2$, однозначно определяет последовательности $\{v^k\}$ шагов и $\{u^k\}$ приближений, и для соответствующих экстраполированных приближений выполняется оценка (2.23) при $k \rightarrow \infty$;
- метод LPN с подзадачей (1.5) определяет последовательности $\{v^k\}$ шагов и $\{u^k\}$ приближений, и для любых таких последовательностей, для соответствующих экстраполированных приближений выполняется оценка (2.25) при $k \rightarrow \infty$.

Полученные в этом разделе результаты существенно используют линейно-квадратичный характер отображения Φ . За пределами линейно-квадратичного случая оценка (2.24) для метода LM с $\theta = 2$, и даже для NM, может нарушаться, в том числе и при $u \in \ker A$, и эффект от экстраполяции для MN и методов LM и LPN при этом может выравниваться, что будет продемонстрировано в следующем разделе.

3. Глобализованные алгоритмы и численные результаты

Алгоритм 1 ниже, глобализующий сходимость NM, использует страховочные шаги градиентного метода для функции $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$,

$$\varphi(u) = \frac{1}{2} \|\Phi(u)\|^2, \quad (3.1)$$

градиент которой имеет вид

$$\varphi'(u) = (\Phi'(u))^\top \Phi(u). \quad (3.2)$$

Такие шаги используются в тех случаях, когда направление NM либо не удастся определить, либо оно оказывается «слишком длинным» в смысле нарушения теста (3.3). Этот алгоритм был предложен в [12, алгоритм 3.1], где были установлены свойства его глобальной сходимости.

Алгоритм 1. Фиксируем параметры $C > 0$, $\tau > 0$, $\rho \in (0, 1)$ и $\varkappa \in (0, 1)$. Выбираем $u^0 \in \mathbb{R}^p$ и полагаем $k = 0$.

1. Если $\Phi(u^k) = 0$, стоп.
2. Вычисляем v^k как решение линейного уравнения (1.2). Если v^k не удастся вычислить, или v^k нарушает неравенство

$$\|v^k\| \leq \max\{C, 1/\|\Phi(u^k)\|^\tau\}, \quad (3.3)$$

переходим к шагу 4.

3. Полагаем $\alpha = 1$. Если выполняется неравенство

$$\|\Phi(u^k + \alpha v^k)\| \leq (1 - \rho\alpha)\|\Phi(u^k)\|, \quad (3.4)$$

полагаем $\alpha_k = \alpha$ и переходим к шагу 6. В противном случае заменяем α на $\varkappa\alpha$ и проверяем снова неравенство (3.4) до тех пор, пока оно не выполнится, после чего полагаем $\alpha_k = \alpha$ и переходим к шагу 6.

4. Полагаем $v^k = -\varphi'(u^k)$, где функция φ определена в (3.1) (см. (3.2)). Если $v^k = 0$, стоп.
5. Полагаем $\alpha = 1$. Если выполняется неравенство Армихо

$$\varphi(u^k + \alpha v^k) \leq \varphi(u^k) - \rho \alpha \|v^k\|^2, \quad (3.5)$$

полагаем $\alpha_k = \alpha$. В противном случае заменяем α на $\varkappa \alpha$ и проверяем снова неравенство (3.5) до тех пор, пока оно не выполнится, после чего полагаем $\alpha_k = \alpha$.

6. Полагаем $u^{k+1} = u^k + \alpha_k v^k$, увеличиваем k на 1 и переходим к шагу 1.

В отличие от NM, метод LM может быть глобализован «чистым» образом, без каких-либо гибридных элементов и, в частности, без необходимости использования каких-либо страховочных шагов.

Алгоритм 2. Фиксируем параметры $\theta > 0$, $\rho \in (0, 1)$ и $\varkappa \in (0, 1)$. Выбираем $u^0 \in \mathbb{R}^p$ и полагаем $k = 0$.

1. Если $\Phi(u^k) = 0$, стоп.
2. Полагаем $\sigma(u^k) = \|\Phi(u^k)\|^\theta$, и вычисляем v^k как решение линейного уравнения (1.4). Если $v^k = 0$, стоп.
3. Полагаем $\alpha = 1$. Если выполняется неравенство

$$\varphi(u^k + \alpha v^k) \leq \varphi(u^k) - \frac{1}{2} \rho \sigma(u^k) \alpha \|v^k\|^2, \quad (3.6)$$

где функция φ определена в (3.1), полагаем $\alpha_k = \alpha$. В противном случае заменяем α на $\varkappa \alpha$ и проверяем снова неравенство (3.6) до тех пор, пока оно не выполнится, после чего полагаем $\alpha_k = \alpha$.

4. Полагаем $u^{k+1} = u^k + \alpha_k v^k$, увеличиваем k на 1 и переходим к шагу 1.

Алгоритм 2 (с $\theta = 2$) и анализ его глобальной сходимости были предложены в [16]; см. также [7, алгоритм 6.2, теорема 6.2].

Метод LPN также допускает «чистую» глобализацию сходимости, разработанную в [17, алгоритм 1, теорема 3.1]. Алгоритм 3 ниже основан на одномерном поиске для функции $f : \mathbb{R}^p \rightarrow \mathbb{R}$,

$$f(u) = \|\Phi(u)\|_\infty. \quad (3.7)$$

Алгоритм 3. Фиксируем параметры $\rho \in (0, 1)$ и $\varkappa \in (0, 1)$. Выбираем $u^0 \in \mathbb{R}^p$ и полагаем $k = 0$.

1. Если $\Phi(u^k) = 0$, стоп.
2. Вычисляем (v^k, γ_k) как решение задачи линейного программирования (1.5), в которой используется ∞ -норма. Если для величины

$$\Delta_k = -f(u^k)(1 - \gamma_k f(u^k)),$$

где функция f определена в (3.7), выполняется $\Delta_k = 0$, стоп.

3. Полагаем $\alpha = 1$. Если выполняется неравенство

$$f(u^k + \alpha v^k) \leq f(u^k) + \rho \alpha \Delta_k, \quad (3.8)$$

полагаем $\alpha_k = \alpha$. В противном случае заменяем α на $\kappa \alpha$ и проверяем снова неравенство (3.8) до тех пор, пока оно не выполнится, после чего полагаем $\alpha_k = \alpha$.

4. Полагаем $u^{k+1} = u^k + \alpha_k v^k$, увеличиваем k на 1 и переходим к шагу 1.

В экспериментах, результаты которых представлены ниже, алгоритм 3 был также снабжен всеми предложенными в [17, разд. 5] модификациями, улучшающими его поведение, кроме немоного одномерного поиска. В процедурах одномерного поиска использовались параметры $\rho = 0.01$ и $\kappa = 0.5$. В алгоритме 1 использовались $C = 10^7$ и $\tau = 2$.

Запуск завершался с успехом в случае выполнения для некоторого $k \leq 100$ условия

$$\|\Phi(u^k)\| \leq 10^{-8},$$

либо условия

$$\min\{\|\Phi(u^k)\|, \|\Phi(u^{k-1} + 2v^{k-1})\|\} \leq 10^{-8} \quad (3.9)$$

в случае использования экстраполяции. В остальных случаях запуск считался неудачным, как и при выполнении для некоторого k условия

$$\|(\Phi'(u^k))^T \Phi(u^k)\| \leq 10^{-20}$$

для алгоритмов 1 и 2, и

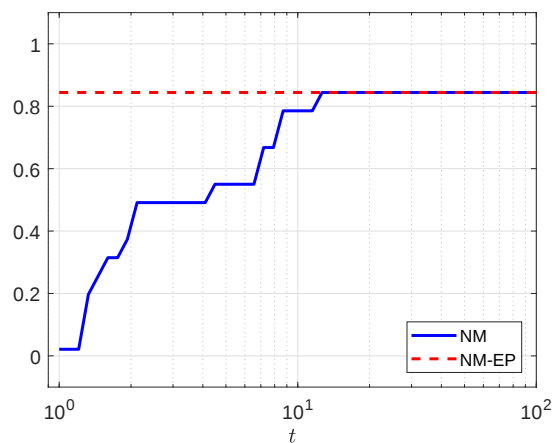
$$|\Delta_k| \leq 10^{-16}$$

для алгоритма 3, либо условия $\alpha \|v^k\| \leq 10^{-16}$ в процессе одномерного поиска на соответствующих шагах алгоритмов.

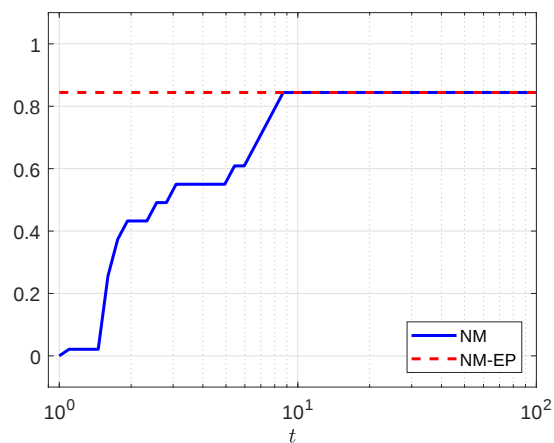
В алгоритме 2 правило выбора параметра регуляризации было модифицировано следующим образом, исключая слишком большие значения этого параметра вдали от решений: $\sigma(u) = \min\{1, \|\Phi(u)\|^\theta\}$, причем использовалось значение $\theta = 2$.

Эксперимент проводился на всех задачах из тестового набора MGH [18], для которых известны точные решения, за исключением трех задач с линейными отображениями Φ . Некоторые из выбранных таким образом задач имеют больше уравнений, чем переменных, и в таких случаях лишние уравнения удалялись (всегда линейные, за исключением теста Beale, в котором выбор уравнения для удаления не влияет существенным образом на поведение алгоритмов). Информация о полученном в результате тестовом наборе содержится в [12, таблица 1]. Далее, выбранные задачи трансформировались посредством техники из [19] таким образом, чтобы известное решение $\bar{u}$ оставалось решением модифицированного уравнения, но при этом выполнялось $\text{rank } \Phi'(\bar{u}) = p - 1$ (за исключением трех задач Powell singular, Extended Powell singular, и Variably dimensioned, которые имеют известные особые решения в своей оригинальной форме). Задачи Variably dimensioned и Brown almost-linear использовались в двух версиях: с $p = 10$ и с $p = 500$.

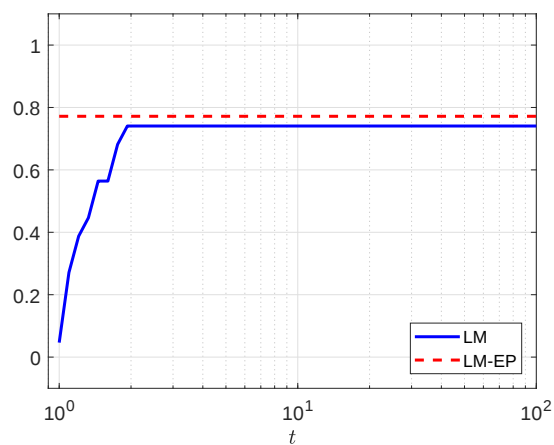
Запуски сравниваемых алгоритмов осуществлялись из 100 одинаковых случайных начальных точек, равномерно распределенных в l_∞ -шаре (гиперкубе) радиуса 1 с центром в решении.



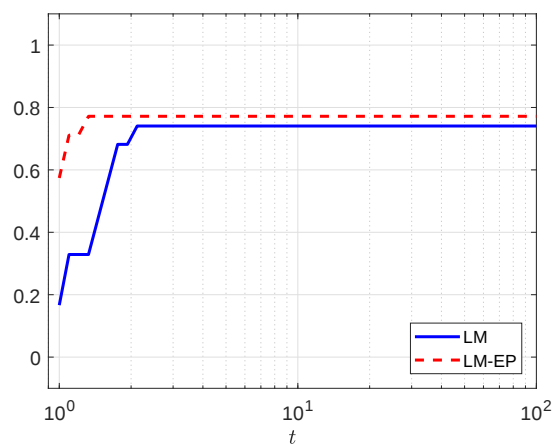
(a) Алгоритм 1, по количеству итераций



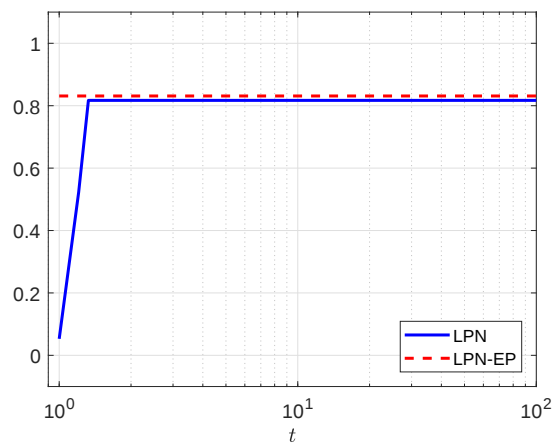
(b) Алгоритм 1, по времени запуска



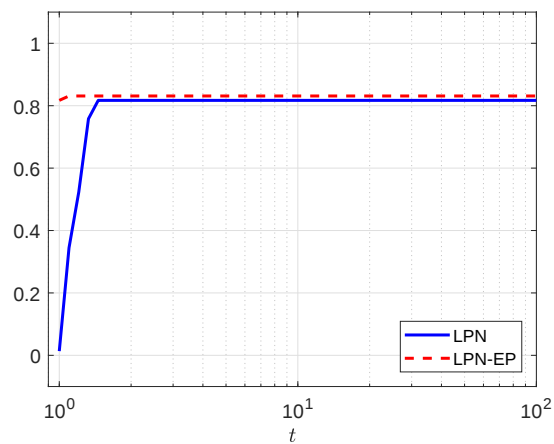
(c) Алгоритм 2, по количеству итераций



(d) Алгоритм 2, по времени запуска

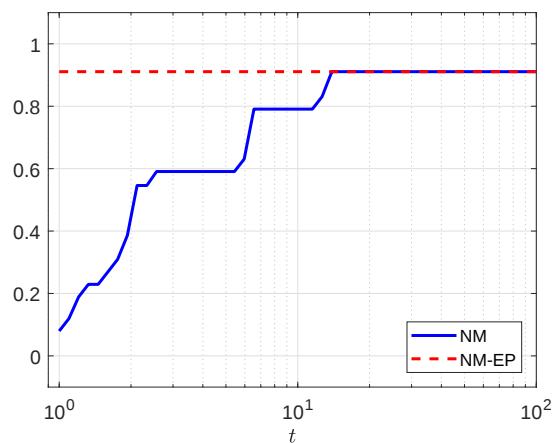


(e) Алгоритм 3, по количеству итераций

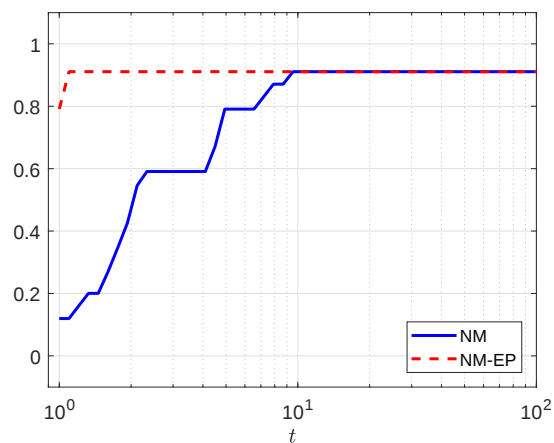


(f) Алгоритм 3, по времени запуска

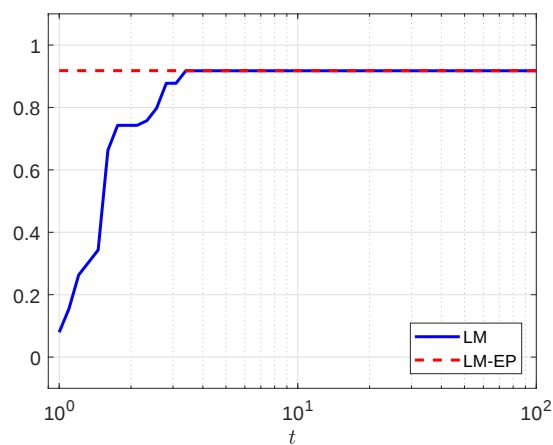
Рис. 2. Сравнение на наборе MGH



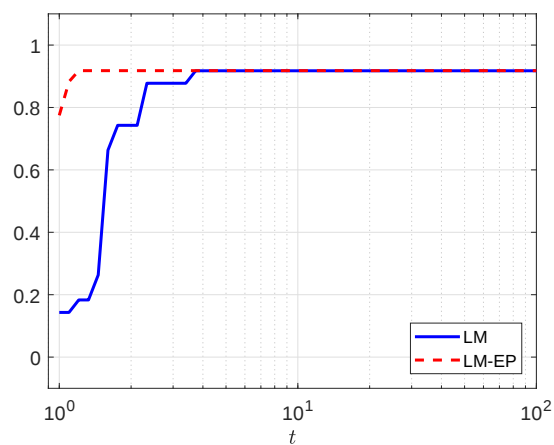
(a) Алгоритм 1, по количеству итераций



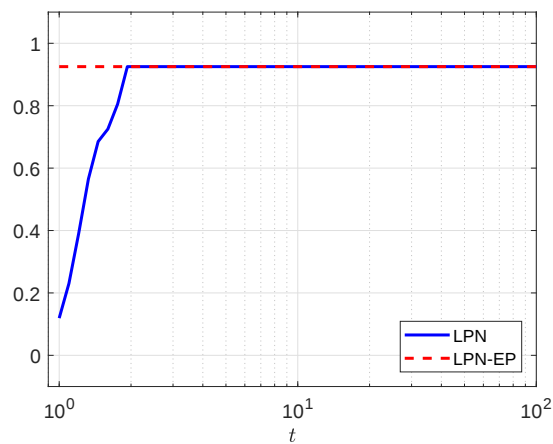
(b) Алгоритм 1, по времени запуска



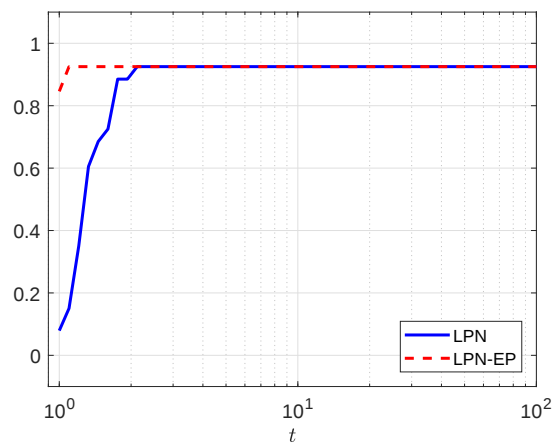
(c) Алгоритм 2, по количеству итераций



(d) Алгоритм 2, по времени запуска



(e) Алгоритм 3, по количеству итераций



(f) Алгоритм 3, по времени запуска

Рис. 3. Сравнение на наборе Misc

Результаты представлены в форме так называемых performance profiles, изначально предложенных в [20], и адаптированных в [21] на случай многократных запусков для каждой тестовой задачи, с использованием средних характеристик эффективности и процентов успешных запусков. На рис. 2 для каждого алгоритма показан график функции, значение которой в t по существу соответствует доле задач в наборе, для которых средний показатель эффективности для данного алгоритма был хуже наилучшего не более, чем в t раз. При этом считается, что результат любого неудачного запуска в бесконечное число раз хуже любого другого. Значения функции на правом конце (при больших t) соответствует доле успешных запусков алгоритма, а на левом — доле задач, для которых его средний показатель эффективности был наилучшим. В данном сравнении в роли показателей эффективности использовались количество итераций и время запуска.

Представленные результаты демонстрируют, что эффект от экстраполяции для метода LM слабее, чем для NM, а для метода LPN слабее, чем для LM.

Эксперимент на тестовом наборе MGN был дополнен аналогичным экспериментом на наборе Misc(ellaneous), состоящим из небольших систем нелинейных уравнений с особыми решениями, собранных из различных источников (первые 11 тестов взяты из [22, разд. 5]). А именно, набор состоит из следующих задач.

Misc 1. Это уравнение (1.1) из примера 2.1, с $p = 1$, $\Phi(u) = u^2$; изолированное особое решение $\bar{u} = 0$.

Misc 2. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1, 2u_2^2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 3. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2 - u_2^2, u_1 u_2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 4. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 + u_2, -u_1 - u_2 + u_1 u_2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 5. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2, u_2^2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 6. Уравнение (1.1) с $p = 2$, $\Phi(u) = (2(u_1 - u_2^2), u_2^2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 7. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1(u_1^2 + u_2), u_2(1 + u_2))$; изолированное особое решение $\bar{u} = (0, 0)$ (имеются также неособые решения $(0, -1)$, $(1, -1)$, $(-1, -1)$).

Misc 8. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 + u_2^2, 3u_1 u_2 / 2 + u_2^2(1 + u_2))$; изолированное особое решение $\bar{u} = (0, 0)$ (имеется также неособое решение $(-4, 2)$).

Misc 9. Уравнение (1.1) с $p = 3$, $\Phi(u) = (u_1 + u_2 + u_3 - 1, u_1^3 / 5 + u_2^2 / 2 - u_3 + u_3^2 / 2 + 1/2, u_1 + u_2 + u_3^2 / 2 - 1/2)$; изолированное особое решение $\bar{u} = (0, 0, 1)$ (имеется также другое особое решение $(-5/2, 5/2, 1)$).

Misc 10. Уравнение (1.1) с $p = 3$, $\Phi(u) = (u_1 + u_1 u_2 + u_2^2, -2u_1 + u_1^2 + u_2^2, u_1 + u_3^2)$; изолированное особое решение $\bar{u} = (0, 0, 0)$.

Misc 11 (дискретизованное H -уравнение Чандрасекара). Уравнение (1.1) с выбираемым значением p (в экспериментах использовалось $p = 5$),

$$\Phi_i(u) = u_i - \frac{i}{2n} u_i \sum_{j=1}^n \frac{u_j}{i+j} - 1, \quad i \in \{1, \dots, p\};$$

для $p = 5$ приближенное изолированное особое решение

$$\bar{u} = (1.359753, 1.688205, 2.005894, 2.318350, 2.627810).$$

Misc 12. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 + au_2^2/2, u_2^2/2)$ с вещественным параметром a (в экспериментах использовалось $a = \sqrt{15}$); изолированное особое решение $\bar{u} = (0, 0)$.

Misc 13. Это уравнение (1.1) из примера 2.2, с $p = 2$, $\Phi(u) = (u_1 + au_2^2/2, u_1u_2 + u_2^2/2)$ с вещественным параметром a (в экспериментах использовалось $a = 1$, как и в примере 2.2); изолированное особое решение $\bar{u} = (0, 0)$ (имеется также неособое решение $(-1/(2a), 1/a)$).

Misc 14. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2 + u_2^3, u_1u_2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 15. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 + u_1u_2 + u_2^2, u_1^2 - 2u_1 + u_2^2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 16. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2 - u_2, u_1^2 + u_2^2)$; изолированное особое решение $\bar{u} = (0, 0)$.

Misc 17. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2 - u_2^2, 3u_1^2 - 3u_2^2)$; неизолированное решение $\bar{u} = (0, 0)$.

Misc 18. Уравнение (1.1) с числом переменных $p = 5$, и с числом уравнений, равным 4, $\Phi(u) = (u_1 + u_2 + u_3^2 + u_4^2 + u_5^2 - 2, u_1 - u_2 + u_3^2 + u_4^2 + u_5^2, -u_3^2 + u_4^2 + u_5^2, u_3^2 + u_4^2 - u_5^2)$; неизолированное решение $\bar{u} = (1, 1, 0, 0, 0)$. Для этой задачи NM работает как метод Гаусса–Ньютона, использующий решение минимальной нормы итерационного уравнения (1.2).

Misc 19. Уравнение (1.1) с $p = 3$ из [1, приложение] с параметрами $k = 1$, $m = 1$; изолированное особое решение $\bar{u} = (0, 0, 0)$.

Misc 20. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1^2 + u_2^2 - 2, e^{u_1-1} + u_2^2 - 2)$; неособое решение $\bar{u} = (1, -1)$, которое трансформировалось в особое посредством упомянутой выше в связи с набором MGN процедуры из [19] (у исходного уравнения имеется также другое неособое решение $(1, 1)$).

Misc 21. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 - 10, u_1u_2 - 5000)$; неособое решение $\bar{u} = (5000, 10)$, которое трансформировалось в особое процедурой из [19].

Misc 22. Уравнение (1.1) с $p = 2$, $\Phi(u) = (e^{u_1+u_2^2} - 1, u_1 + u_2 - \sin(3(u_1 + u_2)))$; изолированное особое решение $\bar{u} = (0, 0)$.

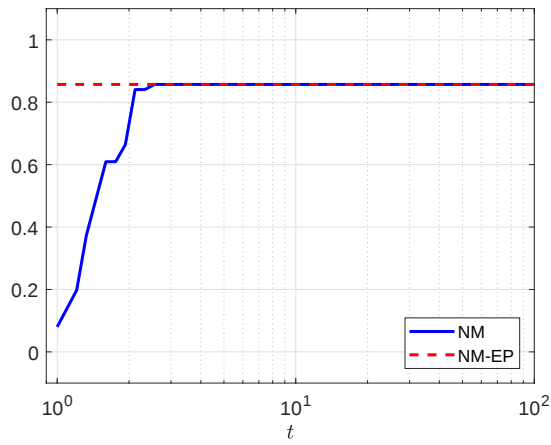
Misc 23. Уравнение из Misc 22, к решению которого дополнительно применялась процедура из [19].

Misc 24. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1, 1 - u_1/(u_1 + 0.1) + 2u_2^2)$; неособое решение $\bar{u} = (5000, 10)$, которое трансформировалось в особое процедурой из [19].

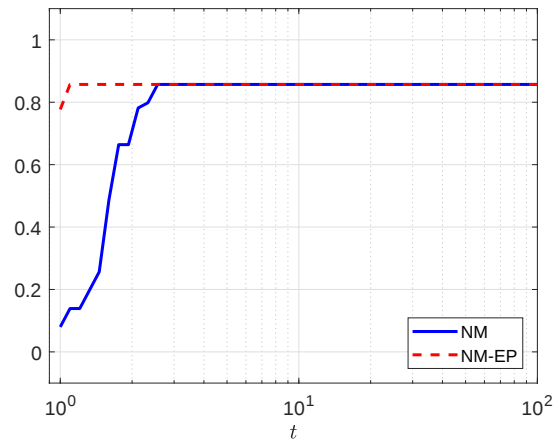
Misc 25. Уравнение (1.1) с $p = 2$, $\Phi(u) = (u_1 + u_2^2, 2(u_1 - 1)u_2)$; неособое решение $\bar{u} = (0, 0)$, которое трансформировалось в особое процедурой из [19].

Результаты эксперимента на наборе Misc представлены на рис. 3, и выводы из них аналогичны выводам для MGN: эффект от экстраполяции для NM больше, чем для метода LM, а для метода LM больше, чем для метода LPN.

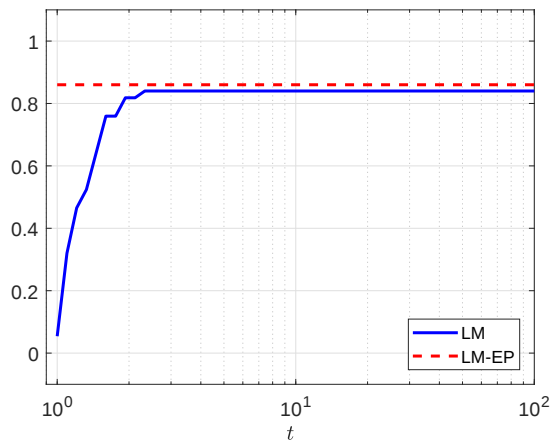
Дальнейший эксперимент был направлен на демонстрацию того, что за пределами линейно-квадратичного случая эффект от экстраполяции для NM и методов LM и LPN выравнивается, что согласуется с рассмотрениями в разд. 3. Для этого из наборов MGN и Misc были отобраны задачи, в которых отображение Φ не является линейно-квадратичным, а именно, Freudenstein and Roth, Beale, Helical valley, Gulf research and development, Box three-dimensional, Biggs EXP6, Trigonometric, Brown almost-linear;



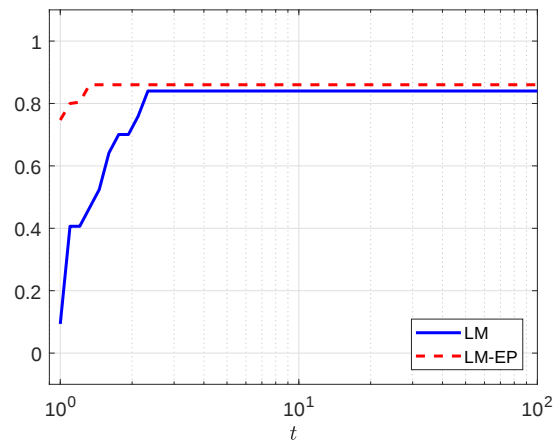
(a) Алгоритм 1, по количеству итераций



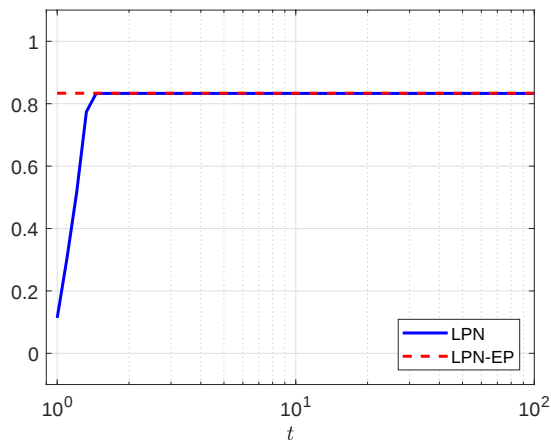
(b) Алгоритм 1, по времени запуска



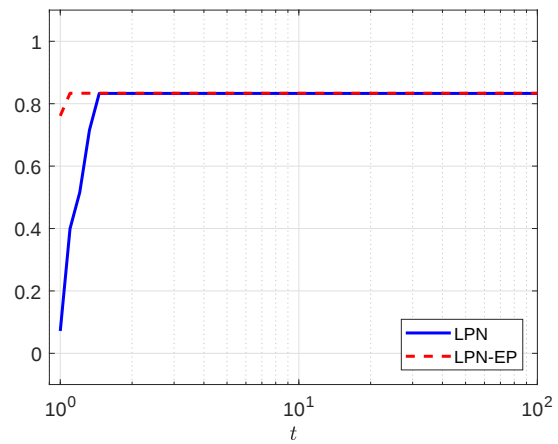
(c) Алгоритм 2, по количеству итераций



(d) Алгоритм 2, по времени запуска



(e) Алгоритм 3, по количеству итераций



(f) Алгоритм 3, по времени запуска

Рис. 4. Сравнение на задачах, не являющихся линейно-квадратичными

Misc 7-9, Misc 14, Misc 20, Misc 22-24. Результаты на рис. 4 показывают, что на данной выборке тестов сравниваемые алгоритмы имеют схожий эффект от экстраполяции, хотя для метода LPN эффект все же меньше, чем для других алгоритмов.

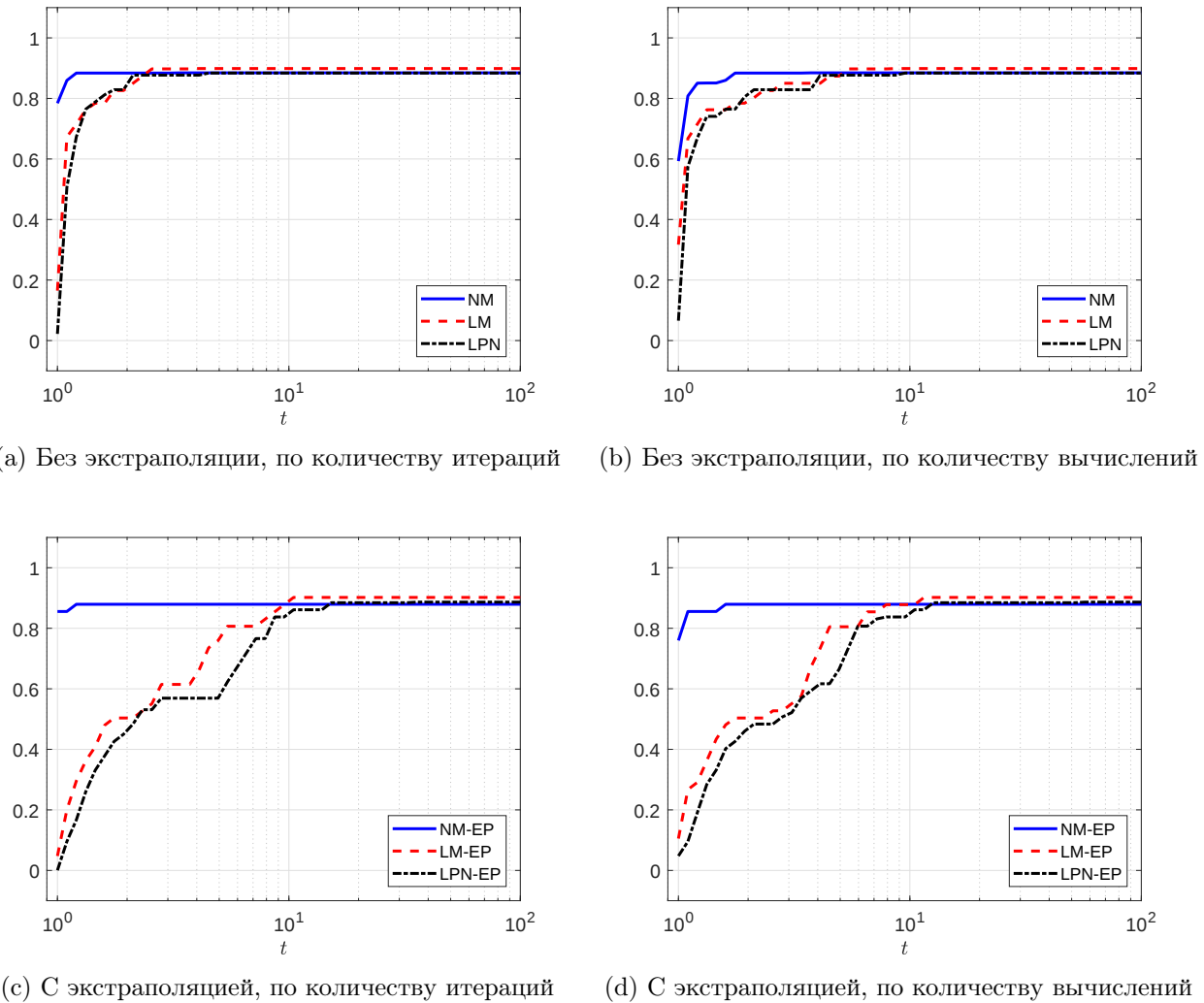


Рис. 5. Сравнение на полном тестовом наборе

На рис. 5 приведенные выше результаты дополнены сравнением трех алгоритмов 1, 2 и 3 между собой. Сравнение по времени здесь не приводится, поскольку метод LPN для рассматриваемых здесь уравнений без дополнительных ограничений неизбежно проигрывает по этому показателю, так как решение его подзадач (задач линейного программирования) является гораздо более затратным, чем решение подзадач NM и метода LM (систем линейных уравнений). Вместо этого приводится сравнение по другой важной характеристике, а именно, по количеству вычислений Φ . Важное наблюдение состоит в том, что по этой характеристике методы с экстраполяцией могут быть существенно менее эффективны, чем по количеству итераций, поскольку на каждой итерации требуют по крайней мере одного лишнего вычисления Φ для проверки условия останова (3.9).

Для алгоритмов с одномерным поиском эффект от экстраполяции во многом определяется асимптотическим принятием полного шага, поскольку эта техника основана на характере сходимости к особым решениям именно полношаговых ньютоновских методов.

Таблица 1. Процент среднего количества последних полных шагов

Тест	NM	LM	LPN
Rosenbrock	98.67	49.70	70.87
Freudenstein and Roth	99.44	98.95	99.75
Brown badly scaled	32.14	NaN	100.00
Beale	99.08	100.00	99.93
Helical valley	95.77	33.95	49.63
Gulf research and development	24.05	13.64	17.39
Box three-dimensional	90.71	28.46	44.85
Powell singular	100.00	100.00	100.00
Wood	99.23	94.97	98.16
Biggs EXP6	83.20	4.20	6.91
Extended Rosenbrock	93.48	26.68	87.02
Extended Powell singular	100.00	100.00	100.00
Variably dimensioned, $p = 10$	100.00	100.00	99.82
Variably dimensioned, $p = 500$	100.00	100.00	62.91
Trigonometric	NaN	53.00	NaN
Brown almost-linear, $p = 10$	98.24	100.00	95.34
Brown almost-linear, $p = 500$	100.00	98.27	65.41
Misc 1	100.00	100.00	100.00
Misc 2	100.00	100.00	100.00
Misc 3	100.00	100.00	100.00
Misc 4	98.65	100.00	100.00
Misc 5	100.00	100.00	100.00
Misc 6	100.00	100.00	100.00
Misc 7	97.89	99.74	99.18
Misc 8	98.57	100.00	100.00
Misc 9	97.75	100.00	98.77
Misc 10	98.50	100.00	100.00
Misc 11	100.00	100.00	100.00
Misc 12	100.00	100.00	100.00
Misc 13	96.75	99.33	99.50
Misc 14	99.86	100.00	100.00
Misc 15	98.43	100.00	100.00
Misc 16	97.94	100.00	100.00
Misc 17	87.50	100.00	100.00
Misc 18	99.76	100.00	100.00
Misc 19	89.61	83.42	86.06
Misc 20	89.69	55.38	79.58
Misc 21	100.00	NaN	13.21
Misc 22	98.07	100.00	99.07
Misc 23	13.29	18.83	13.12
Misc 24	96.56	52.57	57.29
Misc 25	76.92	73.04	84.40

В этих экспериментах единственный тест, на котором иногда наблюдались случаи неполного последнего шага (для всех трех алгоритмов), — это Misc 23, в котором $\Phi'(\bar{u}) = 0$, но 2-регулярности не имеет места ни по какому направлению. Таблица 1 содержит информацию о проценте среднего количества последних полных шагов на один успешный запуск, от общего количества итераций.

Интересное наблюдение состоит в том, что по времени алгоритм 2 показывает лучший результат на примерно 70% задач (по указанным выше причинам эти результаты здесь не приводятся), а по количеству вычислений Φ и особенно по количеству итераций такой процент гораздо ниже; см. рис. 5a, 5b. Возможное объяснение состоит в том, что по средним значениям принимаемых алгоритмами параметров длины шага алгоритм 2 также оказывается лучшим на сравнимой доле задач, как и по среднему количеству последних полных шагов; см. таблицу 1.

В заключение кратко опишем результаты эксперимента с методом LM, в котором в правиле для параметра регуляризации $\sigma(u) = \min\{\bar{\sigma}, \|\Phi(u)\|^2\}$ использовались разные значения $\bar{\sigma}$, а именно, наряду с 1, рассматривались значения 10^{-1} , 10^{-5} и 10^{-10} . На данных тестовых наборах версии метода LM с меньшими значениями $\bar{\sigma}$ (что в принципе позволяет алгоритму 2 вдали от решений генерировать и принимать более длинные шаги, оказываются несколько более эффективными, но при этом несколько менее робастными, однако разница невелика. Вместе с тем, выбор $\bar{\sigma} = 1$ значительно проигрывает по времени рассматриваемым альтернативам.

References

- [1] A. Griewank, *Analysis and Modification of Newton's Method at Singularities*, PhD Thesis, Australian National University, Canberra, 1980.
- [2] A. Griewank, "On solving nonlinear equations with simple singularities or nearly singular solutions", *SIAM Review*, **27** (1985), 537–563.
- [3] A. F. Izmailov, M. V. Solodov, *Newton-Type Methods for Optimization and Variational Problems*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, 2014, 573 pp.
- [4] K. Levenberg, "A method for the solution of certain non-linear problems in least squares", *Quarterly of Applied Mathematics*, **2** (1944), 164–168.
- [5] D. W. Marquardt, "An algorithm for least squares estimation of non-linear parameters", *SIAM J.*, **11** (1963), 431–441.
- [6] Дж. Дэннис, Р. Шнабель, *Численные методы безусловной оптимизации и решение нелинейных уравнений*, Мир, М., 1988; англ. ориг.: J. E. Dennis, Jr., R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, 1983.
- [7] A. Fischer, A. F. Izmailov, M. V. Solodov, "The Levenberg–Marquardt method: an overview of modern convergence theories and more", *Computational Optimization and Applications*, **89**:1 (2024), 33–67.
- [8] F. Facchinei, A. Fischer, M. Herrich, "An LP-Newton method: Nonsmooth equations, KKT systems, and nonisolated solutions", *Mathematical Programming*, **146** (2014), 1–36.
- [9] А. Ф. Измаилов, А. А. Третьяков, *2-регулярные решения нелинейных задач. Теория и численные методы*, Физматлит, М., 1999. [A. F. Izmailov, A. A. Tret'yakov, *2-Regular Solutions of Nonlinear Problems. Theory and Numerical Methods*, Fizmatlit Publ., Moscow, 1999 (In Russian)].
- [10] A. F. Izmailov, A. S. Kurennoy, M. V. Solodov, "Critical solutions of nonlinear equations: local attraction for Newton-type methods", *Mathematical Programming*, **167**:2 (2018), 355–379.
- [11] A. Fischer, A. F. Izmailov, M. Jelitte, "Behavior of Newton-type methods near critical solutions of nonlinear equations with semismooth derivatives", *Journal of Optimization Theory and*

- Applications*, 2023, New Trends in Optimization, Control, and Their Applications: A Special Issue Dedicated to Boris Sh. Mordukhovich.
- [12] A. Fischer, A. F. Izmailov, M. V. Solodov, “Accelerating convergence of the globalized Newton method to critical solutions of nonlinear equations”, *Computational Optimization and Applications*, **78**:1 (2021), 273–286.
 - [13] A. Fischer, A. F. Izmailov, M. V. Solodov, “Unit stepsize for the Newton method close to critical solutions”, *Mathematical Programming*, **187**:1–2 (2021), 697–721.
 - [14] A. Fischer, A. F. Izmailov, M. Jelitte, “Newton-type methods near critical solutions of piecewise smooth nonlinear equations”, *Computational Optimization and Applications*, **80**:2 (2021), 587–615.
 - [15] A. Griewank, “Starlike domains of convergence for Newton’s method at singularities”, *Numerische Mathematik*, **35** (1980), 95–111.
 - [16] A. Fischer, P. K. Shukla, “A Levenberg–Marquardt algorithm for unconstrained multicriteria optimization”, *Operations Research Letters*, **36** (2008), 643–646.
 - [17] A. Fischer, M. Herrich, A. F. Izmailov, M. V. Solodov, “A globally convergent LP-Newton method”, *SIAM J. on Optimization*, **26** (2016), 2012–2033.
 - [18] J. J. Moré, B. S. Garbow, K. E. Hillstom, “Testing unconstrained optimization software”, *ACM Transactions of Mathematical Software*, **7** (1981), 17–41.
 - [19] R. B. Schnabel, P. D. Frank, “Tensor methods for nonlinear equations”, *SIAM J. on Numerical Analysis*, **21** (1984), 815–843.
 - [20] E. D. Dolan, J. J. More, “Benchmarking optimization software with performance profiles”, *Mathematical Programming*, **91** (2002), 201–213.
 - [21] A. F. Izmailov, M. V. Solodov, E. I. Uskov, “Combining stabilized SQP with the augmented Lagrangian algorithm”, *Computational Optimization and Applications*, **62** (2015), 405–429.
 - [22] А. Ф. Измаилов, “О новых реализациях 2-фактор-метода”, *Журнал вычислительной математики и математической физики*, **55**:6 (2015), 933–946; англ. пер.: A. F. Izmailov, “New implementations of the 2-factor method”, *Computational Mathematics and Mathematical Physics*, **55**:6 (2015), 922–934.

Информация об авторах

Измаилов Алексей Ферилович, доктор физико-математических наук, профессор кафедры исследования операций, Московский государственный университет им. М. В. Ломоносова, г. Москва, Российская Федерация. E-mail: izmaf@cs.msu.ru
ORCID: <https://orcid.org/0000-0001-9851-0524>

Усков Евгений Иванович, кандидат физико-математических наук, научный сотрудник, Тамбовский государственный университет им. Г. Р. Державина, г. Тамбов, Российская Федерация. E-mail: euskov@cs.msu.ru
ORCID: <https://orcid.org/0000-0002-3639-0317>

Конфликт интересов отсутствует.

Для контактов:

Измаилов Алексей Ферилович
 E-mail: izmaf@cs.msu.ru

Поступила в редакцию 30.05.2024 г.
 Поступила после рецензирования 10.09.2024 г.
 Принята к публикации 22.11.2024 г.

Information about the authors

Alexey F. Izmailov, Doctor of Physical and Mathematical Sciences, Professor of the Operations Research Department, Lomonosov Moscow State University, Moscow, Russian Federation. E-mail: izmaf@cs.msu.ru
ORCID: <https://orcid.org/0000-0001-9851-0524>

Evgeniy I. Uskov, Candidate of Physical and Mathematical Sciences, Researcher, Derzhavin Tambov State University, Tambov, Russian Federation. E-mail: euskov@cs.msu.ru
ORCID: <https://orcid.org/0000-0002-3639-0317>

There is no conflict of interests.

Corresponding author:

Alexey F. Izmailov
 E-mail: izmaf@cs.msu.ru

Received 30.05.2024
 Reviewed 10.09.2024
 Accepted for press 22.11.2024

SCIENTIFIC ARTICLE

© S. V. Ludkovsky, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-425-439>

Decomposition of modules over generalized Dickson algebras

Sergey V. LUDKOVSKY

MIREA – Russian Technological University

78 Vernadsky Av., Moscow 119454, Russian Federation

Abstract. The article is devoted to modules over generalized Dickson algebras. These algebras are nonassociative and generally can be nonalternative. They compose an important class of algebras and an area in mathematics. Left, right and two-sided modules over generalized Dickson algebras are studied. Their structure and submodules are investigated. Bimodules with involution are scrutinized over generalized Dickson algebras with involution. Such bimodules have specific features caused by involution. Minimal submodules and decomposition of modules are investigated. In particular, cyclic submodules are studied.

Keywords: module, decomposition, generalized Dickson algebra, involution, ring

Mathematics Subject Classification: 15A04, 15A69, 16D70, 17A05, 17A36.

For citation: Ludkovsky S.V. Decomposition of modules over generalized Dickson algebras. *Vestnik Rossiyskikh Universitetov. Matematika = Russian Universities Reports. Mathematics*, 29:148 (2024), 425–439. <https://doi.org/10.20310/2686-9667-2024-29-148-425-439>

НАУЧНАЯ СТАТЬЯ

© Людковский С.В., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-425-439>

УДК 512.553.1+512.553.8+512.54



Разложение модулей над обобщенными алгебрами Диксона

Сергей Викторович ЛЮДКОВСКИЙ

МИРЭА – Российский технологический университет

119454, Российская Федерация, г. Москва, просп. Вернадского, 78

Аннотация. Статья посвящена модулям над обобщенными алгебрами Диксона. Эти алгебры неассоциативны и в общем случае могут быть неальтернативными. Они составляют важный класс алгебр и раздел математики. В работе изучаются левые, правые и двусторонние модули над обобщенными алгебрами Диксона. Исследуется их структура и подмодули. Особое внимание уделено бимодулям с инволюцией над обобщенными алгебрами Диксона с инволюцией. Такие бимодули имеют специфические особенности, вызванные наличием инволюции. Исследуются минимальные подмодули и разложение модулей. В частности, изучаются циклические подмодули.

Ключевые слова: модуль, разложение, обобщенная алгебра Диксона, инволюция, кольцо

Для цитирования: Людковский С.В. Разложение модулей над обобщенными алгебрами Диксона // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 425–439. <https://doi.org/10.20310/2686-9667-2024-29-148-425-439> (In Engl., Abstr. in Russian)

Introduction

Dickson algebras compose a great class of nonassociative algebras (see [1, 2]). They are formed by induction using a doubling procedure of a smashed product (see [3–6] and references therein). This class of algebras is the generalization of the octonion (Cayley) algebra. There are wide-spread applications of Dickson algebras in the theory of Lie groups and algebras (see [7–11]) and their generalizations (see [12]), noncommutative mathematical analysis, noncommutative geometry (see [13, 14]), operator theory (see [15, 16]), PDE (see [17]), elementary particle physics and quantum field theory (see [18]). In the aforementioned areas naturally modules over Dickson algebras are very important, but they are only a little studied.

In this article left, right and two-sided modules over generalized Dickson algebras are studied. They are complicated in comparison with alternative algebras. Specific definitions and notations are given (see Definitions 1.1, 1.2, 1.3, 2.1, 2.2, Remark 1.1), because generalized Dickson algebras are neither associative nor alternative. Structure of modules and submodules over generalized Dickson algebras are investigated. For this purpose auxiliary Lemmas 1.1, 1.2, Corollaries 1.1, 1.2, Examples 1.1 and 2.1 are provided. Dickson algebras possess very important involution property. Therefore bimodules with involution are studied in Section 1. Bimodules with an involution are scrutinized in Theorems 1.1, 1.2, Corollary 1.3. For them necessary and sufficient conditions are elucidated. Identities in them are studied in Proposition 1.1. Subbimodules are investigated in Theorem 1.3 and Corollaries 1.4, 1.5. Relations between left, right and two-sided modules over Dickson algebras are given in Corollary 1.6 and Remark 1.3. Bimodules which are not bimodules with involution also are studied (see Proposition 1.2). Left subbimodules are investigated in Theorem 1.4, Proposition 1.3, Corollary 1.7. In particular, cyclic submodules are studied.

All main results of this paper are obtained for the first time.

1. Modules over generalized Dickson algebras

To avoid misunderstandings we recall necessary definitions and notations in Definition 1.1 and Remark 1.1 (see also [1, 3, 4] and Appendix).

Definition 1.1. Assume that F is an associative commutative and unital ring. Then over F a unital algebra A is considered, which may be generally nonassociative (relative to multiplication $A \times A \rightarrow A$). Assume that A is supplied with a scalar involution $a \mapsto \bar{a}$ so that its norm N and trace T maps have values in F and fulfil conditions:

$$a\bar{a} = N(a)1 \text{ with } N(a) \in F, \quad (1.1)$$

$$a + \bar{a} = T(a)1 \text{ with } T(a) \in F, \quad (1.2)$$

$$T(ab) = T(ba) \quad (1.3)$$

for each a and b in A .

If a scalar $f \in F$ satisfies the condition: $\forall a \in A \quad fa = 0 \Rightarrow a = 0$, then such element f is called cancelable. Using a cancelable scalar f the Dickson doubling procedure provides new algebra $C(A, f)$ over F such that:

$$C(A, f) = A \oplus A\mathbf{l}, \quad (1.4)$$

$$(a + b\mathbf{l})(c + d\mathbf{l}) = (ac - f\bar{d}b) + (da + b\bar{c})\mathbf{l} \text{ and} \quad (1.5)$$

$$\overline{(a + b\mathbf{l})} = \bar{a} - b\mathbf{l} \quad (1.6)$$

for each a and b in A . Then $\mathbf{l}$ is called a doubling generator.

R e m a r k 1.1. From Definition 1.1 identities follow: $\forall a \in A \quad \forall b \in A \quad T(a) = T(a + b\mathbf{l})$ and $N(a + b\mathbf{l}) = N(a) + fN(b)$. The algebra A is embedded into $C(A, f)$ as $A \ni a \mapsto (a, 0)$, where $(a, b) = a + b\mathbf{l}$. It is put by induction $A_n(f_{(n)}) = C(A_{n-1}, f_n)$, where $A_0 = A$, $f_1 = f$, $n = 1, 2, \dots$, $f_{(n)} = (f_1, \dots, f_n)$. Then $A_n(f_{(n)})$ are generalized Dickson algebras, when F is not a field, or Dickson algebras, when F is a field, where $1 \leq n \in \mathbf{N}$.

If the characteristic of F is $\text{char}(F) \neq 2$, then the imaginary part of a Dickson number z is defined by:

$$\text{Im}(z) = z - T(z)/2,$$

hence $N(a) := N_2(a, \bar{a})/2$, where $N_2(a, b) := T(a\bar{b})$.

If the doubling procedure starts from $A = F\mathbf{l} =: A_0$, then $A_1 = C(A, f_1)$ is a $*$ -extension of F .

R e m a r k 1.2. We consider also the following generalizations of the Dickson algebras. Let F be a commutative associative unital ring of characteristic

$$\text{char}(F) \neq 2; \quad (1.7)$$

an algebra B has a structure of a F -bimodule with

$$\begin{aligned} x + y &= y + x, \quad (x + y) + z = x + (y + z), \\ a(a_1x) &= (aa_1)x, \quad (xa_1)a = x(aa_1) \text{ and such that } ax = xa, \end{aligned} \quad (1.8)$$

for each a and a_1 in F , x, y and z in B , B as the F -bimodule is free and isomorphic with the direct sum

$$B \simeq \bigoplus_{j=0}^n Fi_j \quad (1.9)$$

with elements $i_j \in B$ for each $j = 0, \dots, n$, satisfying $T_k i_l = \xi_{k,l} i_l$, where $T_k x = (i_k x) i_k$, $\xi_{k,l} \in F$ for each k, l in $\{0, 1, 2, \dots, n\}$, x in B , where $n > 2$ is a natural number,

$$\xi = (\xi_{k,l})_{k,l=1,\dots,n+1} \quad (1.10)$$

is a $(n+1) \times (n+1)$ matrix having matrix elements $\xi_{k,l}$ such that the corresponding F -linear operator is invertible.

It will frequently be useful also the additional condition

$$i_j i_j = v_j i_0 \quad (1.11)$$

with nonzero cancelable v_j in F possessing an inverse $v_j^{-1} \in F$ for each $j = 0, \dots, n$.

Lemma 1.1. *Let an algebra B satisfy Conditions (1.7)–(1.10) in Remark 1.2. Then there exist F -linear operators $\pi_j : B \rightarrow Fi_j$ which are F -linear combinations of the operators $T_0, \dots, T_n$ for each $j \in \{0, 1, \dots, n\}$ such that $\sum_{j=0}^n \pi_j = \text{id}_B$, where $\text{id}_B(x) = x$ for each $x \in B$.*

P r o o f. From the conditions of this lemma it follows that there exists an inverse operator having matrix $p = \xi^{-1}$ with matrix elements $p_{k,l}$ belonging to F . Then we put $\pi_j(x) = \sum_{k=0}^n p_{j,k} T_k(x)$, consequently, $\pi_j(x) = \sum_{l=0}^n x_l \pi_j(i_l)$, where $x_l \in F$ for each l such that $x = \sum_{l=0}^n x_l i_l$, $x \in B$. Then $\pi_j(i_l) = \sum_{k=0}^n p_{j,k} T_k(i_l)$ hence $\pi_j(i_l) = \sum_{k=0}^n p_{j,k} \xi_{k,l} i_l = \delta_{j,l} i_l$, where $\delta_{j,j} = 1$, $\delta_{j,k} = 0$ for each $k \neq j$. Thus $\pi_j(x) = x_j i_j$. $\square$

Corollary 1.1. *Let the algebra B satisfy conditions (1.7)–(1.11) in Remark 1.2. Then $v_j^{-1} i_j \pi_j(x) = x_j$ for each $x \in B$ and $j = 0, \dots, n$, where $x_l \in F$ for each l such that $x = \sum_{l=0}^n x_l i_l$.*

E x a m p l e 1.1. Assume that F is a (commutative associative) field of characteristic $\text{char}(F) \neq 2$, B satisfies conditions (1.7)–(1.10) in Remark 1.2, $\{i_0, i_1, \dots, i_n\}$ is a basis of B over F , $\det(\xi) \neq 0$. Then there exists an inverse matrix $p = \xi^{-1}$ with matrix elements $p_{k,l}$ belonging to the field F .

In particular, let us choose $B = A_m(f_{(m)})$ such that $2 \leq m \in \mathbf{N}$, where F is the field of characteristic $\text{char}(F) \neq 2$, $f_1 = 1, \dots, f_m = 1$, $n = 2^m - 1$, $A_0 = F$ with the trivial involution (i. e. $a = \bar{a}$ for each $a \in A_0$), $i_0 = 1$, where $1 = 1_B$ is the unit element in B (see Remark 1.1). Then $\bar{x} = x_0 i_0 - x_1 i_1 - \dots - x_n i_n$ for each $x \in B$, where $x_0, \dots, x_n$ denote expansion coefficients belonging to F for x such that $x = x_0 i_0 + x_1 i_1 + \dots + x_n i_n$. Then $T_0(x) = x$, $T_1(x) = -x_0 i_0 - x_1 i_1 + x_2 i_2 + \dots + x_n i_n$, $\dots$, $T_n(x) = -x_0 i_0 + x_1 i_1 + \dots + x_{n-1} i_{n-1} - x_n i_n$, since $i_0 i_k = i_k$, $i_k^2 = -1$, $i_k i_l = -i_l i_k$ and $(i_k i_l) i_k = i_l$ for each $k \neq l$ with $k \geq 1$ and $l \geq 1$. Therefore, $\frac{1}{2-2^m} (T_0 + \dots + T_{2^m-1})(x) = \bar{x}$, consequently, $\pi_0(x) = x_0 i_0 = \frac{1}{2} (T_0 + \frac{1}{2-2^m} (T_0 + \dots + T_{2^m-1}))(x)$. Then $\pi_0(\bar{i}_k x) = x_k i_0$ for each $k \geq 1$, hence $\pi_k(x) = (\pi_0(\bar{i}_k x)) i_k = x_k i_k$. Thus ξ is the invertible matrix.

Lemma 1.2. *Let $A_n = A_n(f_{(n)})$, $A_0 = A$, $2 \leq n \in \mathbf{N}$, where A is the commutative associative unital algebra with the trivial involution over the commutative associative unital ring F of characteristic $\text{char}(F) \neq 2$. Let $i_0 = 1_A$, $i_{2^k-1} = \mathbf{l}_k$ for each $k = 1, \dots, n$, $i_{j_s} = i_{j_{s-1}} \mathbf{l}_{k_s}$ with $j_1 = 2^{k_1-1}$, $j_s = j_{s-1} + 2^{k_s-1}$ for each $s = 2, \dots, p$, $2 \leq p \leq n$, $1 \leq k_1 < \dots < k_p \leq n$, where $\mathbf{l}_p$ denotes the doubling generator $\mathbf{l}$ at the p -th step in Formula (1.5) in Definition 1.1. Then $\{i_j : j = 0, 1, \dots, 2^n - 1\}$ is a family of generators of A_n over A_0 satisfying the identities:*

$$i_j(i_j u) = (i_j i_j) u, \quad (u i_j) i_j = u (i_j i_j), \quad i_j(v i_j) = (i_j v) i_j, \quad T(i_j(i_k v)) = 0 \quad (1.12)$$

for each $u \in A_n$, $v = \bar{v} \in A_n$ and $j = 0, 1, \dots, 2^n - 1$, $1 \leq k \neq j$.

P r o o f. Since the ring F is commutative and associative, then as it is known the left and right F -module structures can be considered as equivalent: $(pp_1)u = p(p_1 u) = p_1(pu) = (p_1 p)u$ for each p, p_1 in F , $u \in A_n$, by putting $L_p = R_p$ on A_n for each $p \in F$, where $L_p u = pu$, $R_p u = up$ (see [9, Ch. 2]). The algebra A_0 is unital, hence A_1 is unital, and by induction A_n is unital according to Formulas (1.4), (1.5) in Definition 1.1. The elements f_k in F are cancelable for each k , consequently, the product $f_{k_1} \dots f_{k_p}$ is nonzero for each $1 \leq k_1 < \dots < k_p \leq n$, $p \geq 2$, since A_n is the unital algebra. For each a_0, a_1 in A_0 by the conditions of this lemma $a_0 a_1 = a_1 a_0$ and $\overline{a_0 a_1} = a_0 a_1$.

Using Formulas (1.4), (1.5) in Definition 1.1 by induction we deduce that for each x in A_n there exist elements $x_0, \dots, x_{2^n-1}$ in A_0 such that $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$. That is, $\{i_j : j = 0, \dots, 2^n - 1\}$ is the family of generators of A_n over A_0 . Therefore,

$$i_j(i_j x) = \sum_{m=0}^{2^n-1} i_j(i_j x_m i_m) = \sum_{m=0}^{2^n-1} (i_j(i_j i_m)) x_m \quad \text{and} \quad (x i_j) i_j = \sum_{m=0}^{2^n-1} x_m ((i_m i_j) i_j). \quad (1.13)$$

From Formula (1.5) in Definition 1.1 we deduce that

$$\begin{aligned} \mathbf{l}_k(\mathbf{l}_k(a_{k-1} + d_{k-1}\mathbf{l}_k)) &= (\mathbf{l}_k\mathbf{l}_k)(a_{k-1} + d_{k-1}\mathbf{l}_k) \text{ and} \\ ((a_{k-1} + d_{k-1}\mathbf{l}_k)\mathbf{l}_k)\mathbf{l}_k &= (a_{k-1} + d_{k-1}\mathbf{l}_k)(\mathbf{l}_k\mathbf{l}_k) \end{aligned} \quad (1.14)$$

for each a_{k-1}, d_{k-1} in A_{k-1} , where $k \geq 1$. Note that $\mathbf{l}_k^2 = -f_k$ for each $k \geq 1$, since A_n is the unital algebra. If $\{\mathbf{l}_{k_1}, \dots, \mathbf{l}_{k_p}\}_{q(p)}$ denotes an ordered product, where $q(p)$ is a vector indicating an order of pairwise multiplication with a corresponding order of brackets in $\mathbf{l}_{k_1}, \dots, \mathbf{l}_{k_p}$, where $1 \leq k_l \leq n$, $k_l \neq k_s$ for each $l \neq s$, l and s in $\{1, \dots, n\}$, $2 \leq p \leq n$, then there exist unique $\eta(k_1, \dots, k_p, q(p)) \in \{1, 2\}$ and $j(k_1, \dots, k_p, q(p)) \in \{1, \dots, 2^n - 1\}$ such that $(-1)^\eta \{\mathbf{l}_{k_1}, \dots, \mathbf{l}_{k_p}\}_{q(p)} = i_j$, where $\eta = \eta(k_1, \dots, k_p, q(p))$, $j = j(k_1, \dots, k_p, q(p))$. Therefore, from (1.14) by induction in $k = 1, \dots, n$ it follows that $i_j(i_j i_m) = (i_j i_j) i_m$ and $(i_m i_j) i_j = i_m (i_j i_j)$ for each m and j . For each $v = \bar{v} \in A_n$ from Formulas (1.5) and (1.6) in Definition 1.1 it follows that $T(\mathbf{l}_k v) = 0$, $\mathbf{l}_k(v\mathbf{l}_k) = (\mathbf{l}_k v)\mathbf{l}_k = \bar{\mathbf{l}}_k(v\mathbf{l}_k)$ for each $k \geq 1$ and by induction $i_j(v i_j) = (i_j v) i_j$ for each j , since $i_0 = 1$. The latter and (1.13) imply (1.12). $\square$

Corollary 1.2. *If the conditions of Lemma 1.2 are satisfied and F is a field, $A_0 = F$, then $\{i_j : j = 0, \dots, 2^n - 1\}$ is a basis in A_n (as in the F -linear space).*

Definition 1.2. Let F be the commutative associative unital ring. Let B be a unital algebra over F with $F \subseteq C_B(B)$. Let M be a unital left $C_B(B)$ -module:

$$b(b_1 u) = (bb_1)u, \quad b(u + v) = bu + bv, \quad (b + b_1)u = bu + b_1 u, \quad u + (u_1 + v) = (u + u_1) + v \quad (1.15)$$

for each u, u_1, v in M , b, b_1 in $C_B(B)$. Let μ_1 be a $C_B(B)$ -bilinear map $\mu_1 : B \times M \rightarrow M$, that is,

$$\mu_1(x, u + v) = \mu_1(x, u) + \mu_1(x, v), \quad \mu_1(x + y, u) = \mu_1(x, u) + \mu_1(y, u), \quad \mu_1(bx, u) = b\mu_1(x, u) \quad (1.16)$$

such that μ_1 is compatible with the left $C_B(B)$ -module structure of M :

$$\mu_1(x, bu) = b\mu_1(x, u) \quad (1.17)$$

for each x, y in B , u, v in M , b in $C_B(B)$. Then M will be called a left B -module. Shortly $\mu_1(x, u)$ can also be denoted by xu . Similarly is defined a right B -module, or a B -bimodule.

For $B = A_n(f_{(n)})$ with $A_0 = A$, $n \geq 2$, where A is the commutative associative unital algebra with the trivial involution over the associative commutative unital ring F of characteristic $\text{char}(F) \neq 2$, if M satisfies conditions (1.15)–(1.17) and

$$i_j(i_j x) = (i_j i_j)x \quad (1.18)$$

for each $x \in M$, $j = 0, \dots, 2^{n-1}$, then M will be called a left $A_n(f_{(n)})$ -module. Symmetrically is defined the right $A_n(f_{(n)})$ -module with condition (1.19) instead of (1.18):

$$(x i_j) i_j = x(i_j i_j) \quad (1.19)$$

for each $x \in M$, $j = 0, \dots, 2^{n-1}$. If M satisfies (1.15)–(1.19) and (1.20):

$$L_b = R_b \text{ on } M \text{ for each } b \in C_B(B), \quad (1.20)$$

then it will be called $A_n(f_{(n)})$ -bimodule and denoted by ${}_B M_B$ or shortly by M , where $B = A_n(f_{(n)})$.

If in the B -bimodule M there exists a $\mathbb{C}_B(B)$ -linear map $J : M \rightarrow M$ such that

$$J(bx) = (Jx)\bar{b} \text{ for each } x \in M \text{ and } b \in B, \quad J^2 = I \quad (1.21)$$

$$\text{and } ((\exists x \in M \quad \forall b \in B \quad bx + J(bx) = 0) \Rightarrow (x = 0)) \quad (1.22)$$

$$\text{and } (i_j y)i_j = i_j(yi_j) \text{ and } i_j(i_k y) + \overline{i_j(i_k y)} = 0 \quad (1.23)$$

for each $y = Jy$ in M and $j \geq 0$, $j \neq k \geq 1$, where $B = A_n(f_{(n)})$, $I : M \rightarrow M$ is the identity map $I = id_M$ on M , $Ix = x$, then M will be called the B -bimodule with the involution J and denoted by ${}_B \hat{M}_B$ or shortly by M . Briefly Jx will also be denoted by $\bar{x}$.

Theorem 1.1. *Let M be the unital $A_n(f_{(n)})$ -bimodule with the involution J , let the subalgebra A_0 over F be commutative associative and with the trivial involution $\bar{a} = a$ for each $a \in A_0$, let also $\text{char}(F) \neq 2$ and f_j possess an inverse element f_j^{-1} in F relative to multiplication for each $j = 1, \dots, n$, where $2 \leq n \in \mathbb{N}$. Then there exists an A_0 -subbimodule M_0 such that ${}_A M_{A_0} = \bigoplus_{j=0}^{2^n-1} i_j M_0$ with $JM_0 = M_0$, and $M_0 = \mathbb{C}_M(A_n(f_{(n)}))$, and there exists an A_0 -linear map $\hat{\pi}_k$ from M onto $i_k M_0$ with $\hat{\pi}_k \circ \hat{\pi}_j = \delta_{k,j} \hat{\pi}_k$, where $\delta_{k,k} = 1$, $\delta_{k,j} = 0$ if $k \neq j$, for each k and j in $\{0, 1, \dots, 2^n - 1\}$.*

P r o o f. By virtue of Lemma 1.2 the Dickson algebra $B = A_n(f_{(n)})$ has the family of generators $\beta_n := \{i_j : j = 0, \dots, 2^n - 1\}$ over A_0 . By the conditions imposed above in Definition 1.2 the algebra A_0 and the ring F are unital such that there is the natural embedding of F into A_0 as $F1_{A_0}$ and hence into B . Therefore, B contains the Dickson subalgebra $\text{alg}_F \beta_n$ over F with generators $i_0, \dots, i_{2^n-1}$.

Note that $F \subseteq \mathbb{C}_B(B)$ and $F i_k \beta_n = F \beta_n$, $F \beta_n i_k = F \beta_n$ for each k , since $F \subseteq \mathbb{C}_{A_0}(A_0)$. As in Remark 1.2 let $T_k u = (i_k u)i_k$ for each $u \in B$, and $\hat{T}_k x = (i_k x)i_k$ for each $x \in M$, $k = 0, \dots, 2^n - 1$. We put $\pi_0(u) = \frac{u+\bar{u}}{2}$ for each $u \in B$, and $\hat{\pi}_0(x) = \frac{x+\bar{x}}{2}$ for each $x \in M$. Let $M_0 = \hat{\pi}_0(M)$, hence $M_0 = \bigcup \{y \in M : \exists x \in M, y = \frac{x+\bar{x}}{2}\}$. On the other hand, $R_b x = L_b x$ for each $b \in \mathbb{C}_B(B)$ and $x \in M$ by Definition 1.2, where $L_b u = bu$, $R_b u = ub$ for each u, b in B . The algebra A_0 is commutative and associative with $a_0 = \bar{a}_0$ for each $a_0 \in A_0$, consequently, $A_0 \subseteq \mathbb{C}_B(B)$ and hence $\overline{a_0 y} = \bar{y} \bar{a}_0 = y a_0 = a_0 y$ for each $y \in M_0$ and $a_0 \in A_0$. Therefore, M_0 is the A_0 -subbimodule in M , since $A_0 \subset B$ and M has also the structure of the A_0 -bimodule ${}_A M_{A_0}$. It follows that $y = \bar{y}$ and $\hat{\pi}_0(y) = y$ for each $y \in M_0$, hence $\hat{\pi}_0 \circ \hat{\pi}_0 = \hat{\pi}_0$, where as usually $(g \circ h)(v) = g(h(v))$ denotes the composition of maps g and h with a variable v of h .

For each $x \in M$ there is the decomposition $x = y + z$ such that $y = \frac{x+\bar{x}}{2}$, $z = \frac{x-\bar{x}}{2}$. Let $M_- := \{z \in M : \bar{z} = -z\}$. Evidently, $M_0 \cap M_- = \{0\}$ and M_- is the A_0 -subbimodule in ${}_A M_{A_0}$ such that $M_0 \oplus M_- = {}_A M_{A_0}$, since $J : M \rightarrow M$ and $J^2 = I$, also $B_0 \cap B_- = \{0\}$, $B_0 = \pi_0(B)$, $B_0 = A_0$, $B_- := \{u \in B : \exists v \in B, u = \frac{v-\bar{v}}{2}\}$.

We put

$$S_n(f_{(n)}) := \bigcup \left\{ f \in F : \exists p \in \{1, \dots, n\} \exists j_1 \in \{1, \dots, n\} \dots \exists j_p \in \{1, \dots, n\} \right. \\ \left. \exists \alpha_1 \in \mathbb{Z} \dots \exists \alpha_p \in \mathbb{Z} \exists v \in \{-1, 1\} : f = v f_{j_1}^{\alpha_1} \dots f_{j_p}^{\alpha_p} \right\},$$

consequently, $i_k^2 = s_k \in S$ for each $k \geq 1$, where $S = S_n(f_{(n)})$. Note that if $y \in M_0$, $z \in M_-$, $b \in B_-$, then $(by - yb) \in M_0$, $(by + yb) \in M_-$, $bz + zb \in M_0$, $bz - zb \in M_-$.

On the other hand, $(i_k y = y i_k) \Leftrightarrow ((i_k y) i_k = y s_k) \Leftrightarrow (i_k (y i_k) = y s_k)$ for each $k \geq 1$, consequently, from conditions (1.21), (1.22), (1.23) in Definition 1.2 it follows that $i_k y = y i_k$ for each $k = 0, \dots, 2^n - 1$, since $s_k \in S \subset F$, $i_0 = 1$. This implies that $i_k y \in M_-$ for each $k \geq 1$. By the A_0 -linearity and Lemma 1.2 this implies that $M_0 \subseteq \text{Com}_M(B)$.

Then we put $\hat{\pi}_j(x) = -s_j^{-1}(\hat{\pi}_0(\bar{b}_j x))i_j$ for each $x \in M$, $j = 1, \dots, 2^n - 1$. Notice that $\hat{\pi}_k(x) = \hat{\pi}_k(y) + \hat{\pi}_k(z) = \hat{\pi}_k(z)$ for each $k \geq 1$ and $x \in M$, where $y \in M_0$, $z \in M_-$, $x = y + z$, $y = \frac{I+J}{2}x$, $z = \frac{I-J}{2}x$, since $i_k y \in M_-$ and $\hat{\pi}_0(M_-) = \{0\}$. Thus $\hat{\pi}_k \circ \hat{\pi}_0 = 0$ for each $k \geq 1$, since $\hat{\pi}_k(M_0) = \{0\}$ and $\hat{\pi}_0(M) = M_0$. From $i_k^2 x = i_k(i_k x) = s_k x$ with $s_k \in S$ for each $k \geq 1$, $\text{char}(F) \neq 2$, $F \subseteq A_0 \subseteq \mathbb{C}_B(B)$ and the conditions (1.15)–(1.19) in Definition 1.2 it follows that $L_{i_k} : M \rightarrow M$ and similarly $R_{i_k} : M \rightarrow M$ are A_0 -linear bijections for each $k = 0, \dots, 2^n - 1$, since M is the unital B -bimodule with involution, since the algebra B is unital, $i_0 = 1_B$. Then we deduce that $\hat{\pi}_k(x) = \frac{-1}{2s_k}([I + J](\bar{b}_k x))i_k = \frac{-1}{2}[s_k^{-1}(\bar{b}_k x)i_k + \bar{x}]$ for each $x \in M$ and $k = 1, \dots, 2^n - 1$. Therefore, $\hat{\pi}_k : M_- \rightarrow M_-$ for each $k \geq 1$, since $z = -\bar{z}$ for each $z \in M_-$. This implies that $\hat{\pi}_0 \circ \hat{\pi}_k(x) = \hat{\pi}_0(\hat{\pi}_k(z)) = 0$ for each $x \in M$ and $k \geq 1$, since $M_0 \cap M_- = \{0\}$, where $z = \frac{I-J}{2}x$. Then we infer that $\hat{\pi}_k \circ \hat{\pi}_k(x) = -\hat{\pi}_k(\bar{x}) = \hat{\pi}_k(x)$ for each $x \in M$ and $k \geq 1$, since $\hat{\pi}_k(x) = \hat{\pi}_k(z)$ with $z = \frac{I-J}{2}x$, since $L_{i_k} R_{i_k} \hat{\pi}_0 = L_{i_k^2} \hat{\pi}_0$, since $\hat{\pi}_0(M) = M_0$. Particularly $\hat{\pi}_k(i_k y) = i_k y$ for each $k \geq 1$ and $y \in M_0$, since $i_k y = y i_k$. This implies that $\hat{\pi}_k(M) = \hat{\pi}_k \circ \hat{\pi}_k(M) = \hat{\pi}_k(i_k M_0) = i_k M_0$, $i_k M_0 = M_0 i_k$ for each $k \geq 0$, since $i_0 = 1_B$. Note that $i_k M_0 \subset M_-$ and $\hat{\pi}_0(\bar{b}_k x) = \hat{\pi}_0(\bar{b}_k z)$ with $z = \frac{I-J}{2}x$ for each $k \geq 1$ and $x \in M$. Thus $\hat{\pi}_0|_{M_0} = \text{id}_{M_0}$, $\hat{\pi}_0|_{M_-} = 0$, where $\text{id}_{M_0}(y) = y$ for each $y \in M_0$. Therefore, $\hat{\pi}_j \circ \hat{\pi}_k = 0$ for each $j \neq k$, since $i_j M_0 \cap i_k M_0 = \{0\}$, since $i_j(i_k y) + i_j(i_k y) = 0$ for each $y = Jy$ in M and $j \geq 0$, $j \neq k \geq 1$, since f_j is invertible relative to multiplication in F for each j .

Then we put $\hat{K} = \sum_{j=0}^{2^n-1} \hat{\pi}_j$ on M , and $K = \sum_{j=0}^{2^n-1} \pi_j$ on B . These operators are idempotent $\hat{K}^2 = \hat{K}$ and $K^2 = K$, since $\hat{\pi}_j \circ \hat{\pi}_k = \delta_{j,k} \hat{\pi}_j$ and $\pi_j \circ \pi_k = \delta_{j,k} \pi_j$ for each $j, k = 0, \dots, 2^n - 1$. Hence $I - \hat{K}$ also is the idempotent operator.

It is known that the minimal subalgebra $A_{(j,k)}$ in $A_n(f_{(n)})$ generated by $\{A_0, i_j, i_k\}$ is associative for each $j, k = 0, \dots, 2^n - 1$, since F and A_0 are commutative and associative by the conditions of this theorem (see [1, 4, 9]). Therefore, $M_{(j,k)} := M_0 \oplus i_j M_0 \oplus i_k M_0 \oplus (i_j i_k) M_0$ is the $A_{(j,k)}$ -subbimodule with involution in M , since $i_k M_0 = M_0 i_k$ for each k , $i_j M_0 \cap i_k M_0 = \{0\}$ for each $j \neq k$, $i_j i_k \in G$, where $G = G_n(f_{(n)}) = \{i_0, \dots, i_{2^n-1}\} \cdot S$, $F \subseteq A_0$.

On the other hand, $\hat{K}y = \hat{\pi}_0 y = y$ and $\hat{K}(i_j y) = \hat{\pi}_j(i_j y) = i_j y$ for each $y \in M_0$ and $j \geq 1$, since $\hat{\pi}_k(i_j y) = \hat{\pi}_k \circ \hat{\pi}_j(i_j y) = 0$ for each $j \neq k$. Then we deduce that $\hat{K}M_- = \bigoplus_{j=1}^{2^n-1} i_j M_0$, since $L_{i_j} : M_0 \rightarrow M_-$ and $\hat{\pi}_j M_- = i_j M_0$ for each $j \geq 1$, since $(\hat{\pi}_j M_-)i_j = M_0$, $M_0 i_j = i_j M_0$. Hence $\hat{\pi}_k(i_k P) = \hat{\pi}_0(P) = \{0\}$, where $P := M \ominus (\bigoplus_{j=0}^{2^n-1} i_j M_0)$. Notice that P is the proper $A_n(f_{(n)})$ -subbimodule with involution in M , that is P satisfies conditions (1.18)–(1.23) in Definition 1.2. On the other side, the condition

$$((\exists x \in M \forall b \in B \quad bx + J(bx) = 0) \Rightarrow (x = 0)) \text{ is equivalent to } \\ ((\exists x \in M \forall j \in \{0, \dots, 2^n - 1\} \quad i_j x + J(i_j x) = 0) \Rightarrow (x = 0)),$$

since for each $b \in B$ there exist $a_0, \dots, a_{2^n-1}$ in A_0 such that $b = a_0 i_0 + \dots + a_{2^n-1} i_{2^n-1}$ by Lemma 1.2. From $\hat{\pi}_0(P) = \{0\}$ and $P_0 = \hat{\pi}_0(P)$ it follows that $P_0 = \{0\}$, consequently, $P_- = \{0\}$, since $P_0 = i_j \hat{\pi}_j(P_-) = \{0\}$ for each $j \geq 1$. Thus $P = \{0\}$, consequently, $\hat{K} = I$ on M and hence ${}_{A_0}M_{A_0} = \bigoplus_{j=0}^{2^n-1} i_j M_0$ and consequently, $M_0 = \text{Com}_M(B)$, where M is considered

as the A_0 -bimodule ${}_{A_0}M_{A_0}$, since $M_- \cap \text{Com}_M(B) = \{0\}$. Analogously ${}_{A_0}B_{A_0} = \bigoplus_{j=0}^{2^n-1} i_j B_0$ and $B_0 = \text{Com}_B(B)$, $B_0 = \pi_0(B)$, where B is considered as the A_0 -bimodule ${}_{A_0}B_{A_0}$.

Therefore, for each $y \in M_0$ and for each $j \neq k$ such that $j \geq 1$ and $k \geq 1$ we infer that $\hat{\pi}_t((i_j y)i_k) = 0$ and $\hat{\pi}_t(i_j(yi_k)) = 0$ for each $t \neq r$, $\hat{\pi}_r((i_j y)i_k) = (i_j y)i_k$ and $\hat{\pi}_r(i_j(yi_k)) = i_j(yi_k)$, where $r \in \{1, \dots, 2^n - 1\}$ is such that $i_j i_k \in i_r S$, since $(i_j y)i_k \in M_{(j,k)-}$, $i_j(yi_k) \in M_{(j,k)-}$. We put $v = i_j(yi_k) - (i_j y)i_k$. Therefore, $v \in L_{i_r} M_0 \subset M_{(j,k)-}$. From Formulas (1.21), (1.22) and (1.23) in Definition 1.2 we deduce that $v = \bar{v}$. Thus $v \in M_0 \cap M_- = \{0\}$, that is $v = 0$. Hence $(i_j y)i_k = i_j(yi_k)$ for each $j \neq k$ such that $j \geq 1$ and $k \geq 1$. For $j = 0$ or $k = 0$ evidently $i_j(yi_k) = (i_j y)i_k$, since $i_0 = 1_B$. Using $M_0 = \text{Com}_M(B)$ and Conditions (1.21), (1.22), (1.23) in Definition 1.2 we infer that $i_j(yi_k) = -(yi_k)i_j$ for each $y \in M_0$ and $j \neq k$ such that $j \geq 1$, $k \geq 1$. Then it is similarly deduced that $i_j(i_k y) = (i_j i_k)y$ and $(yi_k)i_j = y(i_k i_j)$ for each $j \neq k$ in $\{1, \dots, 2^n - 1\}$, $y \in M_0$, since $v + v_1 = 0$ and $\bar{v} = v_1$ with $v = i_j(i_k y) - (i_j i_k)y$, $v_1 = (yi_k)i_j - y(i_k i_j)$, since $v \in M_-$, $v_1 \in M_-$, $M_0 \cap M_- = \{0\}$, since i_j , i_k , $i_j i_k$ belong to B_- . If $j = 0$ or $k = 0$, evidently $i_j(i_k y) = (i_j i_k)y$ and $(yi_k)i_j = y(i_k i_j)$ for each $y \in M_0$. By the A_0 -linearity and Lemma 1.2 we infer that $M_0 \subseteq N_M(B)$, consequently, $\mathbb{C}_M(B) = M_0$. $\square$

Corollary 1.3. *Let the conditions of Theorem 1.1 be satisfied and $n = 3$. Then $b(bx) = (bb)x$, $(bx)b = b(xb)$, $(xb)b = x(bb)$ for each $x \in M$ and $b \in B$.*

Proposition 1.1. *If the conditions of Theorem 1.1 are satisfied and there is some equality with a finite sum like*

$$\sum_{\theta \in \mathbf{S}_m; k_1, \dots, k_m; l} \gamma_{\theta; k_1, \dots, k_m; l} \{d_{\theta(k_1)} \dots d_{\theta(k_m)}\}_{q_{l, \theta}(m)} = 0$$

in $A_n(f_{(n)})$, where $d_{k_j} \in A_n(f_{(n)})$, $\gamma_{\theta; k_1, \dots, k_m; l} \in A_0$ for each k_j , j , l , θ , then there exists a corresponding identity in M .

Proof. For the identity satisfying the conditions of this proposition we use the decomposition ${}_{A_0}M_{A_0} = \bigoplus_{j=0}^{2^n-1} i_j M_0$ and $JM_0 = M_0$, where $M_0 = \mathbb{C}_M(B)$. Then we substitute one of d_{k_j} on $d_{k_j}y$ with an arbitrary fixed nonzero $y \in M_0$ for each additive $\{d_{\theta(k_1)} \dots d_{\theta(k_m)}\}_{q_{l, \theta}(m)}$, where $\mathbf{S}_m$ denotes the symmetric group of $\{1, \dots, m\}$, $\theta : \{1, \dots, m\} \rightarrow \{1, \dots, m\}$ is a bijection for each $\theta \in \mathbf{S}_m$, $q_{l, \theta}(m)$ is a vector indicating an order of pairwise multiplications in $\{\dots\}$. Then it is possible to make sums of such type equalities with multipliers from A_0 . $\square$

This proposition shows that definitions above are natural, because particularly the algebra has also the structure of the module over itself. There may other equivalent definitions be given.

Theorem 1.2. *Assume that F is a commutative associative unital ring, $\text{char}(F) \neq 2$, a unital algebra A_0 over F is associative and commutative with the trivial involution $\bar{a} = a$ for each $a \in A_0$, M_0 is a unital A_0 -bimodule, $B = A_n(f_{(n)})$ is the generalized Dickson algebra, and f_j possess an inverse element f_j^{-1} in F relative to multiplication for each $j = 1, \dots, n$, ${}_{A_0}M_{A_0} = \bigoplus_{j=0}^{2^n-1} i_j M_0$ such that $M_0 = \mathbb{C}_M(B)$, where $n \geq 2$. Then ${}_{A_0}M_{A_0}$ can be supplied with B -bimodule with involution ${}_B\hat{M}_B$ structure.*

Proof. We put $by = yb$, $a(by) = (ab)y$, $a(yb) = (ay)b$, $(ya)b = y(ab)$, $J(by) = \bar{b}y$ for each $y \in M_0$, a and b in B , $J(x + z) = Jx + Jz$ for each x and z in $M = \bigoplus_{j=0}^{2^n-1} i_j M_0$,

since $M_0 = \mathbb{C}_M(B)$ and $i_0 y = y$ for each $y \in M_0$. Therefore, $Jx = x_0 \bar{b}_0 + \dots + x_{2^n-1} \bar{b}_{2^n-1}$ for each $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$ in M with $x_0, \dots, x_{2^n-1}$ in M_0 , consequently, $J^2 = I$ and hence J is the involution on M . In view of Lemma 1.2 the equalities $by = yb$, $a(by) = (ab)y$, $a(yb) = (ay)b$, $(ya)b = y(ab)$ for each $y \in M_0$, a and b in B , supply M with properties (1.15)–(1.23) in Definition 1.2, since the minimal subalgebra $A_{(j,k,l)}$ in $A_n(f_{(n)})$ generated by $\{A_0, i_j, i_k, i_l\}$ is alternative for each j, k, l in $\{0, \dots, 2^n - 1\}$ (see [1, 4, 9]), since F and A_0 are unital, associative and commutative, $\bar{a} = a$ for each $a \in A_0$, since each $x \in M$ has the decomposition $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$ in M with $x_0, \dots, x_{2^n-1}$ in M_0 . $\square$

Definition 1.3. If M is the left B -module (see Definition 1.2), E is a subset in B , D is a subset in B (or in M), then

$$E \cdot D = \bigcup \{ed : e \in E, d \in D\},$$

$$ED = \bigcup \left\{ x = \sum_{k=1}^m e_k d_k : m \in \mathbb{N}, \forall k \ e_k \in E, d_k \in D \right\}$$

denote subsets in B (or in M correspondingly). Then it is put

$$\begin{aligned} E^{(1)} &= E, \quad (E \cdot D)^{(1)} = E \cdot D; \\ \forall n > 1 \quad dE^{(n)} &= E \cdot E^{n-1}, \quad (E \cdot D)^{(n)} = E \cdot (E \cdot D)^{(n-1)}; \\ E^{<1>} &= E, \quad (ED)^{<1>} = ED, \\ \forall n > 1 \quad E^{<n>} &= EE^{n-1}, \quad (ED)^{<n>} = E(ED)^{<n-1>}; \\ E^{(\infty)} &= \bigcup_{n=1}^{\infty} E^{(n)}, \quad d(E \cdot D)^{(\infty)} = \bigcup_{n=1}^{\infty} (E \cdot D)^{(n)}; \\ E^{<\infty>} &= \sum_{n=1}^{\infty} E^{<n>}, \quad (ED)^{<\infty>} = \sum_{n=1}^{\infty} (ED)^{<n>}. \end{aligned}$$

If N is a left B -submodule in M such that $\exists d \in M$, $D = \{d\}$, $E = B$, $N = (B\{d\})^{<\infty>}$, then N is called a cyclic left B -submodule in M generated by d .

Similar notations are for right B -modules or B -bimodules.

If M is the B -bimodule, then

$$\begin{aligned} (E \cdot D)^{(1,1)} &= (E \cdot D) \cup (D \cdot E), \quad (ED)^{<1,1>} = (ED) + (DE), \\ \forall n > 1 \quad (E \cdot D)^{(n,n)} &= (E \cdot (E \cdot D)^{(n-1,n-1)}) \cup ((E \cdot D)^{(n-1,n-1)} \cdot E), \\ (ED)^{<n,n>} &= (E(ED)^{<n-1,n-1>}) + ((ED)^{<n-1,n-1>} E); \\ (E \cdot D)^{(\infty,\infty)} &= \bigcup_{n=1}^{\infty} (E \cdot D)^{(n,n)}, \quad (ED)^{<\infty,\infty>} = \sum_{n=1}^{\infty} (ED)^{<n,n>}. \end{aligned}$$

If N is a B -subbimodule in M such that $\exists d \in M$, $D = \{d\}$, $E = B$, $N = (B\{d\})^{<\infty,\infty>}$, then N is called a cyclic B -subbimodule in M generated by d .

If F is the field and V is an $A_n(f_{(n)})$ -subbimodule with the involution in an $A_n(f_{(n)})$ -bimodule with the involution, then $\dim_F V$ denotes the dimension of V over F .

Theorem 1.3. Let the conditions of Theorem 1.1 be satisfied, $D \subset M$. Then $(BD)^{<m,m>} \subseteq (BD)^{<m+1,m+1>}$ for each $m \geq 1$, $(BD)^{<k,k>} = (BD)^{<4,4>}$ for each $k \geq 4$. Moreover, $(BD)^{<4,4>}$ is the B -subbimodule with involution in M and $(BD)^{<4,4>} = (BD)^{<\infty,\infty>}$, $(BD)^{<\infty,\infty>} = (B\bar{D})^{<\infty,\infty>}$.

P r o o f. The algebra $B = A_n(f_{(n)})$ is unital, the B -bimodule with involution M is unital, $n \geq 2$, by the imposed conditions. Therefore, $(BD)^{<m,m>} \subseteq (BD)^{<m+1,m+1>}$ for each $m \geq 1$.

For each $x \in D$ the element $\hat{\pi}_j(x)$ belongs to $(B(B\{x\}))B$ for each $j = 0, \dots, 2^n - 1$, since

$$\bar{x} = \frac{1}{2-2^n} \left(x + \frac{(\bar{b}_1 x) i_1}{s_1} + \dots + \frac{(\bar{b}_{2^n-1} x) i_{2^n-1}}{s_{2^n-1}} \right)$$

by Theorem 1.1, where s_j is invertible in F (relative to multiplication) for each $j \geq 1$. Evidently, $(B(B\{x\}))B \subseteq (BD)^{<3,3>}$ for each $x \in D$. On the other hand, $R_{i_k} \hat{\pi}_j(x)$ belong to $(BD)^{<4,4>}$ for each j and k in $\{0, \dots, 2^n - 1\}$, $x \in D$, since by Theorem 1.1 $\hat{\pi}_j(x) \in i_j M_0$, each x in M has the decomposition $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$ with $x_0, \dots, x_{2^n-1}$ in M_0 , $\hat{\pi}_j(x) = x_j i_j$, $M_0 = \mathbb{C}_M(B)$. This implies that $(BD)^{<4,4>} = \sum_{k=0}^{2^n-1} i_k V_0$ with $V_0 = \text{span}_F \bigcup \{x_j : \exists x \in D \exists j \in \{0, \dots, 2^n - 1\} x_j = \frac{i_j}{s_j} \hat{\pi}_j(x)\}$, where $\text{span}_F Q$ denotes the F -linear span of a subset Q in M . Certainly, $V_0 \subset M_0$ and consequently, $(BD)^{<4,4>}$ is the B -subbimodule with involution in M and $(BD)^{<4,4>} = (BD)^{<\infty, \infty>}$, $(BD)^{<\infty, \infty>} = (B\bar{D})^{<\infty, \infty>}$. $\square$

Corollary 1.4. *If the conditions of Theorem 1.3 are satisfied and F is the field, then $\dim_F(BD)^{<\infty, \infty>} = 2^n \dim_F V_0$ and $\dim_F V_0 \leq 2^n \text{card}(D)$.*

Corollary 1.5. *Let F be a commutative associative unital ring, $\text{char}(F) \neq 2$, let A_0 be a commutative associative unital algebra over F with trivial involution $a = \bar{a}$ for each $a \in A_0$, $2 \leq n \in \mathbb{N}$, f_j be invertible in F relative to multiplication for each $j = 1, \dots, n+1$. Let also N be an $A_n(f_{(n)})$ -bimodule with involution and N be contained in some $A_{n+1}(f_{(n+1)})$ -bimodule P such that $\mathbb{C}_N(A_n(f_{(n)})) = \mathbb{C}_N(A_{n+1}(f_{(n+1)}))$, then $M = N \oplus (N\mathbf{1}_{n+1})$ is an $A_{n+1}(f_{(n+1)})$ -bimodule with involution and $M_0 = N_0$.*

P r o o f. By virtue of Theorem 1.1 N has the decomposition $N = \bigoplus_{k=0}^{2^n-1} N_0 i_k$ with $N_0 = \mathbb{C}_N(A_n(f_{(n)}))$, hence

$$M = \bigoplus_{j=0}^{2^{n+1}-1} N_0 i_j.$$

From Theorem 1.2 it follows that M is the $A_{n+1}(f_{(n+1)})$ -bimodule with involution and $M_0 = N_0$, since $\mathbb{C}_N(A_n(f_{(n)})) = \mathbb{C}_N(A_{n+1}(f_{(n+1)}))$. $\square$

R e m a r k 1.3. For the generalized Dickson algebra $B = A_n(f_{(n)})$ with $n \geq 2$, there is its unvolutorial algebra $\bar{B}$, which as an F -linear space, is the same, but has the multiplication obtained from B by the following formula: $\bar{a} \diamond \bar{b} = \bar{c}$ with $\bar{c} = ba$ induced from B by the involution operator $Jb = \bar{b}$ for each $\bar{a}, \bar{b}$ in $\bar{B}$, an addition in $\bar{B}$ is induced by that of in B .

Therefore, the left $\bar{B}$ -module ${}_B M$ also has the structure of the right B -module M_B such that $\bar{a} \diamond (\bar{b} \diamond x) = (xb)a$, where $\diamond$ denotes the multiplication of x in M on $\bar{b}$, $\bar{a}$ in $\bar{B}$. Using the tensor product over F and the involutorial algebra $\bar{B}$ instead of the opposite algebra B^{op} one gets the involutorial enveloping algebra $\check{B}^e = B \otimes_F \bar{B}$ instead of the enveloping algebra $B^e = B \otimes_F B^{op}$. Then the left $\check{B}^e$ -module ${}_{\check{B}^e} M$ also has the structure of B -bimodule ${}_B M_B$, but generally it may not have the structure of the B -bimodule with involution ${}_B \hat{M}_B$.

P r o p o s i t i o n 1.2. *Let $B = A_n(f_{(n)})$, $n \geq 2$, where A_0 is the commutative associative unital algebra with trivial involution $\bar{a} = a$ for each $a \in A_0$ over the commutative associative unital ring F , $\text{char}(F) \neq 2$, f_j is invertible in F relative to multiplication for each $j = 1, \dots, n$. Then there exist B -bimodules which are not B -bimodules with involution.*

P r o o f. Take $A_{n+p}(f_{(n+p)})$ with $n \geq 2$ and $p \geq 1$, with f_j invertible in F for each $j = n+1, \dots, n+p$. Then $M = A_{n+p}(f_{(n+p)})$ has the structure of the B -bimodule ${}_B M_B$, but it is not the B -bimodule with involution by Theorems 1.1 and 1.2. That is, this M does not satisfy conditions (1.21), (1.22), (1.23) in Definition 1.2. $\square$

Theorem 1.4. *Let ${}_B N$ the left B -module with $B = A_n(f_{(n)})$, $n \geq 2$, where A_0 is the commutative associative unital algebra with trivial involution $\bar{a} = a$ for each $a \in A_0$ over the commutative associative unital ring F , $\text{char}(F) \neq 2$, f_j is invertible in F relative to multiplication for each $j = 1, \dots, n$. Let $D \subset N$, $N \subseteq M$, where M has the structure of the B -bimodule with involution ${}_B \hat{M}_B$. Then $(BD)^{<m>} = (BD)^{<1>}$ for each $1 < m \leq \infty$ and $(BD)^{<1>}$ is the left B -submodule in ${}_B N$.*

P r o o f. In view of Lemma 1.2 $(BD)^{<1>}$ is the A_0 -linear span $\text{span}_{A_0} Q$ of the family $Q = \{i_j x : x \in D, j \in \{0, \dots, 2^n - 1\}\}$. By virtue of Theorem 1.1 each element x in M has the decomposition $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$ with $x_0, \dots, x_{2^n-1}$ belonging to M_0 , that is $x = \beta[x]$, where $\beta = \beta_n = (i_0, \dots, i_{2^n-1})$, $[x]^t = (x_0, \dots, x_{2^n-1})$, U^t denotes a transposed matrix of a matrix U . Consequently, $i_j x = x_0(i_j i_0) + \dots + x_{2^n-1}(i_j i_{2^n-1})$ for each $j \in \{0, \dots, 2^n - 1\}$, since $M_0 = \mathbb{C}_M(B)$.

On the other hand, $\{i_{j_1} \dots i_{j_m}\}_{q(m)} \in G$ for each $j_1, \dots, j_m$ in $\{0, \dots, 2^n - 1\}$, $2 \leq m \in \mathbb{N}$, where $G = G_n(f_{(n)})$, where $q(m)$ is a vector indicating an order of pairwise multiplications in $\{\dots\}$. Note that $s i_j G = G$ for each $j \in \{0, \dots, 2^n - 1\}$ and $s \in S$, where $S = S_n(f_{(n)})$.

Notice also that $S \subset F \subseteq A_0$. On the other side, $A_0(A_0\{x\}) = A_0\{x\}$ and $A_0(A_0\{b\}) = A_0\{b\}$ for each $x \in M$, $b \in B$. For each j, k, l in $\{0, \dots, 2^n - 1\}$ the minimal subalgebra $A_{(j,k,l)}$ in B generated by $\{A_0, i_j, i_k, i_l\}$ is alternative (see [1, 4, 9]). Therefore, $i_j(i_k i_l) + i_k(i_j i_l) = 0$ for each $j \neq k$ with $j \geq 1$ and $k \geq 1$, l in $\{0, \dots, 2^n - 1\}$, since $(i_j + i_k)((i_j + i_k)i_l) = ((i_j + i_k)(i_j + i_k))i_l$, $i_j i_k + i_k i_j = 0$, $i_j(i_j i_l) = (i_j i_j)i_l$. Then $i_k \beta = U_k \beta$ with $2^n \times 2^n$ matrix U_k with entries in S for each k . From this and Conditions (1.1)–(1.4) in Definition 1.2, and $x = \beta[x]$ for each $x \in {}_B \hat{M}_B$, it follows that $\text{span}_{A_0} Q = (BD)^{<1>}$ and $\text{span}_{A_0} Q = \text{span}_{A_0}(GD)$, $\text{span}_{A_0}(GD) = \text{span}_{A_0}(G \text{span}_{A_0}(GD))$, since $D \subset {}_B \hat{M}_B$, $S \subset A_0$. It implies that $(BD)^{<2>} = (BD)^{<1>}$. By induction this gives $(BD)^{<m+1>} = (BD)^{<m>}$ for each $2 \leq m \in \mathbb{N}$, hence $(BD)^{<\infty>} = (BD)^{<1>}$.

Certainly $(BD)^{<\infty>}$ is the left B -submodule in ${}_B N$, consequently, $(BD)^{<1>}$ is the left B -submodule in ${}_B N$. $\square$

Corollary 1.6. *Let the conditions of Theorem 1.4 be satisfied. Then $\overline{(BD)^{<\infty>}} = (\bar{D}B)^{<\infty>}$.*

P r o p o s i t i o n 1.3. *Let the conditions of Theorem 1.1 be satisfied with $A_0 = F$, where F is a field, $\text{char}(F) \neq 2$. Let either $M_0 = F^m$ and $m \in \mathbb{N}$, or M_0 be a F -linear space such that $M_0 \ominus Fy$ be isomorphic with M_0 for each $y \in M_0$. Then for each $x \in M$ there exist an invertible F -linear operator $V : M \rightarrow M$ and $b \in B$ and $y \in M_0$ such that $Vx = by$.*

P r o o f. If $x = 0$ the assertion of this theorem is evident. For $x \neq 0$ in M there is the decomposition $x = x_0 i_0 + \dots + x_{2^n-1} i_{2^n-1}$ with $x_0, \dots, x_{2^n-1}$ in M_0 such that there exists $k \in \{0, \dots, 2^n - 1\}$ with $x_k \neq 0$. So it is possible to choose such marked k . If $M_0 = F^m$, then it has a basis $e_1, \dots, e_m$ as the F -linear space. Therefore, for each $0 \neq x_j \in M_0$ there exists an invertible F -linear operator V_{x_j} on M_0 such that $V_{x_j} x_j = x_k$. If M_0 is the F -linear space such that $M_0 \ominus Fy$ is isomorphic with M_0 for each $y \in M_0$, then for each

$0 \neq x_j \in M_0$ there exists an invertible F -linear operator V_{x_j} on M_0 such that $V_{x_j}x_j = x_k$. We put $V = \bigoplus_{j=0}^{2^n-1} \hat{V}_{x_j}$, where $V_{x_l} = id_{M_0}$ if $x_l = 0$ or if $l = k$, where $\hat{V}_{x_j} : M_0 i_j \rightarrow M_0 i_j$, $\hat{V}_{x_j}(y i_j) = (V_{x_j}(y))i_j = i_j(V_{x_j}(y))$ for each $y \in M_0$. In view of Theorem 1.1 $M_0 = \mathbb{C}_M(B)$, hence it is naturally $V(bI) = (bI)V$ for each b in F , where $I = id_M$. Therefore, V is the left and right F -linear operator on M such that V is invertible on M , since $\mathbb{C}_B(B) = F$ in the considered case $A_0 = F$ with $n \geq 2$. This implies that $Vx = by$ with $y = x_k$ and $b = \sum_{j \in \Lambda_x} i_j$, where $\Lambda_x = \{j \in \{0, \dots, 2^n - 1\} : x_j \neq 0\}$. $\square$

Corollary 1.7. *Let B be the division alternative algebra, let M be a B -bimodule with involution satisfying the conditions of Theorem 1.1, $x = by$ with $y \in M_0$, $b \in B$. Then*

$$(B\{x\})^{<\infty>} = (\{x\}B)^{<\infty>} = (B\{x\})^{<\infty, \infty>}.$$

1.1. Conclusion

The results of this paper can be used for further studies of a structure of modules over nonassociative algebras, operator theory in modules over Dickson algebras, their applications to PDEs, mathematical physics, quantum field theory, their applications in other sciences, etc.

This can be used for analysis and solution of PDEs utilized in gas dynamics and high energy density physics, hydrodynamics, particularly, describing tidal deformations and the gravitational potential of the planet [17, 19–21].

It is worth to mention, that spectral theory of operators over Dickson algebras and particularly Cayley algebras was studied in [15–17]. Therefore, using the results obtained in this article, it will be important to investigate further operator theory in modules over generalized Dickson algebras, theory of factors for nonassociative analogs of C^* -algebras, analogs of direct integrals for them, applications in coding theory [22], etc.

2. Appendix

Definition 2.1. Let X be an algebra over a ring F , let M be a X -bimodule and $B \subseteq X$. We put

$$\begin{aligned}
 Com_M(B) &:= \{x \in M : \forall b \in B, xb = bx\}; \\
 N_{M,l}(B) &:= \{x \in M : \forall b \in B, \forall c \in B, (xb)c = x(bc)\}; \\
 N_{M,m}(B) &:= \{x \in M : \forall b \in B, \forall c \in B, (bx)c = b(xc)\}; \\
 N_{M,r}(B) &:= \{x \in M : \forall b \in B, \forall c \in B, (bc)x = b(cx)\}; \\
 N_M(B) &:= N_{M,l}(B) \cap N_{M,m}(B) \cap N_{M,r}(B) \text{ and} \\
 \mathbb{C}_M(B) &:= Com_M(B) \cap N_M(B).
 \end{aligned}$$

Then $Com_M(B)$, $N_M(B)$, and $\mathbb{C}_M(B)$ are called a commutant, a nucleus and a centralizer correspondingly of the X -bimodule M relative to a subset B in X . Instead of $Com_M(X)$, $N_M(X)$, or $\mathbb{C}_M(X)$ it will be also written shortly Com_M , N_M , or $\mathbb{C}_M$ correspondingly.

A left (or right) X -module M is also denoted by ${}_X M$ (or M_X correspondingly), similarly for bimodules.

Example 2.1. Particularly over the real field $F = A_0 = \mathbf{R}$ for $A_r(f_{(r)})$, $2 \leq r$, up to normalization of the doubling generator $\mathbf{l}_k$ on k -th step, a scalar $f_k \in \{-1, 1\}$ can be chosen for each $k = 1, 2, \dots$ (see Definition 1.1 and Remark 1.1). Frequently $\bar{a}$ is also denoted by a^* or $\tilde{a}$.

Definition 2.2. Let N and M be two left B -modules (see Definition 1.2). A map $T : N \rightarrow M$ we call a left B -quasi-linear operator, if it is additive:

$$T(v + w) = T(v) + T(w)$$

and left $C_B(B)$ -homogeneous:

$$T(av) = aT(v)$$

for each $a \in C_B(B)$, v and $w \in N$.

Evidently, each left B -quasi-linear operator is left $C_B(B)$ -linear. Similarly right B -quasi-linear operators for right B -modules are defined. If N and M are B -bimodules and a map $T : N \rightarrow M$ is left and right B -quasi-linear, then T will be called a B -quasi-linear operator.

If for left B -modules N and M the operator T is additive and

$$T(bv) = bT(v)$$

for each $b \in B$, v in N , then it will be called left B -linear. Analogously right B -linear operators for right B -modules are defined. If N and M are B -bimodules and a map $T : N \rightarrow M$ is left and right B -linear, then T will be called a B -linear operator.

The operator left or right B -quasi-linear (or left or right B -linear) $T : M \rightarrow M$ is called invertible if there exists a left or right B -quasi-linear (or left or right B -linear correspondingly) operator $V : M \rightarrow M$ such that $TV = I$ and $VT = I$, where $I = id_M$, where $id_M(x) = x$ for each $x \in M$. Then V is called an inverse operator of T and also denoted by T^{-1} .

References

- [1] L. E. Dickson, *The Collected Mathematical Papers*, Cambridge Library Collection – Mathematics, 1–5, Chelsea Publishing Co., New York, 1975.
- [2] K. A. Zhevlakov, A. M. Slin'ko, I. P. Shestakov, A. I. Shirshov, *Rings that are Nearly Associative*, Academic Press, New York, 1982.
- [3] R. B. Brown, "On generalized Cayley–Dickson algebras", *Pacific Journal of Mathematics*, **20**:3 (1967), 415–422.
- [4] J. C. Baez, "The octonions", *Bull. Amer. Math. Soc.*, **39**:2 (2002), 145–205.
- [5] R. D. Schafer, *An Introduction to Nonassociative Algebras*, Academic Press, New York, 1966.
- [6] S. V. Ludkovski, "Completions and homomorphisms of infinite dimensional Cayley–Dickson algebras", *Linear Multilinear Algebra*, **69**:11 (2021), 2040–2049.
- [7] D. Allcock, "Reflection groups and octave hyperbolic plane", *J. Algebra*, **213**:2 (1998), 467–498.
- [8] A. Belov, L. Bokut, L. Rowen, J.-T. Yu, "The Jacobian conjecture, together with Specht and Burnside-type problems", *Automorphisms and birational affine geometry*, Springer Proceedings in Mathematics and Statistics, **79**, ed. I. Cheltsov, 2014, 249–285.
- [9] N. Bourbaki, *Algebra*, Springer, Berlin, 2022.
- [10] N. Jacobson, "Cayley numbers and normal simple Lie algebras of type G ", *Duke Math. J.*, **5** (1939), 775–783.
- [11] M. Goto, F. D. Grosshans, *Semisimple Lie Algebras*, Marcel Dekker, Inc., New York, 1978.
- [12] P. Eakin, A. Sathaye, "On automorphisms and derivations of Cayley–Dickson algebras", *J. Algebra*, **129**:2 (1990), 263–278.
- [13] H. B. Lawson, M.-L. Michelson, *Spin Geometry*, Princ. Univ. Press, Princeton, 1989.
- [14] C. Culbert, "Cayley–Dickson algebras and loops", *J. Gener. Lie Theory Appl.*, **1**:1 (2007), 1–17.

- [15] S. V. Ludkovsky, “ C^* -algebras of meta-invariant operators in modules over Cayley–Dickson algebras”, *Southeast Asian Bull. Math.*, **39**:5 (2015), 625–684.
- [16] S. V. Ludkovsky, W. Sprössig, “Spectral representations of operators in Hilbert spaces over quaternions and octonions”, *Complex Var. Elliptic Equ.*, **57**:12 (2012), 1301–1324.
- [17] S. V. Ludkovsky, W. Sprössig, “Spectral theory of super-differential operators of quaternion and octonion variables”, *Adv. Appl. Clifford Algebr.*, **21**:1 (2011), 165–191.
- [18] F. Gürsey, C.-H. Tze, *On the Role of Division, Jordan and Related Algebras in Particle Physics*, World Scientific Publ. Co., Singapore, 1996.
- [19] A. V. Shatina, A. S. Borets, “A mathematical model of the gravitational potential of the planet taking into account tidal deformations”, *Russian Technol J.*, **12**:2 (2024), 77–89.
- [20] S. V. Ludkovsky, “Integration of vector Sobolev type PDE over octonions”, *Complex Var. Elliptic Equ.*, **61**:7 (2016), 1014–1035.
- [21] S. V. Ludkovsky, “Integration of vector hydrodynamical partial differential equations over octonions”, *Complex Var. Elliptic Equ.*, **58**:5 (2013), 579–609.
- [22] V. N. Markov, A. V. Mikhalev, A. A. Nechaev, “Nonassociative algebraic structures in cryptography and coding”, *J. Math. Sci. (N. Y.)*, **245**:2 (2020), 178–196.

Information about the author

Sergey V. Ludkovsky, Doctor of Physical and Mathematical Sciences, Professor of Applied Mathematics Department, MIREA – Russian Technological University, Moscow, Russian Federation. E-mail: sludkowski@mail.ru
ORCID: <https://orcid.org/0000-0002-4733-8151>

Received 29.07.2024
Reviewed 15.10.2024
Accepted for press 22.11.2024

Информация об авторе

Людковский Сергей Викторович, доктор физико-математических наук, профессор кафедры прикладной математики, МИРЭА – Российский технологический университет, г. Москва, Российская Федерация. E-mail: sludkowski@mail.ru
ORCID: <https://orcid.org/0000-0002-4733-8151>

Поступила в редакцию 29.07.2024 г.
Поступила после рецензирования 15.10.2024 г.
Принята к публикации 22.11.2024 г.

НАУЧНАЯ СТАТЬЯ

© Пименов В.Г., Лекомцев А.В., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-440-454>

УДК 519.633



Компактная схема для решения супердиффузионного уравнения с несколькими переменными запаздываниями

Владимир Германович ПИМЕНОВ^{1,2}, Андрей Валентинович ЛЕКОМЦЕВ¹

¹ ФГАОУ ВО «Уральский федеральный университет им. первого Президента России Б.Н. Ельцина»

620002, Российская Федерация, г. Екатеринбург, ул. Мира, 19

² ФГБУН «Институт математики и механики им. Н.Н. Красовского»

Уральского отделения Российской академии наук

620108, Российская Федерация, г. Екатеринбург, ул. Софьи Ковалевской, 16

Аннотация. Рассматривается супердиффузионное уравнение с дробными производными Рисса по пространству с несколькими переменными запаздываниями. Производится дискретизация задачи. По времени конструируется аналог разностного метода Кранка–Николсона с кусочно-линейной интерполяцией для учета эффекта переменного запаздывания и с экстраполяцией продолжением для того, чтобы неявность метода стала конечномерной. По пространству конструируется аналог компактной схемы со специальной заменой дробных производных Рисса дробными центральными разностями. В результате метод сводится к решению на каждом шаге времени системы линейных алгебраических уравнений с симметричной и положительно определенной главной матрицей. Изучается порядок малости относительно шагов дискретизации по времени Δ и пространству h невязки метода без интерполяции и с интерполяцией, он равен $O(\Delta^2 + h^4)$. Основной результат состоит в доказательстве того, что метод сходится с порядком $O(\Delta^2 + h^4)$ в энергетической и компактной норме послыонного вектора погрешности. Приводятся результаты тестовых примеров для супердиффузионных уравнений с постоянным и переменным запаздываниями. Вычислимые порядки сходимости по каждому шагу дискретизации в примерах оказались близки к теоретически полученным порядкам сходимости по соответствующим шагам дискретизации.

Ключевые слова: уравнение супердиффузии, несколько переменных запаздываний, компактная схема, кусочно-линейная интерполяция

Для цитирования: Пименов В.Г., Лекомцев А.В. Компактная схема для решения супердиффузионного уравнения с несколькими переменными запаздываниями // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 440–454.

<https://doi.org/10.20310/2686-9667-2024-29-148-440-454>

SCIENTIFIC ARTICLE

© V. G. Pimenov, A. V. Lekomtsev, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-440-454>

A compact scheme for solving a superdiffusion equation with several variable delays

Vladimir G. PIMENOV^{1,2}, Andrei V. LEKOMTSEV¹

¹ Ural Federal University named after the first President of Russia B. N. Yeltsin

19 Mira St., Yekaterinburg 620002, Russian Federation

² N. N. Krasovskii Institute of Mathematics and Mechanics

of the Ural Branch of the Russian Academy of Sciences

16 S. Kovalevskaya St., Yekaterinburg 620108, Russian Federation

Abstract. A superdiffusion equation with Riesz fractional derivatives with respect to space with several delay variables is considered. The problem is discretized. For this purpose, an analog of the Crank–Nicolson difference method with piecewise linear interpolation to account for the effect of variable delay and with extrapolation by continuation is constructed in time so that the implicitness of the method becomes finite-dimensional. An analog of a compact scheme with a special replacement of Riesz fractional derivatives by fractional central differences is constructed in space. As a result, the method is reduced to solving a system of linear algebraic equations with symmetric and positive-definite main matrix at each time step. The order of smallness with respect to the discretization time-steps Δ and space-steps h of the residual of the method without interpolation and with interpolation is studied; it is equal to $O(\Delta^2 + h^4)$. The main result consists in proving that the method converges with the order $O(\Delta^2 + h^4)$ in the energy and compact norm of the layered error vector. The results of test examples for superdiffusion equations with constant and variable delay are presented. The computable orders of convergence for each discretization step in the examples turned out to be close to the theoretically obtained orders of convergence for the corresponding discretization steps.

Keywords: superdiffusion equation, several variable delays, compact scheme, piecewise linear interpolation

Mathematics Subject Classification: 65N06, 34A08.

For citation: Pimenov V.G., Lekomtsev A.V. A compact scheme for solving a superdiffusion equation with several variable delays. *Vestnik Rossiyskikh Universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 440–454.

<https://doi.org/10.20310/2686-9667-2024-29-148-440-454> (In Russian, Abstr. in Engl.)

Введение

Уравнения в частных производных с запаздывающим аргументом широко применяются в математическом моделировании (см., например, [1, 2]), также в последнее время все чаще в математических моделях используются дробные производные. Численные алгоритмы решения уравнений с частными производными с эффектом запаздывания изучались, например, в [3, 4]. Литература по численным методам решения уравнений с дробными производными огромна, отметим лишь работы [5–8], в которых рассматривались численные методы решения уравнения супердиффузионного типа, т. е. с дробной производной по пространственной переменной порядка от 1 до 2. Для супердиффузионных уравнений с функциональным эффектом запаздывания в работе [9] построен и исследован численный метод порядка $\Delta^2 + h$, где Δ — шаг разбиения по времени, а h — шаг разбиения по пространству, а в работе [10] построен и исследован численный метод порядка $\Delta^2 + h^2$. В работе [11] на основе процедуры экстраполяции Ричардсона был построен метод порядка $\Delta^3 + h^3$, однако при жестком условии пропорциональности шагов Δ и h .

Целью данной работы было построение для супердиффузионного уравнения с эффектом запаздывания численного метода порядка $\Delta^2 + h^4$, основанного на идеях компактной разностной схемы. Компактная схема изучалась для разных дробных уравнений без запаздывания во многих работах, в частности, для субдиффузионных уравнений в [12, 13], для дробного по времени и пространству уравнения Блоха–Торри в [14], для субдиффузионных уравнений с постоянным запаздыванием в [15].

В данной работе для супердиффузионного уравнения с производной Рисса и эффектом запаздывания строится аналог метода Кранка–Никольсона по временной переменной и аналог компактной схемы для дробной пространственной переменной. Основным результатом работы — доказательство сходимости в энергетической норме с порядком $\Delta^2 + h^4$. Техника доказательства потребовала ограничить эффект запаздывания до уравнений с несколькими переменными запаздываниями. В силу переменности запаздываний, существенным элементом алгоритма является интерполяция и экстраполяция с заданными свойствами для функции-предыстории численной модели, в работе применяется кусочно-линейная интерполяция с экстраполяцией продолжением.

1. Постановка задачи и основные предположения

Рассмотрим уравнение супердиффузионного типа с несколькими переменными запаздываниями

$$\frac{\partial u(x, t)}{\partial t} = d \frac{\partial^\alpha u}{\partial |x|^\alpha} + f(x, t, u(t - \tau_1(t)), u(t - \tau_2(t)), \dots, u(t - \tau_K(t))), \quad (1.1)$$

где $0 \leq t \leq T$, $0 \leq x \leq X$ — независимые переменные, $u(x, t)$ — искомая функция решения, $d > 0$ — коэффициент супердиффузии, величины запаздывания удовлетворяют ограничениям $0 \leq \tau_k(t) \leq \tau$ для всех $k = 1, \dots, K$ и $t \in [0, T]$. Дробная производная Рисса по пространству порядка α , $1 < \alpha < 2$, определяется формулой

$$\frac{\partial^\alpha u(x, t)}{\partial |x|^\alpha} = -\frac{1}{2\cos(\alpha\pi/2)\Gamma(2-\alpha)} \frac{\partial^2}{\partial x^2} \int_0^x \frac{u(\xi, t)}{|x-\xi|^{\alpha-1}} d\xi.$$

Заданы граничные условия

$$u(0, t) = 0, \quad u(X, t) = 0, \quad 0 \leq t \leq T, \quad (1.2)$$

и начальные условия

$$u(x, s) = \varphi(x, s), \quad 0 \leq x \leq X, \quad -\tau \leq s \leq 0. \quad (1.3)$$

Будем предполагать, что решение $u(x, t)$ задачи (1.1)–(1.3) существует и единственно. Кроме того, при доказательстве сходимости численных методов будем предполагать необходимую гладкость решения $u(x, t)$.

Дополнительно будем предполагать, что функция $f(x, t, u^1, u^2, \dots, u^K)$ липшицева по u^k , т. е. существует постоянная L_f такая, что для всех $x \in [0, X]$, $t \in [0, T]$, $u^1, u^2, \dots, u^K$, $v^1, v^2, \dots, v^K$ выполняется

$$|f(x, t, u^1, u^2, \dots, u^K) - f(x, t, v^1, v^2, \dots, v^K)| \leq L_f \sum_{k=1}^K |u^k - v^k|. \quad (1.4)$$

2. Дискретизация. Метод Кранка–Никольсон с кусочно-линейной интерполяцией и дробным оператором центральных разностей

Введем шаг по времени $\Delta = \frac{\tau}{M_0}$, где M_0 — натуральное, и пусть $M = \lfloor \frac{T}{\Delta} \rfloor$. Введем точки (узлы по времени) $t_j = j\Delta$, $j = -M_0, \dots, M$. Полуцелые узлы будем обозначать $t_{j+\frac{1}{2}} = t_j + \frac{\Delta}{2}$.

Разобьем отрезок $[0, X]$ на части, введя шаг по пространству $h = \frac{X}{N}$, где N — целое. Введем точки (узлы по пространству) $x_i = ih$, $i = 0, \dots, N$. Аппроксимацию функции $u(x_i, t_j)$ в узлах будем обозначать U_j^i .

При всяком фиксированном $i = 0, \dots, N$ введем дискретную предысторию к моменту t_j , $j = 0, \dots, M$: $\{U_k^i\}_j = \{U_k^i, j - M_0 \leq k \leq j\}$.

О п р е д е л е н и е 2.1. Назовем отображение $I : \{U_k^i\}_j \rightarrow U_j^i(\cdot) \in C[t_j - \tau, t_{j+\frac{1}{2}}]$ оператором интерполяции-экстраполяции дискретной предыстории.

О п р е д е л е н и е 2.2. Будем говорить, что оператор интерполяции-экстраполяции имеет *порядок погрешности* p на точном решении, если существуют константы C_1 и C_2 такие, что для всех i, j и $t \in [t_j - \tau, t_{j+\frac{1}{2}}]$ выполняется неравенство

$$|U_j^i(t) - u(x_i, t)| \leq C_1 \max_{j-M_0 \leq k \leq j} |U_k^i - u(x_i, t_k)| + C_2 \Delta^p.$$

Рассмотрим кусочно-линейную интерполяцию $I(\{U_k^i\}_j) = U_j^i(\cdot)$, задаваемую соотношениями

$$U_j^i(t_j + s) = \frac{1}{\Delta}((t_k - t_j - s)U_{k-1}^i + (t_j + s - t_{k-1})U_k^i), \quad (2.1)$$

$$t_{k-1} \leq t_j + s \leq t_k, \quad j - M_0 \leq k \leq j,$$

с экстраполяцией продолжением

$$U_j^i(t_j + s) = \frac{1}{\Delta}((-s)U_{j-1}^i + (\Delta + s)U_j^i), \quad t_j \leq t_j + s \leq t_{j+\frac{1}{2}}. \quad (2.2)$$

Кусочно-линейная интерполяция с экстраполяцией продолжением имеет порядок 2, если точное решение $u(x, t)$ дважды непрерывно дифференцируемо по t на промежутке $[-\tau, T]$ (см. [16, с. 97]). Кусочно-линейная интерполяция обладает следующим свойством.

Пусть $u(t)$ и $v(t)$ — результаты кусочно-линейной интерполяции по узлам $\{u_k\}_j$ и $\{v_k\}_j$ соответственно и $t \in [t_{k-1}, t_k]$, $j - M_0 \leq k \leq j$. Тогда

$$|u(t) - v(t)| \leq |u_{k-1} - v_{k-1}| + |u_k - v_k|. \quad (2.3)$$

Экстраполяция продолжением обладает следующим свойством. Пусть $u(t)$ и $v(t)$ являются результатами экстраполяции продолжением кусочно-линейной интерполяции по узлам $\{u_k\}_j$ и $\{v_k\}_j$ соответственно и $t \in [t_j, t_{j+\frac{1}{2}}]$. Тогда

$$|u(t) - v(t)| \leq \frac{1}{2}|u_{j-1} - v_{j-1}| + \frac{3}{2}|u_j - v_j|. \quad (2.4)$$

Оба эти свойства проверяются по определению кусочно-линейной интерполяции и экстраполяции продолжением. Объединяя (2.3) и (2.4), получаем, что если $u(t)$ и $v(t)$ результаты кусочно-линейной интерполяции с экстраполяцией продолжением по узлам $\{u_k\}_j$ и $\{v_k\}_j$ соответственно и $t \in [t_j - \tau, t_{j+\frac{1}{2}}]$, то

$$|u(t) - v(t)| \leq \frac{3}{2}(|u_{k-1} - v_{k-1}| + |u_k - v_k|), \quad (2.5)$$

где $t \in [t_{k-1}, t_k]$ в случае интерполяции и $k = j$ в случае экстраполяции. Назовем это свойство *квазилипшицевостью* с константой $3/2$.

Для аппроксимации дробной производной Рисса выберем метод дробных центральных разностей [7]. Пусть

$$g_k^\alpha = \frac{(-1)^k \Gamma(\alpha + 1)}{\Gamma(\alpha/2 - k + 1) \Gamma(\alpha/2 + k + 1)}, \quad k = 0, \mp 1, \mp 2, \dots$$

Тогда, если функция $u(x, t)$ пять раз непрерывно дифференцируема по x на отрезке $[0, X]$, то (см. [7])

$$\frac{\partial^\alpha u}{\partial |x|^\alpha} = -h^{-\alpha} \sum_{k=-\frac{X-x}{h}}^{\frac{x}{h}} g_k^\alpha u(x - kh, t) + O(h^2), \quad |O(h^2)| \leq C_3 h^2.$$

Методом Кранка–Николсон назовем неявную разностную схему

$$\begin{aligned} \frac{U_{j+1}^i - U_j^i}{\Delta} = & -\frac{1}{2} dh^{-\alpha} \sum_{k=-N+i+1}^{i-1} g_k^\alpha U_{j+1}^{i-k} - \frac{1}{2} dh^{-\alpha} \sum_{k=-N+i+1}^{i-1} g_k^\alpha U_j^{i-k} \\ & + f(x_i, t_{j+\frac{1}{2}}, U_j^i(t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), U_j^i(t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, U_j^i(t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}}))), \\ & i = 1, \dots, N-1, \quad j = 0, \dots, M-1, \end{aligned} \quad (2.6)$$

с начальными условиями

$$U_j^i(t) = \varphi(x_i, t), \quad -\tau \leq t \leq 0, \quad i = 0, \dots, N,$$

и граничными условиями

$$U_j^0 = 0, \quad U_j^N = 0, \quad j = 0, \dots, M,$$

где $U_j^i(t)$ — результат действия оператора кусочно-линейной интерполяции (2.1) с экстраполяцией (2.2). Обозначим

$$\delta_x^\alpha U_j^i = h^{-\alpha} \sum_{k=-N+i+1}^{i-1} g_k^\alpha U_j^{i-k},$$

$$\tilde{f}_{j+\frac{1}{2}}^i = f(x_i, t_{j+\frac{1}{2}}, U_j^i(t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), U_j^i(t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, U_j^i(t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}}))).$$

Тогда метод (2.6) может быть записан в краткой форме

$$\frac{U_{j+1}^i - U_j^i}{\Delta} = -\frac{1}{2} d\delta_x^\alpha U_{j+1}^i - \frac{1}{2} d\delta_x^\alpha U_j^i + \tilde{f}_{j+\frac{1}{2}}^i.$$

Этот метод был изучен в работе [10] в более общем случае функционального запаздывания и двух пространственных переменных. Как следует из результатов работы [10], метод устойчив и сходится с порядком $\Delta^2 + h^2$. Цель этой статьи состоит в построении алгоритма более высокого порядка сходимости на основе метода (2.6).

3. Компактная схема

Введем следующие операторы

$$\delta_x^2 U_j^i = \frac{U_j^{i-1} - 2U_j^i + U_j^{i+1}}{h^2}, \quad i = 1, \dots, N-1,$$

$$HU_j^i = U_j^i + \frac{\alpha h^2}{24} \delta_x^2 U_j^i, \quad i = 1, \dots, N-1.$$

Сконструируем метод, который по аналогии с подобными методами для других уравнений будем называть *компактной схемой*:

$$\frac{HU_{j+1}^i - HU_j^i}{\Delta} = -\frac{1}{2} d\delta_x^\alpha U_{j+1}^i - \frac{1}{2} d\delta_x^\alpha U_j^i + H\tilde{f}_{j+\frac{1}{2}}^i. \quad (3.1)$$

Перепишем метод в виде

$$(H + \frac{\Delta}{2} d\delta_x^\alpha) U_{j+1}^i = (H - \frac{\Delta}{2} d\delta_x^\alpha) U_j^i + \Delta H \tilde{f}_{j+\frac{1}{2}}^i. \quad (3.2)$$

Рассмотрим главную матрицу $A = H + \frac{\Delta}{2} d\delta_x^\alpha$ этой системы. Распишем матрицы

$$H = \begin{pmatrix} 1 - \frac{\alpha}{12} & \frac{\alpha}{24} & \dots & 0 & 0 & 0 & \dots & 0 \\ \frac{\alpha}{24} & 1 - \frac{\alpha}{12} & \dots & 0 & 0 & 0 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \frac{\alpha}{24} & 1 - \frac{\alpha}{12} & \frac{\alpha}{24} & \dots & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 - \frac{\alpha}{12} \end{pmatrix},$$

$$\delta_x^\alpha = \begin{pmatrix} g_0 & g_{-1} & \dots & g_{2-i} & g_{1-i} & g_{-i} & \dots & g_{2-N} \\ g_1 & g_0 & \dots & g_{3-i} & g_{2-i} & g_{1-i} & \dots & g_{3-N} \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ g_{i-1} & g_{i-2} & \dots & g_{-1} & g_0 & g_1 & \dots & g_{i+1-N} \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ g_{N-2} & g_{N-3} & \dots & g_{N-i} & g_{N-i-1} & g_{N-i-2} & \dots & g_0 \end{pmatrix}.$$

Матрица H симметричная и положительно определенная. В силу свойств коэффициентов $g_0 > 0$, $g_{-k} = g_k < 0$, $k \neq 0$, $\sum_{k=-\infty}^{\infty} g_k = 0$, матрица δ_x^α также симметричная и положительно определенная. Откуда следует, что матрица A — симметричная и положительно определенная, т. е. система (3.2) однозначно разрешима.

4. Невязка компактной схемы

О п р е д е л е н и е 4.1. Невязкой без интерполяции метода (3.1) назовем

$$\psi_j^i = H \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{\Delta} + \frac{1}{2} d\delta_x^\alpha u(x_i, t_{j+1}) + \frac{1}{2} d\delta_x^\alpha u(x_i, t_j) - H f_{j+\frac{1}{2}}^i,$$

$$f_{j+\frac{1}{2}}^i = f(x_i, t_{j+\frac{1}{2}}, u(x_i, t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), u(x_i, t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, u(x_i, t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}}))).$$

Лемма 4.1. Если точное решение $u(x, t)$ шесть раз непрерывно дифференцируемо по x и по t , причем смешанные частные производные до шестого порядка непрерывны, функция f дважды непрерывно дифференцируема по x , то невязка без интерполяции метода (3.1) представима в виде

$$\psi_j^i = O(\Delta^2 + h^4), \quad |O(\Delta^2 + h^4)| \leq C_4(\Delta^2 + h^4).$$

Д о к а з а т е л ь с т в о. Разложим по формуле Тейлора величины, входящие в определение невязки без интерполяции метода (3.1) в окрестности точки с координатами $(x_i, t_{j+\frac{1}{2}})$. Для краткости будем опускать аргумент $(x_i, t_{j+\frac{1}{2}})$ у функции $u = u(x_i, t_{j+\frac{1}{2}})$ и ее производных. Тогда

$$\begin{aligned} \psi_j^i &= H \left(\frac{u}{\Delta} + \frac{1}{2} u'_t + \frac{\Delta}{4} u''_{tt} + O(\Delta^2) - \frac{u}{\Delta} + \frac{1}{2} u'_t - \frac{\Delta}{4} u''_{tt} + O(\Delta^2) \right) \\ &\quad + \frac{d}{2} \delta_x^\alpha \left(u + \frac{\Delta}{2} u'_t + O(\Delta^2) + u - \frac{\Delta}{2} u'_t + O(\Delta^2) \right) - H f_{j+\frac{1}{2}}^i \\ &= H u'_t + d \delta_x^\alpha u - H f_{j+\frac{1}{2}}^i + O(\Delta^2). \end{aligned} \quad (4.1)$$

Умножим уравнение (1.1) при $u = u(x_i, t_{j+\frac{1}{2}})$ на H , получим

$$H u'_t = d H \frac{\partial^\alpha u}{\partial |x|^\alpha} + H f_{j+\frac{1}{2}}^i. \quad (4.2)$$

Выполняется (см. [14])

$$H \frac{\partial^\alpha u}{\partial |x|^\alpha} = -\delta_x^\alpha u + O(h^4). \quad (4.3)$$

Тогда из (4.1), (4.2), (4.3) вытекает утверждение леммы. $\square$

О п р е д е л е н и е 4.2. Невязкой метода (3.1) с кусочно-линейной интерполяцией и экстраполяцией продолжением назовем величину

$$\hat{\psi}_j^i = H \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{\Delta} + \frac{1}{2} d\delta_x^\alpha u(x_i, t_{j+1}) + \frac{1}{2} d\delta_x^\alpha u(x_i, t_j) - H \hat{f}_{j+\frac{1}{2}}^i, \quad (4.4)$$

$$\hat{f}_{j+\frac{1}{2}}^i = f(x_i, t_{j+\frac{1}{2}}, \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}}))),$$

где $\hat{u}(x_i, t_j + s)$ — результат кусочно-линейной интерполяции с экстраполяцией продолжением точного решения:

$$\begin{aligned}\hat{u}(x_i, t_j + s) &= \frac{1}{\Delta}((t_k - t_j - s)u(x_i, t_{k-1}) + (t_j + s - t_{k-1})u(x_i, t_k)), \\ t_{k-1} &\leq t_j + s \leq t_k, \quad j - M_0 \leq k \leq j, \\ \hat{u}(x_i, t_j + s) &= \frac{1}{\Delta}((-s)u(x_i, t_{j-1}) + (\Delta + s)u(x_i, t_j)), \quad t_j \leq t_j + s \leq t_{j+\frac{1}{2}}.\end{aligned}$$

Лемма 4.2. При условиях леммы 4.1 невязка метода (3.1) с кусочно-линейной интерполяцией и экстраполяцией продолжением представима в виде

$$\hat{\psi}_j^i = O(\Delta^2 + h^4), \quad |O(\Delta^2 + h^4)| \leq C_5(\Delta^2 + h^4).$$

Д о к а з а т е л ь с т в о. Невязка без интерполяции и невязка с кусочно-линейной интерполяцией и экстраполяцией продолжением связаны соотношением

$$\hat{\psi}_j^i = \psi_j^i + Hf_{j+\frac{1}{2}}^i - H\hat{f}_{j+\frac{1}{2}}^i. \quad (4.5)$$

В силу определения оператора H , липшицевости (1.4) функции f и свойств оператора кусочно-линейной интерполяцией с экстраполяцией продолжением получаем

$$\begin{aligned}|H(f_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i)| &\leq \max_{i=1, \dots, N-1} |f_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i| \\ &\leq \max_{i=1, \dots, N-1} L_f \sum_{k=1}^K |u(x_i, t_{j+\frac{1}{2}} - \tau_k(t_{j+\frac{1}{2}})) - \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_k(t_{j+\frac{1}{2}}))| \leq KL_f C_2 \Delta^2.\end{aligned}$$

Учитывая эту оценку, а также утверждение леммы 4.1, из (4.5) вытекает утверждение леммы 4.2. $\square$

5. Сходимость компактной схемы

Рассмотрим векторное пространство размерности $N - 1$, в котором введем скалярное произведение: если $u = (u^1, u^2, \dots, u^{N-1})$ и $v = (v^1, v^2, \dots, v^{N-1})$, то

$$(u, v) = \sum_{i=1}^{N-1} h u^i v^i.$$

Введем энергетическую норму

$$\|u\|^2 = (u, u),$$

компактную норму и норму предыстории погрешности

$$\|u\|_H^2 = (u, Hu), \quad \|\{u_k\}_j\|_\tau = \max_{j-M_0 \leq k \leq j} \|u_k\|_H.$$

Выполняются оценки (см. [12])

$$\frac{2}{3}\|u\|^2 \leq \|u\|_H^2 \leq \|u\|^2. \quad (5.1)$$

Справедливы следующие вспомогательные оценки.

Лемма 5.1.

$$\|Hu\| \leq \|u\|. \quad (5.2)$$

Д о к а з а т е л ь с т в о. Наряду с вектором $u = (u^1, u^2, \dots, u^{N-1})$ введем векторы смещения $u^- = (0, u^1, \dots, u^{N-2})$ и $u^+ = (u^2, u^3, \dots, 0)$. По определению энергетической нормы выполняются соотношения $\|u^-\| \leq \|u\|$, $\|u^+\| \leq \|u\|$.

Вектор Hu представим в виде

$$Hu = u + \frac{\alpha h^2}{24} \delta_x^2 u = u + \frac{\alpha h^2}{24} \frac{u^- - 2u + u^+}{h^2} = \frac{1}{24}(\alpha u^- + (24 - 2\alpha)u + \alpha u^+),$$

тогда

$$\|Hu\| \leq \frac{1}{24}(\alpha \|u^-\| + (24 - 2\alpha)\|u\| + \alpha \|u^+\|) \leq \frac{1}{24}(\alpha \|u\| + (24 - 2\alpha)\|u\| + \alpha \|u\|) = \|u\|.$$

□

Лемма 5.2. Если $a_1, a_2, \dots, a_K$ — произвольные числа, то

$$\left(\sum_{k=1}^K a_k\right)^2 \leq 2^{K-1} \sum_{k=1}^K a_k^2.$$

Д о к а з а т е л ь с т в о. Проводится индукцией по K . □

Введем погрешность $\varepsilon_j^i = u(x_i, t_j) - U_j^i$ и послойную векторную погрешность метода (3.1) как вектор ε_j с координатами ε_j^i , $i = 1, \dots, N-1$.

Теорема 5.1. Пусть выполняются условия леммы 4.1. Тогда

$$\|\varepsilon_j\| \leq C(\Delta^2 + h^4).$$

Д о к а з а т е л ь с т в о. Используя (3.1) и (4.4), выпишем уравнения для координат вектора погрешности

$$H \frac{\varepsilon_{j+1}^i - \varepsilon_j^i}{\Delta} = -\frac{1}{2} d\delta_x^\alpha \varepsilon_{j+1}^i - \frac{1}{2} d\delta_x^\alpha \varepsilon_j^i + H(\tilde{f}_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i) + \hat{\psi}_j^i. \quad (5.3)$$

Введем векторы $\tilde{f}_{j+\frac{1}{2}} = (\tilde{f}_{j+\frac{1}{2}}^1, \tilde{f}_{j+\frac{1}{2}}^2, \dots, \tilde{f}_{j+\frac{1}{2}}^{N-1})$, $\hat{f}_{j+\frac{1}{2}} = (\hat{f}_{j+\frac{1}{2}}^1, \hat{f}_{j+\frac{1}{2}}^2, \dots, \hat{f}_{j+\frac{1}{2}}^{N-1})$, $\hat{\psi}_{j+\frac{1}{2}} = (\hat{\psi}_{j+\frac{1}{2}}^1, \hat{\psi}_{j+\frac{1}{2}}^2, \dots, \hat{\psi}_{j+\frac{1}{2}}^{N-1})$ и запишем (5.3) в векторном виде

$$H \frac{\varepsilon_{j+1} - \varepsilon_j}{\Delta} = -\frac{1}{2} d\delta_x^\alpha \varepsilon_{j+1} - \frac{1}{2} d\delta_x^\alpha \varepsilon_j + H(\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}) + \hat{\psi}_j.$$

Умножим это уравнение скалярно на вектор $\varepsilon_{j+\frac{1}{2}} = \frac{1}{2}(\varepsilon_j + \varepsilon_{j+1})$, получим

$$\left(H \frac{\varepsilon_{j+1} - \varepsilon_j}{\Delta}, \varepsilon_{j+\frac{1}{2}}\right) = -(d\delta_x^\alpha \varepsilon_{j+\frac{1}{2}}, \varepsilon_{j+\frac{1}{2}}) + (H(\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}), \varepsilon_{j+\frac{1}{2}}) + (\hat{\psi}_j, \varepsilon_{j+\frac{1}{2}}). \quad (5.4)$$

Левая часть этого уравнения преобразуется к виду

$$\left(H \frac{\varepsilon_{j+1} - \varepsilon_j}{\Delta}, \varepsilon_{j+\frac{1}{2}}\right) = \frac{1}{2\Delta}(\|\varepsilon_{j+1}\|_H^2 - \|\varepsilon_j\|_H^2). \quad (5.5)$$

Первое слагаемое в правой части (5.4) в силу положительности d и положительной определенности δ_x^α оценивается следующим образом

$$-(d\delta_x^\alpha \varepsilon_{j+\frac{1}{2}}, \varepsilon_{j+\frac{1}{2}}) \leq 0. \quad (5.6)$$

Оценим второе слагаемое в правой части (5.4), учитывая (5.2) и (5.1),

$$\begin{aligned} (H(\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}), \varepsilon_{j+\frac{1}{2}}) &\leq \|H(\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}})\| \|\varepsilon_{j+\frac{1}{2}}\| \\ &\leq \|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\| \frac{3}{2} \|\varepsilon_{j+\frac{1}{2}}\|_H \leq \|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\| \frac{3}{4} (\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H). \end{aligned} \quad (5.7)$$

Оценим компоненты вектора $\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}$, учитывая липшицевость (1.4) функции f

$$\begin{aligned} |\tilde{f}_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i| &= |f(x_i, t_{j+\frac{1}{2}}, U_j^i(t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), U_j^i(t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, U_j^i(t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}}))) \\ &\quad - f(x_i, t_{j+\frac{1}{2}}, \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_1(t_{j+\frac{1}{2}})), \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_2(t_{j+\frac{1}{2}})), \dots, \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_K(t_{j+\frac{1}{2}})))| \\ &\leq L_f \sum_{l=1}^K |U_j^i(t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}})) - \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}}))|. \end{aligned} \quad (5.8)$$

Учитывая свойство квазилипшицевости (2.5) оператора кусочно-линейной интерполяции с экстраполяцией продолжением, получаем

$$|U_j^i(t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}})) - \hat{u}(x_i, t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}}))| \leq \frac{3}{2} (|\varepsilon_{k(l)-1}| + |\varepsilon_{k(l)}|), \quad (5.9)$$

где $t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}}) \in [t_{k(l)-1}, t_{k(l)}]$ или $k(l) = j$ в случае $t_{j+\frac{1}{2}} - \tau_l(t_{j+\frac{1}{2}}) > t_j$. Таким образом, из (5.8) и (5.9) следует

$$|\tilde{f}_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i| \leq L_f \frac{3}{2} \sum_{l=1}^K (|\varepsilon_{k(l)-1}| + |\varepsilon_{k(l)}|).$$

По определению энергетической нормы получаем

$$\|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\|^2 = \sum_{i=1}^{N-1} h |\tilde{f}_{j+\frac{1}{2}}^i - \hat{f}_{j+\frac{1}{2}}^i|^2 \leq (L_f \frac{3}{2})^2 \sum_{i=1}^{N-1} h (\sum_{l=1}^K (|\varepsilon_{k(l)-1}| + |\varepsilon_{k(l)}|))^2.$$

Из полученного соотношения, используя оценку леммы 5.2, следует

$$\begin{aligned} \|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\|^2 &\leq (L_f \frac{3}{2})^2 \sum_{i=1}^{N-1} h 2^{2K-1} \sum_{l=1}^K (|\varepsilon_{k(l)-1}|^2 + |\varepsilon_{k(l)}|^2) \\ &= (L_f \frac{3}{2})^2 2^{2K-1} \sum_{l=1}^K (\|\varepsilon_{k(l)-1}\|^2 + \|\varepsilon_{k(l)}\|^2). \end{aligned}$$

Извлекая корень, получаем

$$\|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\| \leq L_f \frac{3}{2} 2^{K-\frac{1}{2}} \left(\sum_{l=1}^K (\|\varepsilon_{k(l)-1}\|^2 + \|\varepsilon_{k(l)}\|^2) \right)^{\frac{1}{2}} \leq 3L_f 2^K \sum_{l=1}^K (\|\varepsilon_{k(l)-1}\| + \|\varepsilon_{k(l)}\|).$$

Оценивая каждое слагаемое суммы в правой части этого соотношения, получаем неравенство

$$\|\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}\| \leq 3L_f 2^{K+1} K \max_{j-M_0 \leq k \leq j} \|\varepsilon_k\|. \quad (5.10)$$

Из (5.7) и (5.10) вытекает оценка второго слагаемого в правой части (5.4)

$$\begin{aligned} (H(\tilde{f}_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}), \varepsilon_{j+\frac{1}{2}}) &\leq 9L_f 2^{K-1} K (\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H) \max_{j-M_0 \leq k \leq j} \|\varepsilon_k\| \\ &\leq 27L_f 2^{K-2} K (\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H) \max_{j-M_0 \leq k \leq j} \|\varepsilon_k\|_H. \end{aligned} \quad (5.11)$$

Оценим третье слагаемое в правой части (5.4). Имеем

$$(\hat{\psi}_j, \varepsilon_{j+\frac{1}{2}}) \leq \|\hat{\psi}_j\| \|\varepsilon_{j+\frac{1}{2}}\|. \quad (5.12)$$

Согласно лемме 4.2 выполняется

$$\|\hat{\psi}_j\|^2 = \sum_{i=1}^{N-1} h |\hat{\psi}_j|^2 \leq C_5^2 (\Delta^2 + h^4)^2 X,$$

то есть

$$\|\hat{\psi}_j\| \leq C_5 \sqrt{X} (\Delta^2 + h^4). \quad (5.13)$$

Учитывая оценку

$$\|\varepsilon_{j+\frac{1}{2}}\| \leq \frac{3}{2} \|\varepsilon_{j+\frac{1}{2}}\|_H \leq \frac{3}{4} (\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H),$$

из (5.12) и (5.13) получаем

$$(\hat{\psi}_j, \varepsilon_{j+\frac{1}{2}}) \leq \frac{3}{4} C_5 \sqrt{X} (\Delta^2 + h^4) (\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H). \quad (5.14)$$

Подставляя (5.5), (5.6), (5.11) и (5.14) в (5.4) и, сокращая на $(\|\varepsilon_{j+1}\|_H + \|\varepsilon_j\|_H)$, получаем

$$\frac{1}{2\Delta} (\|\varepsilon_{j+1}\|_H - \|\varepsilon_j\|_H) \leq 27L_f 2^{K-2} K \max_{j-M_0 \leq k \leq j} \|\varepsilon_k\|_H + \frac{3}{4} C_5 \sqrt{X} (\Delta^2 + h^4)$$

или

$$\|\varepsilon_{j+1}\|_H \leq \|\varepsilon_j\|_H + \Delta 27L_f 2^{K-1} K \max_{j-M_0 \leq k \leq j} \|\varepsilon_k\|_H + \Delta \frac{3}{2} C_5 \sqrt{X} (\Delta^2 + h^4). \quad (5.15)$$

Обозначим для краткости константы

$$A = 27L_f 2^{K-1} K, \quad B = \frac{3}{2} C_5 \sqrt{X}$$

и запишем следующую из (5.15) оценку для предыстории погрешности

$$\|\{\varepsilon_k\}_{j+1}\|_\tau \leq (1 + A\Delta) \|\{\varepsilon_k\}_j\|_\tau + B\Delta(\Delta^2 + h^4). \quad (5.16)$$

Из (5.16) выводится оценка

$$\|\{\varepsilon_k\}_j\|_\tau \leq \frac{B}{A} e^{AT} (\Delta^2 + h^4)$$

для всех $j = 1, \dots, M$. Откуда следует заключение теоремы. $\square$

6. Численные эксперименты

Пример 6.1. Рассмотрим тестовое уравнение с постоянным сосредоточенным запаздыванием:

$$\begin{aligned} \frac{\partial u(x, t)}{\partial t} = & d \frac{\partial^\alpha u}{\partial |x|^\alpha} + 3t^2 x^4 (1-x)^4 + u(x, t - \tau(t)) - (t - 0.5)^3 x^4 (1-x)^4 \\ & + \frac{t^3 d}{2 \cos \frac{\pi\alpha}{2}} \left[\frac{\Gamma(9)}{\Gamma(9-\alpha)} (x^{8-\alpha} + (1-x)^{8-\alpha}) - \frac{4\Gamma(8)}{\Gamma(8-\alpha)} (x^{7-\alpha} + (1-x)^{7-\alpha}) \right. \\ & + \frac{6\Gamma(7)}{\Gamma(7-\alpha)} (x^{6-\alpha} + (1-x)^{6-\alpha}) - \frac{4\Gamma(6)}{\Gamma(6-\alpha)} (x^{5-\alpha} + (1-x)^{5-\alpha}) \\ & \left. + \frac{\Gamma(5)}{\Gamma(5-\alpha)} (x^{4-\alpha} + (1-x)^{4-\alpha}) \right], \end{aligned}$$

где $x \in [0, 1]$, $t \in [0, 1]$, $\tau(t) = 0.5$, $d = 100$.

Заданы начальные

$$u(x, t) = t^3 x^4 (1-x)^4, \quad x \in [0, 1], \quad t \in [-0.5, 0],$$

и граничные условия

$$u(0, t) = u(1, t) = 0, \quad t \in [0, 1].$$

Точным решением является функция

$$u(x, t) = t^3 x^4 (1-x)^4.$$

Введем следующие обозначения:

$$E_H(\Delta, h) = \max_{0 \leq k \leq M} \|u_k - U_k\|_H,$$

$$u_k = (u(x_1, t_k), u(x_2, t_k), \dots, u(x_{N-1}, t_k)), \quad U_k = (U_k^1, U_k^2, \dots, U_k^{N-1}).$$

Метод (3.1) тестировался для различных шагов Δ и h . Были получены вычислительные порядки сходимости $order_\Delta = \log_2(\frac{E_H(2\Delta, h)}{E_H(\Delta, h)})$, $order_h = \log_2(\frac{E_H(\Delta, 2h)}{E_H(\Delta, h)})$ по Δ и h соответственно.

В таблице 1 показан вычислительный порядок сходимости по h .

Таблица 1

Зависимость погрешностей и вычислительных порядков сходимости от шагов h

α	Δ	h	$E_H(\Delta, h)$	$order_h$
1,1	1/4000	1/10	0,0000031166	
1,1	1/4000	1/20	0,0000001791	4,1211
1,1	1/4000	1/40	0,0000000103	4,1201
1,5	1/4000	1/10	0,0000051913	
1,5	1/4000	1/20	0,0000003065	4,0821
1,5	1/4000	1/40	0,0000000176	4,1222
1,9	1/4000	1/10	0,0000086378	
1,9	1/4000	1/20	0,0000005535	3,9640
1,9	1/4000	1/40	0,0000000341	4,0207

Таблица 2 показывает вычислительный порядок сходимости по Δ .

Таблица 2

Зависимость погрешностей и вычислительных порядков сходимости от шагов Δ

α	h	Δ	$E_H(\Delta, h)$	$order_\Delta$
1,1	1/1000	1/10	0,0000160075	
1,1	1/1000	1/20	0,0000039999	2,0007
1,1	1/1000	1/40	0,0000009999	2,0001
1,5	1/1000	1/10	0,0000160223	
1,5	1/1000	1/20	0,0000040036	2,0007
1,5	1/1000	1/40	0,0000010009	2,0000
1,9	1/1000	1/10	0,0000160299	
1,9	1/1000	1/20	0,0000040062	2,0005
1,9	1/1000	1/40	0,0000010014	2,0002

Пример 6.2. Рассмотрим тестовое уравнение с переменным запаздыванием:

$$\begin{aligned} \frac{\partial u(x, t)}{\partial t} = & d \frac{\partial^\alpha u}{\partial |x|^\alpha} + \alpha(t+1)^{\alpha-1} x^4 (1-x)^4 + u(x, t - \tau(t)) - \left(\frac{2(t+1)}{3}\right)^\alpha x^4 (1-x)^4 \\ & + \frac{(t+1)^\alpha d}{2 \cos \frac{\pi\alpha}{2}} \left[\frac{\Gamma(9)}{\Gamma(9-\alpha)} (x^{8-\alpha} + (1-x)^{8-\alpha}) - \frac{4\Gamma(8)}{\Gamma(8-\alpha)} (x^{7-\alpha} + (1-x)^{7-\alpha}) \right. \\ & + \frac{6\Gamma(7)}{\Gamma(7-\alpha)} (x^{6-\alpha} + (1-x)^{6-\alpha}) - \frac{4\Gamma(6)}{\Gamma(6-\alpha)} (x^{5-\alpha} + (1-x)^{5-\alpha}) \\ & \left. + \frac{\Gamma(5)}{\Gamma(5-\alpha)} (x^{4-\alpha} + (1-x)^{4-\alpha}) \right], \end{aligned}$$

где $x \in [0, 1]$, $t \in [0, 1]$, $\tau(t) = \frac{t+1}{3}$, $d = 1$.

Заданы начальные и граничные условия

$$u(x, t) = (t+1)^\alpha x^4 (1-x)^4, \quad x \in [0, 1], \quad t \in [-1/3, 0].$$

$$u(0, t) = u(1, t) = 0, \quad t \in [0, 1].$$

$u(x, t) = (t+1)^\alpha x^4 (1-x)^4$ — точное решение.

Метод (3.1) тестировался для различных шагов Δ и h . В таблице 3 показан вычислительный порядок сходимости по h .

Таблица 3

Зависимость погрешностей и вычислительных порядков сходимости от шагов h

α	Δ	h	$E_H(\Delta, h)$	$order_h$
1,1	1/4000	1/10	0,0000064160	
1,1	1/4000	1/20	0,0000003705	4,1141
1,1	1/4000	1/40	0,0000000216	4,1004
1,5	1/4000	1/10	0,0000139590	
1,5	1/4000	1/20	0,0000008242	4,0821
1,5	1/4000	1/40	0,0000000473	4,1231
1,9	1/4000	1/10	0,0000307599	
1,9	1/4000	1/20	0,0000019696	3,9651
1,9	1/4000	1/40	0,0000001213	4,0213

Таблица 4 показывает вычислительный порядок сходимости по Δ .

Таблица 4

Зависимость погрешностей и вычислительных порядков сходимости от шагов Δ

α	h	Δ	$E_H(\Delta, h)$	$order_\Delta$
1,1	1/1000	1/10	0,0000001679	
1,1	1/1000	1/20	0,0000000419	2,0026
1,1	1/1000	1/40	0,0000000105	1,9966
1,5	1/1000	1/10	0,0000015431	
1,5	1/1000	1/20	0,0000003845	2,0048
1,5	1/1000	1/40	0,0000000960	2,0019
1,9	1/1000	1/10	0,0000045433	
1,9	1/1000	1/20	0,0000011142	2,0277
1,9	1/1000	1/40	0,0000002775	2,0055

References

- [1] J. Wu, *Theory and Application of Partial Functional Differential Equations*, Springer-Verlag, New York, 1996.
- [2] A. D. Polyanin, V. G. Sorokin, A. I. Zhurov, *Delay Ordinary and Partial Differential Equations*, CRC Press, Boca Raton, 2023.
- [3] Z. Kamont, K. Kropielnicka, “Implicit difference methods for evolution functional differential equations”, *Numerical Analysis and Applications*, **4:4** (2011), 294–308.
- [4] В. Г. Пименов, *Разностные методы решения уравнений в частных производных с наследственностью*, Изд-во Уральского университета, Екатеринбург, 2014. [V. G. Pimenov, *Difference Methods for Solving Partial Differential Equations with Heredity*, Publishing House of the Ural University, Yekaterinburg, 2014 (In Russian)].
- [5] M. M. Meerschaert, C. Tadjeran, “Finite difference approximations for two-sided space-fractional partial differential equations”, *Applied Numerical Mathematics*, **56:1** (2006), 80–90.
- [6] C. Tadjeran, M. M. Meerschaert, H. P. Scheffler, “A second-order accurate numerical approximation for the fractional diffusion equation”, *Journal of Computational Physics*, **213:1** (2006), 205–213.
- [7] C. Celik, M. Duman, “Crank–Nicolson method for the fractional diffusion equation with the Riesz fractional derivative”, *Journal of Computational Physics*, **231:4** (2012), 1743–1750.
- [8] W. Tian, H. Zhou, W. Deng, “A class of second order difference approximations for solving space fractional diffusion equations”, *Mathematics of Computation*, **84** (2015), 1703–1727.
- [9] V. G. Pimenov, A. S. Hendy, “A fractional analog of Crank–Nicolson method for the two sided space fractional partial equation with functional delay”, *Ural Mathematical Journal*, **2:1** (2016), 48–57.
- [10] M. Ibrahim, V. G. Pimenov, “Crank–Nicolson scheme for two-dimensional in space fractional diffusion equations with functional delay”, *Izv. IMI UdGU*, **57** (2021), 128–141.
- [11] В. Г. Пименов, А. Б. Ложников, “Асимптотическое разложение погрешности численного метода для решения супердиффузионного уравнения с функциональным запаздыванием”, Тр. ИММ УрО РАН, **30**, №2, 2024, 138–151. [V. G. Pimenov, A. B. Lozhnikov, “Asymptotic expansion of the error of a numerical method for solving a superdiffusion equation with functional delay”, *Trudy Inst. Mat. i Mekh. UrO RAN*, **30**, no. 2, 2024, 138–151 (In Russian)].
- [12] G. H. Gao, Z. Z. Sun, “A compact finite difference scheme for the fractional sub-diffusion equations”, *Journal of Computational Physics*, **230:3** (2011), 586–595.
- [13] A. S. Hendy, J. E. Macias–Diaz, “A discrete Gronwall inequality and energy estimates in the analysis of a discrete model for a nonlinear time-fractional heat equation”, *Mathematics*, **8:9** (2020), 1–15.

- [14] H. Sun, Z. Z. Sun, G. H. Gao, “Some high order difference schemes for the space and time fractional Bloch–Torrey equations”, *Applied Mathematics and Computation*, **281** (2016), 356–380.
- [15] L. Li, B. Zhou, X. Chen, Z. Wang, “Convergence and stability of compact finite difference method for nonlinear time fractional reaction–diffusion equations with delay”, *Applied Mathematics and Computation*, **337** (2018), 144–152.
- [16] А. В. Ким, В. Г. Пименов, *i-Гладкий анализ и численные методы решения функционально-дифференциальных уравнений*, Регулярная и хаотическая динамика, Москва–Ижевск, 2004. [A. V. Kim, V. G. Pimenov, *i-Smooth Analysis and Numerical Methods for Solving Functional Differential Equations*, Regular and Chaotic Dynamics, Moscow–Izhevsk, 2004 (In Russian)].

Информация об авторах

Пименов Владимир Германович, доктор физико-математических наук, заведующий кафедрой вычислительной математики и компьютерных наук, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина; ведущий научный сотрудник, Институт математики и механики им. Н. Н. Красовского, г. Екатеринбург, Российская Федерация. E-mail: v.g.pimenov@urfu.ru

ORCID: <https://orcid.org/0000-0002-4042-6079>

Лекомцев Андрей Валентинович, кандидат физико-математических наук, доцент кафедры вычислительной математики и компьютерных наук, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Российская Федерация.

E-mail: avlekomtsev@urfu.ru

ORCID: <https://orcid.org/0000-0003-0176-1306>

Конфликт интересов отсутствует.

Для контактов:

Лекомцев Андрей Валентинович

E-mail: avlekomtsev@urfu.ru

Поступила в редакцию 07.10.2024 г.

Поступила после рецензирования 18.11.2024 г.

Принята к публикации 22.11.2024 г.

Information about the authors

Vladimir G. Pimenov, Doctor of Physics and Mathematics, Head of the Computational Mathematics and Computer Science Department, Ural Federal University named after the first President of Russia B. N. Yeltsin; Leading Researcher, N. N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russian Federation. E-mail: v.g.pimenov@urfu.ru

ORCID: <https://orcid.org/0000-0002-4042-6079>

Andrei V. Lekomtsev, Candidate of Physics and Mathematics, Associate Professor of the Computational Mathematics and Computer Science Department, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russian Federation.

E-mail: avlekomtsev@urfu.ru

ORCID: <https://orcid.org/0000-0003-0176-1306>

There is no conflict of interests.

Corresponding author:

Andrei V. Lekomtsev

E-mail: avlekomtsev@urfu.ru

Received 07.10.2024

Reviewed 18.11.2024

Accepted for press 22.11.2024

НАУЧНАЯ СТАТЬЯ

© Сумин В.И., Сумин М.И., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-455-484>

УДК 517.9



Регуляризация классических условий оптимальности в задачах оптимизации линейных распределенных систем вольтеррова типа с поточечными фазовыми ограничениями

Владимир Иосифович СУМИН^{1,2}, Михаил Иосифович СУМИН^{1,2}

¹ ФГБОУ ВО «Тамбовский государственный университет им. Г. Р. Державина»

392000, Российская Федерация, г. Тамбов, ул. Интернациональная, 33

² ФГАОУ ВО «Нижегородский государственный университет им. Н. И. Лобачевского»

603950, Российская Федерация, г. Нижний Новгород, пр-т Гагарина, 23

Аннотация. Рассматривается регуляризация классических условий оптимальности (КУО) — принципа Лагранжа (ПЛ) и принципа максимума Понтрягина (ПМП) — в выпуклой задаче оптимального управления с сильно выпуклым целевым функционалом и с поточечными фазовыми ограничениями типа равенства и неравенства. Управляемая система задается линейным функционально-операторным уравнением II рода общего вида в пространстве L_2^s , основной оператор правой части уравнения предполагается квазинильпотентным. Получение регуляризованных КУО основано на методе двойственной регуляризации. Основное предназначение регуляризованных ПЛ и ПМП — устойчивое генерирование минимизирующих приближенных решений (МПР) в смысле Дж. Варги. Регуляризованные КУО: 1) формулируются как теоремы существования в исходной задаче МПР с одновременным конструктивным представлением этих решений; 2) выражаются в терминах регулярных классических функций Лагранжа и Гамильтона–Понтрягина; 3) «преодолевают» свойства некорректности КУО и дают регуляризирующие алгоритмы для решения оптимизационных задач. Статья продолжает серию работ авторов по регуляризации КУО для ряда канонических задач оптимального управления линейными распределенными системами вольтеррова типа. В качестве приложения полученных в работе «абстрактных результатов» в ее заключительной части рассматривается регуляризация КУО в конкретной оптимизационной задаче с поточечными фазовыми ограничениями типа равенства и неравенства для управляемой системы с запаздыванием.

Ключевые слова: выпуклое оптимальное управление, поточечные фазовые ограничения, распределенная система, функционально-операторное уравнение вольтеррова типа, условия оптимальности, регуляризация, двойственность

Благодарности: Результаты, представленные в разделах 1, 2, получены за счет гранта Российского научного фонда (проект № 23-11-20020, <https://rscf.ru/project/23-11-20020/>); результаты, представленные в разделе 3, получены за счет гранта Министерства образования и науки Тамбовской области (проект № 2-ФП-2023).

Для цитирования: Сумин В.И., Сумин М.И. Регуляризация классических условий оптимальности в задачах оптимизации линейных распределенных систем вольтеррова типа с поточечными фазовыми ограничениями // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 455–484. <https://doi.org/10.20310/2686-9667-2024-29-148-455-484>

SCIENTIFIC ARTICLE

© V. I. Sumin, M. I. Sumin, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-455-484>

Regularization of classical optimality conditions in optimization problems of linear distributed Volterra-type systems with pointwise state constraints

Vladimir I. SUMIN^{1,2}, Mikhail I. SUMIN^{1,2}¹ Derzhavin Tambov State University

33 Internatsionalnaya St., Tambov 392000, Russian Federation

² Lobachevskii Nizhnii Novgorod State University

23 Gagarin Ave., Nizhnii Novgorod 603950, Russian Federation

Abstract. The regularization of classical optimality conditions (COCs) — the Lagrange principle (LP) and the Pontryagin maximum principle (PMP) — in a convex optimal control problem with a strongly convex objective functional and with pointwise state constraints of the equality and inequality type is considered. The control system is defined by a linear functional-operator equation of the second kind of general form in the space L_2^s , the main operator of the right-hand side of the equation is assumed to be quasi-nilpotent. Obtaining regularized COCs is based on the dual regularization method. The main purpose of regularized LP and PMP is stable generation of minimizing approximate solutions (MASs) in the sense of J. Warga. Regularized COCs: 1) are formulated as existence theorems of MASs in the original problem with simultaneous constructive representation of these solutions; 2) are expressed in terms of regular classical Lagrange and Hamilton–Pontryagin functions; 3) “overcome” the ill-posedness properties of the COCs and provide regularizing algorithms for solving optimization problems. The article continues a series of works by the authors on the regularization of the COCs for a number of canonical problems of optimal control of linear distributed systems of the Volterra type. As an application of the “abstract results” obtained in the work, the final part considers the regularization of the COCs in a specific optimization problem with pointwise state constraints of the equality and inequality type for a control system with delay.

Keywords: convex optimal control, pointwise state constraints, distributed system, functional-operator equation of the Volterra type, optimality conditions, regularization, duality

Acknowledgements: The results of Sections 1, 2 were obtained within the Russian Science Foundation (project no. 23-11-20020, <https://rscf.ru/project/23-11-20020/>). The results of Section 3 were obtained within the grant of the Ministry of Education and Science of the Tambov region (project no. 2-ФП-2023).

Mathematics Subject Classification: 49K20 47A52 35R25 49N15 90C46.

For citation: Sumin V.I., Sumin M.I. Regularization of classical optimality conditions in optimization problems of linear distributed Volterra-type systems with pointwise state constraints. *Vestnik Rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, 29:148 (2024), 455–484. <https://doi.org/10.20310/2686-9667-2024-29-148-455-484> (In Russian, Abstr. in Engl.)

Введение

Данная статья продолжает серию работ авторов [1–6] по регуляризации классических условий оптимальности (КУО) — принципа Лагранжа (ПЛ) и принципа максимума Понтрягина (ПМП) — для ряда канонических задач оптимального управления линейными распределенными системами вольтеррова типа. Системами вольтеррова типа мы называем управляемые системы, которые могут быть описаны линейными функциональными (иначе, функционально-операторными) уравнениями II рода общего вида с квазинильпотентным основным линейным оператором правой части. Подобным свойством обладают, прежде всего, различного рода вольтерровы операторы. Заметим, что название «вольтерровы операторы» (операторы типа Вольтерра) присваивалось разными авторами различным классам операторов со сходными свойствами (используются также названия: причинные операторы, наследственные операторы и др.), начиная с известных работ L. Tonelli (1929) и А. Н. Тихонова (1938); см., например, краткий обзор определений вольтерровых операторов [7, Дополнение], а также [6, 8]. В случае линейных операторов эти определения так или иначе связаны со свойством квазинильпотентности: либо это свойство включено в само определение вольтеррова оператора (см., например, [9, с. 10]), либо при естественных условиях следует из этого определения (см., например, цепочечный признак квазинильпотентности [8, теорема 2], опирающийся на определение [10] функционального оператора, «вольтеррова на системе множеств», являющееся многомерным обобщением определения А. Н. Тихонова). Поэтому такие уравнения можно назвать функциональными уравнениями вольтеррова типа.

К уравнениям вольтеррова типа естественным образом (обращением главной части) сводятся самые разнообразные начально-краевые задачи для различных уравнений с частными производными (гиперболических, параболических, интегро-дифференциальных, систем таких уравнений, уравнений с запаздываниями разного рода и др., см. разнообразные конкретные примеры в [7, глава 2], обзоры в [7, 11]). Это позволило в настоящей статье получить регуляризованные ПЛ и ПМП единообразно для широкого класса распределенных оптимизационных задач. При этом существенным образом используется предложенное нами ранее понятие равностепенной квазинильпотентности семейства операторов [12] (историю вопроса см. в [11]).

В статьях [1, 2] рассматривались различные варианты регуляризации КУО для задач с сильно выпуклыми целевыми функционалами и с функциональными ограничениями типа равенства и неравенства. В свою очередь, в [3, 4] аналогичные вопросы изучались применительно к задачам также с сильно выпуклыми целевыми функционалами, но с операторным (т. е. задаваемым оператором с бесконечномерным образом) ограничением равенством и функциональными ограничениями-неравенствами. Наконец, в [5, 6] регуляризация КУО проводилась для задач такого же вида, как и в [1–4], но в ситуации, когда целевой функционал выпуклой задачи может быть и не сильно выпуклым. «Абстрактные результаты» по регуляризации КУО в [1–6] подробно «расшифровывались» для ряда конкретных задач оптимального управления распределенными системами вольтеррова типа (рассматривались, в частности, оптимизационные задачи, связанные с системой уравнений с запаздыванием, с интегродифференциальным уравнением типа уравнения переноса и с гиперболической системой дифференциальных уравнений первого порядка).

В настоящей работе рассматривается регуляризация КУО в выпуклых задачах оптимизации линейных распределенных систем вольтеррова типа, но, в отличие от [1–6],

в качестве операторных ограничений выступают поточечные фазовые ограничения типа равенства и неравенства. В качестве приложения полученных в работе «абстрактных результатов» в ней рассматривается регуляризация КУО в конкретной оптимизационной задаче с поточечными фазовыми ограничениями типа равенства и неравенства для управляемой системы с запаздыванием. Как известно [13–16], задачи с поточечными фазовыми ограничениями вот уже на протяжении более шести десятилетий привлекают постоянное внимание специалистов в области оптимизации сосредоточенных и распределенных систем. Сами же поточечные фазовые ограничения, с точки зрения оптимизационной теории, традиционно относятся к числу наиболее сложных операторных ограничений.

Регуляризация КУО, т. е. регуляризация ПМП и ПЛ, в выпуклых задачах оптимального управления и математического программирования на базе основанного на двойственности подхода к регуляризации [17] была впервые предложена в работах [18, 19]. Указанная регуляризация в теории условной оптимизации, в том числе и в задачах оптимального управления, является «естественной процедурой» ввиду свойств некорректности КУО [19–22]. Говоря о некорректности КУО, мы, прежде всего, имеем в виду такие их свойства, как невыполнимость и неустойчивость [19–22]. Напомним, что о невыполнимости КУО мы говорим тогда, когда известно, что в той или иной задаче на условный экстремум они не могут быть записаны. В свою очередь, неустойчивость КУО мы понимаем в том смысле, что неустойчиво по отношению к возмущению исходных данных экстремальной задачи ведут себя допустимые элементы, удовлетворяющие всем составляющим КУО соотношениям. Анализ различных примеров некорректности [19–22] дают основание рассматривать КУО как математические объекты с присущими им от природы свойствами некорректности. В этом случае центральная роль, как показано в [19, 20, 22], в задачах условной оптимизации естественным образом неизбежно переходит от классического понятия оптимального элемента к понятию минимизирующего приближенного решения (МПР) в смысле Дж. Варги [23], см. ниже определение этого понятия в разделе 2.3 (в математическом программировании такие последовательности часто называют оптимальными обобщенными планами [24]). Такой подход был реализован в [19, 21, 22] (в [22] для обозначения МПР использовалось словосочетание «обобщенная минимизирующая последовательность») применительно к выпуклым задачам на условный экстремум с операторным ограничением-равенством и конечным числом функциональных ограничений-неравенств. Тот же подход был реализован и в наших уже цитированных выше работах [1–6] применительно к регуляризации КУО для канонических задач оптимизации линейных распределенных систем вольтеррова типа. Полученные в указанных работах результаты позволяют выделить следующие важные, на наш взгляд, особенности получаемых регуляризованных КУО в выпуклых задачах условной оптимизации. Можно утверждать, что регуляризованные КУО: 1) при условии конечности значения исходной задачи формулируются как теоремы существования в ней МПР, состоящих из минималей регулярных функционалов Лагранжа, двойственные переменные для которых генерируются в соответствии с выбранной процедурой регуляризации двойственной задачи; 2) обобщают КУО, приводят к ним «в пределе» и сохраняют общую структуру классических аналогов; 3) являются условиями оптимальности, выраженными в секвенциальной форме в терминах регулярных функционалов Лагранжа; 4) «преодолевают» свойства некорректности классических аналогов и представляют собою универсальные МПР-образующие (регуляризирующие) алгоритмы в смысле [25] для решения задач на условный экстремум.

Ранее регуляризация КУО для выпуклых задач оптимального управления с поточечными фазовыми ограничениями типа равенства и неравенства рассматривалась в наших работах в случае управляемых систем, связанных с обыкновенными [18] и параболическими дифференциальными уравнениями [26]. Рассматриваемая ниже управляемая система вольтеррова типа своей общностью существенно отличается от рассмотренных ранее в наших работах [18, 26] управляемых систем.

Статья состоит из введения и трех основных разделов, первый из которых посвящен постановке выпуклой задачи оптимального управления для линейной распределенной системы вольтеррова типа с поточечными фазовыми ограничениями типа равенства и неравенства, формулировке и доказательству важных вспомогательных результатов, определению МПР-образующего алгоритма. Второй раздел посвящен доказательству сходимости алгоритма двойственной регуляризации для рассматриваемой задачи, а также доказательству на этой основе регуляризованных КУО для нее. Наконец, в заключительном третьем разделе подробно обсуждается, как «абстрактные» результаты первых двух разделов могут быть «расшифрованы» применительно к конкретной задаче оптимального управления с поточечными фазовыми ограничениями типа равенства и неравенства для управляемой системы с запаздыванием.

Примем следующие обозначения и соглашения: $\mathbf{R}^n$ — пространство n -векторов-столбцов; $\langle \cdot, \cdot \rangle_n$ и $\|\cdot\|_n$ — евклидовы скалярное произведение и норма в $\mathbf{R}^n$; векторы, если не оговорено противное, считаются столбцами; 0_n — ноль в $\mathbf{R}^n$; $\mathbf{R}_+$ — множество всех положительных чисел; $\mathbf{R}_-$ — множество всех неположительных чисел; $*$ — знак сопряжения и транспонирования; $\Pi \subset \mathbf{R}^n$ — выпуклый компакт с непустой внутренностью, играющий роль основного множества изменения независимых переменных, элементы которого обозначаем через $t \equiv \{t^1, \dots, t^n\}$; $L_p \equiv L_p(\Pi)$ — лебегово пространство со стандартной нормой ($1 \leq p \leq \infty$); $L_p^m \equiv L_p^m(\Pi) \equiv (L_p(\Pi))^m$ ($1 \leq p \leq \infty$); $\|\cdot\|_{p,m}$ — стандартная норма прямого произведения в L_p^m ; $\langle \cdot, \cdot \rangle_{2,m}$ — стандартное скалярное произведение в L_2^m ; $L_{2,-} \equiv \{y \in L_2 : y(t) \leq 0, t \in \Pi\}$; $L_{2,+} \equiv \{y \in L_2 : y(t) \geq 0, t \in \Pi\}$; $C \equiv C(\Pi)$ — пространство непрерывных на Π функций со стандартной нормой $\|\cdot\|_C$; $C^r \equiv (C)^r$; $\|\cdot\|_{C,r}$ — стандартная норма прямого произведения в C^r ; если H гильбертово пространство, то скалярное произведение в нем обозначаем через $\langle \cdot, \cdot \rangle_H$, а соответствующую норму как $\|\cdot\|_H$; субдифференцируемость понимается в смысле выпуклого анализа.

1. Постановка задачи оптимального управления

1.1. Базовая оптимизационная задача

Пусть заданы: натуральные числа m, n, q, r, s ; функция $c(t)$, $t \in \Pi$, класса $L_2^m \equiv L_2^m(\Pi)$; линейный ограниченный оператор (ЛОО) $A : L_2^m \rightarrow L_2^m$ с нулевым спектральным радиусом; ЛОО $B : L_2^s \rightarrow L_2^m$. Рассмотрим линейное функциональное уравнение

$$z(t) = A[z](t) + B[u](t) + c(t), \quad t \in \Pi, \quad z \in L_2^m, \quad u \in L_2^s, \quad (1.1)$$

считая $u(\cdot)$ управляющей функцией (управлением). Ввиду квазинильпотентности A уравнение (1.1) имеет для каждого $u(\cdot) \in L_2^s$ единственное в L_2^m решение $z(t)$, $t \in \Pi$, причем

$$z(t) = S[B[u] + c](t), \quad t \in \Pi, \quad u \in L_2^s, \quad (1.2)$$

где $S : L_2^m \rightarrow L_2^m$ — ЛОО — сумма ряда Неймана: $S[y] \equiv \sum_{i=0}^{\infty} A^i[y]$, $y \in L_2^m$. Отвечающее управлению $u(\cdot)$ решение $z(\cdot)$ уравнения (1.1), задаваемое формулой (1.2), обозначим $z_u(\cdot)$.

Чтобы поставить для управляемой системы (1.1) задачу оптимального управления, будем считать, что заданы: функционал $\mathcal{J}[z, u] \equiv K[z] + M[u]$, $z \in L_2^m$, $u \in L_2^s$, где $K : L_2^m \rightarrow \mathbf{R}$ — непрерывный выпуклый функционал, а $M : L_2^s \rightarrow \mathbf{R}$ — непрерывный сильно выпуклый функционал с постоянной сильной выпуклости κ ; ЛОО $\Sigma : L_2^m \rightarrow C^r$ и $\Upsilon : L_2^m \rightarrow C^q$; функции $\omega(\cdot) \in C$ и $\Xi(\cdot) \in C^q$; функция $F(t, y) : \Pi \times \mathbf{R}^r \rightarrow \mathbf{R}$, непрерывная по совокупности переменных, выпуклая по y на $\mathbf{R}^r$ при каждом $t \in \Pi$.

Используя равенство (1.2) как формулу подстановки, зададим на L_2^s функционал $J[u] \equiv \mathcal{J}[z_u, u] \equiv K[z_u] + M[u]$, $u \in L_2^s$, операторы $\mathcal{G}[u](t) \equiv \langle \Xi(t), \Upsilon[z_u](t) \rangle_q$, $t \in \Pi$, $u \in L_2^s$, и $\mathcal{E}[u](t) \equiv F(t, \Sigma[z_u](t))$, $t \in \Pi$, $u \in L_2^s$. Очевидно, что функционал J непрерывен и сильно выпукл на L_2^s с постоянной сильной выпуклости κ , а операторы $\mathcal{G}$ и $\mathcal{E}$ действуют из L_2^s в C .

Пусть $\mathcal{D}$ — выпуклое ограниченное и замкнутое множество пространства L_2^s . Рассмотрим задачу оптимального управления системой (1.1) с минимизируемым целевым функционалом J при поточечных ограничениях типа равенства и неравенства

$$\mathcal{G}[u](t) = \omega(t), \quad t \in \Pi, \quad u \in L_2^s; \quad \mathcal{E}[u](t) \in \mathbf{R}_-, \quad t \in \Pi, \quad u \in L_2^s, \quad (1.3)$$

и множестве допустимых управлений $\mathcal{D}$. Эту задачу символически запишем в виде

$$J[u] \rightarrow \min, \quad (1.3), \quad u \in \mathcal{D}. \quad (1.4)$$

1.2. Точная и приближенные оптимизационные задачи

Задача (1.4) определяется набором исходных данных $\mathbf{f} \equiv \{A, B, c, K, M, \Xi, \Upsilon, \omega, F, \Sigma\}$. Предположим, что точные исходные данные $\mathbf{f}^0 \equiv \{A^0, B^0, c^0, K^0, M^0, \Xi^0, \Upsilon^0, \omega^0, F^0, \Sigma^0\}$ нам не известны, но мы можем оперировать с приближенными исходными данными

$$\mathbf{f}^\delta \equiv \{A^\delta, B^\delta, c^\delta, K^\delta, M^\delta, \Xi^\delta, \Upsilon^\delta, \omega^\delta, F^\delta, \Sigma^\delta\},$$

где δ — меняющийся в некотором фиксированном полуинтервале $(0, \delta_0]$ числовой параметр, характеризующий близость приближенных данных $\mathbf{f}^\delta$ к точным данным $\mathbf{f}^0$ в указанном ниже условиями А, Б и В смысле (положительным значениям параметра δ соответствует приближенная оптимизационная задача вида (1.4) с данными $\mathbf{f}^\delta$, а значению $\delta = 0$ — точная оптимизационная задача вида (1.4) с данными $\mathbf{f}^0$). Таким образом, мы считаем, что при каждом $\delta \in [0, \delta_0]$ существуют следующие объекты: квазинильпотентный ЛОО $A^\delta : L_2^m \rightarrow L_2^m$; ЛОО $B^\delta : L_2^s \rightarrow L_2^m$; функция $c^\delta(\cdot) \in L_2^m$; непрерывный выпуклый функционал $K^\delta : L_2^m \rightarrow \mathbf{R}$; непрерывный сильно выпуклый функционал $M^\delta : L_2^s \rightarrow \mathbf{R}$ с постоянной сильной выпуклости κ ; ЛОО $\Sigma^\delta : L_2^m \rightarrow C^r$ и $\Upsilon^\delta : L_2^m \rightarrow C^q$; функции $\omega^\delta(\cdot) \in C$ и $\Xi^\delta(\cdot) \in C^q$; непрерывная функция $F^\delta(t, y) : \Pi \times \mathbf{R}^r \rightarrow \mathbf{R}$, выпуклая по y на $\mathbf{R}^r$ при каждом $t \in \Pi$.

Считаем, что приближенные исходные данные $\mathbf{f}^\delta$, $\delta \in (0, \delta_0]$, связаны с точными исходными данными $\mathbf{f}^0$ приведенными ниже условиями А, Б, В, Г, Д.

Условие А. Функционалы $K^\delta : L_2^m \rightarrow \mathbf{R}$ и $M^\delta : L_2^s \rightarrow \mathbf{R}$ при любом $\delta \in [0, \delta_0]$ липшицевы на каждом ограниченном множестве пространств L_2^m и L_2^s соответственно, причем

липшицевость равномерна по параметру $\delta \in [0, \delta_0]$, то есть существуют неубывающие функции $\mathbf{N}_1(\cdot), \mathbf{N}_2(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ такие, что для каждого $l > 0$ и любого $\delta \in [0, \delta_0]$

$$|K^\delta[z_1] - K^\delta[z_2]| \leq \mathbf{N}_1(l) \|z_1 - z_2\|_{2,m} \quad \text{при} \quad \|z_1\|_{2,m}, \|z_2\|_{2,m} \leq l,$$

$$|M^\delta[u_1] - M^\delta[u_2]| \leq \mathbf{N}_2(l) \|u_1 - u_2\|_{2,s} \quad \text{при} \quad \|u_1\|_{2,s}, \|u_2\|_{2,s} \leq l \quad (u_1, u_2 \in \mathcal{D}).$$

Функция $F^\delta(t, y) : \Pi \times \mathbf{R}^r \rightarrow \mathbf{R}$ при любом $\delta \in [0, \delta_0]$ равномерно по $t \in \Pi$ на каждом ограниченном в $\mathbf{R}^r$ множестве липшицева по y . Причем липшицевость равномерна по параметру $\delta \in [0, \delta_0]$, то есть существует неубывающая функция $\mathbf{N}_3(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ такая, что для каждого $l > 0$ и любого $\delta \in [0, \delta_0]$

$$|F^\delta(t, y_1) - F^\delta(t, y_2)| \leq \mathbf{N}_3(l) \|y_1 - y_2\|_r \quad \text{при} \quad t \in \Pi, \|y_1\|_r, \|y_2\|_r \leq l. \quad (1.5)$$

Условие Б. Существует постоянная $\mathbf{C}_1 > 0$ такая, что при любом $\delta \in (0, \delta_0]$

$$\|A^\delta - A^0\| \leq \mathbf{C}_1\delta, \quad \|B^\delta - B^0\| \leq \mathbf{C}_1\delta, \quad \|c^\delta - c^0\|_{2,m} \leq \mathbf{C}_1\delta, \quad (1.6)$$

$$\|\Sigma^\delta - \Sigma^0\| \leq \mathbf{C}_1\delta, \quad \|\Upsilon^\delta - \Upsilon^0\| \leq \mathbf{C}_1\delta, \quad \|\omega^\delta - \omega^0\|_C \leq \mathbf{C}_1\delta, \quad \|\Xi^\delta - \Xi^0\|_{C,q} \leq \mathbf{C}_1\delta.$$

Условие В. Существуют неубывающие функции $\mathbf{N}_4(\cdot), \mathbf{N}_5(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ и постоянная $\mathbf{C}_2 > 0$ такие, что для каждого $l > 0$ и любого $\delta \in (0, \delta_0]$

$$|K^\delta[z] - K^0[z]| \leq \mathbf{N}_4(l)\delta, \quad |M^\delta[u] - M^0[u]| \leq \mathbf{C}_2\delta \quad \text{при} \quad \|z\|_{2,m} \leq l, \quad u \in \mathcal{D}, \quad (1.7)$$

$$|F^\delta(t, y) - F^0(t, y)| \leq \mathbf{N}_5(l)\delta \quad \text{при} \quad t \in \Pi, \|y\|_r \leq l. \quad (1.8)$$

Чтобы сформулировать условие Г, воспользуемся следующим введенным нами ранее (историю вопроса см., например, в [11]) понятием равностепенной квазинильпотентности. Пусть $\mathbf{B}$ — банахово пространство, X — некоторое множество, $\{G(\xi)[\cdot] : \mathbf{B} \rightarrow \mathbf{B}\}_{\xi \in X}$ — семейство зависящих от параметра $\xi \in X$ квазинильпотентных ЛОО (напомним: квазинильпотентность ЛОО $G(\xi)[\cdot] : \mathbf{B} \rightarrow \mathbf{B}$ означает, что $\sqrt[k]{\|G(\xi)\|^k} \rightarrow 0$ при $k \rightarrow \infty$). Назовем семейство операторов $\{G(\xi)\}_{\xi \in X}$ *равностепенно квазинильпотентным*, если $\sup_{\xi \in X} \sqrt[k]{\|G(\xi)\|^k} \rightarrow 0$ при $k \rightarrow \infty$.

Условие Г. Семейство $\{A^\delta : L_2^m \rightarrow L_2^m\}_{\delta \in [0, \delta_0]}$ равностепенно квазинильпотентно.

При любом $\delta \in [0, \delta_0]$ управляемое функциональное уравнение

$$z(t) = A^\delta[z](t) + B^\delta[u](t) + c^\delta(t), \quad t \in \Pi, \quad z \in L_2^m, \quad u \in L_2^s, \quad (1.9)$$

имеет для каждого $u \in L_2^s$ единственное в классе L_2^m решение $z(t)$, $t \in \Pi$, причем

$$z(t) = S^\delta[B^\delta[u] + c^\delta](t), \quad t \in \Pi, \quad u \in L_2^s, \quad (1.10)$$

где $S^\delta : L_2^m \rightarrow L_2^m$ — ЛОО — сумма ряда Неймана: $S^\delta[y] \equiv \sum_{i=0}^{\infty} (A^\delta)^i[y]$, $y \in L_2^m$. Отвечающее управлению $u \in L_2^s$ и задаваемое формулой (1.10) решение $z(\cdot)$ уравнения (1.9) будем обозначать $z_u^\delta(\cdot)$, $\delta \in [0, \delta_0]$. При любом $\delta \in [0, \delta_0]$ мы имеем набор ограничений

$$\mathcal{G}^\delta[u](t) = \omega^\delta(t), \quad t \in \Pi, \quad u \in L_2^s; \quad \mathcal{E}^\delta[u](t) \in \mathbf{R}_-^1, \quad t \in \Pi, \quad u \in L_2^s, \quad (1.11)$$

где $\mathcal{G}^\delta[u](t) \equiv \langle \Xi^\delta(t), \Upsilon^\delta[z_u^\delta](t) \rangle_q$, $t \in \Pi$, $u \in L_2^s$, а $\mathcal{E}^\delta[u](t) \equiv F^\delta(t, \Sigma^\delta[z_u^\delta](t))$, $t \in \Pi$, $u \in L_2^s$, и задачу оптимального управления

$$J^\delta[u] \rightarrow \min, \quad (1.11), \quad u \in \mathcal{D}, \quad (OC^\delta)$$

в которой $J^\delta[u] \equiv \mathcal{J}^\delta[z_u^\delta, u] \equiv K^\delta[z_u^\delta] + M^\delta[u]$, $u \in L_2^s$. Задачу (OC^0) назовем *точной задачей*, а задачи (OC^δ) , $\delta \in (0, \delta_0]$, — *приближенными задачами* оптимального управления.

Условия Б и Г дают такое свойство семейства операторов $\{A^\delta\}_{0 \leq \delta \leq \delta_0}$ [4, лемма 1].

Лемма 1.1. *Существует $\mathcal{K} > 0$ такое, что $\|S^\delta - S^0\| \leq \mathcal{K} \|A^\delta - A^0\|$, $\delta \in [0, \delta_0]$.*

Лемма 1.1 позволяет легко вывести из условий А, Б, В, Г и ограниченности $\mathcal{D}$ существование постоянных C_1 , C_2 и функции $N_1(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ таких, что для $\delta \in [0, \delta_0]$

$$\|z_u^\delta\|_{2,m} \leq C_1, \quad \|z_u^\delta - z_u^0\|_{2,m} \leq C_2\delta, \quad u \in \mathcal{D}, \quad (1.12)$$

$$|J^\delta[u] - J^0[u]|, \|\mathcal{G}^\delta[u] - \mathcal{G}^0[u]\|_C, \|\mathcal{E}^\delta[u] - \mathcal{E}^0[u]\|_C \leq C_2\delta, \quad u \in \mathcal{D}, \quad (1.13)$$

$$|J^\delta[u_1] - J^\delta[u_2]|, \|\mathcal{G}^\delta[u_1] - \mathcal{G}^\delta[u_2]\|_C, \|\mathcal{E}^\delta[u_1] - \mathcal{E}^\delta[u_2]\|_C \leq N_1(l) \|u_1 - u_2\|_{2,s}, \\ \|u_1\|_{2,s}, \|u_2\|_{2,s} \leq l \quad (u_1, u_2 \in \mathcal{D}).$$

Предполагаем, что выполняется следующее условие, дополнительно связывающее ограничения (1.11) с управляемой системой (1.9).

Условие Д. При любом $\delta \in [0, \delta_0]$ слабая сходимость в пространстве L_2^s последовательности $\{u^j\}_{j=1}^\infty \subset \mathcal{D}$ к некоторому элементу u_* влечет за собой поточечную на Π сходимость последовательностей $\{\Sigma^\delta[z_{u^j}^\delta](\cdot)\}_{j=1}^\infty$ и $\{\Upsilon^\delta[z_{u^j}^\delta](\cdot)\}_{j=1}^\infty$ соответственно к $\Sigma^\delta[z_{u_*}^\delta](\cdot)$ и $\Upsilon^\delta[z_{u_*}^\delta](\cdot)$.

Лемма 1.2. *При любом $\delta \in [0, \delta_0]$ операторы $\mathcal{G}^\delta$ и $\mathcal{E}^\delta$ слабо непрерывны как операторы из L_2^s в L_2 , то есть преобразуют слабо сходящуюся к некоторому элементу в L_2^s последовательность в последовательность, сходящуюся в норме L_2 к образу этого элемента.*

Доказательство. Из условия Д следует, что при любом $\delta \in [0, \delta_0]$ слабая в L_2^s сходимость последовательности $\{u^j\}_{j=1}^\infty \subset \mathcal{D}$ к некоторому элементу u_* влечет за собой поточечную на Π сходимость $\{\mathcal{G}^\delta[u^j](\cdot)\}_{j=1}^\infty$ и $\{\mathcal{E}^\delta[u^j](\cdot)\}_{j=1}^\infty$ к $\mathcal{G}^\delta[u_*](\cdot)$ и $\mathcal{E}^\delta[u_*](\cdot)$ соответственно. Так как из-за ограниченности $\mathcal{D}$ в L_2^s указанные поточечно сходящиеся последовательности равномерно ограничены, то они сходятся соответственно к $\mathcal{G}^\delta[u_*](\cdot)$ и $\mathcal{E}^\delta[u_*](\cdot)$ и в норме L_2 . $\square$

1.3. Обобщенная нижняя грань и минимизирующие приближенные решения

Так как $L_{2,-}$ выпукло и замкнуто в гильбертовом пространстве L_2 , то любая точка v пространства имеет единственную проекцию на $L_{2,-}$; расстояние в L_2 от точки v до ее проекции на $L_{2,-}$ обозначим через $\rho(v)$ ($\rho(v) = \min_{w \in L_{2,-}} \|v - w\|_{2,1}$). Проекция точки $v(\cdot) \in L_2$ на $L_{2,-}$ есть $\overline{v(\cdot)}$, где $\overline{v(\cdot)}$ — «срезка нулем сверху» функции $v(\cdot)$ ($\overline{v(\cdot)}$ совпадает с $v(\cdot)$ там, где $v(\cdot)$ отрицательна, и равна нулю там, где $v(\cdot)$ неотрицательна); соответственно, $\rho(v) = \|\underline{v}\|_{2,1}$, где $\underline{v(\cdot)}$ — «срезка нулем снизу» функции $v(\cdot)$.

Пусть

$$\mathcal{D}^{\delta,\epsilon} \equiv \left\{ u \in \mathcal{D} : \|\mathcal{G}^\delta[u] - \omega^\delta\|_{2,1} \leq \epsilon, \quad \rho(\mathcal{E}^\delta[u]) \leq \epsilon \right\}, \quad \delta \in [0, \delta_0], \quad \epsilon \geq 0,$$

и $\mathcal{D}^0 \equiv \mathcal{D}^{0,0}$. Элементарными выкладками проверяется, что при любых $\delta \in [0, \delta_0]$, $\epsilon \geq 0$ множество $\mathcal{D}^{\delta,\epsilon}$ выпукло и замкнуто в L_2^s ; из ограниченности $\mathcal{D}$ следует ограниченность $\mathcal{D}^{\delta,\epsilon}$; вообще говоря, $\mathcal{D}^{\delta,\epsilon}$ может быть и пустым. Положим

$$\beta_0 = \inf_{u \in \mathcal{D}^0} J^0[u], \quad \text{если } \mathcal{D}^0 \neq \emptyset; \quad \beta_0 = +\infty, \quad \text{если } \mathcal{D}^0 = \emptyset.$$

Величину β_0 можно назвать классической нижней гранью задачи (OC^0) . Положим

$$\beta_\epsilon \equiv \inf_{u \in \mathcal{D}^{0,\epsilon}} J^0[u], \quad \text{если } \mathcal{D}^{0,\epsilon} \neq \emptyset; \quad \beta_\epsilon \equiv +\infty, \quad \text{если } \mathcal{D}^{0,\epsilon} = \emptyset.$$

Обобщенной нижней гранью задачи (OC^0) назовем предельную величину

$$\beta \equiv \beta_{+0} \equiv \lim_{\epsilon \rightarrow +0} \beta_\epsilon.$$

Ввиду непрерывности и выпуклости функционала J^0 , ограниченности, замкнутости и выпуклости множества $\mathcal{D}^0$ задача (OC^0) заведомо разрешима, если $\mathcal{D}^0 \neq \emptyset$ [27, теорема 8.2.8]. А так как функционал J^0 сильно выпуклый, то задача (OC^0) может иметь не более одного решения [27, теорема 8.2.10]. Таким образом, если $\mathcal{D}^0 \neq \emptyset$, задача (OC^0) имеет ровно одно решение, будем обозначать его через u^0 (в этом случае $J^0(u^0) = \min_{u \in \mathcal{D}^0} J^0(u) = \beta_0$).

Непосредственно из определения обобщенной нижней грани β следует, что $\beta \leq \beta_0$. Покажем, что наша задача (OC^0) такова, что имеет место равенство $\beta = \beta_0$.

Теорема 1.1. *Величины β и β_0 совпадают. Если они конечны, то задача (OC^0) разрешима, и ее решение единственно.*

Доказательство. При $\beta = +\infty$ равенство $\beta = \beta_0$ следует из неравенства $\beta \leq \beta_0$. Рассмотрим случай $\beta < +\infty$. При $\epsilon > 0$ множество $\mathcal{D}^{0,\epsilon}$ непусто, так как $\beta_\epsilon \leq \beta < +\infty$. Зафиксируем некоторое $\epsilon^0 > 0$. Так как функционал J^0 непрерывен и сильно выпукл на L_2^s , а множество $\mathcal{D}^{0,\epsilon}$ при $\epsilon > 0$ выпукло и замкнуто в L_2^s , то при любом $\epsilon \in (0, \epsilon^0]$ существует единственный элемент $u^\epsilon \in \mathcal{D}^{0,\epsilon}$ такой, что $J^0[u^\epsilon] = \min_{u \in \mathcal{D}^{0,\epsilon}} J^0[u] = \beta_\epsilon$ [27, теорема 8.2.10].

Множество $\mathcal{D}$ выпукло, ограничено и замкнуто, а потому и слабо компактно в L_2^s [27, теорема 8.2.3]. То есть существуют $u_* \in \mathcal{D}$ и стремящаяся к нулю последовательность $\{\epsilon_i\}_{i=1}^\infty \subset (0, \epsilon^0)$ такие, что $\{u^{\epsilon_i}\}_{i=1}^\infty$ слабо в L_2^s сходится к u_* . Функционал J^0 непрерывен и выпукл, а потому и слабо полунепрерывен снизу на $\mathcal{D}$ [27, теорема 8.2.7]. Значит существует подпоследовательность $\{\epsilon_{i_j}\}_{j=1}^\infty$ (обозначим ее $\{\sigma_j\}_{j=1}^\infty$), для которой

$$J^0[u_*] \leq \liminf_{i \rightarrow \infty} J^0[u^{\epsilon_i}] = \lim_{j \rightarrow \infty} J^0[u^{\sigma_j}] = \beta. \quad (1.14)$$

В силу условия Д слабая в L_2^s сходимость $\{u^{\sigma_j}\}_{j=1}^\infty$ к u_* влечет за собой поточечную на Π сходимость последовательностей $\{\Sigma^0[z_{u^{\sigma_j}}](\cdot)\}_{j=1}^\infty$ и $\{\Upsilon^0[z_{u^{\sigma_j}}](\cdot)\}_{j=1}^\infty$ соответственно к $\Sigma^0[z_{u_*}](\cdot)$ и $\Upsilon^0[z_{u_*}](\cdot)$. Отсюда следует поточечная на Π сходимость $\{\mathcal{G}^0[u^{\sigma_j}](\cdot)\}_{j=1}^\infty$

и $\left\{ \underline{\mathcal{E}^0[u^{\sigma_j}]}(\cdot) \right\}_{j=1}^{\infty}$ к $\mathcal{G}^0[u_*](\cdot)$ и $\underline{\mathcal{E}^0[u_*]}(\cdot)$. Так как указанные поточечно сходящиеся последовательности равномерно ограничены, то можем записать предельные соотношения $\|\mathcal{G}^0[u^{\sigma_j}] - \omega^0\|_{2,1} \rightarrow \|\mathcal{G}^0[u_*] - \omega^0\|_{2,1}$, $\|\underline{\mathcal{E}^0[u^{\sigma_j}]}\|_{2,1} \rightarrow \|\underline{\mathcal{E}^0[u_*]}\|_{2,1}$ при $j \rightarrow \infty$. Ввиду $u^\epsilon \in \mathcal{D}^{0,\epsilon}$ получаем

$$\|\mathcal{G}^0[u_*] - \omega^0\|_{2,1} = 0, \quad \|\underline{\mathcal{E}^0[u_*]}\|_{2,1} = 0.$$

То есть $u_* \in \mathcal{D}^0$. Это означает, что $\mathcal{D}^0$ непусто и β_0 конечно.

Так как $\mathcal{D}^0 \subset \mathcal{D}^{0,\epsilon}$, то $u_* \in \mathcal{D}^{0,\epsilon}$, $0 < \epsilon \leq \epsilon^0$. То есть $J^0(u_*) \geq J^0(u^\epsilon) = \min_{u \in \mathcal{D}^{0,\epsilon}} J^0(u) = \beta_\epsilon$. Но $\beta_\epsilon \rightarrow \beta$ при $\epsilon \rightarrow 0$ и поэтому $J^0(u_*) \geq \beta$, что вместе с (1.14) дает $J^0(u_*) = \beta$. Имеем: $\beta_0 = \inf_{u \in \mathcal{D}^0} J^0(u) \leq J^0(u_*) = \beta$. А так как $\beta \leq \beta_0$, то получаем $\beta_0 = \min_{u \in \mathcal{D}^0} J^0(u) = J^0(u_*) = \beta$, то есть $\beta = \beta_0$, u_* — решение задачи (OC⁰).

Итак, если одна из величин β , β_0 конечна, то они совпадают, множество $\mathcal{D}^0$ непусто, а задача (OC⁰) разрешима. Разрешимость, как уже было отмечено, может быть лишь однозначной. $\square$

Следуя [23], введем следующее определение.

О п р е д е л е н и е 1.1. Назовем последовательность $u^i \in \mathcal{D}$, $i = 1, 2, \dots$, минимизирующим приближенным решением (МПР) задачи (OC⁰), если $J^0[u^i] \rightarrow \beta$ при $i \rightarrow \infty$ и $u^i \in \mathcal{D}^{0,\epsilon^i}$ ($i = 1, 2, \dots$) для некоторой сходящейся к нулю последовательности чисел $\epsilon^i > 0$, $i = 1, 2, \dots$.

Укажем условия (теорема 1.2), при которых МПР обязательно так или иначе сходятся к решению задачи (OC⁰). Нам поможет, в частности, следующее утверждение.

Лемма 1.3. Пусть X — выпуклое замкнутое множество гильбертова пространства H , а функция $f(\cdot) : X \rightarrow \mathbf{R}$ выпукла, полунепрерывна снизу и субдифференцируема на X . Тогда, если она является сильно выпуклой на X с постоянной сильной выпуклости κ , то

$$f(z) \geq f(y) + \langle \partial f(y), z - y \rangle_H + \kappa \|z - y\|_H^2, \quad z, y \in X. \quad (1.15)$$

Д о к а з а т е л ь с т в о. Воспользуемся тем, что общее определение сильной выпуклости [27, гл. 8, с. 650] аналогично определению сильной выпуклости функции конечного числа переменных. Аналог для случая гильбертова пространства утверждения [27, лемма 4.3.1] позволяет сказать, что f сильно выпукла на X с постоянной $\kappa > 0$ тогда и только тогда, когда найдется выпуклая функция $g(\cdot) : X \rightarrow \mathbf{R}$ такая, что $f(z) = g(z) + \kappa \|z\|_H^2$, $z \in X$. При этом (см. [28, следствие 4.3]) функции f и g субдифференцируемы в точках X одновременно и

$$\partial f(y) = \partial g(y) + 2\kappa y, \quad y \in X. \quad (1.16)$$

Таким образом, выпуклая функция $g(z) \equiv f(z) - \kappa \|z\|_H^2$, $z \in X$, субдифференцируема на X , т. е.

$$g(z) \geq g(y) + \langle \partial g(y), z - y \rangle_H, \quad z, y \in X. \quad (1.17)$$

Осталось заметить, что из соотношений (1.16), (1.17) и очевидного тождества $\|z\|_H^2 = \|y\|_H^2 + \langle 2y, z - y \rangle_H + \|z - y\|_H^2$, $z, y \in X$, следует (1.15). $\square$

Теорема 1.2. Если $\beta < +\infty$, то любое МПР слабо в L_2^s сходится к решению u^0 задачи (OC^0) . А если еще и функционал J^0 субдифференцируем в точках $\mathcal{D}$, то всякое МПР сходится к u^0 сильно в L_2^s .

Доказательство. Пусть $\beta < +\infty$. По теореме 1.1 задача (OC^0) имеет единственное решение u^0 . Пусть $\{u^i\}_{i=1}^\infty$ — некоторое МПР задачи (OC^0) . Последовательность $\{u^i\}_{i=1}^\infty$ ограничена в L_2^s вместе с $\mathcal{D}$. То есть она слабо компактна в L_2^s [27, теорема 8.2.3] и у нее существует подпоследовательность, слабо сходящаяся в L_2^s к некоторому элементу $u_* \in \mathcal{D}$. Пусть $\{u^{\sigma_j}\}_{j=1}^\infty$ — указанная подпоследовательность. Воспользовавшись выкладками из доказательства теоремы 1.1, получаем, что элемент u_* совпадает с решением u^0 задачи (OC^0) . Тем самым мы доказали существование у каждого МПР задачи (OC^0) подпоследовательности (а она тоже есть МПР), слабо в L_2^s сходящейся к решению задачи.

Покажем теперь, что любое МПР задачи (OC^0) слабо в L_2^s сходится к решению u_0 . Воспользуемся тем, что слабая сходимости в L_2^s эквивалентна сходимости в слабой норме (см. [23, теорема I.3.11]), обозначим ее $\|\cdot\|_w$. Действуя от противного, предположим, что некоторое МПР $\{u^i\}_{i=1}^\infty$ задачи (OC^0) не сходится слабо в L_2^s к u_0 . Это означает, что существует $\nu > 0$ и подпоследовательность $\{u^{i_j}\}_{j=1}^\infty$ такие, что $\|u^{i_j} - u^0\|_w \geq \nu$ ($j = 1, 2, \dots$). Так как последовательность $\{u^{i_j}\}_{j=1}^\infty$ сама есть МПР вместе с $\{u^i\}_{i=1}^\infty$, то по доказанному выше из нее можно выделить подпоследовательность, сходящуюся слабо в L_2^s (а значит и в слабой норме $\|\cdot\|_w$) к решению u^0 задачи (OC^0) . Получили противоречие.

Предположим теперь, что функционал J^0 субдифференцируем в точках $\mathcal{D}$. Пусть $\{u^i\}_{i=1}^\infty$ — некоторое МПР задачи (OC^0) . По лемме 1.3, ввиду сильной выпуклости J^0 , для любого $i = 1, 2, \dots$ справедливо неравенство

$$J^0[u^i] - J^0[u^0] - \langle \partial J^0(u^0), u^i - u^0 \rangle \geq \kappa (\|u^i - u^0\|_{2,s})^2.$$

Так как по доказанному последовательность $\{u^i\}_{i=1}^\infty$ слабо в L_2^s сходится к решению u^0 задачи (OC^0) , а по теореме 1.1 $J^0[u^i] \rightarrow \beta = \beta_0 = J^0[u^0]$ при $i \rightarrow \infty$, то из выписанного неравенства следует, что $\|u^i - u^0\|_{2,s} \rightarrow 0$ при $i \rightarrow \infty$. $\square$

1.4. Регуляризирующий и МПР-образующий операторы

Следуя нашей работе [25], введем определение регуляризирующего оператора в задаче с поточечными фазовыми ограничениями (OC^0) .

Определение 1.2. Зависящий от δ оператор $R(\cdot, \delta)$, ставящий в соответствие каждому набору исходных данных f^δ элемент $R(f^\delta, \delta) \equiv u^\delta \in \mathcal{D}$, удовлетворяющий условиям (1.6), (1.7), (1.8), называется регуляризирующим в задаче (OC^0) , если

$$J^0[u^\delta] \rightarrow \beta, \quad \|\mathcal{G}^0[u^\delta] - \omega^0\|_{2,1} \rightarrow 0, \quad \rho(\mathcal{E}^0[u^\delta]) \rightarrow 0 \quad \text{при} \quad \delta \rightarrow 0.$$

Далее, по аналогии с [25], введем «порождаемое» этим определением понятие МПР-образующего оператора (алгоритма) в задаче (OC^0) .

Определение 1.3. Пусть $\delta^i \in (0, \delta_0)$, $i = 1, 2, \dots$, — сходящаяся к нулю последовательность положительных чисел. Зависящий от δ^i , $i = 1, 2, \dots$, оператор $R(\cdot, \delta^i)$, ставящий в соответствие каждому набору исходных данных f^{δ^i} элемент $R(f^{\delta^i}, \delta^i) \equiv u^{\delta^i} \in \mathcal{D}$, называется МПР-образующим в задаче (OC^0) , если последовательность u^{δ^i} , $i = 1, 2, \dots$, есть МПР в этой задаче.

2. Двойственная регуляризация и условия существования МПР в выпуклой задаче оптимального управления с поточечными фазовыми ограничениями и сильно выпуклым функционалом качества

В этом разделе сформулируем и докажем условия существования МПР (отдельно необходимые и достаточные) в задаче (OC^0) с одновременным конструктивным устойчивым представлением конкретного экземпляра МПР (напомним, что вместо точной задачи (OC^0) в нашем распоряжении имеется семейство задач (OC^δ) с возмущенными исходными данными, зависящими от параметра $\delta > 0$, характеризующего ошибку задания этих данных по сравнению с точными данными). Указанные условия существования можно трактовать и как условия устойчивого конструирования МПР в задаче (OC^0) . Будем придерживаться общей схемы рассуждений, использованной при получении аналогичных результатов в работах [17, 19].

2.1. Двойственная регуляризация в выпуклой задаче оптимизации с поточечными фазовыми ограничениями и необходимые условия существования МПР

В данном разделе опишем процедуру двойственной регуляризации [17, 19] для задачи оптимального управления с поточечными фазовыми ограничениями (OC^0) и получаемые с ее помощью необходимые условия существования МПР в этой задаче.

2.1.1. Двойственная задача. Введем для задачи (OC^δ) , $\delta \in [0, \delta_0]$, регулярный функционал Лагранжа

$$L^\delta(u, \lambda, \mu) \equiv J^\delta[u] + \langle \lambda, \mathcal{G}^\delta[u] - \omega^\delta \rangle_{2,1} + \langle \mu, \mathcal{E}^\delta[u] \rangle_{2,1}, \quad u \in \mathcal{D}, \quad \lambda \in L_2, \quad \mu \in L_2,$$

и соответствующую двойственную задачу

$$V^\delta(\lambda, \mu) \equiv \inf_{u \in \mathcal{D}} L^\delta(u, \lambda, \mu) \rightarrow \sup, \quad \{\lambda, \mu\} \in L_2 \times L_{2,+}. \quad (2.1)$$

Функционал V^δ определен (конечен) в каждой точке $\{\lambda, \mu\} \in L_2 \times L_2$ и вогнут. Поэтому он является локально липшицевым на $L_2 \times L_2$, так как благодаря ограниченности $\mathcal{D}$ ограничен на любом ограниченном множестве [28, теорема 2.1]. Ввиду выпуклости и замкнутости в L_2^s множества $\mathcal{D}$, непрерывности и сильной выпуклости на $\mathcal{D}$ функционала $J^\delta(u)$, здесь $\inf_{u \in \mathcal{D}} L^\delta(u, \lambda, \mu) = \min_{u \in \mathcal{D}} L^\delta(u, \lambda, \mu)$ для каждой пары $\{\lambda, \mu\} \in L_2 \times L_{2,+}$, причем минимум достигается в единственной точке

$$u^\delta[\lambda, \mu] \equiv \operatorname{argmin} \{L^\delta(u, \lambda, \mu), u \in \mathcal{D}\}.$$

Из неравенства $|V^\delta(\lambda, \mu) - V^0(\lambda, \mu)| \leq \sup_{u \in \mathcal{D}} |L^\delta(u, \lambda, \mu) - L^0(u, \lambda, \mu)|$ и условий А, Б, В и Г простыми выкладками, используя лемму 1.1 и ограниченность $\mathcal{D}$, получаем, что при некотором $\tilde{C} > 0$

$$|V^\delta(\lambda, \mu) - V^0(\lambda, \mu)| \leq \tilde{C}\delta(1 + \|\lambda\|_{2,1} + \|\mu\|_{2,1}), \quad \{\lambda, \mu\} \in L_2 \times L_{2,+}, \quad \delta \in (0, \delta_0]. \quad (2.2)$$

При любых $\delta \in [0, \delta_0]$, $\alpha > 0$ сильно вогнутый функционал

$$R^{\delta,\alpha}(\lambda, \mu) \equiv V^\delta(\lambda, \mu) - \alpha\|\lambda\|_{2,1}^2 - \alpha\|\mu\|_{2,1}^2, \quad \{\lambda, \mu\} \in L_2 \times L_{2,+},$$

достигает на $L_2 \times L_{2,+}$ максимума в одной точке

$$\{\lambda^{\delta,\alpha}, \mu^{\delta,\alpha}\} \equiv \operatorname{argmax} \{R^{\delta,\alpha}(\lambda, \mu), \{\lambda, \mu\} \in L_2 \times L_{2,+}\}.$$

Пусть $\alpha(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ — некоторая функция, удовлетворяющая условию

$$\frac{\delta}{\alpha(\delta)} \rightarrow 0, \quad \alpha(\delta) \rightarrow 0 \quad \text{при} \quad \delta \rightarrow 0. \quad (2.3)$$

Общие свойства метода стабилизации [27, гл. 8, § 4] позволяют сказать, что в случае разрешимости двойственной задачи (2.1) при $\delta = 0$ оценка (2.2) и условие согласования (2.3) влекут за собой сходимость

$$\|\lambda^{\delta,\alpha(\delta)} - \lambda^0\|_{2,1} + \|\mu^{\delta,\alpha(\delta)} - \mu^0\|_{2,1} \rightarrow 0, \quad \delta \rightarrow 0, \quad (2.4)$$

где $\{\lambda^0, \mu^0\}$ — минимальное по норме решение указанной двойственной задачи.

2.1.2. Супердифференциал вогнутого целевого функционала двойственной задачи. Далее нам потребуется формула для супердифференциала вогнутого функционала значений $V^\delta(\cdot, \cdot)$, который по определению равен взятому с обратным знаком субдифференциалу выпуклого функционала $-V^\delta(\cdot, \cdot)$. Чтобы выписать эту формулу, обратимся сначала к общей ситуации, рассмотренной в [17].

Пусть $\mathcal{Y}$, $\mathcal{H}$ — некоторые гильбертовы пространства, $\mathcal{P}$ — ограниченное замкнутое множество в $\mathcal{Y}$, $I_0(\cdot) : \mathcal{P} \rightarrow \mathbf{R}$ — липшицев функционал, $I_1(\cdot) : \mathcal{P} \rightarrow \mathcal{H}$ — липшицев оператор. Считаем I_0 и I_1 равномерно ограниченными на $\mathcal{P}$. Определим составной функционал (его естественно назвать функционалом Лагранжа)

$$L(y, \nu) \equiv I_0(y) + \langle \nu, I_1(y) \rangle_{\mathcal{H}}, \quad y \in \mathcal{P}, \quad \nu \in \mathcal{H},$$

и соответствующий (двойственный) функционал значений

$$V(\nu) \equiv \inf_{y \in \mathcal{P}} L(y, \nu), \quad \nu \in \mathcal{H}.$$

Супердифференциал вогнутого функционала значений V задается следующей формулой, при записи которой используется обобщенный градиент Кларка [29, § 2.1].

Лемма 2.4. *Супердифференциал вогнутого функционала V в точке $\nu \in \mathcal{H}$ равен*

$$\partial V(\nu) = \partial_C V(\nu) = \overline{\operatorname{co}} \left\{ w - \lim_{i \rightarrow \infty} I_1(y^i) : y^i \in \mathcal{P}, \quad L(y^i, \nu) \rightarrow \inf_{y \in \mathcal{P}} L(y, \nu), \quad i \rightarrow \infty \right\},$$

где $\partial_C V(\nu)$ — обобщенный градиент Кларка функционала V в точке ν , $\overline{\operatorname{co}} X$ — замыкание выпуклой оболочки множества X , предел $w - \lim$ понимается в смысле слабой сходимости в пространстве $\mathcal{H}$. Другими словами, супердифференциал $\partial V(\nu)$ есть замыкание выпуклой оболочки совокупности всех слабых пределов всевозможных последовательностей $I_1(y^i)$, $i = 1, 2, \dots$, когда последовательность $y^i \in \mathcal{P}$, $i = 1, 2, \dots$, является минимизирующей для функционала $L(y, \nu)$, $y \in \mathcal{P}$.

Д о к а з а т е л ь с т в о этой леммы см. в [17, лемма 2].

Применим лемму 2.4 к функционалу $V^\delta(\lambda, \mu)$, $\{\lambda, \mu\} \in L_2 \times L_2$. В этом случае $\mathcal{Y} = L_2^s$, $\mathcal{H} = L_2 \times L_2$, $\mathcal{P} = \mathcal{D}$, $V(\nu) = V^\delta(\lambda, \mu)$, $I_0(\cdot) = J^\delta[\cdot]$, $I_1(\cdot) = \{\mathcal{G}^\delta[\cdot] - \omega^\delta, \mathcal{E}^\delta[\cdot]\}$.

Лемма 2.5. Супердифференциал вогнутого функционала V^δ в точке $\{\lambda, \mu\} \in L_2 \times L_2$ равен

$$\partial V^\delta(\lambda, \mu) = \partial_C V^\delta(\lambda, \mu) = \overline{\text{co}} \left\{ w - \lim_{i \rightarrow \infty} \{ \mathcal{G}^\delta[u^i] - \omega^\delta, \mathcal{E}^\delta[u^i] \} : u^i \in \mathcal{D}, \right. \\ \left. L^\delta(u^i, \lambda, \mu) \rightarrow \inf_{u \in \mathcal{D}} L^\delta(u, \lambda, \mu), i \rightarrow \infty \right\},$$

где $\partial_C V^\delta(\lambda, \mu)$ — обобщенный градиент Кларка функционала V^δ в точке $\{\lambda, \mu\}$, а предел $w - \lim$ понимается в смысле слабой сходимости в пространстве $L_2 \times L_2$.

В свою очередь, лемма 2.5, в силу слабой непрерывности операторов $\mathcal{G}^\delta[\cdot]$, $\mathcal{E}^\delta[\cdot]$ (см. лемму 1.2) и единственности для любой пары $\{\lambda, \mu\} \in L_2 \times L_{2,+}$ минимали $u^\delta[\lambda, \mu]$, преобразуется к следующему виду.

Лемма 2.6. Супердифференциал вогнутого функционала V^δ в точке $\{\lambda, \mu\} \in L_2 \times L_{2,+}$ является одноточечным множеством и равен

$$\partial V^\delta(\lambda, \mu) = \partial_C V^\delta(\lambda, \mu) = \{ \mathcal{G}^\delta[u^\delta[\lambda, \mu]] - \omega^\delta, \mathcal{E}^\delta[u^\delta[\lambda, \mu]] \},$$

где $\partial_C V^\delta(\lambda, \mu)$ — обобщенный градиент Кларка функционала V^δ в точке $\{\lambda, \mu\}$.

Доказательство. Утверждение леммы вытекает непосредственно из леммы 2.5 и выпуклости задачи (OC^δ) с сильно выпуклым функционалом цели. Действительно, в силу единственности точки $u^\delta[\lambda, \mu]$, минимизирующей сильно выпуклый функционал $L^\delta(u, \lambda, \mu)$ на $\mathcal{D}$, и сильной сходимости к этой точке любой минимизирующей последовательности в этой задаче минимизации (при $\{\lambda, \mu\} \in L_2 \times L_{2,+}$), обобщенный градиент Кларка $\partial_C V^\delta(\lambda, \mu)$ состоит из единственного элемента. Другими словами, в силу равенства $\partial_C(-V^\delta(\lambda, \mu)) = -\partial_C V^\delta(\lambda, \mu)$ и совпадения субдифференциала $\partial(-V^\delta(\lambda, \mu))$ с обобщенным градиентом $\partial_C(-V^\delta(\lambda, \mu))$ [29, предложения 2.2.7, 2.3.1], взятый с обратным знаком субдифференциал выпуклого функционала $-V^\delta(\lambda, \mu)$, то есть супердифференциал вогнутого функционала $V^\delta(\lambda, \mu)$, в каждой точке $\{\lambda, \mu\} \in L_2 \times L_{2,+}$ состоит лишь из одного элемента $\{ \mathcal{G}^\delta[u^\delta[\lambda, \mu]] - \omega^\delta, \mathcal{E}^\delta[u^\delta[\lambda, \mu]] \}$. $\square$

2.1.3. Метод двойственной регуляризации. В данном разделе опишем метод двойственной регуляризации [17, 19] применительно к задаче оптимального управления с поточечными фазовыми ограничениями (OC^0) и докажем соответствующую теорему сходимости этого метода. Считаем при этом, что задача (OC^0) разрешима. Это при сделанных предположениях об исходных данных равносильно существованию в ней МПР (см. теоремы 1.1, 1.2).

Для описания и обоснования процедуры двойственной регуляризации воспользуемся следующей леммой, частным случаем более общего утверждения [30, следствие 4.3.6(c)].

Лемма 2.7. Пусть H — гильбертово пространство, $f : H \rightarrow \mathbf{R} \cup \{+\infty\}$ — собственная выпуклая полунепрерывная снизу функция с $\text{dom } f \equiv \{z \in H : f(z) < +\infty\} = H$, $\Omega \subset H$ — замкнутое выпуклое множество. Тогда $x \in \Omega$ — точка минимума f на Ω в том и только в том случае, когда $0 \in \partial f(x) + N_\Omega(x)$, где $N_\Omega(x)$ — конус нормалей в смысле выпуклого анализа ко множеству Ω в точке x .

Заметим, что в условиях леммы 2.7 (см. [30, предложение 4.1.4])

$$N_{\Omega}(x) = \{p \in H : \langle p, y - x \rangle_H \leq 0, y \in \Omega\}.$$

Применим лемму 2.7, взяв $H = L_2 \times L_2$, $\Omega = L_2 \times L_{2,+}$, а в качестве f выпуклую функцию $f(\lambda, \mu) = -V^{\delta}(\lambda, \mu) + \alpha\|\lambda\|_{2,1}^2 + \alpha\|\mu\|_{2,1}^2$, $\{\lambda, \mu\} \in L_2 \times L_2$. С учетом леммы 2.6, позволяющей выписать субдифференциал выбранной функции f , критерий минимума леммы 2.7 дает следующее вариационное неравенство

$$\begin{aligned} \langle \{\mathcal{G}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^{\delta}, \mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]]\} - 2\alpha(\delta)\{\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}\}, \\ \{\lambda', \mu'\} - \{\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}\} \rangle_{L_2 \times L_2} \leq 0, \quad \{\lambda', \mu'\} \in L_2 \times L_{2,+}. \end{aligned} \quad (2.5)$$

Следствием вариационного неравенства (2.5) являются соотношения

$$\mathcal{G}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^{\delta} = 2\alpha(\delta)\lambda^{\delta,\alpha(\delta)}; \quad (2.6)$$

$$\mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) = 2\alpha(\delta)\mu^{\delta,\alpha(\delta)}(t) \text{ при п.в. } t \in \Pi \text{ таких, что } \mu^{\delta,\alpha(\delta)}(t) > 0; \quad (2.7)$$

$$\mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) \leq 0 \text{ при п.в. } t \in \Pi \text{ таких, что } \mu^{\delta,\alpha(\delta)}(t) = 0. \quad (2.8)$$

Из соотношений (2.6)–(2.8), в свою очередь, получаем:

$$\langle \mathcal{G}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^{\delta}, \lambda^{\delta,\alpha(\delta)} \rangle_{2,1} = 2\alpha(\delta)\|\lambda^{\delta,\alpha(\delta)}\|^2 \geq 0. \quad (2.9)$$

$$\langle \mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] , \mu^{\delta,\alpha(\delta)} \rangle_{2,1} = 2\alpha(\delta)\|\mu^{\delta,\alpha(\delta)}\|^2 \geq 0.$$

Из (2.7) следует также, что если $\mu^{\delta,\alpha(\delta)}(t) > 0$ для некоторого t , принадлежащего множеству полной меры в $\{t \in \Pi : \mu^{\delta,\alpha(\delta)}(t) > 0\}$, то

$$\mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) - 2\alpha(\delta)\mu^{\delta,\alpha(\delta)}(t) = 0, \quad \mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t)\mu^{\delta,\alpha(\delta)}(t) > 0 \quad (2.10)$$

и, значит, при п. в. $t \in \Pi$ таких, что $\mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) < 0$, выполняется равенство $\mu^{\delta,\alpha(\delta)}(t) = 0$. Из (2.7) и (2.10) получаем одновременно, что

$$\mu^{\delta,\alpha(\delta)}(t)\mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) \geq 0 \text{ при п. в. } t \in \Pi.$$

Кроме того, из (2.9) получаем равенство и неравенство

$$\begin{aligned} \langle \{\mathcal{G}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^{\delta}, \mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]]\} , \{\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}\} \rangle_{L_2 \times L_2} \\ = 2\alpha(\delta)(\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2) \geq 0. \end{aligned} \quad (2.11)$$

Далее, так как

$$\begin{aligned} L^{\delta}(u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}], \lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}) &\equiv J^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] \\ &+ \langle \lambda^{\delta,\alpha(\delta)}, \mathcal{G}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^{\delta} \rangle_{2,1} + \langle \mu^{\delta,\alpha(\delta)}, \mathcal{E}^{\delta}[u^{\delta}[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] \rangle_{2,1} \\ &\leq L^{\delta}(u^0, \lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}) \equiv J^{\delta}[u^0] + \langle \lambda^{\delta,\alpha(\delta)}, \mathcal{G}^{\delta}[u^0] - \omega^{\delta} \rangle_{2,1} + \langle \mu^{\delta,\alpha(\delta)}, \mathcal{E}^{\delta}[u^0] \rangle_{2,1} \\ &\leq L^{\delta}(u^0, \lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}) \equiv J^{\delta}[u^0] + \langle \lambda^{\delta,\alpha(\delta)}, \mathcal{G}^{\delta}[u^0] - \omega^{\delta} \rangle_{2,1} + \langle \mu^{\delta,\alpha(\delta)}, \mathcal{E}^{\delta}[u^0] - \mathcal{E}^0[u^0] \rangle_{2,1} \\ &\leq J^0[u^0] + [J^{\delta}[u^0] - J^0[u^0]] + \|\lambda^{\delta,\alpha(\delta)}\|_{2,1} \|\mathcal{G}^{\delta}[u^0] - \omega^{\delta}\|_{2,1} + \|\mu^{\delta,\alpha(\delta)}\|_{2,1} \|\mathcal{E}^{\delta}[u^0] - \mathcal{E}^0[u^0]\|_{2,1}, \end{aligned}$$

то в силу оценок (1.6), (1.13) и равенства (2.11) получаем после ряда элементарных преобразований оценку

$$\begin{aligned} 2\alpha(\delta) (\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2) &\leq C_3\delta (\|\lambda^{\delta,\alpha(\delta)}\|_{2,1} + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}) + J^0(u^0) + C_3\delta - \min_{u \in \mathcal{D}} J^\delta(u) \\ &\leq C_4\delta \sqrt{\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2} + J^0(u^0) + C_3\delta - \min_{u \in \mathcal{D}} J^\delta(u) \quad \text{при } \delta \in (0, \delta_0), \end{aligned}$$

или

$$2\alpha(\delta) (\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2) - C_4\delta \sqrt{\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2} - J^0(u^0) - C_3\delta + \min_{u \in \mathcal{D}} J^\delta(u) \leq 0,$$

где C_3, C_4 — не зависящие от δ положительные постоянные. Отсюда находим, что

$$\sqrt{\|\lambda^{\delta,\alpha(\delta)}\|_{2,1}^2 + \|\mu^{\delta,\alpha(\delta)}\|_{2,1}^2} \leq \frac{C_4\delta + \sqrt{(C_4\delta)^2 - 8\alpha(\delta)K(\delta)}}{4\alpha(\delta)} \quad \text{при } \delta \in (0, \delta_0], \quad (2.12)$$

где $K(\delta) \equiv \min_{u \in \mathcal{D}} J^\delta(u) - J^0(u^0) - C_3\delta$, следствием чего является предельное соотношение

$$\alpha(\delta) \|\{\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}\}\|_{L_2 \times L_2} \rightarrow 0 \quad \text{при } \delta \rightarrow 0. \quad (2.13)$$

Кроме того, из соотношений (2.6)–(2.8), оценок (1.5), (1.6), (1.7), (1.8), (1.12) и оценки (2.12), в свою очередь, следуют оценки (в которых положительные постоянные C_5, C_6, C_7 не зависят от $\delta \in (0, \delta_0]$):

$$\begin{aligned} \|\mathcal{G}^0[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^0\|_{2,1} &\leq \|\mathcal{G}^\delta[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^\delta\|_{2,1} + C_5\delta \\ &\leq 2\alpha(\delta) \|\lambda^{\delta,\alpha(\delta)}\|_{2,1} + C_5\delta \leq C_4\delta + \sqrt{(C_4\delta)^2 - 8\alpha(\delta)K(\delta)} + C_5\delta, \end{aligned} \quad (2.14)$$

$$\begin{aligned} \mathcal{E}^0[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) &\equiv F^0(t, \Sigma^0[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^0](t)) = F^0(t, \Sigma^0[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^0](t)) \\ &\quad - F^\delta(t, \Sigma^0[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^0](t)) + F^\delta(t, \Sigma^0[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^0](t)) \\ &\quad - F^\delta(t, \Sigma^\delta[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^\delta](t)) + F^\delta(t, \Sigma^\delta[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^\delta](t)) \\ &\leq C_6\delta + C_7\delta + F^\delta(t, \Sigma^\delta[z_{u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]}^\delta](t)) \\ &\leq C_6\delta + C_7\delta + 2\alpha(\delta)\mu^{\delta,\alpha(\delta)}(t) \quad \text{при п. в. } t \in \Pi, \end{aligned} \quad (2.15)$$

$$2\alpha(\delta) \|\mu^{\delta,\alpha(\delta)}\|_{2,1} \leq C_4\delta + \sqrt{(C_4\delta)^2 - 8\alpha(\delta)K(\delta)} \quad \text{при } \delta \in (0, \delta_0]. \quad (2.16)$$

Из неравенств (2.14), (2.15), (2.16) выводим, что

$$\mathcal{G}^0[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]] - \omega^0 \rightarrow 0 \quad \text{при } \delta \rightarrow 0, \quad (2.17)$$

и для любого $\delta \in (0, \delta_0]$ существует функция $\phi_\delta(\cdot) \in L_{2,+}$ такая, что

$$\mathcal{E}^0[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](t) \leq \phi_\delta(t), \quad t \in \Pi, \quad \text{причем } \|\phi_\delta(\cdot)\|_{2,1} \rightarrow 0 \quad \text{при } \delta \rightarrow 0. \quad (2.18)$$

Неравенство в (2.18) означает, что

$$(\mathcal{E}^0[u^\delta[\lambda^{\delta,\alpha(\delta)}, \mu^{\delta,\alpha(\delta)}]](\cdot) - \phi_\delta(\cdot)) \in L_{2,-}.$$

Одновременно для всех $u \in \mathcal{D}$ выполняется неравенство

$$\begin{aligned} J^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] + \langle \lambda^{\delta, \alpha(\delta)}, \mathcal{G}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - \omega^\delta \rangle_{2,1} + \langle \mu^{\delta, \alpha(\delta)}, \mathcal{E}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \rangle_{2,1} \\ \leq J^\delta [u] + \langle \lambda^{\delta, \alpha(\delta)}, \mathcal{G}^\delta [u] - \omega^\delta \rangle_{2,1} + \langle \mu^{\delta, \alpha(\delta)}, \mathcal{E}^\delta [u] \rangle_{2,1}, \end{aligned}$$

из которого, с учетом оценок (1.13), следует, что

$$\begin{aligned} J^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] + \langle \lambda^{\delta, \alpha(\delta)}, \mathcal{G}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - \omega^\delta \rangle_{2,1} + \langle \mu^{\delta, \alpha(\delta)}, \mathcal{E}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \rangle_{2,1} \\ \leq J^\delta [u^0] + \langle \lambda^{\delta, \alpha(\delta)}, \mathcal{G}^\delta [u^0] - \omega^\delta \rangle_{2,1} + \langle \mu^{\delta, \alpha(\delta)}, \mathcal{E}^\delta [u^0] \rangle_{2,1} \\ \leq J^0 [u^0] + C_8 (\delta \|\lambda^{\delta, \alpha(\delta)}\|_{2,1} + \delta \|\mu^{\delta, \alpha(\delta)}\|_{2,1} + \delta), \quad (2.19) \end{aligned}$$

где $C_8 > 0$ — постоянная, не зависящая от $\delta \in (0, \delta_0]$. Учитывая ограниченность $\mathcal{D}$, получаем, что при всех $\delta \in (0, \delta_0]$

$$\begin{aligned} J^0 [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \leq J^0 [u^0] + (J^0 [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - J^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]]) \\ + C_8 \delta (\|\lambda^{\delta, \alpha(\delta)}\|_{2,1} + \|\mu^{\delta, \alpha(\delta)}\|_{2,1} + 1) \leq J^0 [u^0] + C_9 (\delta \|\lambda^{\delta, \alpha(\delta)}\|_{2,1} + \delta \|\mu^{\delta, \alpha(\delta)}\|_{2,1} + \delta), \end{aligned}$$

где $C_9 > 0$ не зависит от $\delta \in (0, \delta_0]$. Отсюда, в силу условия согласования (2.3) и соотношения (2.13), следует: для любого $\delta \in (0, \delta_0]$ существует число $\tilde{\phi}_\delta \geq 0$ такое, что

$$J^0 [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \leq J^0 [u^0] + \tilde{\phi}_\delta, \quad \text{причем} \quad \tilde{\phi}_\delta \rightarrow 0 \quad \text{при} \quad \delta \rightarrow 0. \quad (2.20)$$

Итак, с учетом ограниченности $\mathcal{D}$ построено семейство зависящих от $\delta \in (0, \delta_0]$ элементов $u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]$, удовлетворяющих оценкам (2.14), (2.15), (2.20) и, стало быть, такое, что выполняются соотношения (2.17), (2.18) и одновременно

$$J^0 [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \rightarrow \min_{u \in \mathcal{D}^0} J^0 [u] = \beta = \beta_0 = J^0 [u^0], \quad \delta \rightarrow 0. \quad (2.21)$$

Соотношения (2.17), (2.18), (2.21) в соответствии с определением 1.2 позволяют сказать, что оператор $R(\cdot, \delta)$, ставящий в соответствие каждому набору исходных данных $\mathbf{f}^\delta$, удовлетворяющих оценкам (1.13), элемент $u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}] \in \mathcal{D}$, является регуляризирующим в задаче (OC^0) . При этом в соответствии с теоремой 1.2 любая последовательность $u^{\delta^k} [\lambda^{\delta^k, \alpha(\delta^k)}, \mu^{\delta^k, \alpha(\delta^k)}]$, $k = 1, 2, \dots$, $\delta^k \rightarrow 0$, $k \rightarrow \infty$, сходится слабо к решению задачи (OC^0) , а в случае субдифференцируемости J^0 на $\mathcal{D}$ сходится сильно, то есть $R(\mathbf{f}^\delta, \delta) \equiv u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}] \rightarrow u^0$, $\delta \rightarrow 0$.

Наконец, следствием оценки (2.19), условия (2.3), предельного соотношения (2.13), ограниченности $\mathcal{D}$ и соотношения (2.21) является предельное соотношение

$$\langle \{ \lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)} \}, \{ \mathcal{G}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - \omega^\delta, \mathcal{E}^\delta [u^\delta [\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \} \rangle_{L_2 \times L_2} \rightarrow 0, \quad \delta \rightarrow 0. \quad (2.22)$$

Равенство (2.11) и соотношение (2.22) позволяют сказать, что

$$\alpha(\delta) \left\| \{ \lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)} \} \right\|_{L_2 \times L_2}^2 \rightarrow 0 \quad \text{при} \quad \delta \rightarrow 0. \quad (2.23)$$

Покажем, что выполняется и предельное соотношение

$$\lim_{\delta \rightarrow 0} V^0(\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}) = \sup_{\{\lambda, \mu\} \in L_2 \times L_{2,+}} V^0(\lambda, \mu) = \min_{u \in \mathcal{D}^0} J^0 [u] = J^0 [u^0]. \quad (2.24)$$

Заметим сначала, что благодаря соотношениям (2.21), (2.22) можем записать

$$\lim_{\delta \rightarrow 0} V^\delta(\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}) = J^0[u^0],$$

откуда в силу оценки (2.2), предельных соотношений (2.13) и условия (2.3) получаем

$$V^0(\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}) \rightarrow J^0[u^0] \quad \text{при } \delta \rightarrow 0,$$

что и означает, в силу неравенства $\sup_{\{\lambda, \mu\} \in L_2 \times L_{2,+}} V^0(\lambda, \mu) \leq \min_{u \in \mathcal{D}^0} J^0[u] = J^0[u^0]$, справедливость предельного соотношения (2.24).

Таким образом, как результат проведенных выше рассуждений, мы можем сформулировать следующую теорему сходимости метода двойственной регуляризации в задаче (OC^0) . Эта теорема сходимости является следствием сделанного в начале раздела 2.1.3 предположения о существовании решения задачи (OC^0) . По этой причине сформулированный в ней результат можно трактовать и как необходимое условие оптимальности управления u^0 , выраженное в секвенциальной форме.

Теорема 2.3. Пусть задача (OC^0) имеет решение u^0 . Вне зависимости от того, существует или нет вектор Куна–Таккера в задаче (OC^0) , или, другими словами, разрешима или нет двойственная к ней задача, имеют место соотношения

$$\alpha(\delta) \left\| \{ \lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)} \} \right\|_{L_2 \times L_2} \rightarrow 0 \quad \text{при } \delta \rightarrow 0, \quad (2.25)$$

$$J^0[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \rightarrow \min_{u \in \mathcal{D}^0} J^0[u] = J^0[u^0] \quad \text{при } \delta \rightarrow 0,$$

$$\mathcal{G}^0[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - \omega^0 \rightarrow 0 \quad \text{при } \delta \rightarrow 0,$$

$$\mathcal{E}^0[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]](t) \leq \phi_\delta(t), \quad t \in \Pi; \quad \|\phi_\delta\|_{2,1} \rightarrow 0 \quad \text{при } \delta \rightarrow 0,$$

$$\langle \{ \lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)} \}, \{ \mathcal{G}^\delta[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] - \omega^\delta, \mathcal{E}^\delta[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]] \} \rangle_{L_2 \times L_2} \rightarrow 0 \quad \text{при } \delta \rightarrow 0,$$

где $\phi_\delta(\cdot) \in L_{2,+}$ и неравенство означает включение $(\mathcal{E}^0[u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]](\cdot) - \phi_\delta(\cdot)) \in L_{2,-}$.

Элементы $u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}]$ при $\delta \rightarrow 0$ слабо сходятся к u^0 , а в случае субдифференцируемости J^0 на $\mathcal{D}$ сходимость является сильной. Таким образом, зависящий от δ оператор $R(\cdot, \delta)$, ставящий в соответствие каждому набору исходных данных f^δ , удовлетворяющих оценкам (1.6), (1.7), (1.8), элемент $R(f^\delta, \delta) \equiv u^\delta[\lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)}] \in \mathcal{D}$, является регуляризующим (в смысле определения 1.2) в задаче (OC^0) . Кроме того выполняется предельное соотношение (2.24), а в случае разрешимости двойственной к (OC^0) задачи и предельное соотношение (2.4).

З а м е ч а н и е 2.1. Теорема 2.3 сформулирована с учетом «опоры» на предельное соотношение (2.13). Если же это предельное соотношение заменить более «сильным» предельным соотношением (2.23), то мы получаем другой вариант теоремы сходимости метода двойственной регуляризации в задаче (OC^0) , который отличается от сформулированного выше лишь тем, что используемое в теореме 2.3 предельное соотношение (2.25) заменяется на «аналогичное» предельное соотношение $\alpha(\delta) \left\| \{ \lambda^{\delta, \alpha(\delta)}, \mu^{\delta, \alpha(\delta)} \} \right\|_{L_2 \times L_2}^2 \rightarrow 0, \delta \rightarrow 0$. Такой вариант теоремы сходимости используется при рассмотрении задач условной минимизации с неограниченными множествами допустимых элементов $\mathcal{D}$ [20–22].

2.1.4. Необходимые условия существования МПР. Как следствие теоремы 2.3 сформулируем, наконец, необходимые условия существования МПР в задаче (OC^0) .

Теорема 2.4. Пусть $\{\delta^k\}_{k=1}^\infty \subset (0, \delta_0]$ — фиксированная стремящаяся к нулю последовательность чисел. Вне зависимости от разрешимости или неразрешимости двойственной к (OC^0) задачи справедливо следующее. Для того, чтобы в задаче (OC^0) существовало МПР (в случае существования оно сходится слабо к решению задачи), необходимо существование стремящейся к нулю последовательности чисел $\{\epsilon^k\}_{k=1}^\infty \subset [0, \infty)$, а также последовательности двойственных переменных $\{\lambda^k, \mu^k\} \in L_2 \times L_{2,+}$, $k = 1, 2, \dots$, таких, что $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0$ при $k \rightarrow \infty$ и выполняются предельные соотношения

$$u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}^{\delta^k, \epsilon^k}, \quad \epsilon^k \rightarrow 0, \quad k \rightarrow \infty, \quad (2.26)$$

$$\left\langle \{\lambda^k, \mu^k\}, \{\mathcal{G}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] - \omega^{\delta^k}, \mathcal{E}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]]\} \right\rangle_{L_2 \times L_2} \rightarrow 0, \quad k \rightarrow \infty. \quad (2.27)$$

Одновременно с предельным соотношением $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0$, $k \rightarrow \infty$, и предельными соотношениями (2.26), (2.27) выполняется и предельное соотношение

$$\lim_{k \rightarrow \infty} V^0(\lambda^k, \mu^k) = \sup_{\{\lambda, \mu\} \in L_2 \times L_{2,+}} V^0(\lambda, \mu) = \min_{u \in \mathcal{D}^0} J^0[u] = J^0[u^0]. \quad (2.28)$$

Если некоторое МПР существует, то необходимо найдется и последовательность двойственных переменных $\{\lambda^k, \mu^k\}$, $k = 1, 2, \dots$, с указанными выше свойствами, генерируемая методом двойственной регуляризации теоремы 2.3, то есть $\{\lambda^k, \mu^k\} = \{\lambda^{\delta^k, \alpha(\delta^k)}, \mu^{\delta^k, \alpha(\delta^k)}\}$, $k = 1, 2, \dots$; при разрешимости двойственной задачи указанная последовательность сходится к минимальному по норме решению этой задачи. Соответствующая последовательность $u^{\delta^k}[\lambda^k, \mu^k]$, $k = 1, 2, \dots$, будет МПР; в случае субдифференцируемости J^0 на $\mathcal{D}$ это МПР сходится сильно при $k \rightarrow \infty$ к решению u^0 задачи (OC^0) . То есть оператор $R(\cdot, \delta^k)$, ставящий в соответствие при таком выборе $\{\lambda^k, \mu^k\}$, $k = 1, 2, \dots$, каждому набору исходных данных f^{δ^k} , удовлетворяющих оценкам (1.6), (1.7), (1.8) при $\delta = \delta^k$, элемент $u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}$, $k = 1, 2, \dots$, является МПР-образующим в задаче (OC^0) .

З а м е ч а н и е 2.2. Заметим, что ввиду ограниченности $\mathcal{D}$ соотношения (2.26) выполняются тогда и только тогда, когда $u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}^{0, \tilde{\epsilon}^k}$ для некоторой последовательности сходящихся к нулю неотрицательных чисел $\tilde{\epsilon}^k$, $k = 1, 2, \dots$.

З а м е ч а н и е 2.3. С учетом замечания 2.1 другой вариант необходимых условий существования МПР в задаче (OC^0) мы получаем, если используемое в теореме 2.4 предельное соотношение $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0$, $k \rightarrow \infty$, заменить на «аналогичное» предельное соотношение $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2}^2 \rightarrow 0$, $k \rightarrow \infty$. При этом, как легко заметить, первое из указанных двух предельных соотношений является следствием второго.

2.2. Достаточные условия существования МПР в выпуклой задаче оптимального управления с поточечными фазовыми ограничениями

Каждому из двух вариантов необходимых условий существования МПР, о которых идет речь в замечании 2.3, подходит свой вариант указанных достаточных условий существования МПР. Так как предельное соотношение

$$\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0, \quad k \rightarrow \infty,$$

можно трактовать как менее жесткое по сравнению с предельным соотношением

$$\delta^k \left\| \{\lambda^k, \mu^k\} \right\|_{L_2 \times L_2}^2 \rightarrow 0, \quad k \rightarrow \infty,$$

то ниже приводится формулировка и доказательство варианта достаточных условий, соответствующего первому из двух указанных предельных соотношений. Соответствующий второму предельному соотношению вариант достаточных условий применяется при регуляризации КУО в задачах с неограниченными множествами допустимых элементов $\mathcal{D}$ (см. [20–22]).

Справедлива следующая теорема, устанавливающая достаточные условия существования МПР в задаче (OC^0) .

Теорема 2.5. Пусть $\{\delta^k\}_{k=1}^\infty \subset (0, \delta_0]$ — фиксированная стремящаяся к нулю последовательность чисел. Вне зависимости от разрешимости или неразрешимости двойственной к (OC^0) задачи справедливо следующее. Для того, чтобы в задаче (OC^0) существовало МПР (в случае существования оно будет сходиться слабо к решению задачи), достаточно существования стремящейся к нулю последовательности чисел $\{\epsilon^k\}_{k=1}^\infty \subset [0, \infty)$, а также последовательности двойственных переменных $\{\lambda^k, \mu^k\} \in L_2 \times L_{2,+}$, $k = 1, 2, \dots$, таких, что $\delta^k \left\| \{\lambda^k, \mu^k\} \right\|_{L_2 \times L_2} \rightarrow 0$, $k \rightarrow \infty$, и выполняются соотношения (2.26), (2.27). При этом последовательность $u^{\delta^k}[\lambda^k, \mu^k]$, $k = 1, 2, \dots$, является искомым МПР. Это МПР сильно сходится к решению u^0 задачи (OC^0) в случае субдифференцируемости J^0 на $\mathcal{D}$. Другими словами, оператор $R(\cdot, \delta^k)$, ставящий в соответствие каждому набору исходных данных f^{δ^k} , удовлетворяющих оценкам (1.6), (1.7), (1.8) при $\delta = \delta^k$, элемент $R(f^{\delta^k}, \delta^k) \equiv u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}$, является МПР-образующим в задаче (OC^0) . Одновременно с предельным соотношением $\delta^k \left\| \{\lambda^k, \mu^k\} \right\|_{L_2 \times L_2} \rightarrow 0$, $k \rightarrow \infty$, и соотношениями (2.26), (2.27) выполняется и предельное соотношение (2.28).

Доказательство. Прежде всего, напомним (см. раздел 1.2), что функционал J^0 слабо полунепрерывен снизу на $\mathcal{D}$; операторы $\mathcal{G}^0, \mathcal{E}^0$ слабо непрерывны на $\mathcal{D}$, то есть переводят слабо сходящиеся в L_2^s последовательности управлений из $\mathcal{D}$ в сильно сходящиеся в L_2 последовательности их образов. Указанные свойства, ввиду включения $u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}^{\delta^k, \epsilon^k}$ и ограниченности последовательности $u^{\delta^k}[\lambda^k, \mu^k]$, $k = 1, 2, \dots$, обеспечивают непустоту множества $\mathcal{D}^0$, а следовательно, и разрешимость (однозначную) исходной задачи (OC^0) . Далее, так как точка $u^{\delta^k}[\lambda^k, \mu^k]$ минимизирует функционал $L^{\delta^k}(\cdot, \lambda^k, \mu^k)$, при $u \in \mathcal{D}$ можем записать

$$\begin{aligned} J^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] + \left\langle \{\lambda^k, \mu^k\}, \{\mathcal{G}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] - \omega^{\delta^k}, \mathcal{E}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]]\} \right\rangle_{L_2 \times L_2} \\ \leq J^{\delta^k}[u] + \left\langle \{\lambda^k, \mu^k\}, \{\mathcal{G}^{\delta^k}[u] - \omega^{\delta^k}, \mathcal{E}^{\delta^k}[u]\} \right\rangle_{L_2 \times L_2}. \end{aligned}$$

В силу условий теоремы (см. предельное соотношение (2.27)) отсюда следует, что при любом $u \in \mathcal{D}$ существует стремящаяся к нулю последовательность $\{\psi^k\}_{k=1}^\infty$ неотрицательных чисел такая, что

$$J^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] \leq J^{\delta^k}[u] + \left\langle \{\lambda^k, \mu^k\}, \{\mathcal{G}^{\delta^k}[u] - \omega^{\delta^k}, \mathcal{E}^{\delta^k}[u]\} \right\rangle_{L_2 \times L_2} + \psi^k, \quad k = 1, 2, \dots$$

Положим здесь $u = u^0$ и используем ограниченность $\mathcal{D}$, а также условие согласования

$\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0, \quad k \rightarrow \infty$. Можем записать, используя оценки (1.6), (1.13),

$$\begin{aligned} J^0[u^{\delta^k}[\lambda^k, \mu^k]] &\leq J^0[u^0] + |J^{\delta^k}[u^0] - J^0[u^0]| + |J^0[u^{\delta^k}[\lambda^k, \mu^k]] - J^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]]| \\ &\quad + \left| \left\langle \lambda^k, \mathcal{G}^{\delta^k}[u^0] - \mathcal{G}^0[u^0] \right\rangle_{2,1} \right| + \left| \left\langle \lambda^k, \omega^0 - \omega^{\delta^k} \right\rangle_{2,1} \right| + \left| \left\langle \mu^k, \mathcal{E}^{\delta^k}[u^0] - \mathcal{E}^0[u^0] \right\rangle_{2,1} \right| + \psi^k \\ &\leq J^0[u^0] + C_{10}\delta^k + C_{11}\delta^k\|\lambda^k\|_{2,1} + C_{12}\delta^k\|\mu^k\|_{2,1} + \psi^k, \quad k = 1, 2, \dots, \end{aligned}$$

где постоянные $C_{10}, C_{11}, C_{12} > 0$ не зависят от k . Получаем: существует стремящаяся к нулю последовательность $\{\tilde{\psi}^k\}_{k=1}^{\infty}$ неотрицательных чисел такая, что

$$J^0[u^{\delta^k}[\lambda^k, \mu^k]] \leq J^0[u^0] + \tilde{\psi}^k, \quad k = 1, 2, \dots$$

Так как одновременно имеем включения $u^{\delta^k}[\lambda^k, \mu^k] \in \mathcal{D}^{\delta^k, \epsilon^k}$, $k = 1, 2, \dots$, а значит, в силу ограниченности $\mathcal{D}$, при некоторой стремящейся к нулю последовательности $\{\tilde{\epsilon}^k\}_{k=1}^{\infty}$ неотрицательных чисел и включения $u^{\delta^k}[\lambda^k, \mu^k] \in D^{0, \tilde{\epsilon}^k}$, $k = 1, 2, \dots$, то последовательность $u^{\delta^k}[\lambda^k, \mu^k]$, $k = 1, 2, \dots$, является МПР в задаче (OC^0) . Далее, принимая во внимание снова предельное соотношение (2.27) и полученное предельное соотношение $J^0[u^{\delta^k}[\lambda^k, \mu^k]] \rightarrow J^0[u^0]$, $k \rightarrow \infty$, можем записать соотношение:

$$\begin{aligned} V^{\delta^k}(\lambda^k, \mu^k) &= J^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] \\ &\quad + \left\langle \{\lambda^k, \mu^k\}, \{\mathcal{G}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]] - \omega^{\delta^k}, \mathcal{E}^{\delta^k}[u^{\delta^k}[\lambda^k, \mu^k]]\} \right\rangle_{L_2 \times L_2} \rightarrow J^0[u^0], \quad k \rightarrow \infty, \end{aligned}$$

а следовательно, и предельное соотношение (2.28) в силу оценки (2.2) и предельного соотношения $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0, \quad k \rightarrow \infty$. $\square$

2.3. Регуляризованные принцип Лагранжа и принцип максимума Понтрягина в выпуклой задаче оптимального управления с поточечными фазовыми ограничениями

2.3.1. Регуляризованный принцип Лагранжа. «Объединив» теоремы 2.4, 2.5, сформулируем, наконец, регуляризованные принцип Лагранжа и принцип максимума Понтрягина в форме критериев существования МПР в выпуклой задаче оптимального управления с поточечными фазовыми ограничениями. Регуляризованный принцип Лагранжа (ввиду регулярности функций Лагранжа его можно называть также регуляризованной теоремой Куна–Таккера) можно записать в следующем виде.

Теорема 2.6. Пусть $\{\delta^k\}_{k=1}^{\infty} \subset (0, \delta_0]$ — фиксированная стремящаяся к нулю последовательность чисел. Вне зависимости от разрешимости или неразрешимости двойственной к (OC^0) задачи справедливо следующее. Для того, чтобы в задаче (OC^0) существовало МПР (в случае существования оно будет слабо сходиться к точному решению задачи), необходимо и достаточно, чтобы существовала последовательность двойственных переменных $\{\lambda^k, \mu^k\} \in L_2 \times L_{2,+}$, $k = 1, 2, \dots$, такая, что $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0, \quad k \rightarrow \infty$, и выполнялись предельные соотношения (2.26), (2.27). При этом последовательность $u^{\delta^k}[\lambda^k, \mu^k]$, $k = 1, 2, \dots$, является искомым МПР, сходящимся сильно к решению u^0 задачи (OC^0) в случае субдифференцируемости J^0 на $\mathcal{D}$. Другими словами, оператор

$R(\cdot, \delta^k)$, ставящий в соответствие каждому набору исходных данных $\mathbf{f}^{\delta^k}$, удовлетворяющих оценкам (1.6), (1.7), (1.8) при $\delta = \delta^k$, элемент $R(\mathbf{f}^{\delta^k}, \delta^k) \equiv u^{\delta^k}[\lambda^{\delta^k, \alpha(\delta^k)}, \mu^{\delta^k, \alpha(\delta^k)}] \in \mathcal{D}$, является МПР-образующим в задаче (OC^0) .

Одновременно с предельным соотношением $\delta^k \|\{\lambda^k, \mu^k\}\|_{L_2 \times L_2} \rightarrow 0$, $k \rightarrow \infty$, и соотношениями (2.26), (2.27) выполняется и предельное соотношение (2.28). В качестве последовательности $\{\lambda^k, \mu^k\}$, $k = 1, 2, \dots$, может быть взята генерируемая методом двойственной регуляризации теоремы 2.3 последовательность $\{\lambda^{\delta^k, \alpha(\delta^k)}, \mu^{\delta^k, \alpha(\delta^k)}\}$, $k = 1, 2, \dots$, сходящаяся к минимальному по норме решению двойственной задачи в случае разрешимости последней.

2.3.2. О минимизации функции Лагранжа. Ключевой задачей процедуры двойственной регуляризации процесса приближенного решения задачи (OC^0) является задача минимизации функционала Лагранжа $L^\delta(u, \lambda, \mu)$, $\{\lambda, \mu\} \in L_2 \times L_{2,+}$, задачи (OC^δ)

$$L^\delta(u, \lambda, \mu) \rightarrow \min, \quad u \in \mathcal{D}, \quad (2.29)$$

решение которой мы обозначили через $u^\delta[\lambda, \mu]$. Для упрощения изложения предположим, что при каждом $\delta \in (0, \delta_0]$ функционалы $K^\delta[z] : L_2^m \rightarrow \mathbf{R}$, $M^\delta[u] : L_2^s \rightarrow \mathbf{R}$ дифференцируемы по Фреше, а функция $F^\delta(t, y) : \Pi \times \mathbf{R}^r \rightarrow \mathbf{R}$ дифференцируема по y на $\mathbf{R}^r$ при каждом $t \in \Pi$. Тогда при каждом $\delta \in (0, \delta_0]$ дифференцируемы по Фреше на L_2^s оператор $\mathcal{E}^\delta[\cdot]$ (рассматриваем его как оператор из L_2^s в L_2) и функционал Лагранжа $L^\delta(\cdot, \lambda, \mu)$. В этом случае решение $u^\delta[\lambda, \mu]$ задачи (2.29) удовлетворяет критерию минимума

$$L_u^{\delta/} (u^\delta[\lambda, \mu], \lambda, \mu) [u - u^\delta[\lambda, \mu]] \geq 0, \quad u \in \mathcal{D}, \quad (2.30)$$

где $L_u^{\delta/}(\bar{u}, \lambda, \mu)[\cdot]$ — производная Фреше функционала $L^\delta(u, \lambda, \mu)$ по переменной u в точке $\bar{u} \in L_2^s$ при фиксированных λ, μ . Пусть $\Psi^\delta[\bar{u}, \lambda, \mu](\cdot) \in L_2^s$ — функция Рисса функционала $L_u^{\delta/}(\bar{u}, \lambda, \mu)[\cdot] \in (L_2^s)^*$. Критерий (2.30) можно записать в виде

$$\langle \Psi^\delta[u^\delta[\lambda, \mu], \lambda, \mu], u - u^\delta[\lambda, \mu] \rangle_{2,s} \geq 0, \quad u \in \mathcal{D}. \quad (2.31)$$

Найдем представление функции $\Psi^\delta[\bar{u}, \lambda, \mu](t)$, $t \in \Pi$, в терминах задачи (OC^δ) , $\delta > 0$.

Непосредственно вычисляя, получаем

$$L_u^{\delta/}(\bar{u}, \lambda, \mu)[v] = \langle \phi^\delta[\bar{u}], v \rangle_{2,s} + \langle \psi^\delta[\bar{u}, \lambda, \mu], B^\delta[v] \rangle_{2,m}, \quad v \in L_2^s, \quad \bar{u} \in L_2^s, \quad (2.32)$$

здесь $\phi^\delta[\bar{u}](\cdot) \in L_2^s$ — функция Рисса функционала $M_u^{\delta/}(\bar{u}) \in (L_2^s)^*$, а $\psi^\delta[\bar{u}, \lambda, \mu](\cdot) \in L_2^m$ — функция, которая задается формулой

$$\psi^\delta[\bar{u}, \lambda, \mu](\cdot) = (S^\delta)^* [\phi_1^\delta[\bar{u}] + \phi_2^\delta[\bar{u}, \lambda] + \phi_3^\delta[\bar{u}, \mu]](\cdot), \quad (2.33)$$

где $\phi_1^\delta[\bar{u}](\cdot)$, $\phi_2^\delta[\bar{u}, \lambda](\cdot)$ и $\phi_3^\delta[\bar{u}, \mu](\cdot)$ — функции Рисса принадлежащих классу $(L_2^m)^*$ функционалов $K_z^{\delta/}(z_u^\delta)$, $(Y^\delta)^*[\lambda \Xi^\delta]$ и $(\Sigma^\delta)^*[\mu(\cdot)(F_y^{\delta/}(\cdot, \Sigma^\delta[z_u^\delta])(\cdot))]$ соответственно.

Так как $(S^\delta)^* \equiv ((E - A^\delta)^{-1})^* = ((E - A^\delta)^*)^{-1} = (E - (A^\delta)^*)^{-1}$, где E — единичный оператор в L_2^m , то определяемая формулой (2.33) и входящая в (2.32) функция $\psi^\delta[\bar{u}, \lambda, \mu]$ есть (единственное в L_2^m) решение уравнения

$$\psi(t) - (A^\delta)^*[\psi](t) = \phi_1^\delta[\bar{u}](t) + \phi_2^\delta[\bar{u}, \lambda](t) + \phi_3^\delta[\bar{u}, \mu](t), \quad t \in \Pi, \quad \psi \in L_2^m. \quad (2.34)$$

Перепишав (2.32) в виде $L_u^{\delta/}(\bar{u}, \lambda, \mu)[v] = \langle (B^\delta)^* [\psi^\delta[\bar{u}, \lambda, \mu]] + \phi^\delta[\bar{u}], v \rangle_{2,s}$, $v \in L_2^s$, найдем искомое представление функции $\Psi^\delta[\bar{u}, \lambda, \mu]$ в терминах задачи (OC^δ) :

$$\Psi^\delta[\bar{u}, \lambda, \mu](t) = (B^\delta)^* [\psi^\delta[\bar{u}, \lambda, \mu]](t) + \phi^\delta[\bar{u}](t), \quad t \in \Pi, \quad (2.35)$$

в котором $\psi^\delta[\bar{u}, \lambda, \mu]$ — единственное в L_2^m решение уравнения (2.34).

2.3.3. Случай ограниченных управлений. Рассмотрим задачу (OC^0) в ситуации, когда допустимые управления принимают значения из некоторого ограниченного замкнутого и выпуклого множества $U \subset \mathbf{R}^s$, то есть $\mathcal{D} \equiv \{u(\cdot) \in L_\infty^s : u(t) \in U, t \in \Pi\}$. В этом случае получаем из (2.31) критерий минимума функционала Лагранжа в виде следующего линейризованного поточечного принципа максимума.

Лемма 2.8. *Функция $\bar{u}(\cdot) \in \mathcal{D}$ есть решение задачи (2.29) тогда и только тогда, когда*

$$\langle \Psi^\delta[\bar{u}, \lambda, \mu](t), \bar{u}(t) \rangle_s = \max_{w \in U} \langle \Psi^\delta[\bar{u}, \lambda, \mu](t), w \rangle_s \quad \text{при н.в. } t \in \Pi, \quad (2.36)$$

где $\Psi^\delta[\bar{u}, \lambda, \mu]$ задается формулой (2.35), в которой $\psi^\delta[\bar{u}, \lambda, \mu]$ — решение (2.34).

Д о к а з а т е л ь с т в о. Необходимость для решения $\bar{u}$ задачи (2.29) условия (2.36) доказывается простейшим игольчатым варьированием, а достаточность получается стандартным применением теоремы А. А. Ляпунова (см., например, [14, § 2.4, § 8.2]). $\square$

Обозначим через $U_m^\delta[\lambda, \mu]$ множество всех управлений из $\mathcal{D}$, удовлетворяющих (при сформулированных выше дополнительных условиях дифференцируемости) принципу максимума леммы 2.8. В нашем случае, благодаря сильной выпуклости целевого функционала, множество $U_m^\delta[\lambda, \mu]$ состоит из одного элемента, обозначим его через $u_m^\delta[\lambda, \mu]$. Очевидно, что $u_m^\delta[\lambda, \mu] = u^\delta[\lambda, \mu]$. То есть непосредственно из теоремы 2.6 и леммы 2.8 получаем следующий регуляризованный поточечный принцип максимума для задачи (OC^0) .

Теорема 2.7. *При сформулированных выше дополнительных условиях дифференцируемости все утверждения теоремы 2.6 останутся справедливыми, если в них $u^{\delta^k}[\lambda^k, \mu^k]$ заменить везде на $u_m^{\delta^k}[\lambda^k, \mu^k]$.*

3. Пример. Оптимизационная задача с поточечными фазовыми ограничениями для управляемой системы с запаздыванием

Естественный переход от начально-краевой задачи к эквивалентному ей функциональному уравнению II рода вольтеррова типа осуществляется с помощью обращения главной части задачи. Разнообразные конкретные примеры начально-краевых задач (для параболических, гиперболических, интегродифференциальных уравнений с частными производными и систем таких уравнений, различных уравнений с запаздывающим аргументом и др.), которые допускают эквивалентное описание с помощью функциональных уравнений вольтеррова типа, можно найти, например, в [7] (см. также обзор в [11]). Из множества самых различных подобных начально-краевых задач мы для иллюстрации изложенной выше теории выбрали начальную задачу для системы с запаздыванием. В примере выписываются те основные конструкции, которые и участвуют в формулировке регуляризованных КУО (формирующая критерий минимума функционала Лагранжа функция, сопряженное уравнение, ...). Сформулировать с их помощью соответствующие регуляризованные КУО — конкретные реализации сформулированных выше теорем уже не сложно.

Пусть $n = 1$; $\Pi = [0, 1]$; $\rho \in (0, 1)$ — фиксированное число; $\eta \in \mathbf{R}^m$ — фиксированный вектор; $a(\cdot), b(\cdot) \in L_2^{m \times m}$, $d(\cdot) \in L_\infty^{m \times s}$, $\xi(\cdot) \in C^m[-\rho, 0]$ — фиксированные функции. Рассмотрим начальную задачу для линейной управляемой системы дифференциальных уравнений с запаздывающим аргументом ($x(\cdot)$ — m -вектор-функция)

$$\dot{x} = a(t)x(t) + b(t)x(t - \rho) + d(t)u(t), \quad t \in [0, 1]; \quad (3.1)$$

$$x(t) = \xi(t), \quad t \in [-\rho, 0]; \quad x(0) = \eta, \quad (3.2)$$

где $u(\cdot) \in L_2^s$ — управление. Решение начальной задачи (3.1), (3.2) понимаем как решение в смысле «почти всюду» из пространства $(W_2^1[0, 1])^m$ абсолютно непрерывных на отрезке $[0, 1]$ функций с квадратично суммируемыми первыми производными, рассматривая первое из условий (3.2) как требуемое в (3.1) условие доопределения $x(t)$ слева от $t = 0$: $x(t - \rho) = \xi(t - \rho)$ при $t - \rho < 0$. Приведем задачу (3.1), (3.2) к эквивалентному уравнению вида (1.1), показав тем самым, что каждому $u(\cdot) \in L_2^s$ отвечает единственное в классе W функций $x(\cdot) \in (W_2^1[0, 1])^m$, удовлетворяющих второму условию (3.2), решение этой задачи. Для этого сделаем в (3.1), (3.2) замену по формуле

$$x(t) = \eta + \int_0^t z(\zeta) d\zeta, \quad t \in [0, 1], \quad (3.3)$$

устанавливающей взаимно однозначное соответствие между классом W функций $x(\cdot)$ и пространством L_2^m функций $z(\cdot)$ (такое преобразование задачи (3.1), (3.2) естественно назвать обращением главной части этой задачи). «Подставляя» (3.3) в (3.1) (с учетом при $t \in [0, \rho)$ первого условия (3.2)), получаем

$$z(t) = a(t)\eta + a(t) \int_0^t z(\zeta) d\zeta + b(t)\eta + b(t) \int_0^{t-\rho} z(\zeta) d\zeta + d(t)u(t), \quad t \in [\rho, 1], \quad (3.4)$$

$$z(t) = a(t)\eta + a(t) \int_0^t z(\zeta) d\zeta + b(t)\xi(t - \rho) + d(t)u(t), \quad t \in [0, \rho]. \quad (3.5)$$

Положим $\chi(t) \equiv \{\xi(t - \rho), t \in [0, \rho]; \eta, t \in [\rho, 1]\}$, $t \in [0, 1]$; $\Lambda_1[z](t) \equiv \int_0^t z(\zeta) d\zeta$, $\Lambda_2[z](t) \equiv \{0_m, t \in [0, \rho]; \int_0^{t-\rho} z(\zeta) d\zeta, t \in (\rho, 1]\}$, $z(\cdot) \in L_2^m$. Запишем (3.4), (3.5) в виде

$$\begin{aligned} z(t) &= a(t) \{\eta + \Lambda_1[z](t)\} + b(t) \{\chi(t) + \Lambda_2[z](t)\} + d(t)u(t) \\ &\equiv \{a(t)\Lambda_1[z](t) + b(t)\Lambda_2[z](t)\} + d(t)u(t) + \{a(t)\eta + b(t)\chi(t)\}, \quad t \in \Pi. \end{aligned} \quad (3.6)$$

Уравнение (3.6) и есть уравнение вида (1.1), эквивалентное начальной задаче (3.1), (3.2). Здесь $\Pi \equiv [0, 1]$; $A[z](t) \equiv a(t)\Lambda_1[z](t) + b(t)\Lambda_2[z](t)$, $z(\cdot) \in L_2^m$, $t \in \Pi$ (квазинильпотентность оператора $A[\cdot] : L_2^m \rightarrow L_2^m$ легко проверяется, например, с помощью цепочечного признака [8, теорема 2], см. также [11, с. 269]); $B[u](t) \equiv d(t)u(t)$, $u(\cdot) \in L_2^s$, $t \in \Pi$; $c(t) \equiv a(t)\eta + b(t)\chi(t)$, $t \in \Pi$. Если $x(\cdot) \in W$ — решение задачи (3.1), (3.2) при некотором $u(\cdot) \in L_2^s$, то связанная с $x(\cdot)$ формулой (3.3) функция $z(\cdot) \in L_2^m$ есть решение уравнения (3.6) при том же $u(\cdot)$. И наоборот, если $z(\cdot) \in L_2^m$ — решение уравнения (3.6) при данном $u(\cdot) \in L_2^s$, то функция $x(\cdot)$, связанная с $z(\cdot)$ формулой (3.3), есть решение класса W задачи (3.1), (3.2) при этом $u(\cdot)$. Отвечающие управлению $u(\cdot) \in L_2^s$ решения задачи (3.1)–(3.2) и уравнения (3.6) обозначим через x_u и z_u соответственно.

Пусть задано следующее: функции $Q(\cdot) \in C^m$, $\gamma(\cdot) \in C$; непрерывная выпуклая функция $\mathcal{F}(\cdot) : \mathbf{R}^m \rightarrow \mathbf{R}$; выпуклое ограниченное и замкнутое множество $\mathcal{D}$ пространства L_2^s . Рассмотрим задачу оптимального управления системой (3.1), (3.2) с минимизируемым целевым функционалом $\mathcal{J}[u] \equiv \|u\|_{2,s}^2$, $u \in L_2^s$, при ограничениях

$$\langle Q(t), x(t) \rangle_m = \gamma(t) \quad (t \in \Pi), \quad \mathcal{F}(x(t)) \leq 0 \quad (t \in \Pi), \quad x(\cdot) \in (W_2^1[0, 1])^m, \quad (3.7)$$

и множестве допустимых управлений $\mathcal{D}$. Эту задачу символически запишем в виде

$$\mathcal{J}[u] \rightarrow \min, \quad (3.1), \quad (3.2), \quad (3.7), \quad u \in \mathcal{D}. \quad (3.8)$$

Сделав в задаче (3.8) замену (3.3), получим следующую эквивалентную задачу оптимизации управляемой системы (3.6):

$$\mathcal{J}[u] \rightarrow \min, \quad \mathcal{P}[z_u](t) = \gamma(t) - \langle Q(t), \eta \rangle_m \quad (t \in \Pi), \quad \mathcal{F}(\eta + \Lambda_1[z_u](t)) \leq 0 \quad (t \in \Pi), \quad u \in \mathcal{D},$$

где $\mathcal{P}[z](t) \equiv \langle Q(t), \Lambda_1[z](t) \rangle_m$, $z \in L_2^m$ ($t \in \Pi$). Это задача (1.4) с $n = 1$; $\Pi = [0, 1]$; $J[u] = \mathcal{J}[u]$ ($K[z] \equiv 0$, $M[u] = \mathcal{J}[u]$); $\mathcal{G}[u] = \mathcal{P}[z_u]$ ($\Xi(t) = Q(t)$, $\Upsilon[z](t) = \Lambda_1[z](t)$, $q = m$); $\omega(t) = \gamma(t) - \langle Q(t), \eta \rangle_m$; $\mathcal{E}[u](t) = \mathcal{F}(\eta + \Lambda_1[z_u](t))$ ($F(t, y) = \mathcal{F}(\eta + y)$, $\Sigma[z](t) = \Lambda_1[z](t)$, $r = m$).

Пусть $\mathbf{f} \equiv \{\eta, a, b, d, \xi; Q, \gamma; \mathcal{F}\}$ — набор данных задачи (3.8), которые подвергаются возмущению, и точный набор $\mathbf{f}^0 \equiv \{\eta^0, a^0, b^0, d^0, \xi^0; Q^0, \gamma^0; \mathcal{F}^0\}$ нам не известен, но можно оперировать с приближенными наборами $\mathbf{f}^\delta \equiv \{\eta^\delta, a^\delta, b^\delta, d^\delta, \xi^\delta; Q^\delta, \gamma^\delta; \mathcal{F}^\delta\}$, $\delta \in (0, \delta_0]$ ($\delta_0 > 0$ фиксировано), которые связаны с набором $\mathbf{f}^0$ следующими условиями 3.1–3.3.

У с л о в и е 3.1. Функция $\mathcal{F}^\delta(\cdot) : \mathbf{R}^m \rightarrow \mathbf{R}$ липшицева на каждом ограниченном множестве, причем липшицевость равномерна по параметру $\delta \in [0, \delta_0]$, то есть существует неубывающая функция $\mathcal{L}_1(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ такая, что для любых $l > 0$ и $\delta \in [0, \delta_0]$ выполнено $|\mathcal{F}^\delta(y_1) - \mathcal{F}^\delta(y_2)| \leq \mathcal{L}_1(l) \|y_1 - y_2\|_m$ при $\|y_1\|_m, \|y_2\|_m \leq l$.

У с л о в и е 3.2. Существует постоянная $\mathcal{C} > 0$ такая, что при любом $\delta \in (0, \delta_0]$ величины $\|\eta^\delta - \eta^0\|_m$, $\|a^\delta - a^0\|_{2,m \times m}$, $\|b^\delta - b^0\|_{2,m \times m}$, $\|d^\delta - d^0\|_{\infty, m \times s}$, $\|\xi^\delta - \xi^0\|_{C^m[-\rho, 0]}$, $\|\gamma^\delta - \gamma^0\|_C$, $\|Q^\delta - Q^0\|_{C, m}$ не превосходят величины $\mathcal{C}\delta$.

У с л о в и е 3.3. Существует неубывающая функция $\mathcal{L}_2(\cdot) : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ такая, что для каждого $l > 0$ и любого $\delta \in [0, \delta_0]$ имеем $|\mathcal{F}^\delta(y) - \mathcal{F}^0(y)| \leq \mathcal{L}_2(l) \delta$ при $\|y\|_m \leq l$.

При любом $\delta \in [0, \delta_0]$ мы имеем управляемую начальную задачу

$$\dot{x} = a^\delta(t)x(t) + b^\delta(t)x(t - \rho) + d^\delta(t)u(t), \quad t \in [0, 1]; \quad (3.9)$$

$$x(t) = \xi^\delta(t), \quad t \in [-\rho, 0]; \quad x(0) = \eta^\delta, \quad (3.10)$$

набор ограничений

$$\langle Q^\delta(t), x(t) \rangle_m = \gamma^\delta(t) \quad (t \in \Pi), \quad \mathcal{F}^\delta(x(t)) \leq 0 \quad (t \in \Pi), \quad x(\cdot) \in (W_2^1[0, 1])^m, \quad (3.11)$$

и задачу оптимизации системы (3.9)–(3.10) с минимизируемым функционалом цели $\mathcal{J}[u]$ при ограничениях (3.11) и множестве допустимых управлений $\mathcal{D}$. Эту задачу оптимального управления символически запишем в виде

$$\mathcal{J}[u] \rightarrow \min, \quad (3.9), \quad (3.10), \quad (3.11), \quad u \in \mathcal{D}. \quad (3.12)$$

Пусть $x_u^\delta(\cdot)$ — отвечающее управлению $u(\cdot)$ решение начальной задачи (3.9), (3.10).

Сделав в задаче (3.12) замену $x(t) = \eta^\delta + \int_0^t z(\zeta)d\zeta$, $t \in [0, 1]$, соответствующую обращению главной части начальной задачи (3.9), (3.10), получим эквивалентную задачу оптимизации управляемой системы

$$z(t) = \{a^\delta(t)\Lambda_1[z](t) + b^\delta(t)\Lambda_2[z](t)\} + d^\delta(t)u(t) + \{a^\delta(t)\eta^\delta + b^\delta(t)\chi^\delta(t)\}, \quad t \in \Pi, \quad (3.13)$$

где $\chi^\delta(t) \equiv \{\xi^\delta(t - \rho), t \in [0, \rho]; \eta^\delta, t \in [\rho, 1]\}$, $t \in \Pi$. Эту задачу запишем в виде

$$\mathcal{J}[u] \rightarrow \min, \\ \mathcal{P}^\delta[z_u^\delta](t) = \gamma^\delta(t) - \langle Q^\delta(t), \eta^\delta \rangle_m \quad (t \in \Pi), \quad \mathcal{F}^\delta(\eta^\delta + \Lambda_1[z_u^\delta](t)) \leq 0 \quad (t \in \Pi), \quad u \in \mathcal{D}, \quad (3.14)$$

где $\mathcal{P}^\delta[z](t) \equiv \langle Q^\delta(t), \Lambda_1[z](t) \rangle_m$, $z \in L_2^m$ ($t \in \Pi$), z_u^δ — отвечающее u решение (3.13).

Положив $A^\delta[z](t) \equiv a^\delta(t)\Lambda_1[z](t) + b^\delta(t)\Lambda_2[z](t)$, $t \in \Pi$, $z \in L_2^m$; $B^\delta[u](t) \equiv d^\delta(t)u(t)$, $t \in \Pi$, $u \in L_2^s$; $c^\delta(t) \equiv a^\delta(t)\eta^\delta + b^\delta(t)\chi^\delta(t)$, $t \in \Pi$, переписываем (3.13) в форме (1.9). Задача (3.14) имеет вид (OC^δ) , здесь: $\mathcal{J}^\delta[u] = \mathcal{J}[u]$ ($K^\delta[z] \equiv 0$, $M^\delta[u] = \mathcal{J}[u]$); $\mathcal{G}^\delta[u] = \mathcal{P}^\delta[z_u^\delta] \equiv \langle Q^\delta(t), \Lambda_1[z_u^\delta](t) \rangle_m$ ($\Xi^\delta(t) = Q^\delta(t)$, $\Upsilon^\delta[z](t) = \Lambda_1[z](t)$, $q = m$); $\omega^\delta(t) = \gamma^\delta(t) - \langle Q^\delta(t), \eta^\delta \rangle_m$; $\mathcal{E}^\delta[u](t) = \mathcal{F}^\delta(\eta^\delta + \Lambda_1[z_u^\delta](t))$ ($F^\delta(t, y) = \mathcal{F}^\delta(\eta^\delta + y)$, $\Sigma^\delta[z](t) = \Lambda_1[z](t)$, $r = m$).

При сделанных относительно семейства задач (3.12), $\delta \in [0, \delta_0]$, предположениях семейство задач (3.14), $\delta \in [0, \delta_0]$, удовлетворяет условиям А–Д. Действительно, выполнение условий А, Б и В проверяется элементарными выкладками, исходя из предположений в условиях 3.1, 3.2, условия 3.2 и условиях 3.2, 3.3 соответственно. Выполнение условия Г следует из цепочечного признака равностепенной квазинильпотентности [11, теорема 2].

Для проверки выполнения условия Д заметим, что для каждого $u \in L_2^s$ при $t \in \Pi$, используя обозначение $\iota(t, \zeta) \equiv \{1, 0 \leq \zeta \leq t; 0, t < \zeta \leq 1\}$, $0 \leq t \leq 1$, $0 \leq \zeta \leq 1$, получаем

$$\Upsilon^\delta[z_u^\delta](t) = \Sigma^\delta[z_u^\delta](t) = \Lambda_1[z_u^\delta](t) = \int_0^1 \iota(t, \zeta) z_u^\delta(\zeta) d\zeta = \int_0^1 \iota(t, \zeta) S^\delta[B^\delta[u] + c^\delta](\zeta) d\zeta.$$

Рассмотрим сначала случай $m = s = 1$. Переходя к сопряженным операторам, получаем

$$\Upsilon^\delta[z_u^\delta](t) = \Sigma^\delta[z_u^\delta](t) = \int_0^1 B^{\delta*} S^{\delta*}[\iota(t, \cdot)](\zeta) u(\zeta) d\zeta + \int_0^1 S^{\delta*}[\iota(t, \cdot)](\zeta) c^\delta(\zeta) d\zeta, \quad t \in \Pi, \quad u \in L_2,$$

что, очевидно, означает выполнение условия Д. Проверка выполнения Д в общем случае аналогична — можно воспользоваться матричным представлением S и B .

Предположив, что функции $\mathcal{F}^\delta(\cdot)$, $\delta \in (0, \delta_0]$, дифференцируемы, можем выписать для нашего примера критерии (2.31) и (2.36) решения задачи (2.29). Прямые вычисления дают: $\phi_1^\delta[\bar{u}](t) \equiv 0$, $\phi_2^\delta[\bar{u}, \lambda](t) \equiv \int_t^1 \lambda(\xi) Q^\delta(\xi) d\xi$, $\phi_3^\delta[\bar{u}, \mu](t) \equiv \int_t^1 \mu(\xi) \left(\mathcal{F}_y^{\delta'}(x_u^\delta(\xi)) \right)^* d\xi$, $t \in \Pi$. По определению сопряженного оператора

$$(A^\delta)^*[\psi](t) \equiv \Lambda_1^*[(a^\delta)^* \psi](t) + \Lambda_2^*[(b^\delta)^* \psi](t), \quad t \in \Pi, \quad \psi \in L_2^m; \\ \Lambda_1^*[y](t) \equiv \int_t^1 y(\zeta) d\zeta, \\ \Lambda_2^*[y](t) \equiv \left\{ \int_{t+\rho}^1 y(\zeta) d\zeta, 0 \leq t \leq 1 - \rho; 0_m, 1 - \rho < t \leq 1 \right\}, \quad t \in \Pi, \quad y \in L_2^m.$$

То есть уравнение (2.34) можно записать в виде

$$\begin{aligned} \psi(t) - \int_t^1 (a^\delta(\zeta))^* \psi(\zeta) d\zeta &= h^\delta[\bar{u}, \lambda, \mu](t), \quad 1 - \rho \leq t \leq 1; \\ \psi(t) - \int_t^{1-\rho} (a^\delta(\zeta))^* \psi(\zeta) d\zeta - \int_{t+\rho}^1 (b^\delta(\zeta))^* \psi(\zeta) d\zeta &= h^\delta[\bar{u}, \lambda, \mu](t), \quad 0 \leq t \leq 1 - \rho, \end{aligned} \quad (3.15)$$

где $h^\delta[\bar{u}, \lambda, \mu](t) \equiv \int_t^1 \{ \lambda(\xi) Q^\delta(\xi) + \mu(\xi) (\mathcal{F}_y^{\delta/}(x_{\bar{u}}^\delta(\xi)))^* \} d\xi$, $t \in \Pi$. Формирующая критерии (2.31) и (2.36), которым удовлетворяет решение $u^\delta[\lambda, \mu]$ задачи (2.29), функция $\Psi^\delta[\bar{u}, \lambda, \mu]$ задается формулой

$$\Psi^\delta[\bar{u}, \lambda, \mu](t) \equiv (d^\delta(t))^* \psi^\delta[\bar{u}, \lambda, \mu](t) + 2\bar{u}(t), \quad t \in \Pi,$$

где $\psi^\delta[\bar{u}, \lambda, \mu]$ — решение сопряженного уравнения (3.15). Это уравнение вольтеррова типа. Единственное в L_2^m решение $\psi^\delta[\bar{u}, \lambda, \mu]$ уравнения (3.15) абсолютно непрерывно на Π и принадлежит классу $(W_2^1(\Pi))^m$. Уравнение (3.15) эквивалентно системе уравнений

$$\dot{\psi} + (a^\delta(t))^* \psi(t) = -\lambda(t) Q^\delta(t) - \mu(t) (\mathcal{F}_y^{\delta/}(x_{\bar{u}}^\delta(t)))^*, \quad 1 - \rho \leq t \leq 1; \quad \psi(1) = 0; \quad (3.16)$$

$$\begin{aligned} \dot{\psi} + (a^\delta(t))^* \psi(t) + (b^\delta(t + \rho))^* \psi(t + \rho) \\ = -\lambda(t) Q^\delta(t) - \mu(t) (\mathcal{F}_y^{\delta/}(x_{\bar{u}}^\delta(t)))^*, \quad 0 \leq t \leq 1 - \rho; \end{aligned} \quad (3.17)$$

$$\psi(1 - \rho) = \int_{1-\rho}^1 \left\{ (a^\delta(\zeta))^* \psi(\zeta) + \lambda(\zeta) Q^\delta(\zeta) + \mu(\zeta) (\mathcal{F}_y^{\delta/}(x_{\bar{u}}^\delta(\zeta)))^* \right\} d\zeta, \quad (3.18)$$

состоящей из задачи Коши (3.16) для обыкновенного дифференциального уравнения, рассматриваемого на отрезке $1 - \rho \leq t \leq 1$, с условием Коши в точке $t = 1$, и начальной задачи (3.17), (3.18) для рассматриваемого на отрезке $0 \leq t \leq 1 - \rho$ дифференциального уравнения с опережением (3.17), в которой условие (3.18) играет роль условия Коши в точке $t = 1 - \rho$; требуемое в (3.17) и (3.18) доопределение функции ψ справа от точки $t = 1 - \rho$ обеспечивается задачей Коши (3.16).

References

- [1] В. И. Сумин, М. И. Сумин, “Регуляризованные классические условия оптимальности в итерационной форме для выпуклых задач оптимизации распределенных систем вольтеррова типа”, *Вестник Удмуртского университета. Математика. Механика. Компьютерные науки*, **31**:2 (2021), 265–284. [V. I. Sumin, M. I. Sumin, “Regularized classical optimality conditions in iterative form for convex optimization problems for distributed Volterra-type systems”, *Vestnik Udmurtskogo universiteta. Matematika. Mekhanika. Komp'yuternye nauki*, **31**:2 (2021), 265–284 (In Russian)].
- [2] В. И. Сумин, М. И. Сумин, “Регуляризация классических условий оптимальности в задачах оптимального управления линейными распределенными системами вольтеррова типа”, *Журн. вычисл. матем. и матем. физ.*, **62**:1 (2022), 45–70; англ. пер.: V. I. Sumin, M. I. Sumin, “Regularization of the classical optimality conditions in optimal control problems for linear distributed systems of Volterra type”, *Comput. Math. Math. Phys.*, **62**:1 (2022), 42–65.
- [3] В. И. Сумин, М. И. Сумин, “О регуляризации принципа Лагранжа в задачах оптимизации линейных распределенных систем вольтеррова типа с операторными ограничениями”, *Известия Института математики и информатики Удмуртского государственного университета*, **59** (2022), 85–113. [V. I. Sumin, M. I. Sumin, “On regularization of the Lagrange principle in the optimization problems for linear distributed Volterra type systems with operator constraints”, *Izvestiya Instituta Matematiki i Informatiki Udmurtskogo Gosudarstvennogo Universiteta*, **59** (2022), 85–113 (In Russian)].

- [4] В. И. Сумин, М. И. Сумин Об итеративной регуляризации принципа Лагранжа в выпуклых задачах оптимального управления распределенными системами вольтеррова типа с операторными ограничениями, *Дифференциальные уравнения*, **58**:6 (2022), 795–812; англ. пер.: V. I. Sumin, M. I. Sumin, “On the iterative regularization of the Lagrange principle in convex optimal control problems for distributed systems of the Volterra type with operator constraints”, *Differ. Equ.*, **58**:6 (2022), 791–809.
- [5] В. И. Сумин, М. И. Сумин, “Регуляризация классических условий оптимальности в задачах оптимизации линейных систем вольтеррова типа с функциональными ограничениями”, *Вестник российских университетов. Математика*, **28**:143 (2023), 298–325. [V. I. Sumin, M. I. Sumin, “Regularization of classical optimality conditions in optimization problems for linear Volterra-type systems with functional constraints”, *Vestnik Rossiyskikh Universitetov. Matematika = Russian Universities Reports. Mathematics*, **28**:143 (2023), 298–325 (In Russian)].
- [6] В. И. Сумин, М. И. Сумин О регуляризации классических условий оптимальности в выпуклых задачах оптимизации систем вольтеррова типа с операторными ограничениями, *Дифференциальные уравнения*, **60**:2 (2024), 237–259; англ. пер.: V. I. Sumin, M. I. Sumin, “On regularization of classical optimality conditions in convex optimization problems for Volterra-type systems with operator constraints”, *Differ. Equ.*, **60**:2 (2024), 227–246.
- [7] В. И. Сумин, *Функциональные вольтерровы уравнения в теории оптимального управления распределенными системами*, Изд-во Нижегородского государственного университета, Нижний Новгород, 1992. [V. I. Sumin, *Functional Volterra Equations in the Theory of Optimal Control of Distributed Systems*, Publishing House of Nizhnii Novgorod State University, Nizhnii Novgorod, 1992 (In Russian)].
- [8] В. И. Сумин, А. В. Чернов, “Операторы в пространствах измеримых функций: вольтеровость и квазинильпотентность”, *Дифференциальные уравнения*, **34**:10 (1998), 1402–1411; англ. пер.: V. I. Sumin, A. V. Chernov, “Operators in the spaces of measurable functions: the Volterra property and quasinilpotency”, *Differ. Equ.*, **34**:10 (1998), 1403–1411.
- [9] И. Ц. Гохберг, М. Г. Крейн, *Теория вольтерровых операторов в гильбертовом пространстве и ее приложения*, Наука, М., 1967; англ. пер.: I. Ts. Gokhberg, M. G. Krein, *Theory and Applications of Volterra Operators in Hilbert Space*, American Mathematical Society, Providence, 1970.
- [10] В. И. Сумин, “Функционально-операторные вольтерровы уравнения в теории оптимального управления распределенными системами”, *Докл. АН СССР*, **305**:5 (1989), 1056–1059; англ. пер.: V. I. Sumin, “Volterra functional-operator equations in the theory of the optimal control of distributed systems”, *Sov. Math., Dokl.*, **39**:2 (1989), 374–378.
- [11] В. И. Сумин, “Управляемые вольтерровы функциональные уравнения и принцип сжимающих отображений”, *Тр. Ин-та математики и механики УрО РАН*, **25**:1 (2019), 262–278. [V. I. Sumin, “Controlled Volterra functional equations and the contraction mapping principle”, *Trudy Inst. Mat. Mekh. UrO RAN*, **25**:1 (2019), 262–278 (In Russian)].
- [12] В. И. Сумин, “Управляемые вольтерровы функциональные уравнения в лебеговых пространствах”, *Вестн. Нижегород. ун-та. Математическое моделирование и оптимальное управление*, 1998, № 2(19), 138–151. [V. I. Sumin, “Controlled Volterra functional equations in Lebesgue spaces”, *Vestn. Nizhegorod. un-ta. Matematicheskoe modelirovanie i optimal'noe upravlenie*, 1998, № 2(19), 138–151 (In Russian)].
- [13] Р. В. Гамкредидзе, “Оптимальные процессы управления при ограниченных фазовых координатах”, *Изв. АН СССР. Сер. матем.*, **24**:3 (1960), 315–356. [R. V. Gamkrelidze, “Optimal control processes for bounded phase coordinates”, *Izv. Akad. Nauk SSSR Ser. Mat.*, **24**:3 (1960), 316–356 (In Russian)].
- [14] А. Д. Иоффе, В. М. Тихомиров, *Теория экстремальных задач*, Наука, М., 1974; англ. пер.: A. D. Ioffe, V. M. Tikhomirov, *Theory of Extremal Problems*, North-Holland Publishing Company, Amsterdam–New York–Oxford, 1979.
- [15] А. В. Арутюнов, *Условия экстремума. Анормальные и вырожденные задачи*, Факториал, М., 1997; англ. пер.: A. V. Arutyunov, *Optimality conditions: Abnormal and Degenerate Problems*, Kluwer Academic Publishers, Dordrecht–Boston–London, 2000.
- [16] А. А. Милютин, А. В. Дмитрук, Н. П. Осмоловский, *Принцип максимума в оптимальном управлении*, Центр прикладных исследований мехмата МГУ, М., 2004. [A. A. Milyutin, A. V. Dmitruk, N. P. Osmolovsky, *The Maximum Principle in Optimal Control*, Center for Applied Research of the Faculty of Mechanics and Mathematics of Moscow State University, Moscow, 2004 (In Russian)].

- [17] М. И. Сумин, “Регуляризация в линейно выпуклой задаче математического программирования на основе теории двойственности”, *Журн. вычисл. матем. и матем. физ.*, **47**:4 (2007), 602–625; англ. пер.: M. I. Sumin, “Duality-based regularization in a linear convex mathematical programming problem”, *Comput. Math. Math. Phys.*, **47**:4 (2007), 579–600.
- [18] М. И. Сумин, “Параметрическая двойственная регуляризация для задачи оптимального управления с поточечными фазовыми ограничениями”, *Журн. вычисл. матем. и матем. физ.*, **49**:12 (2009), 2083–2102; англ. пер.: M. I. Sumin, “Parametric dual regularization for an optimal control problem with pointwise state constraints”, *Comput. Math. Math. Phys.*, **49**:12 (2009), 1987–2005.
- [19] М. И. Сумин, “Регуляризованная параметрическая теорема Куна–Таккера в гильбертовом пространстве”, *Журн. вычисл. матем. и матем. физ.*, **51**:9 (2011), 1594–1615; англ. пер.: M. I. Sumin, “Regularized parametric Kuhn–Tucker theorem in a Hilbert space”, *Comput. Math. Math. Phys.*, **51**:9 (2011), 1489–1509.
- [20] М. И. Сумин, “Регуляризованные принцип Лагранжа и принцип максимума Понтрягина в оптимальном управлении и обратных задачах”, *Тр. Ин-та математики и механики УрО РАН*, **25**:1 (2019), 279–296. [M. I. Sumin, “Regularized Lagrange principle and Pontryagin maximum principle in optimal control and in inverse problems”, *Trudy Inst. Mat. Mekh. UrO RAN*, **25**:1 (2019), 279–296 (In Russian)].
- [21] М. И. Сумин, “Принцип Лагранжа и его регуляризация как теоретическая основа устойчивого решения задач оптимального управления и обратных задач”, *Вестник российских университетов. Математика*, **26**:134 (2021), 151–171. [M. I. Sumin, “Lagrange principle and its regularization as a theoretical basis of stable solving optimal control and inverse problems”, *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **26**:134 (2021), 151–171 (In Russian)].
- [22] М. И. Сумин, “О некорректных задачах, экстремальных функционала Тихонова и регуляризованных принципах Лагранжа”, *Вестник российских университетов. Математика*, **27**:137 (2022), 58–79. [M. I. Sumin, “On ill-posed problems, extremals of the Tikhonov functional and the regularized Lagrange principles”, *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **27**:137 (2022), 58–79 (In Russian)].
- [23] Дж. Варга, *Оптимальное управление дифференциальными и функциональными уравнениями*, Наука, М., 1977; англ. ориг.: J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [24] Е. Г. Гольштейн, *Теория двойственности в математическом программировании и ее приложения*, Наука, М., 1971. [E. G. Golshtein, *Duality Theory in Mathematical Programming and Its Applications*, Nauka Publ., Moscow, 1971 (In Russian)].
- [25] М. И. Сумин, “О регуляризации классических условий оптимальности в выпуклых задачах оптимального управления”, *Тр. Ин-та математики и механики УрО РАН*, **26**:2 (2020), 252–269. [M. I. Sumin, “On regularization of the classical optimality conditions in convex optimal control problems”, *Trudy Inst. Mat. Mekh. UrO RAN*, **26**:2 (2020), 252–269 (In Russian)].
- [26] M. I. Sumin, “Regularization of Pontryagin maximum principle in optimal control of distributed systems”, *Ural Math. J.*, **2**:2 (2016), 72–86.
- [27] Ф. П. Васильев, *Методы оптимизации: В 2-х кн.*, МЦНМО, М., 2011. [F. P. Vasil’ev, *Optimization Methods: in 2 books*, MCCME, Moscow, 2011 (In Russian)].
- [28] Ж.-П. Обен, *Нелинейный анализ и его экономические приложения*, Мир, М., 1988; франц. ориг.: J.-P. Aubin, *L’analyse non Lineaire et ses Motivations Economiques*, Masson, Paris–New York, 1984.
- [29] Ф. Кларк, *Оптимизация и негладкий анализ*, Наука, М., 1988; англ. ориг.: F. H. Clarke, *Optimization and Nonsmooth Analysis*, A Wiley–Interscience Publication John Wiley and Sons, New York–Chichester–Brisbane–Toronto–Singapore, 1983.
- [30] Ж.-П. Обен, И. Экланд, *Прикладной нелинейный анализ*, Мир, М., 1988; англ. ориг.: J.-P. Aubin, I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.

Информация об авторах

Сумин Владимир Иосифович, доктор физико-математических наук, ведущий научный сотрудник, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов; профессор, Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород, Российская Федерация. E-mail: v_sumin@mail.ru
ORCID: <https://orcid.org/0000-0002-7479-2181>

Сумин Михаил Иосифович, доктор физико-математических наук, ведущий научный сотрудник, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов; профессор, Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород, Российская Федерация. E-mail: m.sumin@mail.ru
ORCID: <https://orcid.org/0000-0002-3700-6428>

Конфликт интересов отсутствует.

Для контактов:

Сумин Михаил Иосифович
E-mail: m.sumin@mail.ru

Поступила в редакцию 05.09.2024 г.
Поступила после рецензирования 18.11.2024 г.
Принята к публикации 22.11.2024 г.

Information about the authors

Vladimir I. Sumin, Doctor of Physics and Mathematics, Leading Researcher, Derzhavin Tambov State University, Tambov; Professor, Lobachevskii Nizhnii Novgorod State University, Nizhnii Novgorod, Russian Federation. E-mail: v_sumin@mail.ru
ORCID: <https://orcid.org/0000-0002-7479-2181>

Mikhail I. Sumin, Doctor of Physics and Mathematics, Leading Researcher, Derzhavin Tambov State University, Tambov; Professor, Lobachevskii Nizhnii Novgorod State University, Nizhnii Novgorod, Russian Federation. E-mail: m.sumin@mail.ru
ORCID: <https://orcid.org/0000-0002-3700-6428>

There is no conflict of interests.

Corresponding author:

Mikhail I. Sumin
E-mail: m.sumin@mail.ru

Received 05.09.2024
Reviewed 18.11.2024
Accepted for press 22.11.2024

SCIENTIFIC ARTICLE

© R. Chakar, S. Dehilis, W. Merchela, H. Guebbai, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-485-493> $\rho - F$ -contraction fixed point theoremRanda CHAKAR¹, Sofiane DEHILIS¹,
Wassim MERCHELA^{2,3}, Hamza GUEBBAI³¹ Laboratory of Dynamical Systems and Control

Larbi Ben M'Hidi University

BP. 358 Constantine's Route, 04000 Oum El Bouaghi, Algeria

² Mustapha Stambouli University

BP. 305 Sheikh El Khaldi Av., 29000 Mascara, Algeria

³ Laboratory of Applied Mathematics and Modeling

8 May 1945 University

BP. 401, 24000 Guelma, Algeria

Abstract. In this paper, we study the question of conditions for the existence and uniqueness of a fixed point of a mapping over a complete metric space. We first discuss the concepts of F -contraction and F^* -contraction in fixed point theory. These concepts, developed respectively by Wardowski and Piri with Kumam, have catalyzed significant research in various metric spaces. We then propose a generalization of these concepts, $\rho - F$ -contraction and $\rho - F^*$ -contraction, and demonstrate its effectiveness in ensuring the existence and uniqueness of fixed points. This new approach provides greater flexibility by including a function ρ that modulates the contraction, extending the applicability of F - and F^* -contractions. We conclude the paper with an example of a mapping that is a $\rho - F$ -contraction and a $\rho - F^*$ -contraction, respectively, and has a unique fixed point. However, this mapping does not satisfy the conditions of Wardowski and the conditions of Piri and Kumam.

Keywords: fixed-point, existence, uniqueness, F -contraction, $\rho - F$ -contraction

Mathematics Subject Classification: 47H10, 54E35.

For citation: Chakar R., Dehilis S., Merchela W., Guebbai H. $\rho - F$ -contraction fixed point theorem. *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 485–493. <https://doi.org/10.20310/2686-9667-2024-29-148-485-493>

НАУЧНАЯ СТАТЬЯ

© Чакар Р., Дехилис С., Мерчела В., Геббай Х., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-485-493>

УДК 517.98

**Теорема о неподвижной точке $\rho - F$ -сжатия****Ранда ЧАКАР¹, Софиан ДЕХИЛИС¹,
Вассим МЕРЧЕЛА^{2,3}, Хамза ГЕББАЙ³**¹ Лаборатория динамических систем и управления,
Университет Ларби Бен Мхиди

04000, Алжир, г. Ум Эль-Буаги, П.Я. 358 маршрут Константина

² Университет Мустафы Стамбули

29000, Алжир, г. Маскара, П.Я. 305 Проспект Шейха Эль Халди

³ Лаборатория прикладной математики и моделирования,
Университет 8 мая 1945 г.

24000, Алжир, г. Гельма, П.Я. 401

Аннотация. В работе исследуется вопрос об условиях существования и единственности неподвижной точки отображения полного метрического пространства. Вначале обсуждаются понятия F -сжатия и F^* -сжатия в теории неподвижных точек. Эти понятия, разработанные соответственно Вардовским и Пири совместно с Кумамо, послужили катализатором значительных исследований в различных метрических пространствах. Затем предлагаются обобщения этих понятий — $\rho - F$ -сжатие и $\rho - F^*$ -сжатие, демонстрируется их эффективность в обеспечении существования и единственности неподвижных точек. Этот новый подход обеспечивает большую гибкость за счет использования функции ρ , которая регулирует сжатие, расширяя возможности применения F - и F^* -сжатий. Завершает статью пример отображения, являющегося $\rho - F$ -сжатием и $\rho - F^*$ -сжатием и имеющего единственную неподвижную точку. При этом это отображение не удовлетворяет условиям Вардовского и условиям Пири и Кумама.

Ключевые слова: неподвижная точка, существование, единственность, F -сжатие, $\rho - F$ -сжатие

Для цитирования: Чакар Р., Дехилис С., Мерчела В., Геббай Х. Теорема о неподвижной точке $\rho - F$ -сжатия // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 485–493. <https://doi.org/10.20310/2686-9667-2024-29-148-485-493> (In Engl., Abstr. in Russian)

Introduction

The introduction of two distinct approaches to the concept of F -contraction and F^* -contraction, whereas F^* stands out as an adaptation of the F -contraction [1], developed respectively by Wardowski [2] and Piri with Kumam on [3] and authors on [4], has catalyzed significant research in the field of fixed point theories [5]. These approaches have found extensive applications in a multitude of metric spaces, such as b-metric, conic, partial, and fuzzy spaces, among others [6–10]. Essentially, this concept guarantees the existence of a fixed point $S : X \rightarrow X$, where (X, d) represents a complete metric space. The F and F^* have the possibility of assimilating these two contractions to well-established contractions such as Boyd–Wong [11] and Matkowski [12], for this, we must ensure certain conditions. Once S satisfies the following property, known as the F -contraction mapping:

$$\exists \tau > 0, \forall x, y \in X, Sx \neq Sy, F(d(Sx, Sy)) + \tau < F(d(x, y)),$$

where $F : [0, +\infty) \rightarrow \mathbb{R}$ is assumed to satisfy the following $F[\mathbb{R}]$ conditions: [1]

- (1_F) : F is strictly increasing, i.e, $0 < t < s \Rightarrow F(t) < F(s)$,
- (2_F) : $\lim_{t \rightarrow 0^+} F(t) = -\infty$,
- (3_F) : There exists $k \in (0, 1)$, such that $\lim_{t \rightarrow 0^+} t^k F(t) = 0$,

or we use the $F^*[\mathbb{R}]$ conditions [12], which means that F verifies (1_F) , (2_F) and

- (3_{F^*}) : F is a continuous function in $(0, +\infty)$.

However, while this theory represents great mathematical interest, it has seen numerous different applications and a strong attraction for scientific research. We find that Awais et al [13] establish fixed-point results for F -contractions of Reich type for single-valued and multivalued applications in complete metric spaces. Sahil et al [14] introduced the notion of generalized F -contraction and established fixed point theorems for this type of functions in complete metric spaces. They also explore F -expansions and their applications. We find also Inayat et al [15] establish fixed-point results for generalized F -contractions in complete metric spaces. They generalize and unify several known results in the literature. But, Meir Keeler [16] addresses fixed-point results for a class of contractions in metric spaces. He demonstrates that F -contractions (and F^* -contractions) fall under the category of Meir–Keeler contractions. Many results derived from F -contractions also apply to Meir–Keeler mappings.

Whereas, Zhukovskiy [17] extends the fixed point theory to f -quasimetric spaces, particularly generalizing the concept of F -contraction. Unlike the classic F -contraction proposed by Wardowski, where the parameter τ is constant, the article introduces a generalized version where τ is a variable function that depends on $F(d(x, y))$. As illustrated in [17, example 7], this new formulation allows τ to vary based on the distance between two points x and y , providing greater flexibility in analyzing contractions within nonsymmetric metric spaces. This generalization maintains key results, such as the existence of a unique fixed point and convergence of iterations towards that point, while broadening the scope to f -quasimetric spaces.

Our objective is to introduce a new class of contraction, $\rho - F$ -contraction condition defined, for $\rho : \mathbb{R} \rightarrow \mathbb{R}$, as follows:

$$\forall x, y \in X \quad d(x, y) > 0 \Rightarrow F(d(Sx, Sy)) \leq \rho(F(d(x, y))).$$

This relation only indicates that it can be a generalization of F -contraction with $\rho(t) = t - \tau$, $t \in \mathbb{R}$. These assumptions are as follows:

- (1_R) : ρ is increasing,
- (2_R) : ρ is continuous,
- (3_R) : For all $t \in \mathbb{R}$, $\rho(t) < t$.

In this work, we will demonstrate that the hypotheses $(1_R), (2_R), (3_R)$ are sufficient with the class $F[\mathbb{R}]$ and $F^*[\mathbb{R}]$ to ensure the existence of a fixed point for an operator S .

1. Preliminaries

Our goal is to demonstrate that if S is a $\rho - F$ -contraction, where F is either of class $F[\mathbb{R}]$ or of class $F^*[\mathbb{R}]$ and ρ satisfies $(1_R), (2_R)$ and (3_R) , then S has a unique fixed point. But before beginning all this, we need the result presented by the following lemma.

Lemma 1.1. *Suppose that $\rho : \mathbb{R} \rightarrow \mathbb{R}$ is verifying $(1_R), (2_R)$ and (3_R) , then,*

$$\forall t \in \mathbb{R} \quad \lim_{n \rightarrow +\infty} \rho^n(t) = -\infty.$$

P r o o f. Suppose that there exists $t_0 \in \mathbb{R}$, such that,

$$\lim_{n \rightarrow +\infty} \rho^n(t_0) = M.$$

We define the sequence $\{t_n\}_{n \geq 0}$ by t_0 chosen in $\mathbb{R}$ and $t_{n+1} = \rho(t_n)$, for all $n \geq 0$. This means that $\lim_{n \rightarrow +\infty} t_n = M$. Using (2_R) we get,

$$\lim_{n \rightarrow +\infty} t_{n+1} = \rho\left(\lim_{n \rightarrow +\infty} t_n\right) \Rightarrow M = \rho(M).$$

And this is a contradiction with (3_R) . □

2. $\rho - F^*$ -contraction

In this section, we show that the class of function $\rho[\mathbb{R}]$ is compatible with the $F^*[\mathbb{R}_+]$ to ensure the existence and the unicity for a mapping S through the new condition $\rho - F$ -contraction.

Theorem 2.1. *Let (X, d) be a complete metric space. Let $S : X \rightarrow X$ and for ρ verifying $(1_R), (2_R)$ and (3_R) , $F \in F^*[\mathbb{R}]$,*

$$\forall x, y \in X \quad d(x, y) > 0 \Rightarrow F(d(Sx, Sy)) \leq \rho(F(d(x, y))).$$

Then, S has a unique fixed point $x^\infty \in X$ and

$$\forall x_0 \in X \quad \lim_{n \rightarrow +\infty} S^n x_0 = x^\infty.$$

P r o o f. For $x_0 \in X$, we define the sequence $\{x_n\}_{n \geq 0}$ given by $x_{n+1} = Sx_n$ for all $n \geq 0$. We remark that if exists $n_0 \geq 0$, $x_{n_0} = Sx_{n_0}$ then x_{n_0} will be the fixed point x^∞ . Now, suppose that for all $n \geq 0$, $d(x_n, Sx_n) > 0$. We obtain,

$$F(d(x_n, Sx_n)) = F(d(Sx_{n-1}, Sx_n)) \leq \rho(F(d(x_{n-1}, x_n))).$$

Applying (1_ρ) n -times we find

$$F(d(x_n, Sx_n)) \leq \rho(F(d(x_{n-1}, x_n))) \leq \rho^2(F(d(x_{n-2}, x_{n-1}))) \leq \dots \leq \rho^n(F(d(x_0, x_1))).$$

Then, using Lemma 1.1,

$$\lim_{n \rightarrow +\infty} F(d(x_n, Sx_n)) = \lim_{n \rightarrow +\infty} \rho^n(F(d(x_0, x_1))) = -\infty,$$

and by (2_F) we conclude that

$$\lim_{n \rightarrow +\infty} d(x_n, Sx_n) = 0.$$

Now to obtain the fixed point we have to show that $\{x_n\}_{n \geq 0}$ is a Cauchy sequence.

We suppose the existence of some $\varepsilon > 0$ and two sequences of natural numbers $\{\varphi(n)\}_{n \geq 0}$ and $\{\psi(n)\}_{n \geq 0}$ such that:

$$\forall n \geq 0 \quad \varphi(n) > \psi(n) > n, \quad d(x_{\varphi(n)}, x_{\psi(n)}) \geq \varepsilon, \quad d(x_{\varphi(n)-1}, x_{\psi(n)}) < \varepsilon.$$

We obtain,

$$\begin{aligned} \varepsilon &\leq d(x_{\varphi(n)}, x_{\psi(n)}) \leq d(x_{\varphi(n)}, x_{\varphi(n)-1}) + d(x_{\varphi(n)-1}, x_{\psi(n)}) \\ &\leq d(x_{\varphi(n)}, x_{\varphi(n)-1}) + \varepsilon = d(x_{\varphi(n)-1}, Sx_{\varphi(n)-1}) + \varepsilon. \end{aligned}$$

But,

$$\lim_{n \rightarrow +\infty} d(x_{\varphi(n)-1}, Sx_{\varphi(n)-1}) = 0 \Rightarrow \lim_{n \rightarrow +\infty} d(x_{\varphi(n)}, x_{\psi(n)}) = \varepsilon.$$

On the other hand and using $\lim_{n \rightarrow +\infty} d(x_n, Sx_n) = 0$, we can conclude the existence of $n_0 \geq 0$,

$$\forall n \geq n_0 \quad d(x_{\varphi(n)}, Sx_{\varphi(n)}) < \frac{\varepsilon}{4}, \quad d(x_{\psi(n)}, Sx_{\psi(n)}) < \frac{\varepsilon}{4}.$$

Now, we show that

$$\forall n \geq n_0 \quad d(x_{\varphi(n)+1}, x_{\psi(n)+1}) = d(Sx_{\varphi(n)}, Sx_{\psi(n)}) > 0. \quad (2.1)$$

In fact, suppose $\exists m \geq n_0$, such that

$$d(x_{\varphi(m)+1}, x_{\psi(m)+1}) = 0.$$

Then,

$$\begin{aligned} \varepsilon &\leq d(x_{\varphi(m)}, x_{\psi(m)}) \leq d(x_{\varphi(m)}, Sx_{\varphi(m)}) + d(x_{\varphi(m)+1}, x_{\psi(m)+1}) + d(Sx_{\psi(m)}, x_{\psi(m)}) \\ &\leq \frac{\varepsilon}{4} + 0 + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}. \end{aligned}$$

Which is a contradiction. From (2.1) and using $\rho - F$ -contraction hypothesis, we obtain

$$F(d(Sx_{\varphi(n)}, Sx_{\psi(n)})) \leq \rho(F(d(x_{\varphi(n)}, x_{\psi(n)}))).$$

Then

$$F(\varepsilon) \leq \rho(F(\varepsilon)) < F(\varepsilon).$$

The last contradiction allows us to confirm that $\{x_n\}_{n \in \mathbb{N}}$ is a Cauchy sequence converging to x^∞ . Then

$$d(Sx^\infty, x^\infty) = \lim_{n \rightarrow +\infty} d(Sx_n, x_n) = 0 \Rightarrow x^\infty = Sx^\infty.$$

Suppose now that there exists another fixed point for S denoted y^∞ , i.e. $y^\infty = Sy^\infty$, and $x^\infty \neq y^\infty \Rightarrow d(x^\infty, y^\infty) > 0$. Now,

$$F(d(x^\infty, y^\infty)) = F(d(Sx^\infty, Sy^\infty)) \leq \rho F(d(x^\infty, y^\infty)) < F(d(x^\infty, y^\infty)),$$

which is a contradiction, then S has a unique fixed point x^∞ . $\square$

3. $\rho - F$ -contraction

Our objective now is to show that the family of conditions (ρ_1) , (ρ_2) and (ρ_3) ensure, for a $\rho - F$ -contraction mapping S , a unique fixed point, when F is in the class $F[\mathbb{R}_+]$. To achieve our goal, we add a condition on ρ :

$$\forall \beta > 1 \quad \forall t \in \mathbb{R} \quad \sum_{n \geq P(t)} |\rho^n(t)|^{-\beta} < \infty, \quad (3.1)$$

where $P(t)$ is the smaller natural such that, $\rho^{P(t)}(t) < 0$.

Theorem 3.1. *Let (X, d) be a complete metric space and $S : X \rightarrow X$ be a $\rho - F$ -contraction. Then S has a unique fixed point x^∞ , i. e.*

$$x^\infty = Sx^\infty \quad \text{and} \quad \forall x_0 \in X \quad \lim_{n \rightarrow +\infty} S^n x_0 = x^\infty.$$

P r o o f. For $x_0 \in X$, $n \in \mathbb{N}$, we introduce the positif real sequence $\{a_n\}_{n \geq 0}$ by

$$a_n = d(x_n, Sx_n) = d(x_n, x_{n+1}),$$

where $\{x_n\}_{n \geq 0} \subset X$ is given by $x_0 \in X$ and $x_{n+1} = Sx_n \quad \forall n \geq 0$. If there exists $n_0 \geq 0$ such $a_{n_0} = 0$, then $x_{n_0} = Sx_{n_0}$ is the fixed point.

Now, suppose that $a_n > 0$, for all $n \geq 0$. Using the $\rho - F$ -contraction hypothesis, we obtain

$$F(a_n) \leq \rho(F(a_{n-1})) \leq \rho^2(F(a_{n-2})) \leq \dots \leq \rho^n(F(a_0)).$$

Then,

$$\lim_{n \rightarrow +\infty} F(a_n) = \lim_{n \rightarrow +\infty} \rho^n(F(a_0)) = -\infty,$$

and

$$\lim_{n \rightarrow +\infty} a_n = 0.$$

Let

$$n \geq p(F(a_0)),$$

then

$$a_n^k F(a_n) \leq a_n^k \rho^n(F(a_0)) \leq 0 \Rightarrow \lim_{n \rightarrow +\infty} a_n^k \rho^n(F(a_0)) = 0.$$

We take $n \geq N \geq P(F(a_0))$, N bigger enough such that,

$$|a_n^k \rho(F(a_0))| \leq 1 \Rightarrow a_n \leq |\rho^n(F(a_0))|^{-\frac{1}{k}},$$

then, for all $m > n > N$,

$$d(x_m, x_n) \leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \dots + d(x_{n+1}, x_n) \leq \sum_{n \geq N} |\rho^n(F(a_0))|^{-\frac{1}{k}} \rightarrow 0.$$

Then, $\{x_n\}_{n \geq 0}$ is a Cauchy sequence and it is converging to x^∞ in X . It is clear that x^∞ is a fixed point of S . Suppose now that $y^\infty \in X$ is also a fixed point of S . Then,

$$F(d(x^\infty, y^\infty)) = F(d(Tx^\infty, Ty^\infty)) \leq \rho(F(d(x^\infty, y^\infty))) < F(d(x^\infty, y^\infty)),$$

which is a contraction. Then, S has a unique fixed point. $\square$

4. Illustration

Let $X = \{x_n \mid n = 1, 2, \dots, \infty\}$ and define the metric $d : X \rightarrow X$ as follows: for $m > n$ $d(x_n, x_m) = d(x_m, x_n) = \frac{1}{n}$, and $d(x_n, x_n) = 0$.

Let $S : X \rightarrow X$ be defined by $S(x_n) = x_{n+1}$. Then we have:

$$d(S(x_n), S(x_m)) = \frac{1}{n+1},$$

and

$$\frac{d(S(x_n), S(x_m))}{d(x_n, x_m)} = \frac{n}{n+1} \rightarrow 1, \quad n \rightarrow \infty.$$

Therefore, S is not an ordinary contraction. Let's define the function $F(d) = \ln d$, which satisfies the conditions $F[\mathbb{R}]$ and $F^*[\mathbb{R}]$.

Now, let's compute τ :

$$\tau = -F(d(S(x_n), S(x_m))) + F(d(x_n, x_m)) = -\ln \frac{1}{n+1} + \ln \frac{1}{n} = \ln \frac{n+1}{n} \rightarrow 0, \quad n \rightarrow \infty.$$

This implies that S does not satisfy the conditions of the Wardowski and Piri-Kumam theorems (where $\tau > 0$). Let us show that S satisfies the conditions of the theorems proven in this article.

Now, let the function $\rho(t) = \ln \left(\frac{e^t}{e^t + 1} \right) = t - \ln(e^t + 1)$. This function is increasing as:

$$\rho'(t) = 1 - \frac{e^t}{e^t + 1} = \frac{1}{e^t + 1} > 0.$$

Furthermore, $\rho(t) < t$ for $t > 0$, and ρ is continuous. Thus, ρ verified all conditions (ρ_1) , (ρ_2) and (ρ_3) .

Finally, we have:

$$\rho(F(d(x_n, x_m))) = \rho\left(F\left(\frac{1}{n}\right)\right) = \rho\left(\ln \frac{1}{n}\right) = \ln \left(\frac{e^{\ln \frac{1}{n}}}{e^{\ln \frac{1}{n}} + 1} \right) = \ln \frac{1}{n+1} = F(d(Sx_n, Sx_m)).$$

We have introduced the concept of $\rho - F$ -contraction, which extends the classic F -contraction. This new approach is crucial because it generalizes fixed-point results in metric spaces by incorporating a function ρ that adjusts the contracting behavior of the F -function. This

generalization is important as it not only covers traditional F -contractions but also broader versions like those of Wardowski and Piri–Kumam, offering greater flexibility in constructing unique fixed points and analyzing contractive behaviors in various contexts.

As a future direction, it would be interesting to apply this notion of $\rho - F$ -contraction to f -quasimetric spaces [17], where distances are not necessarily symmetric. This would open up the exploration of new classes of contractive mappings and further validate the effectiveness of this generalized approach in more complex spaces, where distances may be asymmetric or only partially defined.

Acknowledgment: We would like to thank to the editor and reviewer for their great assistance and remarks proposed to improve our paper.

References

- [1] N. Secolean, D. Wardowski, “New fixed point tools in non-metrizable spaces”, *Results in Mathematics*, **72**:3 (2017), 919–935.
- [2] D. Wardowski, “Fixed points of a new type of contractive mappings in complete metric spaces”, *Fixed Point Theory and Applications*, **94** (2012), Article number: 94(2012).
- [3] H. Piri, P. Kumam, “Some fixed point theorems concerning F -contraction in complete metric spaces”, *Fixed Point Theory and Applications*, **210** (2014), Article number: 210(2014).
- [4] E. Karapinar, M. A. Kutbi, H. Piri, D. O’Regan, “Fixed points of conditionally F -contractions in complete metric-like spaces”, *Fixed Point Theory and Applications*, **126** (2015), Article number: 126(2015).
- [5] J. Z. Vujakovi, S. N. Radenović, “On some F -contraction of Piri–Kumam–Dung-type mappings in metric spaces”, *Military Technical Courier*, **68**:4 (2020), 697–714.
- [6] S. Banach, “Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales”, *Fundamenta Mathematicae*, **3**:1 (1922), 133–181.
- [7] M. A. Alghamdi, N. Hussain, P. Salimi, “Fixed point and coupled fixed point theorems on b -metric-like spaces”, *Journal of Inequalities and Applications*, **2013** (2013), Article number: 402.
- [8] S. Janković, Z. Kadelburg, S. Radenović, “On cone metric spaces: A survey”, *Nonlinear Analysis: Theory, Methods & Applications*, **74**:7 (2011), 2591–2601.
- [9] T. G. Bhaskar, V. Lakshmikantham, “Fixed point theorems in partially ordered metric spaces and applications”, *Nonlinear Analysis: Theory, Methods & Applications*, **65**:7 (2006), 1379–1393.
- [10] D. Turkoglu, C. Alaca, Y. J. Cho, C. Yildiz, “Common fixed point theorems in intuitionistic fuzzy metric spaces”, *Journal of Applied Mathematics and Computing*, **22**:1–1 (2006), 411–424.
- [11] I. Arandjelović, Z. Kadelburg, S. Radenović, “Boyd–Wong-type common fixed point results in cone metric spaces”, *Applied Mathematics and Computation*, **217**:17 (2011), 7167–7171.
- [12] C. Bylka, “Fixed point theorems of Matkowski on probabilistic metric spaces”, *Demonstratio Mathematica*, **29**:1 (1996), 159–164.
- [13] A. Awais, M. Nazam, M. Arshad, S. O. Kim, “ F -metric, F -contraction and common fixed-point theorems with applications”, *Mathematics*, **7**:7 (2019), Article number: 586.
- [14] A. Sahil, M. Masta, M. Poonia, “Fixed point theorems for generalized orthogonal F -contraction and F -expansion of Wardowski kind via the notion of ψ -fixed point”, *International Journal of Nonlinear Analysis and Applications*, **14**:2 (2023), 221–231.
- [15] J. Ahmad, J. Al-Rawashdeh, A. Azam, “New fixed point theorems for generalized F -contractions in complete metric spaces”, *Fixed Point Theory and Algorithms for Sciences and Engineering*, **80** (2015), 1–18.
- [16] M. Cvetković, “The relation between F -contraction and Meir–Keeler contraction”, *Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A-Mat.*, **117** (2023), Article number: 39.

- [17] E. S. Zhukovskiy, “The fixed points of contractions of F -quasimetric spaces”, *Siberian Mathematical Journal*, **59**:6 (2018), 1338–1350.

Information about the authors

Randa Chakar, Post-Graduate Student, Laboratory of Dynamical Systems and Control, Larbi Ben M’Hidi University, Oum El Bouaghi, Algeria. E-mail: chaker.doudi24@gmail.com randa.chakar@univ-oeb.dz
ORCID: <https://orcid.org/0009-0009-4079-876X>

Sofiane Dehilis, PhD, Associate Professor of Mathematics and Computer Science Department, Laboratory of Dynamical Systems and Control. Larbi Ben M’Hidi University, Oum El Bouaghi, Algeria. E-mail: dehilissofiane@yahoo.fr
ORCID: <https://orcid.org/0000-0002-8771-5046>

Wassim Merchela, Candidate of Physical and Mathematical Sciences, Associate Professor of Mathematics Department, Mustapha Stambouli University, Mascara; Researcher at the Laboratory of Applied Mathematics and Modeling, 08 May 1945 University, Guelma, Algeria. E-mail: merchela.wassim@gmail.com wassim.merchela@univ-mascara.dz
ORCID: <https://orcid.org/0000-0002-3702-0932>

Hamza Guebbai, Professor, Professor of Mathematics Department, 08 May 1945 University, Guelma, Algeria. E-mail: guebaihamza@yahoo.fr
ORCID: <https://orcid.org/0000-0001-8119-2881>

There is no conflict of interests.

Corresponding author:

Wassim Merchela
E-mail: merchela.wassim@gmail.com

Received 27.07.2024

Reviewed 15.10.2024

Accepted for press 06.11.2024

Информация об авторах

Чакар Ранда, аспирант, лаборатория динамических систем и управления, Университет Ларби Бен Мхиди, г. Ум Эль-Буаги, Алжир. E-mail: chaker.doudi24@gmail.com randa.chakar@univ-oeb.dz
ORCID: <https://orcid.org/0009-0009-4079-876X>

Дехилис Софиан, кандидат наук, доцент кафедры математики и информатики, лаборатория динамических систем и управления. Университет Ларби Бен Мхиди, г. Ум Эль-Буаги, Алжир. E-mail: dehilissofiane@yahoo.fr
ORCID: <https://orcid.org/0000-0002-8771-5046>

Мерчела Вассим, кандидат физико-математических наук, доцент кафедры математики, Университет Мустафы Стамбули, г. Маскара; научный сотрудник лаборатории прикладной математики и моделирования, Университет 8 мая 1945 г., г. Гельма, Алжир. E-mail: merchela.wassim@gmail.com wassim.merchela@univ-mascara.dz
ORCID: <https://orcid.org/0000-0002-3702-0932>

Геббай Хамза, профессор, профессор кафедры математики, Университет 8 мая 1945 г., г. Гельма, Алжир. E-mail: guebaihamza@yahoo.fr
ORCID: <https://orcid.org/0000-0001-8119-2881>

Конфликт интересов отсутствует.

Для контактов:

Мерчела Вассим
E-mail: merchela.wassim@gmail.com

Поступила в редакцию 27.07.2024 г.

Поступила после рецензирования 15.10.2024 г.

Принята к публикации 06.11.2024 г.

SCIENTIFIC ARTICLE

© J. Ettayb, 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-494-516>

On λ -commuting and left (right) pseudospectrum and left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces

Jawad ETTAYB

Regional Academy of Education and Training Casablanca Settat,
Hamman Al-Fatawaki collegiate High School
Road to Berrechid, Had Soualem 26402, Morocco

Abstract. In this paper, we demonstrate some spectral properties of the λ -commuting of continuous linear operators on ultrametric Banach spaces and we introduce and study the operator equations $ASB = S$ and $AS = SB$. We give some properties of these operator equations. Some illustrative examples are provided. On the other hand, we introduce and study the left (right) pseudospectrum and the left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces. We prove that the left pseudospectra associated with various $\varepsilon > 0$ are nested sets and the intersection of all the left pseudospectra is the left spectrum. We give a relationship between the left (right) pseudospectrum and the left (right) condition pseudospectrum. Moreover, many results are proved concerning the left (right) pseudospectrum and the left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces.

Keywords: ultrametric Banach spaces, bounded linear operators, spectrum, left and right pseudospectrum

Mathematics Subject Classification: 47A10, 47S10.

For citation: Ettayb J. On λ -commuting and left (right) pseudospectrum and left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces. *Vestnik rossiyskikh universitetov. Matematika = Russian Universities Reports. Mathematics*, **29**:148 (2024), 494–516. <https://doi.org/10.20310/2686-9667-2024-29-148-494-516>

НАУЧНАЯ СТАТЬЯ

© Эттайб Дж., 2024

<https://doi.org/10.20310/2686-9667-2024-29-148-494-516>

УДК 517.983, 517.984



О λ -коммутировании, левом (правом) псевдоспектре и левом (правом) условном псевдоспектре линейных непрерывных операторов на ультраметрических банаховых пространствах

Джавад ЭТТАЙБ

Региональная академия образования и обучения Касабланка Сеттат,

Университетская средняя школа Хаммана Аль-Фатаваки

26402, Марокко, г. Хад Суалем, дорога в Беррешид

Аннотация. В работе мы демонстрируем некоторые спектральные свойства λ -коммутирования линейных непрерывных операторов в ультраметрических банаховых пространствах, а также изучаем операторные уравнения $ASB = S$ и $AS = SB$. Мы рассматриваем некоторые свойства этих операторных уравнений; приводим иллюстративные примеры. С другой стороны, мы вводим и изучаем левый (правый) псевдоспектр и левый (правый) условный псевдоспектр линейных непрерывных операторов в ультраметрических банаховых пространствах. Мы доказываем, что левые псевдоспектры, связанные с различными $\varepsilon > 0$, являются вложенными множествами, а пересечение всех левых псевдоспектров является левым спектром. Мы выявляем связь между левым (правым) псевдоспектром и левым (правым) условным псевдоспектром. Более того, доказываем еще ряд результатов, касающихся левого (правого) псевдоспектра и левого (правого) условного псевдоспектра линейных непрерывных операторов в ультраметрических банаховых пространствах.

Ключевые слова: ультраметрические банаховы пространства, линейные ограниченные операторы, спектр, левый и правый псевдоспектр

Для цитирования: Эттайб Дж. О λ -коммутировании, левом (правом) псевдоспектре и левом (правом) условном псевдоспектре линейных непрерывных операторов на ультраметрических банаховых пространствах // Вестник российских университетов. Математика. 2024. Т. 29. № 148. С. 494–516. <https://doi.org/10.20310/2686-9667-2024-29-148-494-516> (In Engl., Abstr. in Russian)

1. Introduction and preliminaries

The classical theory of commutators was studied by H. Weyl [1] and J. von Neumann [2] and it played an important role in quantum mechanics [3–5]. In [6], C.R. Putnam collected some properties of the commutation of continuous linear operators in a Hilbert space over the field of complex numbers $\mathbb{C}$. Recently, many researchers studied and explored the operator equation $AS = \lambda SA$ where $\lambda \in \mathbb{C} \setminus \{0\}$, A and S are continuous linear operators on complex Hilbert spaces, see [7–9].

In ultrametric operator theory, the author [10] extended and studied the operator equation of the form $AS = \lambda SA$ where $\lambda \in \mathbb{K} \setminus \{0\}$, A and S are continuous linear operators on ultrametric Banach spaces over $\mathbb{K}$. He presented some spectral properties of λ -commuting operators on ultrametric Banach spaces over $\mathbb{K}$ and he gave an illustrative examples, see [10].

Recently, A. Ammar et al. [11] introduced and studied the pseudospectra of closed linear operators on ultrametric Banach spaces. On the other hand, A. Ammar et al. [12] introduced and studied the condition pseudospectra of continuous linear operators on ultrametric Banach spaces and gave some of its properties.

In [13], the author presented and studied the determinant spectrum, the M -determinant spectrum, and the C -trace pseudospectrum of ultrametric matrix pencils.

There are many studies on pseudospectra and condition pseudospectra of continuous linear operator pencils and λ -commuting of operators in ultrametric operator theory, see [14–17]. In Section 5., we consider the problem of finding the eigenvalues of the generalized eigenvalue problem of the form

$$P(\lambda)x = 0,$$

where $P(\lambda) = \sum_{k=0}^m \lambda^k A_k$, $A_k \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$, $x \in \mathbb{K}^n$ and $\mathcal{M}_n(\mathbb{K})$ is the space of all $n \times n$ matrices over $\mathbb{K}$. I is the identity matrix of $\mathcal{M}_n(\mathbb{K})$. If $C \in \mathcal{M}_n(\mathbb{K})$, the determinant of C is denoted by $\det(C)$ (for details on the space $\mathcal{M}_n(\mathbb{K})$ see [18] and [19]).

Throughout this paper, $\mathbb{Q}_p$ is the field of p -adic numbers, $\mathcal{E}$ is an ultrametric infinite-dimensional Banach space over a complete ultrametric valued field $\mathbb{K}$ with a non-trivial valuation $|\cdot|$ and $\mathcal{L}(\mathcal{E})$ denotes the set of all continuous linear operators on $\mathcal{E}$. Recall that $\mathbb{K}$ is called spherically complete if each decreasing sequence of balls in $\mathbb{K}$ has a non-empty intersection. For more details, see [20]. Let $S \in \mathcal{L}(\mathcal{E})$, $R(S)$, $N(S)$, S^* , $\sigma_p(S)$, $\sigma(S)$ and $\rho(S)$ denote the range, the kernel, the adjoint, the point spectrum, the spectrum and the resolvent set of S respectively [20].

The aim of this paper is to demonstrate some spectral properties of λ -commuting of continuous linear operators on ultrametric Banach spaces and we introduce and study the operator equations $AS = SB$ and $ASB = S$ for some $S \in \mathcal{L}(\mathcal{E})$. Moreover, some illustrative examples are provided. On the other hand, we introduce and study the left (right) pseudo-spectrum and the left (right) condition pseudospectrum of continuous linear operators on ultrametric Banach spaces. We obtain some results related to them. We continue by recalling some preliminaries.

Definition 1.1. [20] A field $\mathbb{K}$ is said to be ultrametric if it is endowed with an absolute value $|\cdot| : \mathbb{K} \rightarrow \mathbb{R}_+$ such that

(i) $|\alpha| = 0$ if, and only if, $\alpha = 0$;

(ii) For all $\lambda, \alpha \in \mathbb{K}$, $|\lambda\alpha| = |\lambda||\alpha|$;

(iii) For each $\lambda, \alpha \in \mathbb{K}$, $|\lambda + \alpha| \leq \max\{|\lambda|, |\alpha|\}$.

Definition 1.2. [20] Let $\mathcal{E}$ be a vector space over $\mathbb{K}$. A mapping $\|\cdot\| : \mathcal{E} \rightarrow \mathbb{R}_+$ is said to be an ultrametric norm if:

- (i) For all $x \in \mathcal{E}$, $\|x\| = 0$ if and only if $x = 0$,
- (ii) For any $x \in \mathcal{E}$ and $\lambda \in \mathbb{K}$, $\|\lambda x\| = |\lambda|\|x\|$,
- (iii) For each $x, y \in \mathcal{E}$, $\|x + y\| \leq \max(\|x\|, \|y\|)$.

Definition 1.3. [20] An ultrametric Banach space is a complete ultrametric normed space.

Example 1.1. [20] Let $c_0(\mathbb{K})$ be the space of all sequences $(x_i)_{i \in \mathbb{N}}$ in $\mathbb{K}$ such that $\lim_{i \rightarrow \infty} x_i = 0$. Then $c_0(\mathbb{K})$ is a vector space over $\mathbb{K}$ and

$$\|(x_i)_{i \in \mathbb{N}}\| = \sup_{i \in \mathbb{N}} |x_i|$$

is an ultrametric norm for which $(c_0(\mathbb{K}), \|\cdot\|)$ is an ultrametric Banach space.

Theorem 1.1. [21] Let $\mathcal{E}$ be an ultrametric Banach space over a spherically complete field $\mathbb{K}$. For each $x \in \mathcal{E}^* = \mathcal{E} \setminus \{0\}$, there exists $x^* \in \mathcal{E}^*$ such that $x^*(x) = 1$ and $\|x^*\| = \|x\|^{-1}$.

Definition 1.4. [20] An ultrametric Banach space $\mathcal{E}$ over $\mathbb{K}$ is called a free Banach space if there is a family $(x_i)_{i \in \mathcal{I}} \in \mathcal{E}$ indexed by a set $\mathcal{I}$ such that all $x \in \mathcal{E}$ is written in a unique fashion as $x = \sum_{i \in \mathcal{I}} \lambda_i x_i$ and $\|x\| = \sup_{i \in \mathcal{I}} |\lambda_i| \|x_i\|$. The family $(x_i)_{i \in \mathcal{I}}$ is called an orthogonal basis for $\mathcal{E}$. If, for each $i \in \mathcal{I}$, $\|x_i\| = 1$, hence $(x_i)_{i \in \mathcal{I}}$ is called an orthonormal basis of $\mathcal{E}$.

Definition 1.5. [20] Let $\omega = (\omega_i)_i$ be a sequence of $\mathbb{K}^* = \mathbb{K} \setminus \{0\}$. We define $\mathcal{E}_\omega$ by

$$\mathcal{E}_\omega = \{x = (x_i)_i : \forall i \in \mathbb{N} \ x_i \in \mathbb{K}, \text{ and } \lim_{i \rightarrow \infty} |\omega_i|^{\frac{1}{2}} |x_i| = 0\},$$

and it is equipped with the norm

$$\forall x \in \mathcal{E}_\omega : x = (x_i)_i \quad \|x\| = \sup_{i \in \mathbb{N}} (|\omega_i|^{\frac{1}{2}} |x_i|).$$

Remark 1.1. [20]

- (i) The space $(\mathcal{E}_\omega, \|\cdot\|)$ is an ultrametric Banach space.
- (ii) If

$$\langle \cdot, \cdot \rangle : \mathcal{E}_\omega \times \mathcal{E}_\omega \longrightarrow \mathbb{K}, \quad (x, y) \mapsto \sum_{i=0}^{\infty} x_i y_i \omega_i,$$

where $x = (x_i)_i$ and $y = (y_i)_i$. Then the space $(\mathcal{E}_\omega, \|\cdot\|, \langle \cdot, \cdot \rangle)$ is called an ultrametric Hilbert space.

- (iii) The orthogonal basis $\{e_i, i \in \mathbb{N}\}$ is called the canonical basis of $\mathcal{E}_\omega$ where for all $i \in \mathbb{N}$, $\|e_i\| = |\omega_i|^{\frac{1}{2}}$.

Remark 1.2. [20] Let $\mathbb{K} = \mathbb{Q}_p$, if $p \equiv 1(\text{mod } 4)$, then $i = \sqrt{-1} \in \mathbb{Q}_p$ and $i^2 = -1$.

Definition 1.6. [20] Let $S \in \mathcal{M}_n(\mathbb{K})$. The spectrum $\sigma(S)$ of S is defined by

$$\sigma(S) = \{\lambda \in \mathbb{K} : S - \lambda I \text{ is not invertible}\}.$$

By Definition 6 of [13] (where $B = I$), we have the following:

Definition 1.7. If $S \in \mathcal{M}_n(\mathbb{K})$ and $\varepsilon > 0$. Then the ε -determinant spectrum $d_\varepsilon(S)$ of S is the following set:

$$d_\varepsilon(S) = \{\lambda \in \mathbb{K} : |\det(S - \lambda I)| \leq \varepsilon\}.$$

From Remark 2 of [13] (where $B = I$), we get

Remark 1.3. Note that for each $S \in \mathcal{M}_n(\mathbb{K})$ and $\varepsilon > 0$, $\sigma(S) \subseteq d_\varepsilon(S)$ and $d_0(S) = \sigma(S)$.

The λ -commuting of operators is defined as follows.

Definition 1.8. [10] Let $A, B \in \mathcal{L}(\mathcal{E})$, A and B are called λ -commuting operators if $AB = \lambda BA$ for some $\lambda \in \mathbb{K}^*$.

Example 1.2. [10] Let $\mathbb{K} = \mathbb{Q}_p$ with $p \equiv 1(\text{mod } 4)$, let A and B be defined on $\mathbb{Q}_p \times \mathbb{Q}_p$ respectively by

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}.$$

Then $AB = -BA$.

Example 1.3. [10] Let $\lambda \in \mathbb{K}^*$, let A and B be defined on $\mathbb{K}^3$ by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \lambda & 0 \\ 1 & \lambda & \lambda^2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then $AB = \lambda BA$.

Example 1.4. [10] Let $\lambda \in \mathbb{K}$ such that $|\lambda| > 1$ and let $A, B \in \mathcal{L}(c_0(\mathbb{K}))$ be given respectively by

$$B(x_1, x_2, x_3, \dots) = \left(\frac{x_1}{\lambda}, \frac{x_2}{\lambda^2}, \frac{x_3}{\lambda^3}, \dots \right) \quad \text{for all } (x_1, x_2, x_3, \dots) \in c_0(\mathbb{K})$$

and

$$\text{for all } n \geq 1, \quad Ae_n = e_{n+1},$$

where $(e_n)_{n \geq 1}$ is a base of $c_0(\mathbb{K})$. Hence $AB = \lambda BA$.

Let $A \in \mathcal{L}(\mathcal{E})$ be given, set $S_\lambda(A) = \{B \in \mathcal{L}(\mathcal{E}) : AB = \lambda BA\}$. We collect some properties of λ -commuting operators.

Proposition 1.1. [10] Let $A \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}$.

- (i) If $B_1, B_2 \in S_\lambda(A)$, hence $B_1 + B_2 \in S_\lambda(A)$ and $B_1 B_2 \in S_{\lambda^2}(A)$;
- (ii) If B is invertible and $\lambda \neq 0$, then $B^{-1} \in S_{\frac{1}{\lambda}}(A)$;
- (iii) $S_\lambda(A)$ is closed in the uniform operator topology;
- (iv) If $AB = \lambda BA$ and $AB \neq 0$, then $Ap(B) = p(\lambda B)A$ where p is a non-constant polynomial.

Proposition 1.2. [10] Let $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{Z}_p$ such that $AB = \lambda BA$, $AB \neq 0$ and $\|B\| < p^{\frac{1}{1-p}}$. Then

$$Ae^B = e^{\lambda B}A.$$

Proposition 1.3. [10] If $A, B \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$, hence $\sigma_p(AB) = \lambda \sigma_p(BA)$.

Proposition 1.4. [10] Let $A, B \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$. Then $\rho(AB) \subset \lambda \rho(BA)$. Furthermore, for any $\mu \in \rho(AB)$, $R(\mu, AB) = \lambda^{-1}R(\lambda^{-1}\mu, BA)$.

Proposition 1.5. [10] Let $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$. Then

- (i) $N(AB) = N(BA)$;
- (ii) $R(AB) = R(BA)$;
- (iii) For all $\mu \in \mathbb{K}$, $N(AB - \mu) = N(BA - \lambda^{-1}\mu)$;
- (iv) For any $\mu \in \mathbb{K}$, $R(AB - \mu) = R(BA - \lambda^{-1}\mu)$.

From Proposition 1.5, we conclude:

Theorem 1.2. [10] If $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$, then

$$\sigma_e(AB) = \lambda \sigma_e(BA).$$

For $A \in \mathcal{L}(\mathcal{E})$, set $r(A) = \lim_{n \rightarrow \infty} \|A^n\|^{\frac{1}{n}}$. We have the following proposition.

Proposition 1.6. [10] Let $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}$ with $|\lambda| = 1$ and $AB = \lambda BA$. Then $r(AB) \leq r(A)r(B)$.

We continue with the following definitions.

Definition 1.9. [20] Let $\mathcal{E}$ be a non-Archimedean Banach space over $\mathbb{K}$ and let $B \in \mathcal{L}(\mathcal{E})$, the spectrum $\sigma(B)$ of B is defined by

$$\sigma(B) = \{\mu \in \mathbb{K} : B - \mu I \text{ is not invertible}\},$$

the resolvent set of B is defined by $\rho(B) = \mathbb{K} \setminus \sigma(B)$.

Definition 1.10. [13] Let $B \in \mathcal{M}_n(\mathbb{K})$, the trace $Tr(B)$ of B is defined by $\sum_{i=1}^n b_{i,i}$ where for each $i \in \{1, \dots, n\}$, $b_{i,i} \in \mathbb{K}$ are diagonal coefficients of B .

Proposition 1.7. [13] Let $B, C \in \mathcal{M}_n(\mathbb{K})$. Then

- (i) For any $\lambda \in \mathbb{K}$, $\text{Tr}(B + \lambda C) = \text{Tr}(B) + \lambda \text{Tr}(C)$,
- (ii) $\text{Tr}(BC) = \text{Tr}(CB)$.

We have:

Definition 1.11. [13] Let $B \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$, the trace pseudospectrum $\text{Tr}_\varepsilon(B)$ of B is given by

$$\text{Tr}_\varepsilon(B) = \sigma(B) \cup \{\lambda \in \mathbb{K} : |\text{Tr}(B - \lambda I)| \leq \varepsilon\}.$$

The trace pseudoresolvent $\text{Tr}\rho_\varepsilon(B)$ of B is defined by

$$\text{Tr}\rho_\varepsilon(B) = \rho(B) \cap \{\mu \in \mathbb{K} : |\text{Tr}(B - \mu I)| > \varepsilon\}.$$

Lemma 1.1. [20] Let $S \in \mathcal{L}(\mathcal{E})$ with $\|S\| < 1$, then $\|(I - S)^{-1}\| \leq 1$.

2. λ -commuting of ultrametric operators

Similar to the proof of Proposition 1.6, we conclude:

Proposition 2.1. Let $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}$ such that $|\lambda| = 1$ and $AB = \lambda BA$. Then $r(BA) \leq r(A)r(B)$.

Question: In Proposition 2.1, if $|\lambda| \neq 1$, does $r(BA) \leq r(A)r(B)$ hold?

Definition 2.1. Suppose that $\|\mathcal{E}\| \subseteq |\mathbb{K}|$. Let $A \in \mathcal{L}(\mathcal{E})$, the approximate spectrum $\sigma_{ap}(A)$ of A is defined by

$$\sigma_{ap}(A) = \{\mu \in \mathbb{K} : \exists (x_n)_{n \in \mathbb{N}} \in \mathcal{E} \forall n \in \mathbb{N} \|x_n\| = 1 \text{ and } \lim_{n \rightarrow \infty} \|(A - \mu I)x_n\| = 0\}.$$

Proposition 2.2. Suppose that $\|\mathcal{E}\| \subseteq |\mathbb{K}|$. If $A, B \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA \neq 0$, then $\sigma_{ap}(AB) = \lambda \sigma_{ap}(BA)$.

Proof. Let $\mu \in \sigma_{ap}(AB)$, then there is $(x_n)_{n \in \mathbb{N}}$ in $\mathcal{E}$ such that for each $n \in \mathbb{N}$, $\|x_n\| = 1$ and $\lim_{n \rightarrow \infty} \|(AB - \mu I)x_n\| = 0$. Since

$$\|(BA - \frac{\mu}{\lambda} I)x_n\| = \frac{\|(AB - \mu I)x_n\|}{|\lambda|}. \quad (2.1)$$

Then $\frac{\mu}{\lambda} \in \sigma_{ap}(BA)$ that is, $\mu \in \lambda \sigma_{ap}(BA)$. Similarly, if $\frac{\mu}{\lambda} \in \sigma_{ap}(BA)$ and using (2.1), we get $\mu \in \sigma_{ap}(AB)$. $\square$

Lemma 2.1. Let $A, B \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{K}$ such that $AB = \lambda BA \neq 0$. Then for any $n \in \mathbb{N}$, $A^n B = \lambda^n B A^n$.

Proof. Since $AB = \lambda BA \neq 0$. Then $A^2 B = \lambda ABA = \lambda^2 B A^2$. One can see that for all $n \in \mathbb{N}$, $A^n B = \lambda^n B A^n$. $\square$

Proposition 2.3. Let $A, B \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{Z}_p$ such that $AB = \lambda BA$, $AB \neq 0$ and $\|A\| < p^{\frac{1}{1-p}}$. Then

$$e^A B = B e^{\lambda A}.$$

P r o o f. By $\|A\| < p^{\frac{1}{1-p}}$, we get e^A and $e^{\lambda A}$ exist. Since $AB = \lambda BA \neq 0$. Using Lemma 2.1, we conclude that $e^A B = B e^{\lambda A}$. $\square$

In the finite-dimensional ultrametric Banach space, we obtain.

P r o p o s i t i o n 2.4. *If $A, B \in \mathcal{M}_n(\mathbb{K})$ are invertible matrices and $\lambda \in \mathbb{K}^*$ such that $AB = \lambda BA$, then $\lambda^n = 1$.*

P r o o f. From $\det(AB) = \lambda^n \det(BA)$ and $\det(BA) = \det(AB)$. We get $\lambda^n = 1$. $\square$

From Proposition 4.2, we have the following:

Corollary 2.1. *If $A, B \in \mathcal{M}_n(\mathbb{K})$ are invertible matrices and $\lambda \in \mathbb{K}^*$ such that $AB = \lambda BA$, then $|\lambda| = 1$.*

P r o p o s i t i o n 2.5. *Let $A, B \in \mathcal{M}_n(\mathbb{K})$ and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$. Then $\mu \in d_\varepsilon(AB)$ if and only if $\frac{\mu}{\lambda} \in d_{\frac{\varepsilon}{|\lambda|^n}}(BA)$.*

P r o o f. From $\det(AB - \mu I) = \det(\lambda BA - \mu I) = \lambda^n \det(BA - \frac{\mu}{\lambda} I)$ for $\mu \in \mathbb{K}$ and $\lambda \in \mathbb{K}^*$. Then $\mu \in d_\varepsilon(AB)$ if and only if $\frac{\mu}{\lambda} \in d_{\frac{\varepsilon}{|\lambda|^n}}(BA)$. $\square$

P r o p o s i t i o n 2.6. *Let $A, B \in \mathcal{M}_n(\mathbb{K})$ and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA$ and $AB \neq 0$. If $\text{tr}(AB) \neq 0$ or $\text{tr}(BA) \neq 0$, then $\lambda = 1$.*

P r o o f. Since $\text{tr}(AB) = \lambda \text{tr}(BA) = \lambda \text{tr}(AB)$ and $(\text{tr}(AB) \neq 0 \text{ or } \text{tr}(BA) \neq 0)$, we get $\lambda = 1$. $\square$

Let $\mathcal{E}$ is a free Banach space over $\mathbb{K}$, we set $\mathcal{L}_0(\mathcal{E}) = \{A \in \mathcal{L}(\mathcal{E}) : A^* \text{ exists}\}$.

P r o p o s i t i o n 2.7. [20] *If $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$, then*

$$(i) \quad (A + \lambda B)^* = A^* + \lambda B^*.$$

$$(ii) \quad (AB)^* = B^* A^*.$$

D e f i n i t i o n 2.2. [20] Let $A \in \mathcal{L}_0(\mathcal{E}_\omega)$. We have

$$(i) \quad A \text{ is said to be selfadjoint if } A^* = A;$$

$$(ii) \quad A \text{ is said to be normal if } A^* A = A A^*;$$

$$(iii) \quad A \text{ is said to be unitary if } A^* A = A A^* = I.$$

The following proposition describes some spectral properties of λ -commuting operators.

P r o p o s i t i o n 2.8. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$ with $AB = \lambda BA \neq 0$. If A is a selfadjoint, then $ABB^* = BB^*A$ and ABB^* is selfadjoint.*

P r o o f. If $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = \lambda BA$, then $(AB)^* = (\lambda BA)^*$. Hence

$$B^* A^* = \lambda A^* B^*. \quad (2.2)$$

Since A is a selfadjoint and by (2.2), we get $B^* A = \lambda A B^*$. On the other hand

$$ABB^* = \lambda BAB^* = \lambda \lambda^{-1} BB^* A = BB^* A,$$

and

$$(ABB^*)^* = (BB^*)^* A^* = BB^* A = ABB^*.$$

$\square$

As the proof of Proposition 2.8, we get the following:

Proposition 2.9. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$ with $AB = \lambda BA \neq 0$. If B is a selfadjoint, then $BAA^* = AA^*B$ and AA^*B is selfadjoint.*

By Proposition 2.8 and Proposition 2.9, we conclude that:

Lemma 2.2. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$ with $AB = \lambda BA \neq 0$. If A and B are selfadjoint operators, then $AB^2 = B^2A$ and $BA^2 = A^2B$.*

Proposition 2.10. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$ with $AB = \lambda BA \neq 0$ and $BA^2 \neq 0$. If A and B are selfadjoint operators, then $\lambda \in \{-1, 1\}$.*

Proof. From $AB = \lambda BA$, we get $A^2B = \lambda^2BA^2$. By Lemma 2.2, we have $A^2B = B^2A = \lambda^2BA^2$. Since $BA^2 \neq 0$, we get $\lambda^2 = 1$. Then $\lambda \in \{-1, 1\}$. $\square$

We give another proof of Proposition 2.10 without the condition $BA^2 \neq 0$.

Proposition 2.11. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}$ with $AB = \lambda BA \neq 0$. If A and B are selfadjoint operators, then $\lambda \in \{-1, 1\}$.*

Proof. From A and B are selfadjoint operators, we get $(AB)^* = (\lambda BA)^*$, hence

$$BA = \lambda AB. \quad (2.3)$$

Using $AB = \lambda BA$ and (2.3), we get $AB = \lambda^2AB$. Hence $\lambda^2 = 1$. Thus $\lambda \in \{-1, 1\}$. $\square$

Proposition 2.12. *Let $A, B \in \mathcal{L}(\mathcal{E}_\omega)$. If there is an unitary operator $U \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = UBA = BAU$, then $AB^2A = BA^2B$.*

Proof. Since $AB = UBA = BAU$, we have

$$AB^2A = UBABA = BAUBA = BAAB = BA^2B.$$

$\square$

Lemma 2.3. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ be selfadjoint operators. If there is an unitary operator $U \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = UBA$. Then*

(i) U and U^* commute with AB ;

(ii) U and U^* commute with BA .

Proof. (i) From $AB = UBA$, we have $BA = ABU^*$. Hence $AB = UBA = UABU^*$. Thus $ABU = UAB$ and $U^*AB = ABU^*$.

(ii) From (i), we get $U^*BA = (ABU)^* = (UAB)^* = BAU^*$. On the other hand $BAU = UU^*BAU = UBAU^*U = UBA$. $\square$

Proposition 2.13. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ be selfadjoint operators. If there is an unitary operator $U \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = UBA$. Then $AB^2A = BA^2B$.*

Proof. From Lemma 2.3, $UBA = BAU$. Therefore, $BA^2B = BAAB = BAUBA = UBABA = ABBA = AB^2A$. $\square$

Lemma 2.4. *Let $A, B \in \mathcal{L}(\mathcal{E})$. If $AB^2 = B^2A$ and $BA^2 = A^2B$, then $AB^2A = BA^2B$.*

P r o o f. One can see that $AB^2A = B^2A^2$ and $BA^2B = B^2A^2$. Thus $AB^2A = BA^2B$. $\square$

Question: Let $A, B \in \mathcal{L}_0(\mathcal{E})$ be selfadjoint operators. Is the converse of Lemma 2.4 hold?

E x a m p l e 2.1. Let $\mathbb{K} = \mathbb{Q}_p$ and let A and B be defined on $\mathbb{Q}_p^2$ by

$$A = \begin{pmatrix} a & 0 \\ b & \lambda a \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

where $\lambda a \in \mathbb{Q}_p \setminus \{0\}$. Then $AB = \lambda BA$.

E x a m p l e 2.2. Let $\lambda \in \mathbb{K}^*$, let A and B be defined on $\mathbb{K}^3$ by

$$A = \begin{pmatrix} a & 0 & 0 \\ b & \lambda a & 0 \\ c & \lambda b & \lambda^2 a \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then $AB = \lambda BA$.

Lemma 2.5. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$. If there is unitary operators $U, V \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = UB^*A^*$ and $BA = VA^*B^*$, then the following statements hold:*

- (i) AB commutes with U and U^* ;
- (ii) BA commutes with V and V^* .

P r o o f. (i) From $AB = UB^*A^*$, we have $B^*A^* = ABU^*$. Hence $AB = UB^*A^* = UABU^*$. Thus $ABU = UAB$ and $U^*AB = ABU^*$.

(ii) By $BA = VA^*B^*$, we get $A^*B^* = BAV^*$. Thus $BA = VBAV^*$. Hence $BAV = VBA$ and $V^*BA = BAV^*$. $\square$

Theorem 2.1. *If $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$ are selfadjoint operators and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA \neq 0$, then AB and BA are normal commuting operators.*

P r o o f. Set $S = AB$, hence $S^* = BA$. From $S = AB = \lambda BA$, we get $S = \lambda S^*$. Then $SS^* = \lambda(S^*)^2 = S^*S$. Thus AB is normal and $SS^* = S^*S$. Hence AB and BA are normal commuting operators. $\square$

Theorem 2.2. *Let $A \in \mathcal{L}_0(\mathcal{E}_\omega)$ and let $B \in \mathcal{L}(\mathcal{E}_\omega)$ and $\lambda \in \mathbb{K}^*$ with $AB = \lambda BA \neq 0$, $A^*B = BA^*$ and $A^*A = I$ with $B \neq 0$. Then $\lambda = 1$.*

P r o o f. From $AB = \lambda BA$, hence $A^*AB = \lambda A^*BA$. Since $A^*B = BA^*$ and $A^*A = I$, we get $B = \lambda A^*BA = \lambda BA^*A = \lambda B$. Hence $(\lambda - 1)B = 0$. Since $B \neq 0$, we get $\lambda = 1$. $\square$

Proposition 2.14. *Let $A, B \in \mathcal{L}_0(\mathcal{E}_\omega)$. If there is unitary operators $U, V \in \mathcal{L}_0(\mathcal{E}_\omega)$ with $AB = UB^*A^*$ and $BA = VA^*B^*$. Then AB and BA are normal.*

P r o o f. By Lemma 2.5, we get $UB^*A^* = B^*A^*U$. Then

$$AB(AB)^* = ABB^*A^* = UB^*A^*B^*A^* = B^*A^*UB^*A^* = B^*A^*AB = (AB)^*AB$$

and

$$(BA)^*BA = A^*B^*BA = BAV^*BA = BABAV^* = BA(BA)^*.$$

$\square$

Definition 2.3. [22] Let $A \in \mathcal{L}(\mathcal{E})$, A is said to be bounded below if for each $x \in \mathcal{E}$, $M\|x\| \leq \|Ax\|$ for some $M > 0$.

We have the following statement.

Theorem 2.3. Let $A, B \in \mathcal{L}(\mathcal{E})$ and $\lambda \in \mathbb{K}^*$ such that $AB = \lambda BA \neq 0$. Then AB is bounded below if and only if A and B are bounded below.

Proof. Suppose that AB is bounded below and $AB = \lambda BA \neq 0$. Then there is $M > 0$ with for each $x \in \mathcal{E}$,

$$M\|x\| \leq \|ABx\| \leq \|A\|\|Bx\|.$$

Hence B is bounded below. Since $M\|x\| \leq \|ABx\|$ for any $x \in \mathcal{E}$, and $AB = \lambda BA \neq 0$, it follows that for every $x \in \mathcal{E}$, $\frac{M}{|\lambda|\|B\|}\|x\| \leq \|Ax\|$. Consequently, A is bounded below.

Conversely, it is easy to see that if A and B are bounded below, then AB is bounded below. $\square$

3. Some properties of ultrametric operator equations

In this section, let $A, B \in \mathcal{L}(\mathcal{E})$. We shall study the operator equations $AS = SB$ and $ASB = S$ for some $S \in \mathcal{L}(\mathcal{E})$. We continue with the following results.

Lemma 3.1. Let $A, B, S \in \mathcal{L}(\mathcal{E})$ such that $AS = SB$ and $ASB = S$. If $A^2 - I$ or $I - B^2$ is invertible, then $S = 0$.

Proof. From $AS = SB$ and $ASB = S$, we have $S = SB^2$. Then $S(I - B^2) = 0$. Since $I - B^2$ is invertible, we conclude that $S = 0$. Similarly, one can see that $(A^2 - I)S = 0$. From $A^2 - I$ is invertible, then $S = 0$. $\square$

Further, $R(S)$ denotes the range of S dense in $\mathcal{E}$, i.e. $\overline{R(S)} = \mathcal{E}$.

Proposition 3.1. Let $A, B, S \in \mathcal{L}(\mathcal{E})$ and $\overline{R(S)} = \mathcal{E}$ such that $AS = SB$ and $ASB = S$. Then $A^2 = I$.

Proof. From $AS = SB$ and $ASB = S$, then $(A^2 - I)S = 0$. Hence $R(S) \subseteq N(A^2 - I)$. Since $\overline{R(S)} = \mathcal{E}$, we get $A^2 = I$. $\square$

One can see the following:

Lemma 3.2. Let $A, B, S \in \mathcal{L}(\mathcal{E})$ with $ASB = S$. If S is one to one, then B is one to one.

Proof. It follows by S is one to one and $N(ASB) = N(S)$. $\square$

Theorem 3.1. Let $A, B \in \mathcal{L}(\mathcal{E})$ such that A is injective and $R(B)$ is dense. If $A^2S = SB^2$ and $A^3S = SB^3$, then $AS = SB$ for some $S \in \mathcal{L}(\mathcal{E})$.

Proof. Set $U = AS$ and $V = SB$. Using $A^2S = SB^2$ and $A^3S = SB^3$. We get $AU = VB$ and $A^2U = VB^2$. Then $A(AU) = AVB = (UB)B$, thus $(AV - VB)B = 0$. By $R(B)$ is dense, we get $B \neq 0$ then $AV = VB$. From $AU = VB = AV$, we get $A(U - V) = 0$. From A is injective, then $U = V$. Hence $AS = SB$. $\square$

Theorem 3.2. *Let $A, B \in \mathcal{L}(\mathcal{E})$ with A is injective and $R(B)$ is dense. If $A^2SB^2 = S$ and $A^3SB^3 = S$, then $ASB = S$ for some $S \in \mathcal{L}(\mathcal{E})$.*

P r o o f. From $A^2SB^2 = S$ and $A^3SB^3 = S$, then $A^2SB^2 = A^3SB^3$. Hence $A^2SB^2 - A^3SB^3 = 0$. Thus $A(A^2SB^2 - ASB)B = 0$. From A is injective and $R(B)$ is dense, we obtain that $A^2SB^2 - ASB = 0$. Thus $ASB = S$. $\square$

4. Left (right) pseudospectrum and left (right) condition pseudospectrum of bounded linear operators on ultrametric Banach spaces

We introduce the following definitions.

D e f i n i t i o n 4.1. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$ and let $A \in \mathcal{L}(\mathcal{E})$.

- (i) A is said to be left invertible if there exists $B \in \mathcal{L}(\mathcal{E})$ such that $BA = I$.
- (ii) A is said to be right invertible if there exists $C \in \mathcal{L}(\mathcal{E})$ such that $AC = I$.

D e f i n i t i o n 4.2. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$. Let $A \in \mathcal{L}(\mathcal{E})$, the left spectrum $\sigma^l(A)$ of A is defined by

$$\sigma^l(A) = \{\lambda \in \mathbb{K} : A - \lambda I \text{ is not left invertible in } \mathcal{L}(\mathcal{E})\}.$$

D e f i n i t i o n 4.3. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$. Let $A \in \mathcal{L}(\mathcal{E})$, the right spectrum $\sigma^r(A)$ of A is defined by

$$\sigma^r(A) = \{\lambda \in \mathbb{K} : A - \lambda I \text{ is not right invertible in } \mathcal{L}(\mathcal{E})\}.$$

D e f i n i t i o n 4.4. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, the left spectrum $\sigma_\varepsilon^l(A)$ of A is defined by

$$\sigma_\varepsilon^l(A) = \sigma^l(A) \cup \{\lambda \in \mathbb{K} : \inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

with the convention $\inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} = \infty$ if $A - \lambda I$ is not left invertible.

D e f i n i t i o n 4.5. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, the right spectrum $\sigma_\varepsilon^r(A)$ of A is defined by

$$\sigma_\varepsilon^r(A) = \sigma^r(A) \cup \{\lambda \in \mathbb{K} : \inf\{\|C_r\| : C_r \text{ a right inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

with the convention $\inf\{\|C_r\| : C_r \text{ a right inverse of } A - \lambda I\} = \infty$ if $A - \lambda I$ is not right invertible.

We obtain the following results.

R e m a r k 4.1. From Definition 4.4 and Definition 4.5, we get

$$\sigma^l(A) \subset \sigma_\varepsilon^l(A) \subset \sigma_\varepsilon(A)$$

and

$$\sigma^r(A) \subset \sigma_\varepsilon^r(A) \subset \sigma_\varepsilon(A).$$

Proposition 4.1. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, we have*

$$(i) \quad \sigma^l(A) = \bigcap_{\varepsilon > 0} \sigma_\varepsilon^l(A) \text{ and } \sigma^r(A) = \bigcap_{\varepsilon > 0} \sigma_\varepsilon^r(A).$$

(ii) *For all ε_1 and ε_2 such that $0 < \varepsilon_1 < \varepsilon_2$, $\sigma^l(A) \subset \sigma_{\varepsilon_1}^l(A) \subset \sigma_{\varepsilon_2}^l(A)$ and $\sigma^r(A) \subset \sigma_{\varepsilon_1}^r(A) \subset \sigma_{\varepsilon_2}^r(A)$.*

Proof. (i) From Definition 4.4, for any $\varepsilon > 0$, $\sigma^l(A) \subset \sigma_\varepsilon^l(A)$. Conversely, if $\lambda \in \bigcap_{\varepsilon > 0} \sigma_\varepsilon^l(A)$, hence for all $\varepsilon > 0$, $\lambda \in \sigma_\varepsilon^l(A)$. If $\lambda \notin \sigma^l(A)$, then

$$\lambda \in \{\lambda \in \mathbb{K} : \inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

taking limits as $\varepsilon \rightarrow 0^+$, we get $\inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} = \infty$. Thus $\lambda \in \sigma^l(A)$. Similarly, we obtain $\sigma^r(A) = \bigcap_{\varepsilon > 0} \sigma_\varepsilon^r(A)$.

(ii) For ε_1 and ε_2 such that $0 < \varepsilon_1 < \varepsilon_2$. Let $\lambda \in \sigma_{\varepsilon_1}^l(A)$, then

$$\inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} > \varepsilon_1^{-1} > \varepsilon_2^{-1},$$

hence $\lambda \in \sigma_{\varepsilon_2}^l(A)$. Similarly, we have $\sigma^r(A) \subset \sigma_{\varepsilon_1}^r(A) \subset \sigma_{\varepsilon_2}^r(A)$. $\square$

Proposition 4.2. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$. Then*

$$\bigcup_{C \in \mathcal{L}(\mathcal{E}) : \|C\| < \varepsilon} \sigma^l(A + C) \subset \sigma_\varepsilon^l(A). \quad (4.1)$$

Proof. If $\lambda \in \bigcup_{C \in \mathcal{L}(\mathcal{E}) : \|C\| < \varepsilon} \sigma^l(A + C)$. We argue by contradiction. Suppose that $\lambda \notin \sigma_\varepsilon^l(A)$, hence $\lambda \notin \sigma^l(A)$ and $\inf\{\|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} \leq \varepsilon^{-1}$, thus $\|CC_l\| < 1$. Let D defined on $\mathcal{E}$ by

$$D = \sum_{n=0}^{\infty} C_l(-CC_l)^n.$$

One can see that D is well-defined and $D = C_l(I + CC_l)^{-1}$. Hence for all $y \in \mathcal{E}$, $D(I + CC_l)y = C_ly$. Set $y = (A - \lambda I)x$, we have for all $x \in \mathcal{E}$,

$$x = D(I + CC_l)(A - \lambda I)x = D(A - \lambda I + CC_l(A - \lambda I))x = D(A - \lambda I + C)x.$$

Hence $A + C - \lambda I$ is left invertible which is contradiction with $\lambda \in \bigcup_{C \in \mathcal{L}(\mathcal{E}) : \|C\| < \varepsilon} \sigma^l(A + C)$. Thus, (4.1) holds. $\square$

Theorem 4.1. *Let $\mathcal{E}$ be an ultrametric Banach space over a spherically complete field $\mathbb{K}$ such that $\|\mathcal{E}\| \subseteq |\mathbb{K}|$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$. Then,*

$$\sigma_\varepsilon^l(A) = \bigcup_{C \in \mathcal{L}(\mathcal{E}) : \|C\| < \varepsilon} \sigma^l(A + C).$$

P r o o f. According to the proposition 4.2 the embedding (inclusion) (4.1) is satisfied.

Conversely, suppose that $\lambda \in \sigma_\varepsilon^l(A)$. We discuss two cases.

First case: If $\lambda \in \sigma^l(A)$, we may set $C = 0$.

Second case: Assume that $\lambda \in \sigma_\varepsilon^l(A)$ and $\lambda \notin \sigma^l(A)$, then for all C_l a left inverse of $A - \lambda I$, we have $\|C_l\| > \frac{1}{\varepsilon}$. Hence, there exists $y \in \mathcal{E} \setminus \{0\}$ such that

$$\frac{\|C_l y\|}{\|y\|} > \frac{1}{\varepsilon}. \quad (4.2)$$

Set $y = (A - \lambda I)x$, then $C_l y = x$. From (4.2), we have $\|(A - \lambda I)x\| < \varepsilon \|x\|$. Since $\|\mathcal{E}\| \subseteq |\mathbb{K}|$, then there exists $c \in \mathbb{K} \setminus \{0\}$ such that $|c| = \|x\|$. Putting $z = c^{-1}x$, then $\|z\| = 1$, hence $\|(A - \lambda I)z\| < \varepsilon$. By Theorem 1.1, there exists $\phi \in \mathcal{E}^*$ such that $\phi(z) = 1$ and $\|\phi\| = \|z\|^{-1} = 1$. Define

$$\text{for all } y \in \mathcal{E}, \quad Cy = -\phi(y)(A - \lambda I)z.$$

Then $C \in \mathcal{L}(\mathcal{E})$ and $\|C\| < \varepsilon$, since for all $y \in \mathcal{E}$,

$$\|Cy\| = \|\phi(y)\| \|(A - \lambda I)z\| < \varepsilon \|y\|.$$

Furthermore, we have $(A - \lambda I + C)z = 0$. Thus $A - \lambda I + C$ is not left invertible. Consequently, $\lambda \in \bigcup_{C \in \mathcal{L}(\mathcal{E}): \|C\| < \varepsilon} \sigma^l(A + C)$. $\square$

We continue with the following definitions.

Definition 4.6. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, the left condition pseudospectrum $\Lambda_\varepsilon^l(A)$ of A is defined by

$$\Lambda_\varepsilon^l(A) = \sigma^l(A) \cup \{\lambda \in \mathbb{K} : \inf\{\|(A - \lambda I)\| \|D_l\| : D_l \text{ a left inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

with the convention $\inf\{\|(A - \lambda I)\| \|D_l\| : D_l \text{ a left inverse of } A - \lambda I\} = \infty$ if $A - \lambda I$ is not left invertible.

Definition 4.7. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, the right condition pseudospectrum $\Lambda_\varepsilon^r(A)$ of A is defined by

$$\Lambda_\varepsilon^r(A) = \sigma^r(A) \cup \{\lambda \in \mathbb{K} : \inf\{\|A - \lambda I\| \|D_r\| : D_r \text{ a right inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

with the convention $\inf\{\|A - \lambda I\| \|D_r\| : D_r \text{ a right inverse of } A - \lambda I\} = \infty$ if $A - \lambda I$ is not right invertible.

We have the following results.

Remark 4.2. From Definition 4.6 and Definition 4.7, we get

$$\sigma^l(A) \subset \Lambda_\varepsilon^l(A) \subset \Lambda_\varepsilon(A)$$

and

$$\sigma^r(A) \subset \Lambda_\varepsilon^r(A) \subset \Lambda_\varepsilon(A).$$

Proposition 4.3. Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$, we have

$$(i) \quad \sigma^l(A) = \bigcap_{\varepsilon > 0} \Lambda_\varepsilon^l(A) \quad \text{and} \quad \sigma^r(A) = \bigcap_{\varepsilon > 0} \Lambda_\varepsilon^r(A).$$

$$(ii) \quad \text{For all } \varepsilon_1 \text{ and } \varepsilon_2 \text{ such that } 0 < \varepsilon_1 < \varepsilon_2, \quad \sigma^l(A) \subset \Lambda_{\varepsilon_1}^l(A) \subset \Lambda_{\varepsilon_2}^l(A) \quad \text{and} \quad \sigma^r(A) \subset \Lambda_{\varepsilon_1}^r(A) \subset \Lambda_{\varepsilon_2}^r(A).$$

P r o o f. (i) From Definition 4.6, for any $\varepsilon > 0$, $\sigma^l(A) \subset \Lambda_\varepsilon^l(A)$. Conversely, if $\lambda \in \bigcap_{\varepsilon > 0} \Lambda_\varepsilon^l(A)$, hence for all $\varepsilon > 0$, $\lambda \in \Lambda_\varepsilon^l(A)$. If $\lambda \notin \sigma^l(A)$, then

$$\lambda \in \{\lambda \in \mathbb{K} : \inf\{\|A - \lambda I\| \|D_l\| : D_l \text{ a left inverse of } A - \lambda I\} > \varepsilon^{-1}\},$$

taking limits as $\varepsilon \rightarrow 0^+$, we get $\inf\{\|A - \lambda I\| \|D_r\| : D_r \text{ a left inverse of } A - \lambda I\} = \infty$. Hence $\lambda \in \sigma^l(A)$. Similarly, we obtain $\sigma^r(A) = \bigcap_{\varepsilon > 0} \Lambda_\varepsilon^r(A)$.

(ii) For ε_1 and ε_2 such that $0 < \varepsilon_1 < \varepsilon_2$. Let $\lambda \in \Lambda_{\varepsilon_1}^l(A)$, then

$$\inf\{\|A - \lambda I\| \|D_l\| : D_l \text{ a left inverse of } A - \lambda I\} > \varepsilon_1^{-1} > \varepsilon_2^{-1},$$

hence $\lambda \in \Lambda_{\varepsilon_2}^l(A)$. Similarly, we have $\sigma^r(A) \subset \Lambda_{\varepsilon_1}^r(A) \subset \Lambda_{\varepsilon_2}^r(A)$. $\square$

P r o p o s i t i o n 4.4. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$ and let $A \in \mathcal{L}(\mathcal{E})$ and for every $\varepsilon > 0$ and $\|A - \lambda I\| \neq 0$. Then,*

$$(i) \quad \lambda \in \Lambda_\varepsilon^l(A) \text{ if, and only if, } \lambda \in \sigma_{\varepsilon\|A-\lambda I\|}^l(A).$$

$$(ii) \quad \lambda \in \sigma_\varepsilon^l(A) \text{ if and only if } \lambda \in \Lambda_{\frac{\varepsilon}{\|A-\lambda I\|}}^l(A).$$

P r o o f. (i) Let $\lambda \in \Lambda_\varepsilon^l(A)$, then $\lambda \in \sigma^l(A)$ or

$$\inf\{\|(A - \lambda I)\| \|C_l\| : C_l \text{ a left inverse of } A - \lambda I\} > \varepsilon^{-1}.$$

Hence $\lambda \in \sigma^l(A)$ or for all C_l a left invertible of $A - \lambda I$, $\|C_l\| > \frac{1}{\varepsilon\|(A-\lambda I)\|}$. Consequently, $\lambda \in \sigma_{\varepsilon\|A-\lambda I\|}^l(A)$. The converse is similar.

(ii) Let $\lambda \in \sigma_\varepsilon^l(A)$, then, $\lambda \in \sigma^l(A)$ or for all C_l a left inverse of $A - \lambda I$, $\|C_l\| > \varepsilon^{-1}$. Thus $\lambda \in \sigma^l(A)$ or for all C_l a left inverse of $A - \lambda I$, $\|(A - \lambda I)\| \|C_l\| > \varepsilon^{-1}\|(A - \lambda I)\|$. Then, $\lambda \in \Lambda_{\frac{\varepsilon}{\|A-\lambda I\|}}^l(A)$.

The converse is similar. $\square$

One can see the following corollary.

Corollary 4.1. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$. If $\alpha, \beta \in \mathbb{K}$ with $\beta \neq 0$, then $\Lambda_\varepsilon^l(\beta A + \alpha I) = \alpha + \beta \Lambda_\varepsilon^l(A)$.*

P r o p o s i t i o n 4.5. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ such that $A \neq \lambda I$ and $C_A = \inf\{\|A - \lambda I\| : \lambda \in \mathbb{K}\}$ and $\varepsilon > 0$. Then $\sigma_\varepsilon^l(A) \subset \Lambda_{\frac{\varepsilon}{C_A}}^l(A)$.*

P r o o f. Let $\mu \in \sigma_\varepsilon^l(A)$, then $\mu \in \sigma^l(A)$ or for all C_l a left inverse of $A - \mu I$, $\|C_l\| > \varepsilon^{-1}$. Since $\|A - \mu I\| \geq C_A > 0$. Then $\mu \in \sigma^l(A)$ or for all C_l a left inverse of $A - \mu I$, $\|A - \mu I\| \|C_l\| > \varepsilon^{-1}C_A$. Hence $\lambda \in \Lambda_{\frac{\varepsilon}{C_A}}^l(A)$. $\square$

Lemma 4.1. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$. If $\lambda \in \Lambda^l(A) \setminus \sigma^l(A)$. Then there exists $x \in \mathcal{E} \setminus \{0\}$ such that $\|(A - \lambda I)x\| < \varepsilon \|A - \lambda I\| \|x\|$.*

P r o o f. If $\lambda \in \Lambda^l(A) \setminus \sigma^l(A)$, then for all C_l a left inverse of $A - \lambda I$, we have

$$\|A - \lambda I\| \|C_l\| > \frac{1}{\varepsilon}.$$

Thus

$$\|C_l\| > \frac{1}{\varepsilon \|A - \lambda I\|}.$$

Then there exists $y \in \mathcal{E} \setminus \{0\}$ such that

$$\frac{\|C_l y\|}{\|y\|} > \frac{1}{\varepsilon \|A - \lambda I\|}. \quad (4.3)$$

Set $y = (A - \lambda I)x$, then $C_l y = x$. From (4.3), we have $\|(A - \lambda I)x\| < \varepsilon \|A - \lambda I\| \|x\|$. $\square$

Theorem 4.2. *Let $\mathcal{E}$ be an ultrametric Banach space over $\mathbb{K}$, let $A \in \mathcal{L}(\mathcal{E})$, $\lambda \in \mathbb{K}$ and $\varepsilon > 0$. If there exists $C \in \mathcal{L}(\mathcal{E})$ with $\|C\| < \varepsilon \|A - \lambda I\|$ and $\lambda \in \sigma^l(A + C)$. Then, $\lambda \in \Lambda_\varepsilon^l(A)$.*

P r o o f. Assume that there exists $C \in \mathcal{L}(\mathcal{E})$ such that

$$\|C\| < \varepsilon \|A - \lambda I\| \text{ and } \lambda \in \sigma^l(A + C).$$

If $\lambda \notin \Lambda_\varepsilon^l(A)$, hence $\lambda \notin \sigma^l(A)$ and for each C_l a left inverse of $A - \lambda I$, $\|A - \lambda I\| \|C_l\| \leq \varepsilon^{-1}$. Consider D defined on $\mathcal{E}$ by

$$D = \sum_{n=0}^{\infty} C_l (-CC_l)^n.$$

Consequently $D = C_l(I + CC_l)^{-1}$. Hence for all $y \in \mathcal{E}$, $D(I + CC_l)y = C_l y$. Put $y = (A - \lambda I)x$, then

$$(\forall x \in \mathcal{E}) \ D(A - \lambda I + C)x = x$$

Then $A - \lambda I + C$ is a left invertible which is a contradiction. Thus $\lambda \in \Lambda_\varepsilon^l(A)$. $\square$

Set $\mathcal{C}_\varepsilon(\mathcal{E}) = \{C \in \mathcal{L}(\mathcal{E}) : \|C\| < \varepsilon \|A - \lambda I\|\}$, we have.

Theorem 4.3. *Let $\mathcal{E}$ be an ultrametric Banach space over a spherically complete field $\mathbb{K}$ such that $\|\mathcal{E}\| \subseteq |\mathbb{K}|$, let $A \in \mathcal{L}(\mathcal{E})$ and $\varepsilon > 0$. Then,*

$$\Lambda_\varepsilon^l(A) = \bigcup_{C \in \mathcal{C}_\varepsilon(\mathcal{E})} \sigma^l(A + C).$$

P r o o f. By Theorem 4.2, we have $\bigcup_{C \in \mathcal{C}_\varepsilon(\mathcal{E})} \sigma^l(A + C) \subset \Lambda_\varepsilon^l(A)$. Conversely, assume that $\lambda \in \Lambda_\varepsilon^l(A)$. If $\lambda \in \sigma^l(A)$, we may put $C = 0$. If $\lambda \in \Lambda_\varepsilon^l(A)$ and $\lambda \notin \sigma^l(A)$. By Lemma 4.1 and $\|\mathcal{E}\| \subseteq |\mathbb{K}|$, there exists $x \in \mathcal{E} \setminus \{0\}$ such that $\|x\| = 1$ and $\|(A - \lambda I)x\| < \varepsilon \|A - \lambda I\|$.

By Theorem 1.1, there is $\varphi \in \mathcal{E}^*$ such that $\varphi(x) = 1$ and $\|\varphi\| = \|x\|^{-1} = 1$. Consider C on $\mathcal{E}$ defined by for all $y \in X$, $Cy = -\phi(y)(A - \lambda I)x$. Hence, $\|C\| < \varepsilon \|A - \lambda I\|$ and $D(C) = \mathcal{E}$. Moreover, for $x \in \mathcal{E} \setminus \{0\}$, $(A - \lambda I + C)x = 0$. Then, $(A - \lambda I + C)$ is not left invertible. Consequently, $\lambda \in \bigcup_{C \in \mathcal{C}_\varepsilon(\mathcal{E})} \sigma^l(A + C)$. $\square$

5. Determinant spectrum of non-Archimedean polynomial pencils

From Proposition 1 and Theorem 2 and Theorem 3 of [13], we get.

Theorem 5.1. *Let $C \in \mathcal{M}_n(\mathbb{K})$. Hence,*

- (i) *If $0 < \varepsilon_1 \leq \varepsilon_2$, $Tr_{\varepsilon_1}(C) \subset Tr_{\varepsilon_2}(C)$,*
- (ii) *If $\beta \in \mathbb{K}$ and $\alpha \in \mathbb{K} \setminus \{0\}$, hence $Tr_{\varepsilon}(\alpha C + \beta I) = \alpha Tr_{\frac{\varepsilon}{|\alpha|}}(C) + \beta$,*
- (iii) *For any $\alpha, \lambda \in \mathbb{K}$, we have $Tr_{\varepsilon}(\alpha I) = \left\{ \lambda \in \mathbb{K} : |\lambda - \alpha| \leq \frac{\varepsilon}{|n|} \right\}$.*

Theorem 5.2. *Let $C, S, A \in \mathcal{M}_n(\mathbb{K})$ and $\varepsilon > 0$. If $S = ACA^{-1}$, then $Tr_{\varepsilon}(S) = Tr_{\varepsilon}(C)$.*

Theorem 5.3. *Let $C \in \mathcal{M}_n(\mathbb{K})$, hence for any $\varepsilon > 0$,*

$$Tr_{\delta}(C) + B_f(0, \frac{\varepsilon}{|n|}) \subseteq Tr_{\gamma}(C)$$

with $\gamma = \max\{\varepsilon, \delta\}$, if $\delta < \varepsilon$, we get

$$Tr_{\delta}(C) + B_f(0, \frac{\varepsilon}{|n|}) \subseteq Tr_{\varepsilon}(C).$$

We have the following example.

Example 5.1. If

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Then for any $\varepsilon > 0$,

$$Tr_{\varepsilon}(A) = \{1, 2\} \cup \{\lambda \in \mathbb{Q}_p : |3 - 2\lambda|_p \leq \varepsilon\}.$$

By Definition 5 of [13], we get.

Definition 5.1. Let $C \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$. Then the ε -trace set $tr_{\varepsilon}(C)$ of C is

$$tr_{\varepsilon}(C) = \{\lambda \in \mathbb{K} : |Tr(C - \lambda I)| \leq \varepsilon\}.$$

From Remark 1, Theorem 4, Proposition 2 and Proposition 3 of [13], we get.

Remark 5.1. For each $\varepsilon > 0$, $tr_{\varepsilon}(C) \subseteq Tr_{\varepsilon}(C)$.

Theorem 5.4. *If $B, C \in \mathcal{M}_n(\mathbb{K})$. Then, for any $\varepsilon > 0$,*

- (i) $tr_{\varepsilon}(BC) = tr_{\varepsilon}(CB)$,
- (ii) $tr_{\varepsilon}(B) + tr_{\varepsilon}(C) \subseteq tr_{\varepsilon}(B + C)$.

Proposition 5.1. *Let $C \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$, if $\lambda, \mu \in tr_{\varepsilon}(C)$ and $\alpha \in \mathbb{K}$ with $|\alpha| \leq 1$. Then $\alpha\mu + (1 - \alpha)\lambda \in tr_{\varepsilon}(C)$.*

Proposition 5.2. *Let $C \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$ with $\|C\| < \varepsilon$. If $\lambda, \mu \in tr_{\varepsilon}(C)$, then $\lambda - \mu \in tr_{\varepsilon}(C)$.*

The following propositions are valid.

Proposition 5.3. *Let $B \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$ and $\varepsilon > 0$. If there exists $C \in \mathcal{M}_n(\mathbb{K})$ with $|Tr(C)| \leq \varepsilon$ and $Tr(B - \lambda I - C) = 0$, then $\lambda \in tr_\varepsilon(B)$.*

Proof. Since $Tr(B - \lambda I - C) = 0$ and $|Tr(C)| \leq \varepsilon$, hence $Tr(B - \lambda I) = Tr(C)$ and $|Tr(C)| \leq \varepsilon$, then $|Tr(B - \lambda I)| = |Tr(C)| \leq \varepsilon$. Hence $\lambda \in tr_\varepsilon(B)$. $\square$

Proposition 5.4. *Let $A \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$ and $\varepsilon > 0$. If there is $C \in \mathcal{M}_n(\mathbb{K})$ with $|Tr(C)| > \varepsilon$ and $Tr(A - \lambda I - C) = 0$, then $\lambda \notin tr_\varepsilon(A)$.*

Proof. Assume that there is $C \in \mathcal{M}_n(\mathbb{K})$ with $|Tr(C)| > \varepsilon$ and $Tr(A - \lambda I - C) = 0$. If $\lambda \in tr_\varepsilon(A)$, thus $|Tr(A - \lambda I)| = |Tr(C)| \leq \varepsilon$ which is contradiction with $|Tr(C)| > \varepsilon$ and $Tr(A - \lambda I - C) = 0$. $\square$

Proposition 5.5. *Let $A \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$ and $\varepsilon > 0$. If $\lambda \notin tr_\varepsilon(A)$, then there exists $C \in \mathcal{M}_n(\mathbb{K})$ with $|Tr(C)| > \varepsilon$ and $Tr(A - \lambda I - C) = 0$.*

Proof. If $\lambda \notin tr_\varepsilon(A)$, hence $|Tr(A - \lambda I)| > \varepsilon$. Set $C = \frac{Tr(A - \lambda I)}{n}I$. Thus $C \in \mathcal{M}_n(\mathbb{K})$ and $|Tr(C)| = |Tr(\frac{Tr(A - \lambda I)}{n}I)| = |\frac{Tr(A - \lambda I)}{n}Tr(I)| = |\frac{Tr(A - \lambda I)}{n}n| = |Tr(A - \lambda I)| > \varepsilon$. $\square$

Proposition 5.6. *Let $A \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$ and $\varepsilon > 0$. If $\lambda \in tr_\varepsilon(A)$, then there exists $C \in \mathcal{M}_n(\mathbb{K})$ with $|Tr(C)| \leq \varepsilon$ and $Tr(A - \lambda I - C) = 0$.*

Proof. If $\lambda \in tr_\varepsilon(A)$, hence $|Tr(A - \lambda I)| \leq \varepsilon$. Set $C = \frac{Tr(A - \lambda I)}{n}I$. Thus $C \in \mathcal{M}_n(\mathbb{K})$ and $|Tr(C)| = |Tr(\frac{Tr(A - \lambda I)}{n}I)| = |\frac{Tr(A - \lambda I)}{n}Tr(I)| = |\frac{Tr(A - \lambda I)}{n}n| = |Tr(A - \lambda I)| \leq \varepsilon$. $\square$

From Definition 6 and Remark 2 of [13], we get.

Definition 5.2. Let $C \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$, the ε -determinant spectrum $d_\varepsilon(C)$ of C is the set

$$d_\varepsilon(C) = \{\lambda \in \mathbb{K} : |\det(C - \lambda I)| \leq \varepsilon\}.$$

Remark 5.2. If $C \in \mathcal{M}_n(\mathbb{K})$, then for any $\varepsilon > 0$, $\sigma(C) \subseteq d_\varepsilon(C)$ and $d_0 = \sigma(C)$.

Using Proposition 4 of [13], we get.

Proposition 5.7. *Let $C \in \mathcal{M}_n(\mathbb{K})$. Then for any $\varepsilon > 0$,*

- (i) $\sigma(C) = \bigcap_{\varepsilon > 0} d_\varepsilon(C)$,
- (ii) *For any $0 < \varepsilon_1 \leq \varepsilon_2$, $d_{\varepsilon_1}(C) \subseteq d_{\varepsilon_2}(C)$.*

We get:

Example 5.2. Let

$$C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Then for any $\varepsilon > 0$, $d_\varepsilon(C) = \{\mu \in \mathbb{Q}_p : |\mu(\mu - 2)|_p \leq \varepsilon\}$.

Example 5.3. Let $a, b \in \mathbb{Q}_p^*$, $c \in \mathbb{Q}_p$ and

$$C = \begin{pmatrix} a & c \\ 0 & b \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Hence for any $\varepsilon > 0$, $d_\varepsilon(C) = \{\lambda \in \mathbb{Q}_p : |a - \lambda|_p |b - \lambda|_p \leq \varepsilon\}$.

E x a m p l e 5.4. Let

$$C = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Thus for any $\varepsilon > 0$, $d_\varepsilon(C) = \{\lambda \in \mathbb{Q}_p : |\lambda^2 - \lambda \text{Tr}(C) + \det(C)|_p \leq \varepsilon\}$.

E x a m p l e 5.5. Let

$$C = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Then for any $\varepsilon > 0$,

$$d_\varepsilon(C) = \{\lambda \in \mathbb{Q}_p : |\lambda^2|_p \leq \varepsilon\}.$$

We have the following propositions.

P r o p o s i t i o n 5.8. *Let $D_1, D_2 \in \mathcal{L}(\mathbb{Q}_p^n)$ be two diagonal operators with for each $i \in \{1, \dots, n\}$, $D_1 e_i = \lambda_i e_i$ and $D_2 e_i = \mu_i e_i$ with $\lambda_i, \mu_i \in \mathbb{Q}_p$, $\lambda_i \neq \lambda_{i+1}$ and $\mu_i \neq \mu_{i+1}$. Then $d_\varepsilon(D_1, D_2) = \{\mu \in \mathbb{Q}_p : |\lambda_1 - \mu\mu_1|_p \dots |\lambda_n - \mu\mu_n|_p \leq \varepsilon\}$.*

P r o o f. For each $i \in \{1, \dots, n\}$, $(D_1 - \lambda D_2)e_i = (\lambda_i - \lambda\mu_i)e_i$ where $(e_j)_{1 \leq j \leq n}$ is a basis of $\mathbb{Q}_p^n$. Hence, $|\det(D_1 - \lambda D_2)|_p = |\lambda_1 - \lambda\mu_1|_p \dots |\lambda_n - \lambda\mu_n|_p$. Consequently for any $\varepsilon > 0$,

$$\begin{aligned} d_\varepsilon(D_1, D_2) &= \{\mu \in \mathbb{Q}_p : |\det(D_1 - \mu D_2)|_p \leq \varepsilon\} \\ &= \{\mu \in \mathbb{Q}_p : |\mu\mu_1 - \lambda_1|_p \dots |\mu\mu_n - \lambda_n|_p \leq \varepsilon\}. \end{aligned}$$

□

P r o p o s i t i o n 5.9. *Let $C \in \mathcal{M}_n(\mathbb{K})$ be invertible and $\lambda \in \mathbb{K} \setminus \{0\}$. Then for any $\varepsilon > 0$,*

$$\lambda \in d_\varepsilon(C) \text{ if and only if } \lambda^{-1} \in d_{\frac{\varepsilon}{|\det(\lambda C)|}}(C^{-1}). \quad (5.1)$$

P r o o f. By virtue of the relation

$$\det(C - \lambda I) = \det(\lambda C(\lambda^{-1} - C^{-1})) = \det(\lambda C) \det(\lambda^{-1} - C^{-1}),$$

where $\lambda \neq 0$, (5.1) is satisfied. □

P r o p o s i t i o n 5.10. *Let $B, C \in \mathcal{M}_n(\mathbb{K})$ with $\det(B) \neq 0$ and $\varepsilon > 0$. Then $d_\varepsilon(BC) = d_\varepsilon(CB)$.*

P r o o f. Since B is invertible, then

$$\begin{aligned} \det(BC - \lambda I) &= \det(B(C - \lambda B^{-1})) = \det(B) \det(C - \lambda B^{-1}) \\ &= \det(C - \lambda B^{-1}) \det(B) = \det(CB - \lambda I). \end{aligned}$$

Then $\lambda \in d_\varepsilon(BC)$ if and only if $\lambda \in d_\varepsilon(CB)$. □

Now, we consider the problem of the eigenvalue of the polynomial pencil given by

$$P(\lambda)x = 0,$$

where $P(\lambda) = \sum_{k=0}^n \lambda^k A_k$ and $A_k \in \mathcal{M}_n(\mathbb{K})$ and $x \in \mathbb{K}^n$, we introduce the determinant pseudospectrum of polynomial pencils. Set $P(\lambda) = \sum_{k=0}^n \lambda^k A_k$ and $A_k \in \mathcal{M}_n(\mathbb{K})$, we have.

Definition 5.3. Let $P(\lambda) \in \mathcal{M}_n(\mathbb{K})$, the resolvent set $\rho(P(\lambda))$ of the polynomial pencil $P(\lambda)$ is

$$\rho(P(\lambda)) = \{\lambda \in \mathbb{K} : P(\lambda) \text{ is invertible}\},$$

the spectrum $\sigma(P(\lambda))$ of $P(\lambda)$ is $\mathbb{K} \setminus \rho(P(\lambda))$.

Definition 5.4. Let $P(\lambda) \in \mathcal{M}_n(\mathbb{K})$, $\varepsilon > 0$, the ε -determinant spectrum $d_\varepsilon(P(\lambda))$ of the polynomial pencil $P(\lambda)$ is defined by

$$d_\varepsilon(P(\lambda)) = \{\lambda \in \mathbb{K} : |\det(P(\lambda))| \leq \varepsilon\}.$$

Remark 5.3. From the Definition 5.4, if $P(\lambda) \in \mathcal{M}_n(\mathbb{K})$, then for any $\varepsilon > 0$, $\sigma(P(\lambda)) \subseteq d_\varepsilon(P(\lambda))$ and $d_0 = \sigma(P(\lambda))$.

Proposition 5.11. If $P(\lambda) \in \mathcal{M}_n(\mathbb{K})$, then

- (i) $\sigma(P(\lambda)) = \bigcap_{\varepsilon > 0} d_\varepsilon(P(\lambda))$,
- (ii) For all $0 < \varepsilon_1 \leq \varepsilon_2$, we have $d_{\varepsilon_1}(P(\lambda)) \subseteq d_{\varepsilon_2}(P(\lambda))$.

Proof. (i) Obvious.

(ii) Let $0 < \varepsilon_1 \leq \varepsilon_2$ and $\lambda \in d_{\varepsilon_1}(P(\lambda))$. Then $|\det(P(\lambda))| \leq \varepsilon_1 \leq \varepsilon_2$. Hence $\lambda \in d_{\varepsilon_2}(P(\lambda))$. $\square$

Example 5.6. Let

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \in \mathcal{M}_2(\mathbb{Q}_p).$$

Set $P(\lambda) = \lambda^2 A^2 - I$. Then for any $\varepsilon > 0$,

$$d_\varepsilon(P(\lambda)) = \{\lambda \in \mathbb{Q}_p : |(2\lambda - 1)(2\lambda + 1)|_p \leq \varepsilon\}.$$

Let $A, B, C \in \mathcal{M}_n(\mathbb{K})$. We consider $P(\lambda) = \lambda^2 A + \lambda B + C$. For all $\lambda \in \rho(P(\lambda))$, $R(\lambda, P) = (\lambda^2 A + \lambda B + C)^{-1}$.

Proposition 5.12. Let $A, B, C \in \mathcal{M}_n(\mathbb{K})$. If the inclusions $\lambda \in \rho(P)$ and $\mu \in \mathbb{K}$ with $\|R(\lambda, P)((\lambda^2 - \mu^2)A + (\lambda - \mu)B)\| < 1$, then $\mu \in \rho(P)$ and $\|R(\mu, P)\| \leq \|R(\lambda, P)\|$.

Proof. Because

$$\begin{aligned} \mu^2 A + \mu B + C &= \lambda^2 A + \lambda B + C - ((\lambda^2 - \mu^2)A + (\lambda - \mu)B) \\ &= (\lambda^2 A + \lambda B + C) (I - R(\lambda, P)((\lambda^2 - \mu^2)A + (\lambda - \mu)B)), \end{aligned}$$

and $\|R(\lambda, P)((\lambda^2 - \mu^2)A + (\lambda - \mu)B)\| < 1$, there is $R(\mu, P)$ and $\|R(\mu, P)\| \leq \|R(\lambda, P)\|$. $\square$

The next theorem presents the perturbation of operators.

Theorem 5.5. Let $\mathcal{E}$ be a non-Archimedean Banach space over $\mathbb{K}$, let $B, C \in \mathcal{L}(\mathcal{E})$. Let $\mu \in \rho(B, C)$ and $\lambda \in \mathbb{K}$ with $|\lambda - \mu| < \|R(\mu, B, C)C\|^{-1}$, then $\lambda \in \rho(B, C)$ and $\|R(\lambda, B, C)\| \leq \|R(\mu, B, C)\|$ where $R(\lambda, B, C) = (B - \lambda C)^{-1}$.

P r o o f. Let $\mu \in \rho(B, C)$, we have:

$$B - \lambda C = B - \mu C + \mu C - \lambda C = (B - \mu C)(I - (\lambda - \mu)R(\mu, B, C)C).$$

Since $\lambda \in \mathbb{K}$ with $|\lambda - \mu| < \|R(\mu, B, C)C\|^{-1}$, by Lemma 1.1,

$$(I - (\lambda - \mu)R(\mu, B, C)C)^{-1} \text{ and } \|(I - (\lambda - \mu)R(\mu, B, C)C)^{-1}\| \leq 1.$$

Then $\lambda \in \rho(B, C)$ and $R(\lambda, B, C) = (I - (\lambda - \mu)R(\mu, B, C)C)^{-1}R(\mu, B, C)$. Thus $\lambda \in \rho(B, C)$ and $\|R(\lambda, B, C)\| \leq \|R(\mu, B, C)\|$. $\square$

From Theorem 5.5, we have.

Corollary 5.1. $\sigma(B, C)$ is closed in $\mathbb{K}$.

From the results of M. Vishik [23] for $C = I$, there is a nonanalytic resolvent of an operator, for that we assume that application $\lambda \in \rho(B, C) \mapsto R(\lambda, B, C) = (B - \lambda C)^{-1}$ is analytic on $\rho(B, C)$.

Theorem 5.6. Let $\mathcal{E}$ be a non-Archimedean Banach space over an algebraically closed field $\mathbb{K}$, let $B, C \in \mathcal{L}(\mathcal{E})$ with $R(\lambda, B, C)$ is analytic on $\rho(B, C)$. Then

$$\frac{d}{d\lambda} R(\lambda, B, C) = R(\lambda, B, C)CR(\lambda, B, C). \quad (5.2)$$

P r o o f. Let $\lambda \in \rho(B, C)$, let $\mu \in \mathbb{K}$ with $|\lambda - \mu| < \|R(\lambda, B, C)C\|^{-1}$, by Theorem 5.5, we have:

$$\begin{aligned} R(\mu, B, C) &= (I - (\mu - \lambda)R(\lambda, B, C)C)^{-1}R(\lambda, B, C) \\ &= \sum_{k=0}^{\infty} ((\mu - \lambda)R(\lambda, B, C)C)^k R(\lambda, B, C). \end{aligned}$$

Then

$$\begin{aligned} &\left\| \frac{R(\mu, B, C) - R(\lambda, B, C)}{\mu - \lambda} - R(\lambda, B, C)CR(\lambda, B, C) \right\| \\ &= \left\| \sum_{k=2}^{\infty} (\mu - \lambda)^{k-1} (R(\lambda, B, C)C)^k R(\lambda, B, C) \right\| \\ &\leq |\lambda - \mu| \sup_{k \geq 2} \|(\mu - \lambda)^{k-2} (R(\lambda, B, C)C)^k R(\lambda, B, C)\|, \end{aligned}$$

hence

$$\lim_{\mu \rightarrow \lambda} \left\| \frac{R(\mu, B, C) - R(\lambda, B, C)}{\mu - \lambda} - R(\lambda, B, C)CR(\lambda, B, C) \right\| = 0.$$

Therefore, (5.2) is satisfied. $\square$

From Theorem 5.7 and $B = I$, we get.

Theorem 5.7. Let $\mathcal{E}$ be a non-Archimedean Banach space over an algebraically closed field $\mathbb{K}$, let $C \in \mathcal{L}(\mathcal{E})$ such that $R(\lambda, C)$ is analytic on $\rho(C)$. Then

$$\frac{d}{d\lambda} R(\lambda, C) = R(\lambda, C)^2.$$

References

- [1] H. Weyl, “Quantenmechanik und gruppentheorie”, *Z. Physik*, **46** (1927), 1–46.
- [2] J. von Neumann, “Die eindeutigkeit der Schrödingerschen operatoren”, *Mathematische Annalen*, **104** (1931), 570–587.
- [3] P. Busch, P. J. Lahti, P. Mittelstaedt, *The Quantum Theory of Measurement*, Springer-Verlag, Berlin, 1996.
- [4] E. B. Davies, *Quantum Theory of Open Systems*, Academic Press, London–New York, 1976.
- [5] S. Gudder, G. Nagy, “Sequential quantum measurements”, *Journal of Mathematical Physics*, **42**:11 (2001), 5212–5222.
- [6] C. R. Putnam, *Commutation Properties of Hilbert Space Operators and Related Topics*, Springer-Verlag, New York, 1967.
- [7] M. Cho, B. P. Duggal, R. Harte, S. Ôta, “Operator equation $AB = \lambda BA$ ”, *International Math. Forum*, **5**:53–56 (2010), 2629–2637.
- [8] C. Cowen, “Commutants and the operator equation $AX = \lambda XA$ ”, *Pacific J. Math.*, **80**:2 (1979), 337–340.
- [9] J. Yang, H. K. Du, “A note on commutativity up to a factor of bounded operators”, *Proc. Amer. Math. Soc.*, **132**:6 (2004), 1713–1720.
- [10] J. Ettayb, “ λ -commuting of bounded linear operators on ultrametric Banach spaces and determinant spectrum of ultrametric matrices”, *Topological Algebra and its Applications*, **11** (2023), Article number: 20230103.
- [11] A. Ammar, A. Bouchekoua, A. Jeribi, “Pseudospectra in a non-Archimedean Banach space and essential pseudospectra in E_ω ”, *Filomat*, **33**:12 (2019), 3961–3976.
- [12] A. Ammar, A. Bouchekoua, N. Lazrag, “The condition ε -pseudospectra on non-Archimedean Banach space”, *Boletín de la Sociedad Matemática Mexicana*, **28**:2 (2022), 1–24.
- [13] J. Ettayb, “Pseudo-spectrum of non-Archimedean matrix pencils”, *Bull. Transilv. Univ. Braşov. Series III: Mathematics and Computer Science*, **4(66)**:1 (2024), 73–86.
- [14] J. Ettayb, “Ultrametric Fredholm operators and approximate pseudospectrum”, *Arab Journal of Mathematical Sciences*, 2024 (to appear).
- [15] J. Ettayb, “ (N, ε) -pseudospectra of bounded linear operators on ultrametric Banach spaces”, *Gulf Journal of Mathematics*, **17**:1 (2024), 12–28.
- [16] J. Ettayb, “Common properties of the operator equations in ultrametric spectral theory”, *Gulf Journal of Mathematics*, **16**:1 (2024), 79–95.
- [17] J. Ettayb, “Condition pseudospectrum of operator pencils on non-archimedean Banach spaces”, 2023, arXiv: [abs/2305.18401](https://arxiv.org/abs/2305.18401).
- [18] R. A. Horn, C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press & Assessment, Cambridge, 1991.
- [19] K. G. Krishna, “Determinant spectrum: A generalization of eigenvalues”, *Funct. Anal. Approx. Comput.*, **10**:2 (2018), 1–12.
- [20] T. Diagana, F. Ramaroson, *Non-Archimedean Operators Theory*, Springer, Cham, 2016.
- [21] A. C. M. van Rooij, *Non-Archimedean Functional Analysis*, Monographs and Textbooks in Pure and Applied Math., **51**, Marcel Dekker, Inc., New York, 1978.
- [22] J. Ettayb, “Some results on non-Archimedean operators theory”, *Sahand Communications in Mathematical Analysis*, **20**:4 (2023), 139–154.
- [23] M. Vishik, “Non-Archimedean spectral theory”, *J. Sov. Math.*, **30** (1985), 2513–2554.

Information about the author

Jawad Ettayb, Doctor of Mathematics, Professor at Hamman Al-Fatawaki collegiate High School, Regional Academy of Education and Training of Casablanca-Settat, Had Soualem, Morocco. E-mail: jawad.ettayb@gmail.com

ORCID: <https://orcid.org/0000-0002-4819-943X>

Received 03.10.2024

Reviewed 18.11.2024

Accepted for press 22.11.2024

Информация об авторе

Эттайб Джавад, доктор математики, профессор университетской средней школы-колледжа Хаммана Аль-Фатаваки, Региональная академия образования и обучения Касабланки-Сеттата, г. Хад-Суалем, Марокко. E-mail: jawad.ettayb@gmail.com

ORCID: <https://orcid.org/0000-0002-4819-943X>

Поступила в редакцию 03.10.2024 г.

Поступила после рецензирования 18.11.2024 г.

Принята к публикации 22.11.2024 г.