Financial Mathematics and Fintech

Zhiyong Zheng *Editor*

Proceedings of the Second International Forum on Financial Mathematics and Financial Technology

OPEN ACCESS

Springer

Financial Mathematics and Fintech

This series addresses the emerging advances in mathematical theory related to finance and application research from all the fintech perspectives. It is a series of monographs and contributed volumes focusing on the in-depth exploration of financial mathematics such as applied mathematics, statistics, optimization, and scientific computation, and fintech applications such as artificial intelligence, block chain, cloud computing, and big data. This series is featured by the comprehensive understanding and practical application of financial mathematics and fintech. This book series involves cutting-edge applications of financial mathematics and fintech in practical programs and companies.

The Financial Mathematics and Fintech book series promotes the exchange of emerging theory and technology of financial mathematics and fintech between academia and financial practitioner. It aims to provide a timely reflection of the state of art in mathematics and computer science facing to the application of finance. As a collection, this book series provides valuable resources to a wide audience in academia, the finance community, government employees related to finance and anyone else looking to expand their knowledge in financial mathematics and fintech.

The key words in this series include but are not limited to:
a) Financial mathematics
b) Fintech
c) Computer science
d) Artificial intelligence
e) Big data

Zhiyong Zheng
Editor

Proceedings of the Second International Forum on Financial Mathematics and Financial Technology

Springer

Editor
Zhiyong Zheng
Renmin University of China
Beijing, China

ISSN 2662-7167 ISSN 2662-7175 (electronic)
Financial Mathematics and Fintech
ISBN 978-981-99-2365-6 ISBN 978-981-99-2366-3 (eBook)
https://doi.org/10.1007/978-981-99-2366-3

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Financial technology is reshaping the financial industry ecology with explosive growth, making China's financial industry constantly achieve breakthroughs on a new runway. The rapid development of FinTech at the same time triggers the indepth integration between mathematics, finance, and advanced technology.

The Second International Academic Forum on Financial Mathematics and Financial Technology was successfully held online on August 13–15, 2021, jointly held by the School of Mathematics of Renmin University of China, the Engineering Research Center of the Ministry of Financial Computing and Digital Engineering, the Statistics and Big Data Research Institute of Renmin University of China, the Blockchain Research Institute of Renmin University of China, the Zhong guancun Internet Finance Research Institute, and the Renmin University Press. Several distinguished scholars engaged in the interdisciplinary research of mathematics, statistics, information technology, and finance delivered excellent speeches and discussed in depth on the bottlenecks faced by emerging technologies such as big data, AI, cloud computing, and blockchain. This forum has provided insightful understandings on the development frontier and research hotspot of financial mathematics and financial technology, and strengthened the contact between our institute and research institutes from home and abroad.

The proceedings emphasize the selected aspects of current and upcoming trends in FinTech, presenting the innovative mathematical models and state-of-the-art technologies, benefiting both scholars and practitioners in pursuing perfect integration of elegant mathematical models and up-to-date data mining technologies in financial market analysis.

Chapter "On the Development of Fintech in Asia" provides the general overview on the Development of Fintech in Asia. Chapter "A Probability Inequality with Application to Lattice Theory" gives a more precise estimation probability of decryption error about GGH public-key encryption scheme based on the Hoeffding inequality. The upper bound probability could be closed to 0 with applicable parameters, which means that the probability of decryption error for the cryptosystem could be sufficiently small. It is also confirmed that the GGH public-key cryptosystem could have high security. Chapter "Robust Identification of Gene-Environment Interactions

Under High-Dimensional Accelerated Failure Time Models" considers censored survival data and adopt a high-dimensional accelerated failure time (AFT) model for robust identification of gene-environment interactions. Chapter "A Novel Approach for Improving Accuracy for Distributed Storage Networks" propose two approaches, periodic self-verification and user verification, to guarantee the reliability of the storage network while improving efficiency in distributed storage. Chapter "Iterative Learning Control Based on Random Variance Reduction Gradient Method" proposed a novel iterative learning control scheme based on stochastic variance reduced gradient (SVRG), which is not only suitable for resolving the incomplete information problem, but also converges efficiently under both strongly convex and non-strongly convex control objectives. Chapter "A Generalization of NTRUEncrypt" first discusses a more general form of the ordinary cyclic code and gives a generalized construction of NTRU based on ideal matrix and q-ary lattice theory. Compared with other variations of NTRU, such as CTRU, GNTRU, QTRU, and BITRU, the extended NTRU cryptosystem is constructed with general ideal matrix rather than some special algebraic structures. Chapter "Cyclic Lattices, Ideal Lattices, and Bounds for the Smoothing Parameter" shows that ideal lattices are actually a special subclass of cyclic lattices, and proves that there is a one-to-one correspondence between cyclic lattices and finitely generated R-modules. Chapter "On the LWE Cryptosystem with More General Disturbance" gives estimation probability of decryption error based on Gaussian disturbances and proves that the decryption error could be sufficiently small. The most salient innovation and contribution is that for any general disturbances, the decryption error could also be small enough. This indicates high security and reliability of LWE-based cryptosystem. In other words, this cryptosystem is secure enough against passive eavesdroppers and could be applied in many kinds of encryption process. Chapter "On the High Dimensional RSA Algorithm—A Public Key Cryptosystem Based on Lattice and Algebraic Number Theory" proves that high-dimensional RSA is a lattice based on public-key cryptosystem, of which would be considered as a new number in the family of post-quantum cryptography. Moreover, the matrix expression of any algebraic number field is also given, which is a new result even in the sense of classical algebraic number theory. Chapter "Central Bank Digital Currency Cross-Border Payment Model Based on Blockchain Technology" combines the time-series model with fiscal science and puts forward a model for the fiscal budget variance of China's national general public budget. Chapter "LLE Based K-Nearest Neighbor Smoothing for scRNA-Seq Data Imputation" proposed LLE-based k-nearest neighbor smoothing for scRNA-seq data imputation where the data is of high dimensionality, sparse and noisy. Chapter "The Application of Time Series Analysis in the Fiscal Budget Variance of China" is about the application of time-series analysis in the fiscal budget variance of China.

We would like to take this opportunity to thank all the participants at the second International Forum on Financial Mathematics and FinTech. We are also pleased to

thank the support of School of Mathematics, Renmin University of China, and Engineering Research Center of Finance Computation and Digital Engineering, Ministry of Education.

Beijing, China Zhiyong Zheng

Contents

On the Development of Fintech in Asia

Liu Yong

Abstract There are five models of fintech development in the world: the technology promotion model represented by the USA, the rule-driven model represented by the UK, the market pull model represented by China, the mixed competition model represented by Japan and Indonesia, and the model of fanning out from point to area represented by South Korea and Israel. In terms of the layout, the transformation of traditional financial hubs has been accelerated, China and the USA have outstanding advantages in fintech, and the Asia-Pacific region has great potential for fintech development. The fintech of China has been promoted to the worlds leading level; Japan boosts the rapid growth of fintech through advantages of backwardness; Singapore gathers innovative resources with a relaxed and inclusive atmosphere; South Korea promotes scale development of fintech industry by fanning out from point to area; India is gradually exerting its potential for fintech development; Israel builds the highland of fintech development through guidance plus service; Indonesia has gradually become a rising star in fintech development in Southeast Asia; Hong Kong promotes the momentum of sound fintech development with government assistance.

Keywords Fifintech development · Asia · Policy and regulatory measures · Digital transformation

1 Overview of Global Fintech Development

In recent years, global fintech has maintained a high speed of development, the adoption rate of fintech has gradually increased, and a large number of fintech unicorn enterprises have emerged. With the application of big data, blockchain, AI, and other technologies in the financial field becoming more and more mature, new models and

L. Yong (✉)
Zhongguancun Internet Finance Institute, Beijing, China
e-mail: liuyong@guopeijigou.com

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_1

industry forms of financial service have come into being. Among them, some application fields have developed more rapidly including digital currency, open banking, digital banking, etc.

1.1 Development Dynamics

Fintech enterprises are growing fast. According to the relevant data, there are 1057 unicorn enterprises in the world now as of November 2021, and fintech unicorns play a decisive role in fintech field with the most amount of enterprises on the list which is 139 and the total valuation is 4.7 trillion yuan, accounting for 19% of the total valuation of unicorn enterprises on the list. From a country perspective, the USA has the largest number of unicorn enterprises in the fintech sector, followed by China. In 2019 Fintech 100 announced by Klynveld Peat Marwick Goerdeler (KPMG), the enterprises in Asia-Pacific region (including Australia and New Zealand) performed brilliantly, with a total of 42 enterprises on the list. As far as payment enterprises were concerned, 27 companies were on the list, which took the lead. As for other categories of companies on the list, there were 19 wealth management companies, 17 insurance companies, 15 lending companies, and 13 companies with relatively comprehensive financial business.

The developing economies represented by Southeast Asia and Latin America have obvious development characteristics in the field of financial science and technology. According to the report of the Future of Southeast Asian fintech by the British consultancy Dealroom, European venture capital company FinchCapital and Indonesian venture capital company MDIVentures, the outbreak of COVID-19 pneumonia has accelerated the digital transformation of fintech in the region, especially in the field of digital payment. Indonesia is expected to become the largest financial technology hub in the region by 2025, with an expected market value of US $130 billion in related fields. According to the global fintech report for the second quarter of 2021 by CB Insights, fintech financing in Latin America has increased at a compound annual growth rate of 57% since 2016, reaching US $4.246 billion by the second quarter of 2021. Among them, the financing amount of fintech companies in Brazil alone accounts for 70% of the total financing in the region.

More and more central banks have begun to actively study the issuance of CBDC (Central Bank Digital Currency), and some countries have even begun to build the underlying infrastructure of CBDC and start the pilot of CBDC technology. As the first country in the world to launch a sovereign digital currency, DC/EP has conducted pilot projects in some domestic cities, commercial banks, and cross-border payments since April 2020, and completed the country's first digital RMB insurance policy in December 2020. In 2020, the Bank of France launched a digital currency pilot project. European and American countries are also unwilling to fall behind. The central banks of Canada, Sweden, the UK, and other countries jointly set up a CBDC group with BIS. In May 2020, the United States released a white chapter on the digital dollar

project (DDP), which introduced in detail the basic architecture, distribution purpose, and potential application scenarios of CBDC in the United States.

The evolution of digital banking is accelerating. As a banking development model which has arisen in recent years, digital banking is an important achievement of digital transformation of banks. Currently, 60% of the worlds banking population is using digital banking through online services and cashless transactions. According to the relevant data in the Nets (an European transaction processing center) report, non-contacting digital wallet transactions increased by more than two-thirds in the first half of 2020 compared with 2019. With the increase of the users of digital banking, the number of digital banks has gradually increased. In 2019, Hong Kong Monetary Authority (HKMA) approved the establishment of 8 virtual banks. Monetary Authority of Singapore (MAS) opened up applications for digital banking licences in 2020. In addition, digital banks in many countries engage in online banking business with traditional banking license or in traditional authorized business forms, such as Monzo Bank and N26 Bank in the UK, aiBank, WeBank, and MYbank, in China.

The world has a deeper understanding of the concept of sustainable development, and the practice scenes of fintech in the field of green finance have increased. From the perspective of application scenarios, the use of fintech tools covers ESG investment and financing, national carbon market trading, green building, green consumption, green agriculture, small and micro enterprises, and other fields. Fintech is widely used in environmental data, ESG data and evaluation, green credit information management system of financial institutions, and other scenarios.

1.2 The Financing Profile

Global fintech investment and financing grew strongly. In 2020, the number of financing transactions reached 3443, and the number of financing transactions in the first three quarters of 2021 was 3549, which has exceeded the total amount of financing in the whole year of last year. The total financing amount of fintech in 2020 was US $48.4 billion, and the amount of financing in the first three quarters of 2021 was US $94.7 billion, nearly twice the total financing amount of last year.

The financing amount of financing projects is mainly concentrated in North America, Asia, and Europe, with a quarter on quarter increase of more than 50%. Among them, North America has the highest amount of total financing, accounting for more than half of the total global investment, reaching the highest in the second quarter of this year, with USD 16.56 billion, followed by Asia, which reached the highest in the third quarter of this year, with $5.9 billion. South America exceeded USD 1 billion for the first time in the second quarter of this year. In Africa and Oceania, the amount of financing is relatively stable and has little change (Fig. 1).

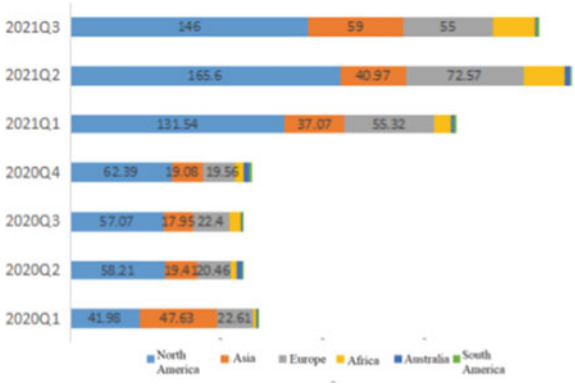

Fig. 1 The amount of financing in fintech in global continents from 2010 to Q3 of 2021

1.3 Regulatory Environment

In recent years, financial management departments in various economies have increasingly improved their regulation on fintech activities, and have promoted the healthy and orderly development of fintech through measures such as continuous monitoring, the establishment of regulators, and the introduction of regulatory policies. On the one hand, financial management departments support the entrance of fintech companies into the market to make up for the current weak links in financial services; on the other hand, countries have set a high threshold for access to financial business to reasonably guard against systemic risks.

The legislative process of data protection has been accelerated. In recent years, with the iterative innovation of new technologies, various business entities are accelerating the development of new data resources, and meanwhile the incurred problems such as data privacy protection are also increasingly valued by various countries. EU countries summarize and improve data legislation in practice: since the second half of 2019, the European Commission (EC) and Council of the European Union have organized each member countrys regulators to submit a law enforcement summary, and they have received 19 law enforcement summary reports from different countries. In September 2020, European Data Protection Board (EDPB) issued Guidelines on the Targeting of Social Media Users (the Draft Guidelines), expounding on the requirements of data protection in social media. At the beginning of 2020, California Consumer Privacy Act (CCPA) of the USA formally came into force and was formally incorporated into Californias judicial system. On October 21, 2020, the Peoples Republic of China released (Draft) and solicited public opinions. It is the first law that specifically stipulates personal information protection. Promulgated, it will become the basic law in the field of personal information protection, the Personal Information Protection Law of the Peoples Republic of China, officially came into force on November 1, 2021.

The innovation of fintech regulation tools has been continuously strengthened. Firstly, some countries have established fintech innovation mechanism. To cite a few

examples, France proposed in March 2021 to establish a European exemption mechanism in regard to blockchain, relax some legal requirements that cannot meet the needs of blockchain development, and it suggested that exempted entities should follow the key principles of financial regulation. Secondly, some countries have further improved the Sandbox Mechanisms. Thirdly, many countries vigorously support the development of RegTech. To cite a few examples, Central Bank of Brazil (CBB) announced in April that Pier, an information integration platform for financial regulators based on blockchain technology, began its operation online, which could help the participating institutions quickly access the latest data of other institutions, thus shortening the data query operation that might have taken a month to several seconds.

The fintech policy system has been continuously improved. Nowadays, countries all over the world gradually realize the potential value of fintech and formulate relevant development strategies and improve relevant policy systems to support the development of fintech. At present, apart from the policies related to AI, blockchain, big data, and other key underlying technologies of fintech, areas such as digital banking, online payment, and encrypted assets are gradually covered. The regulation on the application of fintech has basically realized full coverage, and the fintech policy system is continuously improved.

1.4 The Models of Fintech Development

At present, around the world there are generally five models of fintech development. The first is the Technology Promotion Model represented by the USA, which is characterized by mutual promotion of finance and technology and a win-win relationship between industry and culture. The second is the Rule Driven Model represented by the UK, which is characterized by innovating regulatory methods and boosting industrial development through rules. The third is the Market Pull Model represented by China, which is characterized by accelerated digital transformation and breakthroughs sought in strict regulation. The fourth is the Mixed Competition Model represented by Japan and Indonesia, which is characterized by accelerating the pace of reform and continuous stimulation of potential. The fifth is the Model of Fanning out from Point to Area represented by South Korea and Israel, which is characterized by locating breakthroughs and focusing on tackling key problems (Fig. 2).

1.5 Spatial Layout

In recent years, the fintech hubs represented by Shanghai, Beijing, Shenzhen, Hangzhou, San Francisco (Silicon Valley), New York, London, and Chicago are accelerating their rise based on financial industry and driven by technology. China and the USA have their distinctive advantages in the development of fintech and have become leaders in the development of fintech worldwide. The Asia-Pacific region

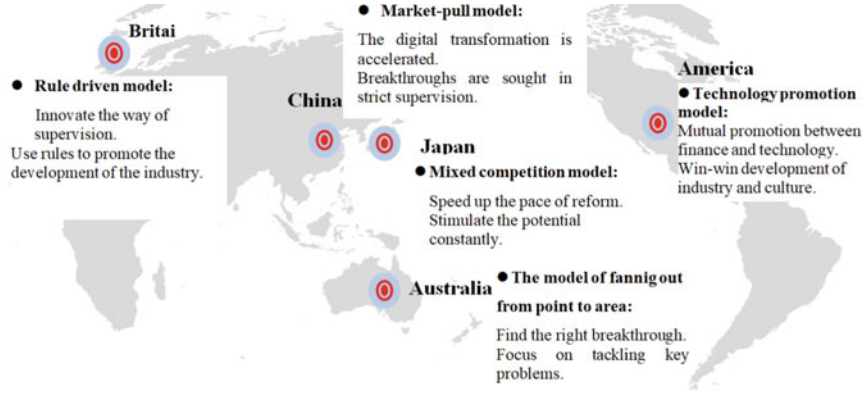

Fig. 2 The models of fintech development around the world

has gradually demonstrated its potential for fintech development and has attracted a large influx of capital, showing its advantage of backwardness.

The transformation of traditional financial hubs has been accelerated and there is great potential for the development of fintech in the Asia-Pacific region. With the comprehensive empowerment and transformation of finance by technology, the transformation of traditional financial hubs having been accelerated and newly emerging financial cities having been upgraded in an all-round way, and a new ecology of regional economy having been created with a strategic height, in the future financial hubs will take fintech as the core competitiveness of cities and compete for the commanding heights of fintech without exception. According to Global Fintech Hub Report 2021, the 9 cities in the first echelon of the global fintech hubs were Beijing, San Francisco (Silicon Valley), New York, Shanghai, Shenzhen, London, Hangzhou, Singapore, and Chicago respectively. These cities are home to the large financial institutions and the headquarters of financial institutions of the country. Most of them have a solid foundation for financial industry. They are currently starting the pace of all-round digital transformation of financial industry supported by technology. From the perspective of fintech experience, developing countries and Asia continue to maintain an overall leading edge. Not only the top 10 cities all located in developing countries in Asia, but also developing countries account for 80% among the top 20 cities for two consecutive years and Asian cities account for 65%.

2 Practice of Fintech Development in Asia

2.1 China—The Fintech Has Been Promoted to the Worlds Leading Level

2.1.1 Development Features: Accelerated Digital Transformation

According to the development stages of technology application in financial industry, the development nodes of Chinas fintech industry are relatively clear. The development of fintech in China can be divided into four stages, as is shown in Fig. 3. China has entered the fintech 4.0 era, when finance and technology develop in a highly integrated way.

The development of fintech industry leap into the front ranks of the world. There are 139 unicorns in China's fintech industry, ranking first in the world. The market scale of China's fintech enterprises is growing steadily. According to the prediction and display of relevant data of the Forward Looking Industry Research Institute, the market scale of China's fintech enterprises is expected to reach 463.1 billion yuan in 2021, an increase of nearly 17% over the previous year. It is expected that the scale of China's fintech market will still achieve stable growth in 2022.

Great progress has been made in technological innovation. From 2015 to the first half of 2019, a total of more than 22,000 enterprises applied for fintech-related patents in China, with a total number of more than 88,000 patents. Among them, big data analysis, interconnection technology, and cloud computing accounted for the highest proportion, while big data, cloud computing, biometric security, and AI maintained relatively smooth and steady growth; blockchain technology performed brilliantly with explosive growth, with the proportion of patents increasing from 0.4% in 2015 to 8.5% in 2019 (Fig. 4).

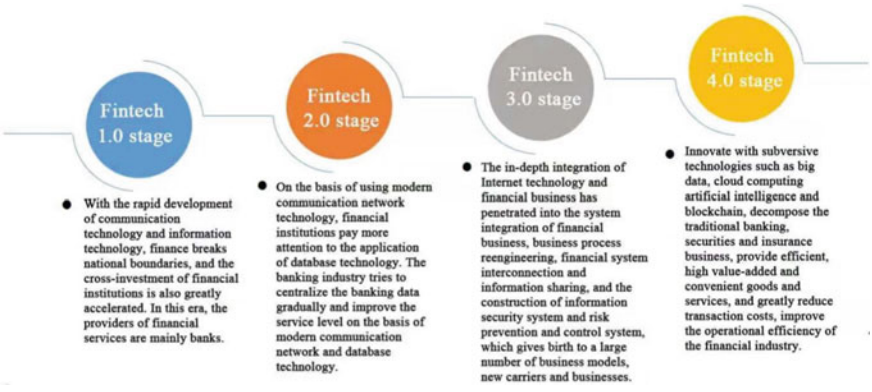

Fig. 3 Development process of Chinas fintech

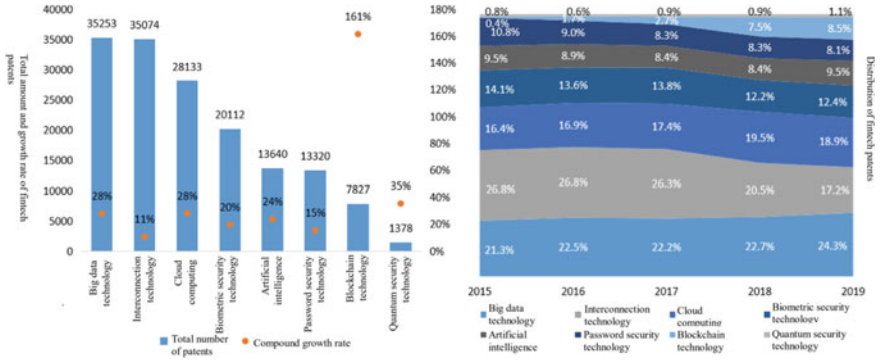

Fig. 4 Fintech patents

There is a shortage of fintech talents. At present, fintech talents are in short supply, and the growth rate is far lower than the development rate of fintech itself. According to 2018 China Fintech Employment Report released by Michael Page (China), 92% of the fintech enterprises interviewed found that China is currently confronting a severe shortage of fintech professional talents, 85% of the employers interviewed said that they encountered recruitment difficulties, and 45% of the employers interviewed said that the greatest difficulty they confronted in recruitment was the difficulty in finding talents that could meet the specific position requirements. According to the survey, the most popular fintech positions were big data position, AI position, and risk management position, accounting for 40%, 32%, and 12%, respectively.

2.1.2 Policies and Regulatory Measures: Finding Breakthroughs in Strict Regulation and Ensuring Steady Development of Data Protection

The top-down design for fintech development has been continuously improved. In August 2019, the people's Bank of China issued the Financial Technology (fintech) Development Plan (2019–2021). The introduction of this programmatic document will build the top-level design of "four beams and eight columns" of financial technology. In December 2021, the central bank issued the Fintech Development Plan (2022–2025), which is the second round of fintech development plan issued by the central bank after the release of the plan in 2019. Compared with the first round of planning, this round of planning will focus on solving the problem of uneven and insufficient development of financial science and technology, with clearer key tasks, clearer development direction, and stronger implementation guarantee. At the same time, the plan puts forward the financial technology development vision of "striving to achieve the leap forward improvement of the overall level and core competitiveness by 2025", which has opened up a broader development space for China's financial technology industry.

The system of fintech supervision rules has been gradually improved. The basic regulatory rules system of fintech is gradually improving. While improving the rule system in a single technical field, it enriches the supervision of business links such as fintech innovative product design, operation mode, and risk control means. In addition, it further complements and improves the regulatory rules for consumer rights and interests protection, personal privacy, and financial information data.

The standardization of fintech has been gradually strengthened. The central bank has issued and implemented technical standards for payment tokenization, payment information protection, acceptance terminal registration management, mobile terminal trusted execution environment, mobile financial client application software, incorporated financial science and technology products into the national unified certification system, and continued to carry out leader activities in the field of point of sale terminals (POS), self-service terminals (ATM), bar code payment acceptance terminals and online banking services.

2.1.3 Layout of Key Fintech Cities: The Cities in East China are Leading, but Each of the Cities has Its Own Characteristics

At present, China is already leading the global fintech. However, there are differences in the development speed and level of fintech among its cities. The overall strength of the cities in east China is relatively strong, the optimized layout of Beijings fintech develops steadily, Shanghai tries to build an international brand of fintech, Shenzhen strives to be the leading role in the development of Guangdong-Hong Kong-Macao Greater Bay Areas fintech, and Hangzhou adopts the strategy of policy plus talents to re-create new vitality for the citys development. Cities such as Chengdu, Chongqing, Guangzhou, Nanjing, and Qingdao are also actively laying out the development of fintech.

2.2 Japan—Boosting the Rapid Growth of Fintech Through Advantages of Backwardness

2.2.1 Development Features: The Advantage of Backwardness in Fintech has Shown

The comprehensive competitive strength lays the foundation for the development of fintech. Japan is the third largest developed country in the world, But its fintech development began relatively late. In 2018, the scale of Japans fintech market reached 214.5 billion yen, and it has been going up all the way. It was expected to reach 572.7 billion yen in 2020, with an average annual growth rate of more than 50% (Fig. 5).

Optimizing cultural soft environment and accelerating the shaping of a non-cash society. According to EY Global Fintech Adoption Index 2019, in terms of the

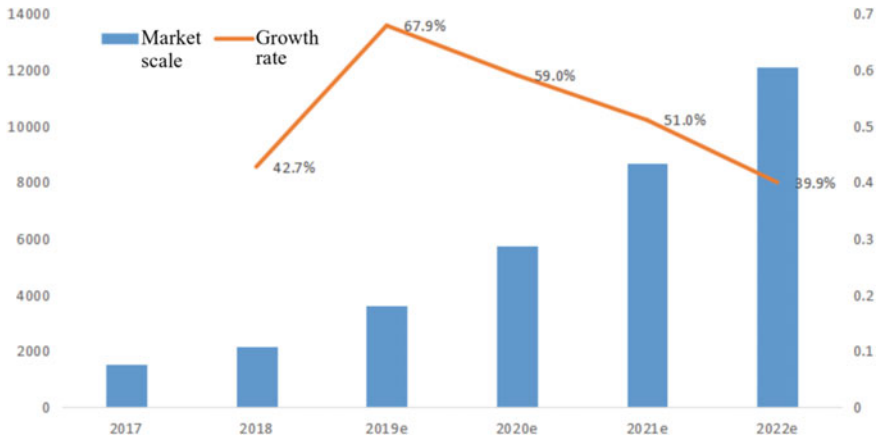

Fig. 5 The scale of Japans fintech market (Unit: 100 million yen, %)

global consumer fintech application index, Japan ranked the lowest in 27 markets, with only 34%. The Japanese government issued Fintech Vision in May 2017, which clearly proposed that it should pay attention to the added value of fintech and focus on improving the adoption rate of electronic payments. After that, the government issued Future Investment Strategy 2017, explicitly proposing to triple the proportion of non-cash payments to more than 40% by Expo Osaka 2025. Since then, the Japanese government has been committed to promoting non-cash payment rebate activities [Consumers would get a rebate of about 2–5% for each non-cash payment], continuously optimizing the cultural environment and accelerating the shaping of a non-cash society.

The commercial configuration of various industries has gradually taken shape. The mobile payment sector has stepped out of the era of barbaric growth and formed a duopoly pattern of Line pay and PayPay, and has nurtured a number of outstanding fintech start-ups on this basis. Japan attaches great importance to the development of blockchain. In 2018, the market size of Japan blockchain reached 8.07 billion yen, and it reached 33.57 billion yen by 2020. With relatively strong development momentum, it saw the emergence of a number of blockchain start-ups with certain strength and characteristics, such as Dobulejump.tokyo and Nayuta Japans regulation on network lending is relatively loose, and network lending and crowdfunding have become an important part of Japans inclusive finance, hence the much rapid development of the industry. In 2014, when Japan amended its financial commodity trading law, crowdfunding suddenly came to the fore. In 2018, the scale of Japans crowdfunding market reached 204.5 billion yen. During the epidemic, many crowdfunding platforms also took on the responsibilities of assisting commercial tenants, etc. Japan is also actively promoting the development of sectors such as personal loan, Robo advising, and supply chain finance. Although they are still in the initial stage, they have great potential for future development (Fig. 6).

Fig. 6 A diagram of the industrial ecology of Japans fintech

2.2.2 Policy and Regulatory Measures: Optimizing the Policy System and Forging Ahead with Determination

In terms of the development policy and regulatory measures in fintech, Japan adopts strict regulation and easing measures at the same time. For the development of some traditional industries, especially in the aspects of digital transformation, the regulation is relatively strict. However, the regulation on sectors such as crowdfunding and network lending is relatively loose, so these sectors can develop rapidly. Strict regulation measures can effectively control the risks in fintech innovation. Moreover, in 2018, JFSA started to implement a sandbox mechanism for financial innovation, allowing financial and insurance products to be put into trial operation within a certain risk range, and steadily promoting healthy and sustainable innovative development. The loose measures in some sectors can stimulate the development vitality of the fintech industry for it to reform in development, maintain stability in progress, and create a safe and controllable development ecology in an all-round way.

2.2.3 Layout of Key Fintech Cities: Tokyo Bay Area Endowed with Good Resources to Push Traditional Financial Institutions on the Way of Reform

Fintech got developed in Japan later than in other developed economies, so it has not yet formed a ubiquitous layout of fintech hubs. Whether according to the fintech hub report released by Global Fintech Hub Federation or the index and list of fintech hubs released by institutions such as Deloitte and Z/Yen Group, the fintech hubs of Japan that may enter the list tend to be Tokyo. Therefore, this chapter focused on

the relevant situation and policy measures of Tokyo as a fintech hub. Tokyo ranks among internationally renowned financial centers together with other international financial centers such as New York financial center and London financial center. Meanwhile, Tokyo is also the capital of Japan and the financial capital of Japan. Since the 1960s, the Japanese government has been planning to build the capital circle of Tokyo, linking Tokyo with several neighboring counties for joint development and construction. At present, Tokyo Bay Area has become one of the worlds eight recognized bay areas.

Since the 1990s, the Japanese government has formulated and promulgated a series of science and technology innovation strategies and policy measures to stimulate the high-speed rise of science and technology innovation level in Tokyo Bay Area. Relying on internationally first-class universities and research institutions, innovative enterprise clusters, and the support of the Japanese governments policy inclination, Tokyo Bay Area has absorbed advanced technology and innovation concepts in its opening to the outside world, vigorously developed advanced scientific and technological productivity, formed a bay area ecological environment conducive to scientific and technological innovation, spawned numerous scientific and technological innovation institutions, and witnessed the emergence of a large number of scientific and technological innovation achievements, making Tokyo Bay Area gradually develop into a world center of innovation with international influence.

2.3 Singapore—Gathering Innovative Resources with a Relaxed and Inclusive Atmosphere

2.3.1 Development Features: An Active Atmosphere for Fintech Innovation

International innovation elements gather and multiple resources converge. Singapore is an international trade hub, an Asian financial center, and a place of strategic importance for technological innovation. Its convenient geographical conditions have facilitated the convergence of financial and technological innovation resources. On the one hand, as a global financial center, Singapore has financial industry as its service industry with the highest added value, with more than 1,200 financial institutions stationed here. On the other hand, Singapores scientific and technological innovation has developed rapidly. In Global Innovation Index 2018 released by WIPO (WIPO), Singapore ranked the fifth, overtaking traditional science and technology powers such as the USA, Germany, Israel, South Korea, and Japan and successfully ranking among the worlds leading science and technology innovation centers.

There are rich forms of activities and strong vitality in fintech. Since 2016, Singapore has been hosting Singapore Fintech Festival (SFF) and Singapore Week of Innovation & Technology (SWITCH). In 2019, SWITCH and SFF merged into SFF X SWITCH for the first time. On June 8, 2020, on the basis of previous experience

of holding activities, Singapore held the MAS Global Fintech Innovation Challenge for the first time by innovating the form of activities. With the theme of Building Defenses, Seizing Opportunities, and Emerging Stronger, the competition had a total bonus of S$1.75 million and comprised two parts: MAS FinTech Awards and MAS Global FinTech Hackcelerator.

Digital banking booms and digital finance accelerates. At present, Singapore is gradually loosening the restriction on the application for digital full bank license. The introduction of digital bank license is the largest banking liberalization in Singapore in the past 20 years. In December 2020, MAS issued a total of 4 digital bank licenses, of which 2 were DFB licenses and another 2 were DWB licenses. The launch of digital banks in Singapore will form competition with traditional banks, but meanwhile it will promote the rapid development of fintech in Singapore.

Actively extending the application scenarios of blockchain technology. Singapore is the friendliest country to the development of blockchain in Southeast Asia and even all over the world. At present, a large number of mature blockchain projects are distributed in sectors such as trading platforms, public blockchains, hosting, cloud storage, infrastructure, consulting, and insurance. Singapore vigorously promotes the application of blockchain technology in financial scenarios. On the one hand, it uses blockchain technology to promote the development of digital payment. On the other hand, it focuses on SME financing and supply chain finance. In addition, it adopts blockchain technology to ameliorate the pain points of service of industrial finance including supply chain finance, etc.

In sound fintech ecology, various subjects jointly pursue interconnected development. Singapores rich and diverse international fintech activities and its open and inclusive innovation environment, etc. have attracted diversified fintech talents to gather here. In August 2020, Singapore established Asian Institute of Digital Finance (AIDF), jointly founded by MAS, National Research Foundation (NRF) of Singapore, and National University of Singapore (NUS), to meet the demand for digital financial services in Asia. The strong community effect has attracted the convergence of talents such as entrepreneurs, domain experts, angel investors, and industry mentors, and it has provided a platform for exchanges and collaboration among entrepreneurs, investors, and financial institutions. Meanwhile, it has attracted high-quality native start-ups of Asian countries such as India and Indonesia to migrate to Singapore, forming a highly open international fintech ecosystem (Fig. 7).

2.3.2 Policy and Regulatory Measures: The Top-Down Design is Optimized and Special Policies are Increased

Perfecting the top-down design of fintech. The Singaporean government has authorized MAS to be the policy subject for the innovation and development of fintech which is fully responsible for the strategic planning, policy framework, and policy coordination of the development of fintech. In order to further promote the coordinated and efficient development of fintech, MAS has established professional fintech management institutions. Firstly, FinTech & Innovation Group (FTIG) was set up,

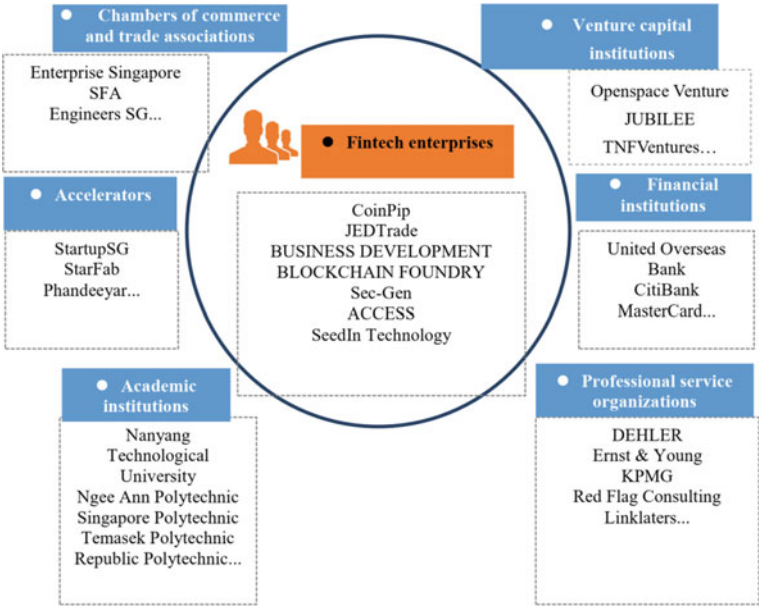

Fig. 7 Fintech ecosystem

which comprises three offices, respectively for payment and technology solutions, technology infrastructure, and technology innovation lab. FTIG invested S$ 225 million to promote Financial Sector Technology & Innovation Scheme (FSTI) and encourage the global financial industry to set up an innovation and research and development center in Singapore. Secondly, Fintech Office was established, which is mainly responsible for three tasks: the first one is to review, correspond to and improve fintech-related subsidy schemes for cross-governmental agencies; the second one is to pay attention to industrial infrastructure and the gap between talent training and manpower demand and put forward target strategies, policies, and programs to enhance the competitiveness of the industry and enterprise organizations; the third one is to manage Singapores fintech brand and marketing strategy through fintech activities and related initiatives and strive to become a global fintech hub.

Formulating special fintech regulatory policies to promote the healthy development of enterprises. The Singaporean government has adopted a multi-pronged approach to promote the development of fintech at home. Some of the measures are universal, including providing a supportive environment for start-ups, adopting a collaborative approach, and attracting foreign investment. Apart from common measures, Singapore has formulated special fintech regulatory policies to guide the development of various segments of fintech, mainly focusing on AI and data analysis (hereinafter referred to as AIDA), blockchain technology, digital assets, payment bill, and open banking.

Promoting the development of RegTech application to reduce risks SGX has launched a new RegTech scheme that can automatically report market irregularities and promote fair trading. At present, Singapore has had representative companies ranking the worlds top 100 in RegTech involving sectors such as compliance management and anti-money laundering. Besides, for local start-ups, the Singaporean government has established a Regulatory Sandbox system to encourage the innovative development of fintech start-ups and nurture and incubate outstanding enterprises.

2.4 South Korea—Promoting Scale Development of Fintech Industry by Fanning Out From Point to Area

2.4.1 Development Features: The Foundation for the Development of Fintech is Solid

The 5G information and communication technology takes the lead. According to the data from South Koreas Ministry of Science & Information and Communication Technology (MSICT), South Korea is the first country in the world to start 5G commercial use. After the official launch of 5G business services on April 3, 2019, the number of users has increased continuously, reaching nearly 10 million by the end of October 2020.

Blockchain technology is developing rapidly. According to the data from Korean Intellectual Property Office (KIPO), a total of 1,301 blockchain patents were registered in South Korea in 2019, a 50-fold increase from 24 in 2015, and the number of the patents increased further in 2020 after the Covid-19 pandemic began.

Forming the mechanism design of two kinds of institutions and three modes of sharing, the big data credit system is refined. South Korea has established a much refined personal and corporate big data credit system. In this system, Korea Federation of Banks (KFB) is the pillar of the credit industry. On this basis, there are three data-sharing modes of credit information service. One is to force financial institutions to submit credit information to KFB, which will then be provided by KFB to private credit reporting companies; the second is to share information within the industry through associations or corporate groups; the third is that credit reporting companies collect other information through commercial contracts. Under the mechanism design of two kinds of institutions and three modes of sharing, the South Korean credit investigation industry can not only quickly and timely collect nationwide credit information, but also ensure that valuable credit information could be legally and fully shared across the whole society.

P2P loan industry develops rapidly. According to the statistics of the South Korean government, the total investment in P2P loans in South Korea increased from 37.3 billion won at the end of 2015 to 2.34 trillion won at the end of 2017, and then rapidly increased to 6.2 trillion won by the end of June 2019. In August 2020, the Law on Financial Industry Related to Online Investment and laws related to user protection

(P2P laws) were officially implemented, which would strengthen the protection of investors and formally set a legal framework for P2P development.

2.4.2 Policy and Regulatory Measures: Strengthen Planning and Launch Fintech Development Strategy

On December 4, 2019, Financial Services Commission (FSC), Republic of Korea announced that it would vigorously promote the large-scale development of the fintech industry, and introduced 8 measures in different sectors, involving improving the current Regulatory Sandbox system, carrying out regulatory reform to promote the development of fintech, loosen the entry restrictions of the financial industry, establishing a regulatory basis for the digital age, developing new growth engine for financial innovation, promoting the investment in fintech and establishing a venture capital ecosystem with private sector investment as the core, assisting fintech enterprises with overseas expansion expanding public support for fintech enterprises.

Preferential taxation is applied to research and development of blockchain technology. A report released by the local Ministry of Strategy and Finance announced the latest tax laws that came into effect in February 2019. And the blockchain has been added to the research and development list that provides tax credits. This means that the companies or enterprises that develop blockchain technology can deduct some taxes from the research and development expenses. The tax reduction depends on the size of the company.

Implement the Regulatory Sandbox plan and accelerate the digital transformation. On April 1, 2019, FSC, Republic of Korea officially launched a fintech Regulatory Sandbox program, thereby hoping to promote competition in South Koreas financial industry and bring more favorable services to consumers. Up to May 2020, FSC had held a total of 14 assessment committee meetings. Through various assessments of business innovation, consumer convenience, and project stability and feasibility, a total of 102 innovative financial services were eventually selected into the Regulatory Sandbox, which obtained exemption from licensing and other regulations.

2.4.3 Layout of Key Fintech Cities: Seoul—Taking Various Measures to Create a Business Environment for Fintech

Seoul is the largest city on the Korean Peninsula and one of the major financial cities in Asia with advantages in economic, technological, and cultural development. Seoul ranks the fifth among the top ten Asian cities in terms of economy, only after Tokyo, Shanghai, Beijing, and Hong Kong. Its economic aggregate accounts for about 23% of that of South Korea. The population of Seoul is about 10.2 million, accounting for 20% of the total population of South Korea. In addition, Seoul accounts for half of Korea in terms of personal income tax, corporate income tax, and bank deposits, and the number of innovative enterprises and graduates from colleges and universities account for 30% of Koreas total. Seoul is South Koreas political and economic

center, which has laid a good foundation for fintech development. In recent years, Seoul has created a good business environment for fintech by holding the fintech week, establishing fintech labs, setting up innovation funds to increase investment in fintech industries, etc.

Holding fintech week to create an innovative atmosphere. South Korea launched the first Korea Fintech Week from May 23 to 25, 2019. The event was held in Seouls Dongdaemun Design Plaza (DDP). It is the first global fintech fair in South Korea, and the FSC hopes to develop it into an important annual fintech event in Asia. In 2019, Korea Fintech Week invited global financial institutions, international organizations, and global fintech companies to discuss relevant policies to help local fintech companies expand ties with local and global investors. In addition, the activity also provided counseling services for college students and young job seekers who are interested in the fintech industry.

Affected by the epidemic, the 2020 Korea Fintech Week was changed to be held online. The FSC said it had attracted more than 170,000 page visitors and received more than 110 million page views. Financial companies and fintech enterprises set up a total of 150 virtual exhibition halls. A total of 35 enterprises participated in the online job fair session, with more than 1,000 job seekers competing for about 80 jobs provided by 21 fintech enterprises.

Launching SEOUL FINTECH LAB. In April 2018, the Seoul municipal government launched the first SEOUL FINTECH LAB, which was mainly aimed at early-stage start-ups. In July 2019, the second fintech lab was launched, which would be targeted at growth-stage start-ups and accommodate approximately 14 start-ups from South Korea, the USA, Hong Kong, and Singapore. The fintech lab will become a key anchor point for South Koreas fintech industry, so as to help South Koreas promising fintech start-ups develop abroad. The Seoul Innovation Growth Fund was established. At the end of 2018, Seoul set up an innovation fund of around 130 billion won (approximately USD 116 million) to be invested exclusively in innovative industries such as blockchain and fintech. In early 2019, the Seoul municipal government announced that the Seoul municipal government would invest 1.2 trillion won (USD 1.07 billion) in start-up companies in the fintech sector through the investment fund by 2022.

2.5 KazakhstanCDigital Transformation Speeds Up the Construction of Central Asian Fintech Hub

2.5.1 Development Features: The Digital Economy is Booming

Central Asia is located between the worlds two largest economies. Proposed by the Belt and Road Initiative, it can be regarded as a bridge connecting Europe and China. As a growing potential market, Kazakhstan has played a key role in the process of Central Asia becoming a global fintech hub.

Although in the beginning the level was relatively low, the digitization process in Kazakhstan is developing rapidly, mainly including: (1) The rapid growth of e-commerce and mobile commerce; (2) The transition from cash payments to non-contact and digital payments; (3) The growth of innovative digital financial products and services. Since the outbreak of COVID-19, these structural changes which had lasted for many years have accelerated, creating a favorable environment for the further development of fintech.

The fast-growing e-commerce market is one of the driving forces behind the development of fintech. Compared with other emerging market countries and developed economies, in Kazakhstan, the e-commerce penetration has a significant upward potential. According to Euromonitors data, the market value of e-commerce in Kazakhstan in 2019 was estimated to be KZT 401.3 billion (USD 1.1 billion), equivalent to 3.4% of the total volume of retail trade, with a CAGR of 33.3% from 2016 to 2019.

Digital payment is developing rapidly. The adoption rates of Internet and mobile phone have increased significantly. According to Ovum (world mobile information service), Kazakhstans total number of smartphones increased from 12.7 million in 2016 to 19.2 million in 2019 and is expected to reach 23.4 million by 2024. Banks are using mobile and Internet banks to provide better financial services to remote and rural areas. Fintech companies have fewer opportunities to cooperate with standard financial sector, but with the increase of mobile Internet adoption rate, they have obtained huge opportunities in areas not covered by traditional financial markets. In 2019, Kazakhstans digital payment amount more than tripled to about USD 34.8 billion Like e-commerce, this trend has been accelerated by the COVID-19 epidemic. In 2019, Kaspi.kz accounted for 83% of the growth of the entire payment market in Kazakhstan and became the largest contributor to Kazakhstans transition to digital payment.

The digital transformation of banking and insurance industry is at the right time. The COVID-19 epidemic has enabled most retail banking activities to be carried out online, which has promoted the development of digital banking. Banking services are rapidly moving from branch-based, product-centric organizations that use traditional technologies to more personalized digital solutions that are consumer-centric and deliver seamlessly. Since January 2019, the citizens of Kazakhstan have been able to use electronic insurance and choose to submit their applications online. Within the conceptual framework of the development of the financial sector in the Republic of Kazakhstan to 2030, it is expected that electronic insurance policy sales will be introduced into the compulsory courses.

2.5.2 Policy and Regulatory Measures: Being Committed to Promoting Financial Innovation in a Wider Range of Areas

Astana Financial Services Authority (AFSA), established on January 1, 2018, is independent from the National Bank of Kazakhstan (NBK) and the financial market supervision and Development Department of Kazakhstan. It is an independent regu-

lator of financial and non-financial services activities of AIFC (Astana international financial center, established on July 5, 2018, is the financial center of Astana, Kazakhstan). In its fintech hub department, AIFC assists relevant companies in developing new products and services in the fintech sector in various ways. One way is to provide acceleration projects where start-ups can work closely with mentors from around the world to develop the necessary capabilities. Another way is for the fintech department of AFSA to provide a legal and regulatory basis for the development of new financial products and technologies and to test them at the fintech lab (Regulatory Sandbox). At present, 30 projects are being tested in Regulatory Sandbox. Currently, more than 125 start-ups work with the fintech department of AIFC. These companies are distributed in different sectors such as payment and mobile wallet, market, credit, AI and machine learning, blockchain, digital identification, network security, and fraud prevention.

The fintech department of AIFC supports venture capital and corporate innovative development with the goal of creating a healthy venture capital ecosystem and expanding opportunities for start-ups in Central Asia and the countries belonging to Commonwealth of Independent States (CIS) to attract investment and perform transactions. In this way, AIFC is creating a comprehensive ecosystem, which covers strengthening regulation, supporting start-ups, helping attract investment, and implementing fintech solutions within enterprises. To address the innovation challenge in the financial sector, AIFC is taking a series of regulatory measures to promote innovation and strengthen the protection of the consumers of financial services/products., including setting up a fintech lab to promote fintech development, introducing a regulatory framework to promote the development of crowdfunding, expanding the framework for the list of regulated and market activities, implementing a series of policies to promote the healthy development of digital asset, creating a looser banking system to strengthen inclusive financing, implementing open API to promote the innovation of digital currency and payment services, providing corporate income tax and value-added tax exemptions for fintech companies optimizing e-commerce regulatory measures, ameliorating the framework to promote venture capital financing, launching Global Financial Innovation Network (GFIN) to promote cross-border regulation and innovation.

2.5.3 Layout of Key Fintech Cities: Nur Sultan and Almaty—Leading the Development of Non-cash Payment

Kazakhstans cities with the most active fintech development are undoubtedly Nur Sultan (the capital) and Almaty (the countrys largest city). As the most densely populated and economically developed cities, they have non-cash payments leading in both quantity and share. Almatys non-cash payments occupied the largest market share: nearly KZT 7 trillion (about USD 16.5 billion). The city also had the highest proportion of non-cash payments, which was 76.8%. Nur Sultan ranked the second with a market share of KZT 2.9 trillion (approximately USD 6.8 billion) in non-cash

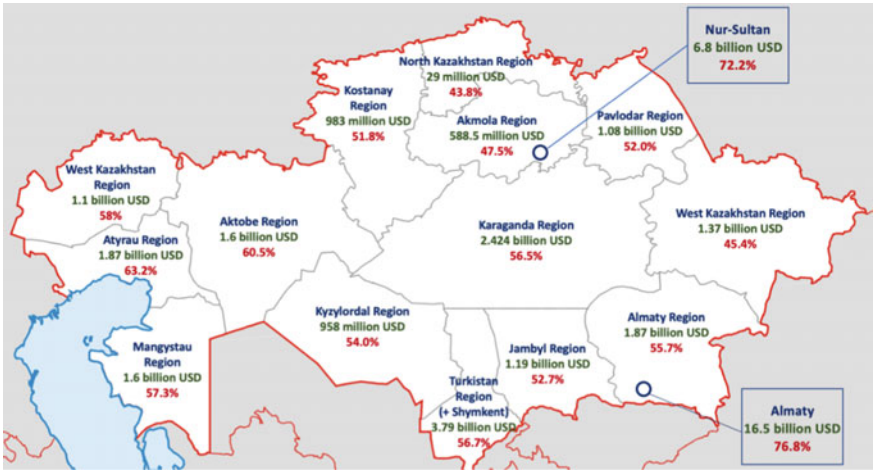

Fig. 8 Market share and proportion of non-cash payment in different cities and regions of Kazakhstan

payments, with the proportion of non-cash payments reaching 72.2%, which also ranked the second (Fig. 8).

Apart from being absolute leaders in the market share of non-cash payments, Almaty and Nur Sultan are also major hubs for fintech start-ups in the country. The countrys largest fintech accelerator, science park, and center are located in the following cities: the AIFC Fintech Hub in Astana, AIFC and Nuris; TechGarden and Most in Almaty.

2.6 India—Potential for Fintech Development Has Been Gradually Exerted

2.6.1 Development Features: Digital Technology Promotes the Innovative Development of Fintech

Indian fintech enterprises as a whole are experiencing the transition from initial stage to growth stage. According to relevant statistics, as of 2019, the number of fintech start-ups in India was second only to the USA, ranking the second in the world. Taking the development of fintech enterprises in several major sub-sectors as an example, there are only dozens of network lending platforms in India at present, which are in the initial stage, and few Indians have experienced online lending investment. Indias credit investigation industry is still in the exploratory stage, with a huge long tail market. Indias crowdfunding industry is in the early stage of development, showing a slow growth trend since 2014. The crowdfunding industry lacks clear regulation,

and SEBI(Securities and Exchange Commission of India), the main regulator, has not yet issued regulatory regulations.

The commercial forms of fintech are constantly improving. Sectors such as payment, loan, wealth management technology, personal finance, insurance, and RegTech have blossomed in an all-round way. Take the sectors of payment and online lending as an example. In the sector of payment, Internet payment covers enterprises with various systems, such as telecom, e-commerce, banks, wallet companies, and other enterprises with different representatives, and some representative payment companies such as Paytm are popular with capital investment. Indian fintech also has great potential in payment, online lending, blockchain, robo advising, inclusive financing, technology-driven integrated banking services, Internet financial security, biometrics, etc.

Digital payment helps India seize the innovation highland of fintech. Digital payment industry has become the core field of accelerating digital capacity building in India and has greatly boosted India to seize the innovative highland of fintech. In 2016, Modi put forward the slogan of Stand Up India, which officially helped the entrepreneurial trend from the height of national policies, with a view to establishing a new ecosystem in the financial scope, and announced the implementation of the Digital India program, initiating the banknote scrapping campaign, and making clear that digital ID cards should be bound to financial services. The banknote scrapping campaign directly boosted Indias fintech industry to the mainstream position, and Indias unique payment infrastructure with a unified payment interface won the trust of the people, which solved the problem of cash-based mode of payment and reduced the financing difficulties of the enterprises. In 2019, India launched the Digital India program, hoping to digitize every offline transaction by unifying the payment industry and e-commerce system. Meanwhile, Mastercard has also launched a project called Team Cashless India in India. This activity would help merchants accept digital payment and improve the coverage of fintech. In addition, the huge development potential of Indias fintech market has also attracted more companies around the world to deploy the Indian fintech market, such as famous Chinese enterprises Alibaba, Tencent, and JD.com and investment institutions Sequoia Capital, Hillhouse Capital, etc. The international influence of the fintech development has been continuously enhanced.

Population advantage lays the foundation for fintech development. India has the second largest population in the world. According to the statistics of the World Bank, as of 2018, the population of India was 1.353 billion. In terms of population structure, the population under 35 years old accounts for 65% of the total, and the population under 25 years old accounts for 50%. In terms of age and proportion, India has a larger number of young people, which is an idealized proportion in the population structure. There are a large number of talents available for fintech development. In the entrepreneurial development of fintech, more young entrepreneurs have the opportunity to start businesses, and there are more long-tailed users of fintech. By 2019, the number of Internet users in India was 560 million, accounting for about 41% of the total population. As a country with big Internet demand only second to China, India has great room for the development and implementation of intelligent

applications. In terms of the adoption rate of fintech, according to public data, as of 2019, the adoption rate of fintech in India had reached 87%.

2.6.2 Fintech Development Policies and Measures: Two-Pronged Approach of Regulation and Publicity to Accelerate Fintech Development

Supervision is the key to the development of fintech. India insists on doing by learning, learning by doing in fintech regulation. In the aspect of fintech regulation, the top-down design is gradually improved, fintech is included in the regulation scope, the Regulatory Sandbox of fintech has been launched, and the popularization rate and adoption rate of fintech are improved by setting up funds, launching fintech publicity activities, etc.

2.6.3 Fintech Development Measures in Key Cities: Mumbai—Sound Financial Foundation and Satisfying Scene Experience

In 2019, Mumbais overall ranking in Global Fintech Hub Index (GFHI) improved by 6 places and it entered the top 20 in the world for the first time. Mumbais fintech industry has improved rapidly, and with 6 highly financed unlisted fintech enterprises such as Freecharge and InCred the number ranked the 16th in the world. Moreover, Mumbai, with its huge population size and excellent population structure (the average age was only 27 years old), had 64% fintech users, ranking the 12th in the world. The advantage of ranking the first in Asia except Chinese cities was glamorous all over the world, making the fintech experience in Mumbai a major advantage.

Actively building fintech into one of the characteristic industries for urban development. Mumbai, as the largest financial center in India, constantly ameliorates its fintech ecology. At present, it has large financial institutions such as HDFC Bank, Kotak Mahindra Bank, ICICI Bank, and State Bank of India, ranking the seventh in the global TOP200 financial institutions by total market value. The digital transformation of traditional finance is accelerated actively, e.g., State Bank of India has launched YONO, a comprehensive life and financial service platform, HDFC Bank launched UltraCash, a mobile payment application program, etc. The rate of utilization of fintech has been increased in an accelerated manner.

2.7 *Israel—Guidance Plus Service to Create a Highland for the Development of Fintech*

2.7.1 Development Features: Technology and International Resources are Transformed Into Fintech Development Advantages

Israel is an internationally recognized innovation powerhouse. The proportion of scientists and engineers engaged in high-tech research and development in Israel is the highest in the world. Among the high-tech companies listed on Nasdaq in the United States, the total number of Israeli companies ranks second. Israel has more than 6,000 technology start-up companies, ranking first in the world. More than 270 multinational companies in the world have set up scientific research centers in Israel. Israel has strong scientific and technological innovation genes and international resources. These resources have laid a solid foundation for fintech innovation and development of Israel. In general, the development of Israeli fintech presents three major development features.

First, the underlying technology shows obvious endowment advantages. Israel is a model of the integration of military and civilian development all over the world. At the same time, the demand for cutting-edge technology and related innovations in the military field are smoothly transmitted to the commercial field. Israel has strong military applications in the fields of security, computer vision, and neuro-language planning. These technological applications are also applied to the development of fintech. Currently, it is the country with the highest usage density of fintech applications. Israel is one of the first countries in the world to adopt blockchain and digital encryption technology. It has many start-up companies with core technologies in the field of blockchain and digital encryption, such as QEDit and DAGLabs. From 2013 to 2019, the amount of financing for fintech and related underlying technologies (artificial intelligence, network security, etc.) was on an upward trend. Especially in the field of artificial intelligence, the amount of financing doubled from USD 1.463 billion in 2017 to USD 3.182 billion in 2019, and the average amount of a single financing increased from USD 7.07 million in 2017 to USD 16.41 million in 2019, increased by more than 100%.

Second, the science and innovation ecology is increasingly perfect. Israel attaches great importance to the formation of the science and innovation ecology. Its government has taken various measures to ensure that scientific research is one of its priorities. It provides security through tax and fee reduction, and at the same time it increases the expenditure in the industry. In the 2020 OECD R&D Intensity Index (the ratio of R&D investment to GDP), Israel continued to maintain its leading position. It is expected that the total future expenditure will continue to increase. According to the latest annual global entrepreneurial ecosystem rankings released by the global entrepreneurial research organization StartupBlink, Israels global ranking has risen by one over last year, ranking third in the world.

Third, the degree of internationalization of fintech is high. Due to geographical restrictions and market restrictions, Israels fintech has attracted international invest-

ment and fintech companies and has exported products and services to the outside world since its emerging. According to a report released by the Israel Venture Capital Data Center in 2020, the participation of foreign investors in Israeli equity investment in the fintech sector increased from 57% in 2018 to 69% in 2019. In the fields of payment, transaction, and digital currency, more than 90% of fintech companies provide international services. In 2019, Israels high-tech industry export continued to grow, reaching a historical record of USD 45.8 billion, accounting for about 46% of Israels total export, increased by 1.2% than that in 2018. Among them, the export of related technology products and services such as fintech and artificial intelligence accounted for a relatively large proportion.

2.7.2 Policies and Regulatory Measures: Guidance but Not Leading, and Strengthening of Communication Between Government and Industries

Israel has many policies and regulatory measures. From the establishment of a regulatory and innovative fintech hub to the establishment of a fintech assistance center, from adjusting the fintech license application process to launching the data sandbox program, the Israeli government basically guides the development of fintech as a market assistant and industrial development guider. Strengthening the communication between the government and the industries and being a partner for the development of fintech are the characteristics of Israeli fintech policies and regulatory measures.

In July 2018, the Israel Securities Authority (hereinafter referred to as ISA) announced the establishment of a regulatory innovation fintech hub, mainly aiming at promoting dialogue between regulators and participants in the fintech industry. In 2019, the Capital Market Authority of Israel joined the GFIN and participated in the global fintech regulatory reform together with international institutions such as the World Bank and the International Monetary Fund, etc. In July 2020, the Israel Securities Authority and the Israel Innovation Authority jointly launched a data sandbox plan for fintech start-ups.

2.7.3 Layout of Key Fintech Cities: Tel Aviv—The Integration of Internal and External Strategies to Promote the Innovation and Development of Fintech

Tel Aviv is Israels second largest city. The city cluster centered on Tel Aviv has become Israels largest metropolitan area and economic hub, and is known as Israels economic capital and technology center. 77% of Israeli start-ups, 81% of investment institutions, 72% of incubators, and 85% of R&D centers are located in Tel Aviv. Tel Aviv owns Israels only stock exchange, Tel Aviv Stock Exchange (TASE), which has become the international headquarters of venture capital companies, scientific research institutions, and a gathering place for high-tech companies. At the same time, Tel Aviv has a relatively complete innovation incubation system and scientific

talents. Among the top 200 global entrepreneurial ecosystem cities, Tel Aviv of Israel ranks seventh. The advantage of focusing on financial innovation with leading technology is obvious. According to the 27th Global Financial Center Index Report (GFCI 27), Tel Aviv ranks 36th.

Tel Avivs fintech development adopts an international strategy. Taking advantage of its own superior innovation environment and developed international capital agglomeration, the products of major fintech companies consider Tel Aviv as an effective test point for product technology, and Tel Aviv will be the first place to test the effects of products and services before international marketing. Tel Aviv Global City Office is used to implement targeted marketing for international fintech customers. While enhancing the citys global media image, various activities are held to meet fintech services and needs, link start-ups and investment capital, as well as implement cross-bank and cross-domain cooperation.

2.8 Indonesia—A Rising Star of Fintech Development in Southeast Asia

2.8.1 Development Features: Fintech is in the Preliminary Development Stage, and Its Potential Continues to be Highlighted

The development of the Internet has certain advantages. According to the 2019 Southeast Asia Digital Economy Report, Indonesia is the country with the largest Internet economy in Southeast Asia. It was even more than quadrupled in 2019, with more than USD 40 billion, and it is expected to reach USD 130 billion in 2025. Internet users in Indonesia are growing rapidly. According to a report released by a global social media marketing company We Are Social and Hootsuite, in January 2020, Internet penetration rate of Indonesia was 64%, with an average annual growth rate of close to 20%. Moreover, Indonesia has crossed the mature development stage of the Internet and moved directly to the development stage of the mobile Internet. In January 2019, there were 356 million mobile phone users in Indonesia, the penetration rate of mobile phones was 133%, and the number of active mobile Internet users reached 142 million.

The fintech industry is developing rapidly. According to a market report by Swiss Global Enterprise, a Swiss export and promotion agency, Indonesias digital financial services revenue is expected to grow significantly at a compound annual growth rate (CAGR) of 34%, and will reach USD 8.6 billion in 2025. The research report Future of Southeast Asia Financial Technology shows that the total valuation of Indonesian fintech companies reached USD 35 billion in 2020, accounting for 32% of that in Southeast Asia.

The online lending and payment industry is booming. The cumulative amount of loans for online lending in Indonesia increased from IDR 2.56 trillion in December 2017 to IDR 102.52 trillion in March 2020, increased by 40 times. According to data

compiled by the Otoritas Jasa Keuangan (OJK), Indonesias total loans from Fintech loans in May 2020 increased by 166.03% on a year-on-year basis. OJK estimates that there are more than 25 million borrower accounts and more than 654,200 entities providing loans. In terms of fintech payment, the total number of electronic money transactions at the end of 2019 reached 5.2 billion, an increase of 79.3% from 2.9 billion in last year. In May 2020, BI had issued licenses to 51 electronic money operators, and the main participants included GoPay, Ovo, Dana, and LinkAja.

2.8.2 Policies and Regulatory Measures: The System Continues to be Improved and the Supervision Continues to be Upgraded

The fintech policy and supervision system have been continuously improved to encourage the development of the industry. Indonesias fintech sector is under the supervision of Bank Indonesia (BI) and the Otoritas Jasa Keuangan (OJK), with the Ministry of Information and Communications of Indonesia playing a supporting role. Bank Indonesia and OJK are responsible for different regulatory fields. Each of them has a supervisory team, and they learn from each other and complement each other (Table 1).

In October 2017, the Otoritas Jasa Keuangan (OJK) issued the 2017–2022 Development Plan, which formulated 10 major policies and implementation plans, and clearly stated that appropriate supervision should be carried out to optimize the development of financial technology. On November 30, 2017, Bank Indonesia issued the Financial Technology Regulatory Regulation No. 19/12/PBI/2017 for the first time, which aimed to regulate fintech behaviors to promote innovation, protect consumers and manage risks so as to maintain a stable currency and financial system and build an efficient, safe and reliable payment system. In the same year, the Bank Indonesia launched a fintech Regulatory Sandbox, allowing fintech companies [Payment system development (including blockchain and distributed ledgers); aggregate payment; Internet investment management and risk control; Internet insurance; credit,

Table 1 Fintech regulatory authorities and their responsibilities

Regulatory authorities	Specific regulatory responsibilities
Bank Indonesia	Electronic wallet, electronic cash, payment gateway, principal, conversion company, card issuer and receiver, clearing office, settlement agent, virtual currency, blockchain, national payment gateway, payment transaction support
Otoritas Jasa Keuangan (OJK)	P2P, crowdfunding. Digital banking, insurtech, capital market fintech, venture capital, online financing, data security, consumer protection
Ministry of Information and Communications	Telecommunications, information technology, fintech related to information technology

financing business, and capital allocation; other financial services (as judged by Bank Indonesia).] in six major sectors to perform a six-month test on their services under the supervision of the Bank Indonesia. On August 16, 2018, OJK, based on the experience of the Bank Indonesia in the fintech Regulatory Sandbox and pre-audit mechanism in the payment field, issued the Digital Financial Innovation Regulation No. 13/POJK.02/2018, which proposed a package of regulations on fintech supervision, established a Regulatory Sandbox system, and filled the blank of Indonesian Bank Regulation No. 19/12/PBI/2017.

Strengthen international fintech cooperation. In October 2018, at the annual meeting of the International Monetary Fund and the World Bank, the Indonesian government promoted the adoption of the FinTech Agenda. In the same year, Chinas Alibaba Cloud announced the establishment of its first data center in Indonesia and officially put it into operation. Since then, various Chinese fintech companies and investors have entered the Indonesian market. In September 2020, the Securities Commission of Malaysia (SC) signed a fintech cooperation agreement with Indonesias Otoritas Jasa Keuangan (OJK) in order to establish a cooperation framework to develop fintech ecosystems in the two markets.

2.8.3 Layout of Key Fintech Cities: Jakarta—The Rapid Development of Fintech with Inclusive Finance as the Core

Jakarta is the capital, the largest city, and the economic center of Indonesia. The Greater Jakarta Region surrounding the surrounding towns is the second largest metropolitan area in the world. Its industry is dominated by finance, accounting for about one-third of the countrys GDP. It has the largest financial and major industrial and commercial institutions in the country. Stock exchanges and futures exchanges are all located in Jakarta. At the same time, Jakarta is also Indonesias fintech hub city. At present, most fintech companies are located in Jakarta (or Greater Jakarta Region), and domestic business customers are basically in the same area.

Jakarta has become the birthplace of fintech companies, the first test site, and the first launch site for products and services. Indonesias first fintech unicorn company OVO was born in Jakarta, and Akselan, the first equity crowdfunding platform, was officially established in Jakarta. Among the numerous fintech companies, more than 70% are engaged in digital financial inclusion business, mainly providing financing and lending services for small and micro enterprises and rural populations.

2.9 *Hong Kong of China—The Government Assists the Strong Development of Fintech*

2.9.1 Development Features: Seek Innovation While Maintaining Stability to Build a Fintech Hub

The enthusiasm for the development of fintech is booming. Hong Kong has a huge financial system, relatively complete financial ecology, and favorable conditions for the development of fintech. The construction and development of the Guangdong-Hong Kong-Macao Greater Bay Area also brings more opportunities for development of fintech in Hong Kong. According to statistics from the Hong Kong Investment Promotion Agency, there are currently more than 600 fintech companies in Hong Kong covering multiple business areas. According to relevant KPMG data and its statistics on global investment and financing, Hong Kong, China ranked ninth in Asia in terms of investment and financing in 2018. Between 2014 and 2018, the total investment of fintech companies operating in Hong Kong amounted to USD 1.1 billion. In the first half of 2019, fintech companies in Hong Kong raised a total of USD 150 million of fund, an increase of 561% on a year-on-year basis. At present, Hong Kongs key application areas of fintech involve mobile payment, cross-border e-commerce payment, securities payment settlement, online financing platform, wealth technology, commercial insurance, etc.

The internationalization characteristics in various fields of fintech are obvious. In terms of payment and settlement, with a wide variety of financial products in Hong Kong, coupled with the opening of Shanghai-Hong Kong Stock Connect, Shenzhen-Hong Kong Stock Connect, and Bond Connect, post-trade processing platforms have huge space for fintech. In terms of wealth management, as an international investment and asset management center, Hong Kong has already applied a large number of technologies in the field of asset management, such as computerized transactions and investments, and still has great potential in automated consulting, big data, and artificial intelligence. In terms of cross-border e-commerce in trade field, it involves payment and exchange in multiple currencies. Hong Kong has the obvious advantages of free currency convertibility and offshore RMB center. Coupled with tax laws and other conditions, Hong Kong is still preferred cross-border e-commerce. Relying on the above advantages, a large number of cross-border payment companies have recently appeared in Hong Kong. In terms of supply chain finance, the blockchain trade financing platform is the construction focus in Hong Kong.

The fintech talent training system is complete. Currently, the main body of cultivating fintech talents in Hong Kong mainly includes two natures, namely, colleges and universities and social organizations. In terms of colleges and universities, many colleges and universities have set up fintech majors at the undergraduate, master, and doctoral levels. The Chinese University of Hong Kong, the University of Hong Kong, City University of Hong Kong, and the Open University of Hong Kong offer fintech major at the undergraduate level; the Chinese University of Hong Kong, Hong Kong University of Science and Technology, Hong Kong Baptist University, etc. offer such

major at the masters level; and the Hong Kong Polytechnic University offers fintech major. In terms of social organizations, organizations such as the Hong Kong Youth Association Continuing Education Center, the Vocational Training Council, the Institute of Financial Technologists of Asia, and the Hong Kong Institute of Bankers have attracted or strengthened the training of fintech talents in online and offline methods by setting up basic fintech courses and issuing certificates.

2.9.2 Fintech Development Policies and Measures: Innovative Support Measures Contribute to Strong Development of Fintech

The government supervision service system is efficient and perfect. Hong Kong has successively established and improved relevant government service systems. First, specialized agencies are established and a Regulatory Sandbox is established. The government of Hong Kong Special Administrative Region has established an Innovation and Technology Bureau to coordinate the development of fintech. At the same time, the Hong Kong Monetary Authority, the Securities and Futures Commission, and the Insurance Regulatory Authority have respectively set up fintech Regulatory Sandboxes to provide companies with a pilot-based regulatory environment for the application of innovative technologies. At the same time, the government has also set up a fast track for Internet insurance sales companies, such as ZhongAn Insurance and other online insurance companies, to apply for licenses. Second, the introduction of corporate resources is strengthened. The Hong Kong Investment Promotion Agency has established a fintech task team to successfully attract 19 fintech companies to settle in Hong Kong and provide assistance to more than 310 fintech companies.

Integrate into the development of the Guangdong-Hong Kong-Macao Greater Bay Area. Hong Kong cooperates with companies such as Tencent, etc., and the Hong Kong Monetary Authority has issued the first batch of third-party payment licenses to them. Through the deployment of WeChat Hong Kong Wallet, Passenger QR Code, We Remit and other products, Tencent has well integrated with the advantages of Hong Kong and Macau based on its accumulated mobile payment capabilities for many years. With the support from all regulatory parties, Tencents E-Pass will give priority to pilot virtual multi-certificate integration in the Guangdong-Hong Kong-Macao Greater Bay Area, which can meet the needs of residents in Guangdong, Hong Kong, and Macau to use a unified digital identity to enjoy multiple services so as to realize the interconnection in Guangdong-Hong Kong-Macao Greater Bay Area.

A Probability Inequality with Application to Lattice Theory

Tian Kun

Abstract Here we mainly provide a probability inequality about GGH public-key encryption scheme. Given a constant σ, we first choose a lattice vector $v \in \mathbb{Z}^n$, and a small error vector e is generated satisfying $|e| \leqslant \sigma$. The ciphertext result c could be computed by the function $f_{B,\sigma}(v, e) = Bv + e$ with a public basis B. To extract the message v, the function $f_{B,\sigma}^{-1}(c) = B^{-1}[c]_R$ will be used based on the private basis R. In this work we produce a bound for the error probability of $v \neq B^{-1}[c]_R$. We also illustrate the way choosing σ such that the error probability is arbitrarily small.

Keywords Probability inequality · Encryption scheme · Lattice

1 Introduction

Given a full-rank lattice $L \subset \mathbb{Z}^n$, we denote the public basis of L by B and private basis of L by R. Both B and R are $n \times n$ invertible matrices. In the GGH public-key encryption scheme, for a plaintext vector $v \in \mathbb{Z}^n$, the random error vector e is chosen by setting the absolute value of each entry no more than a constant σ, where σ is a positive real number. The ciphertext c is computed by $c = f_{B,\sigma}(v, e) = Bv + e \in \mathbb{R}^n$. Using the results of BaBai and some other ones (Ajtai, 1996; Ajtai & Dwork, 1997; Babai, 1986; Coppersmith & Shamir, 1997; Goldreich et al., 1997; Micciancio, 2001; Hoffstein et al., 2017, 1998), we can decipher the plaintext $v = B^{-1}[c]_R$ given B, R and ciphertext c. Here the lattice point $[c]_R$ is obtained by representing c as a linear combination on the columns of R and rounding the coefficients in this linear combination to the nearest integers. The problem is that how σ should be chosen so that we can get a right plaintext v or guarantee a low error probability. We show three theorems to solve this problem. A probability inequality is given to estimate the bound of inversion error probability.

T. Kun (✉)
School of Mathematics, Renmin University of China, Beijing 100872, China
e-mail: tkun19891208@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_2

2 Main Results

Theorem 1 *B is the public basis and R is the private basis of lattice L. $v \in \mathbb{Z}^n$, e is the random error vector, $|e|_\infty \leqslant \sigma$, $c = f_{B,\sigma}(v, e) = Bv + e$. Then $B^{-1}[c]_R = v$ if and only if $[R^{-1}e] = 0$, here $[R^{-1}e]$ denotes the vector in $\mathbb{Z}^n$ which is obtained by rounding each entry in $R^{-1}e$ to the nearest integer.*

Proof Let $T = B^{-1}R$, then

$$B^{-1}[c]_R = B^{-1}[Bv + e]_R = B^{-1}R[R^{-1}(Bv + e)] = T[T^{-1}v + R^{-1}e]$$

since $T = B^{-1}R$ is a unimodular matrix, T^{-1} is also a unimodular matrix. $v \in \mathbb{Z}^n$, so $T^{-1}v \in \mathbb{Z}^n$.

$$B^{-1}[c]_R = T[T^{-1}v + R^{-1}e] = v + T[R^{-1}e]$$

Thus $B^{-1}[c]_R = v$ is equivalent to $T[R^{-1}e] = 0$, and this equality holds if and only if $[R^{-1}e] = 0$.

Remark 1 This theorem gives an equivalent condition to check whether the decryption result is accurate.

Theorem 2 *Let R be the private basis of lattice L. e is the random error vector such that $|e|_\infty \leqslant \sigma$. Suppose the maximum L_1 norm of the rows in R^{-1} is ρ. Then if $\sigma < \frac{1}{2\rho}$, $[R^{-1}e] = 0$ holds.*

Proof Let $R^{-1} = (c_{ij})_{n\times n}$, $R^{-1}e = (a_1, a_2, ..., a_n)^T$, i.e., $a_i = \sum_{j=1}^{n} c_{ij}e_j$, $1 \leqslant i \leqslant n$.

$$|a_i| = |\sum_{j=1}^{n} c_{ij}e_j| \leqslant |e_j||\sum_{j=1}^{n} c_{ij}| \leqslant \sigma\rho < \frac{1}{2}$$

This means that $[R^{-1}e] = 0$.

Remark 2 Theorem 2 shows how σ can be chosen so that no inversion error occurs.

Theorem 3 *Let an $n \times n$ matrix R be the private basis used in the inversion of $f_{B,\sigma}$, and denote the maximum L_∞ norm of the rows in R^{-1} by $\frac{r}{\sqrt{n}}$. Then the probability of inversion errors is bounded by*

$$P\{[R^{-1}e] \neq 0\} \leqslant 2n \cdot exp\left(-\frac{1}{8\sigma^2 r^2}\right),$$

here $e = (e_1, e_2, ..., e_n)^T$ and $e_1, e_2, ..., e_n$ are n independent random variables such that $|e_i| \leqslant \sigma$ and $E(e_i) = 0$ for $1 \leqslant i \leqslant n$.

Lemma 1 *For any non-negative random variable X with finite expectation $E(X)$ and any positive real number μ, we have*

$$P\{X \geqslant \mu\} \leqslant \frac{E(X)}{\mu}.$$

Proof Here we treat X as a random variable of continuous type. For the other situations, the proof is similar. Let $f(x)$ be the probability density function of X. Since $E(X) = \int_0^{+\infty} xf(x)\mathrm{d}x \geqslant \int_\mu^{+\infty} xf(x)\mathrm{d}x \geqslant \int_\mu^{+\infty} \mu f(x)\mathrm{d}x = \mu P\{X \geqslant \mu\}$, then we have $P\{X \geqslant \mu\} \leqslant \frac{E(X)}{\mu}$.

Lemma 2 *Given random variable X satisfying $-a \leqslant X \leqslant a$ with $E(X) = 0$, here $a > 0$. For any real number λ, we have*

$$E(e^{\lambda X}) \leqslant exp\left(\frac{\lambda^2 a^2}{2}\right).$$

Proof For any real number λ, $f(x) = e^{\lambda x}$ is a convex function. Notice that

$$x = \frac{x+a}{2a} \cdot a + \frac{a-x}{2a} \cdot (-a), \quad -a \leqslant x \leqslant a$$

then

$$f(x) \leqslant \frac{x+a}{2a} f(a) + \frac{a-x}{2a} f(-a)$$

$$e^{\lambda x} \leqslant \frac{x+a}{2a} e^{\lambda a} + \frac{a-x}{2a} e^{-\lambda a}$$

$$E(e^{\lambda X}) \leqslant E(\frac{X+a}{2a} e^{\lambda a} + \frac{a-X}{2a} e^{-\lambda a}) = \frac{1}{2}(e^{\lambda a} + e^{-\lambda a})$$

Let $t = \lambda a$, next we prove that $\frac{1}{2}(e^t + e^{-t}) \leqslant \exp(\frac{t^2}{2})$. This inequality is equivalent to

$$\ln \frac{e^t + e^{-t}}{2} \leqslant \frac{t^2}{2}$$

Let $g(t) = \frac{t^2}{2} - \ln \frac{e^t + e^{-t}}{2}$, then $g'(t) = t - \frac{e^t - e^{-t}}{e^t + e^{-t}}$ and $g'(0) = 0$. Since $g''(t) \geqslant 0$, we get $g'(t) \leqslant 0$ if $t \leqslant 0$ and $g'(t) \geqslant 0$ if $t \geqslant 0$. Then $g(t) \geqslant g(0) = 0$ and we complete the proof.

Lemma 3 *Suppose $X_1, X_2, ..., X_n$ are n independent random variables. For $1 \leqslant i \leqslant n$, we have $-a \leqslant X_i \leqslant a$ and $E(X_i) = 0$, here $a > 0$. Let $S_n = \sum_{i=1}^n X_i$, $\varepsilon > 0$, then*

$$P\{|S_n| \geqslant \varepsilon\} \leqslant 2exp(-\frac{\varepsilon^2}{2na^2}).$$

Proof For any $\lambda > 0$, based on Lemma 1, we can get

$$P\{S_n \geqslant \varepsilon\} = P\{e^{\lambda S_n} \geqslant e^{\lambda\varepsilon}\} \leqslant \frac{E(e^{\lambda S_n})}{e^{\lambda\varepsilon}}$$

Since $X_1, X_2, ..., X_n$ are independent random variables, combine with Lemma 2,

$$E(e^{\lambda S_n}) = \prod_{i=1}^{n} E(e^{\lambda X_i}) \leqslant \prod_{i=1}^{n} e^{\frac{\lambda^2 a^2}{2}} = e^{\frac{n\lambda^2 a^2}{2}}$$

$$P\{S_n \geqslant \varepsilon\} \leqslant \frac{E(e^{\lambda S_n})}{e^{\lambda\varepsilon}} \leqslant e^{-\lambda\varepsilon + \frac{n\lambda^2 a^2}{2}}$$

Let $\lambda = \frac{\varepsilon}{na^2}$, therefore, the above inequality becomes to

$$P\{S_n \geqslant \varepsilon\} \leqslant \exp\left(-\frac{\varepsilon^2}{2na^2}\right)$$

In the same way, we can prove that

$$P\{S_n \leqslant -\varepsilon\} \leqslant \exp\left(-\frac{\varepsilon^2}{2na^2}\right)$$

Thus

$$P\{|S_n| \geqslant \varepsilon\} \leqslant 2\exp\left(-\frac{\varepsilon^2}{2na^2}\right)$$

Proof of Theorem 3. Now we can prove Theorem 3 given at first according to Lemma 3.
Let $R^{-1} = (c_{ij})_{n\times n}$, $e = (e_1, e_2, ..., e_n)^T$, here $e_1, e_2, ..., e_n$ are n independent random variables such that $|e_i| \leqslant \sigma$ and $E(e_i) = 0$ for $1 \leqslant i \leqslant n$.
We denote $R^{-1}e = (a_1, a_2, ..., a_n)^T$, i.e., $a_i = \sum_{j=1}^{n} c_{ij}e_j$, $1 \leqslant i \leqslant n$.
Since $|c_{ij}| \leqslant \frac{r}{\sqrt{n}}$ and $|e_j| \leqslant \sigma$, then the random variable $c_{ij}e_j$ is limited to the interval $[-\frac{r\sigma}{\sqrt{n}}, \frac{r\sigma}{\sqrt{n}}]$. Based on Lemma 3,

$$P\{|a_i| \geqslant \frac{1}{2}\} = P\{|\sum_{j=1}^{n} c_{ij}e_j| \geqslant \frac{1}{2}\} \leqslant 2\exp(-\frac{(\frac{1}{2})^2}{2n(\frac{r\sigma}{\sqrt{n}})^2}) = 2\exp(-\frac{1}{8\sigma^2 r^2})$$

$$P\{[R^{-1}e] \neq 0\} \leqslant \sum_{i=1}^{n} P\{|a_i| > \frac{1}{2}\} \leqslant \sum_{i=1}^{n} P\{|a_i| \geqslant \frac{1}{2}\} \leqslant 2n \cdot \exp(-\frac{1}{8\sigma^2 r^2})$$

Thus the inequality in Theorem 3 holds.

Corollary 1 $P\{[R^{-1}e] \neq 0\} < \varepsilon$ *if* $\sigma < \left(2r\sqrt{2\ln\frac{2n}{\varepsilon}}\right)^{-1}$.

Proof $\sigma < \left(2r\sqrt{2\ln\frac{2n}{\varepsilon}}\right)^{-1} \Leftrightarrow 2n \cdot \exp\left(-\frac{1}{8\sigma^2 r^2}\right) < \varepsilon$, from Theorem 3,

$$P\{[R^{-1}e] \neq 0\} \leqslant 2n \cdot \exp\left(-\frac{1}{8\sigma^2 r^2}\right) < \varepsilon$$

Remark 3 Theorem 3 provides a way to estimate the bound of inversion error probability, and Corollary 1 gives a detailed bound for σ based on Theorem 3 to get the error probability no more than a constant ε.

3 Conclusions

In this work we mainly present a probability inequality about GGH public-key encryption scheme. In this scheme, we first take a lattice vector $v \in \mathbb{Z}^n$ and generate a small error vector e such that $|e| \leqslant \sigma$. Given a public basis B, the function $f_{B,\sigma}(v, e) = Bv + e$ computes the ciphertext result c. To decrypt, the private basis R and the function $f^{-1}_{B,\sigma}(c) = B^{-1}[c]_R$ will be used to extract the message v. We give a bound for the error probability of $v \neq B^{-1}[c]_R$ and explain how to choose σ in order to obtain the error probability no more than a given constant ε.

References

Ajtai, M. (1996). Generating hard instances of lattice problems. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing* (pp. 99–108).

Ajtai, M., Dwork, C. (1997). A public-key cryptosystem with worst-case/average-case equivalence. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing* (pp. 284–293).

Babai, L. (1986). On Lovasz lattice reduction and the nearest lattice point problem. *Combinatorica, 6,* 1–13.

Coppersmith, D., & Shamir, A. (1997). Lattice attacks on NTRU. *Advances in Cryptology, 1233,* 52–61.

Goldreich, O., Goldwasser, S., & Halevi, S. (1997). Public-key cryptosystems from lattice reduction problems. *Annual International Cryptology Conference, 1294,* 112–131.

Hoffstein, J., Pipher, J., & Silverman, J. H. (1998). NTRU: a ring based public key cryptosystem. *Algorithmic Number Theory, 1423,* 267–288.

Hoffstein, J., Pipher, J., Schanck, J. M., et al. (2017). Choosing parameters for NTRUEncrypt. *Topics in Cryptology, 10159,* 3–18.

Micciancio, D. (2001). Improving lattice based cryptosystems using the hermite normal form. *International Cryptography and Lattices Conference, 2146,* 126–145.

Robust Identification of Gene-Environment Interactions Under High-Dimensional Accelerated Failure Time Models

Qingzhao Zhang, Hao Chai, Weijuan Liang, and Shuangge Ma

Abstract For complex diseases, beyond the main effects of genetic (G) and environmental (E) factors, gene-environment (G-E) interactions also play an important role. Many of the existing G-E interaction methods conduct marginal analysis, which may not appropriately describe disease biology. Joint analysis methods have been developed, with most of the existing loss functions constructed based on likelihood. In practice, data contamination is not uncommon. Development of robust methods for interaction analysis that can accommodate data contamination is very limited. In this study, we consider censored survival data and adopt an accelerated failure time (AFT) model. An exponential squared loss is adopted to achieve robustness. A sparse group penalization approach, which respects the "main effects, interactions" hierarchy, is adopted for estimation and identification. Consistency properties are rigorously established. Simulation shows that the proposed method outperforms direct competitors. In data analysis, the proposed method makes biologically sensible findings.

Keywords Robust identification · Gene-environment interactions · High-dimensional · Accelerated failure time models

Q. Zhang
School of Economics and the Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China
e-mail: zhangqingzhao@amss.ac.cn

H. Chai · S. Ma (✉)
Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, US
e-mail: shuangge.ma@yale.edu

W. Liang
School of Statistics, Renmin University of China, Beijing, China
e-mail: weijuanliang@yeah.net

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_3

1 Introduction

For many complex diseases, it is essential to identify important risk factors that are associated with prognosis. In the omics era, profiling studies have been extensively conducted. It has been found that, beyond the main effects of genetic (G) and environmental (E) risk factors, gene-environment (G-E) interactions can also have important implications.

Denote T and C as the prognosis and censoring times, respectively. Denote $X = (X_1, \ldots, X_q)^\top$ as the q environmental/clinical variables, and $Z = (Z_1, \ldots, Z_p)^\top$ as the p genetic variables. The existing G-E interaction analysis methods mainly belong to two families. The first family conducts marginal analysis (Hunter, 2005; Shi et al., 2014; Thomas, 2010), under which one or a small number of genes are analyzed at a time. Despite its significant computational simplicity, marginal analysis contradicts the fact that the prognosis of complex diseases is attributable to the *joint* effects of multiple main effects and interactions. The second family of methods, which is biologically more sensible, conducts joint analysis (Liu et al., 2013; Wu et al., 2014; Zhu et al., 2014). Among the existing joint analyses, the regression-based is the most popular and proceeds as follows. Consider the model $T \sim \phi(\alpha_0 + \sum_{j=1}^{q} X_j\alpha_j + \sum_{k=1}^{p} Z_k\beta_k + \sum_{j=1}^{q}\sum_{k=1}^{p} X_j Z_k \gamma_{j,k})$, where model $\phi(\cdot)$ is known up to the regression coefficients α_0, $\{\alpha_j\}_1^q$, $\{\beta_k\}_1^p$, and $\{\gamma_{j,k}\}_1^{q,p}$. Conclusions on the importance of interactions are drawn based on $\{\gamma_{j,k}\}_1^{q,p}$. With the high data dimensionality and demand for the selection of relevant effects, regularized estimation is usually needed.

In the dominating majority of the existing studies, estimation is based on the standard likelihood, which is nonrobust. In practice, data contamination is not uncommon and can be caused by multiple reasons. Many diseases are heterogeneous, and different subtypes behave differently. When the subtype information is accurately available, subtype-specific analysis can be conducted. However, when such information is not or partially available, which is often the case in practice (He et al., 2015), subjects belonging to small subtypes may be viewed as "contamination" to those of the leading subtype. Human errors can also happen. It has been well noted that survival information extracted from medical records is not always reliable (Bowman, 2015; Fall et al., 2008), creating contamination in prognosis distributions. In low-dimensional biomedical studies, it has been well established that even a single contaminated observation can lead to biased model estimation and so false marker identification (Huber & Ronchetti, 2009). Our literature review suggests that in the analysis of G-E interactions, robust methods that can effectively accommodate contamination in prognosis outcomes have been very rare. For marginal interaction analysis, a few robust methods, for example, the multifactor dimensionality reduction (MDR), have been developed. However, they are not directly applicable to joint analysis because of both methodological and computational challenges. As discussed in (Wu & Ma, 2015), a handful of robustness studies have been conducted under high-dimensional settings for joint analysis. However, they are mostly on main effects and not directly applicable to interaction analysis because of the additional complexity caused by the

"main effects, interactions" hierarchy. Most of them adopt the quantile regression technique. Studies under low-dimensional settings suggest that no robust technique can dominate. It is thus desirable to examine alternative robust techniques under high-dimensional settings. In addition, for quite a few existing methods, statistical properties have not been well studied, casting doubts on their validity.

Consider data with a prognosis outcome and both G and E measurements. Our goal is to conduct joint analysis and identify important G-E interactions and main G and E effects. This study advances from the literature in multiple aspects. Specifically, we consider the scenario with possible contamination in the prognosis outcome, which is commonly encountered but little addressed. We adopt an exponential squared loss to achieve robustness. This loss function provides a useful alternative to the popular quantile regression and other robust approaches but has not been well investigated under high-dimensional settings, especially not for interaction analysis. This study also marks a novel extension of the exponential squared loss to accommodate censored survival data. For regularized estimation and selection of relevant effects, we propose adopting a penalization technique, which respects the "main effects, interactions" hierarchy. Significantly advancing from most of the existing studies, consistency properties are rigorously established. Theoretical research for high-dimensional robust methods remains limited. As such, this study may provide valuable insights. With both methodological and theoretical developments, this study is warranted beyond the existing literature.

2 Methods

2.1 Data and Model Settings

For describing prognosis, we adopt the AFT model, which has been the choice of multiple studies with high-dimensional genetic data (Liu et al., 2013; Shi et al., 2014). Compared to alternatives including the Cox model, advantages of the AFT model include intuitive interpretations and low computational cost, which are especially desirable with high-dimensional genetic data. With a slight abuse of notation, still use T and C to denote the logarithms of the event and censoring times, and $\delta = I_{\{T \le C\}}$. The AFT model specifies that

$$T = \alpha_0 + \sum_{j=1}^{q} X_j \alpha_j + \sum_{k=1}^{p} Z_k \beta_k + \sum_{j=1}^{q} \sum_{k=1}^{p} X_j Z_k \gamma_{j,k} + \varepsilon,$$

where ε is the random error. Following Stute (1993, 1996), we assume that T and C are independent, and δ is conditionally independent of $(X^\top, Z^\top)^\top$ given T. Let $W_k = (Z_k, X_1 Z_k, \ldots, X_q Z_k)^\top$ and $b_k = (\beta_k, \gamma_{1,k}, \ldots, \gamma_{q,k})^\top$, which represent all main and interaction effects corresponding to the kth genetic variable.

With n independent subjects, use subscript "i" to denote the ith subject. For subject i, let $y_i = \min\{T_i, C_i\}$ and $\delta_i = I_{\{T_i \le C_i\}}$ be the observed time and event indicator, respectively. Then the ith observation consists of $(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$, and $W_{k,i} = (z_{ik}, x_{i1}z_{ik}, \ldots, x_{iq}z_{ik})^\top$ denoting the ith realization of X, Z, and W_k, respectively. Denote $\mathbf{u}_{i,}^\top = (1, \mathbf{x}_i^\top, W_{1,i}^\top, \ldots, W_{p,i}^\top)$, $\mathbf{U} = (\mathbf{u}_{1,}, \cdots, \mathbf{u}_{n,})^\top$, and $\zeta = (\alpha_0, \ldots, \alpha_q, b_1^\top, \ldots, b_p^\top)^\top$. Without loss of generality, assume that $(y_i, \delta_i, \mathbf{u}_{i,})$'s have been sorted according to y_i's in an ascending manner.

2.2 Robust Estimation and Identification

Consider the scenario where the distribution of ε is not specified, which significantly differs from the existing parametric studies and makes the proposed method more flexible. To motivate the proposed estimation, first consider data without contamination. Stute (1993) developed a weighted least squared estimation approach. Under low-dimensional settings, Stute's estimator is defined as the minimizer of the loss function

$$\sum_{i=1}^{n} \omega_i (y_i - \mathbf{u}_{i,}^\top \zeta)^2.$$

Here the weights $\boldsymbol{\omega} = (\omega_i)_{i=1}^n$ are computed based on the Kaplan-Meier estimation and defined as

$$\omega_1 = \frac{\delta_1}{n}, \omega_i = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_j}, i = 2, \ldots, n.$$

It is noted that Stute's estimator is not necessarily the most efficient. However, under high-dimensional settings, it can be computationally the most convenient with the least squared loss. It can be seen that, if $\omega_i \neq 0$, one contaminated y_i can lead to severely biased model estimation.

Now consider the scenario with possible outliers in the prognosis data. We propose the objective function

$$Q_\theta(\zeta) = \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta). \tag{1}$$

This function has been motivated by the following considerations. Under low-dimensional regression analysis without censoring, (Wang et al., 2013) adopted an exponential squared loss to achieve robustness. The intuition is as follows. For a contaminated subject with the observed y_i deviating from $\mathbf{u}_{i,}^\top \zeta$ (the "predicted" value based on the model), $(y_i - \mathbf{u}_{i,}^\top \zeta)^2$ has a large value. The exponential function down-

weighs such a contaminated observation. The degree of down-weighing is adjusted by θ: when θ gets smaller, the contaminated observations have smaller influence. While sharing certain similar ground as (Wang et al., 2013) and others, the present study has three main challenges/advancements. The first is the high dimensionality, which brings tremendous challenges to theoretical and computational developments. The second is the need to respect the "main effects, interactions" hierarchy (more details below). The third is censoring, to accommodate which we introduce the weight function ω_i motivated by Stute's approach. As the weights are data-dependent, they bring challenges to the establishment of theoretical properties.

When $p \gg n$, regularized estimation is needed. In addition, out of a large number of profiled G factors and G-E interactions, only a few are expected to be associated with prognosis. We adopt penalization for regularized estimation and identification, which has been the choice of a large number of genetic studies, especially recent interaction analyses (Bien et al., 2013; Liu et al., 2013; Shi et al., 2014). Specifically, consider the penalized robust objective function

$$L_{\lambda_1,\lambda_2,\theta}(\zeta) = Q_\theta(\zeta) - \sum_{k=1}^{p} \rho(\|b_k\|; \lambda_1, s) - \sum_{k=1}^{p}\sum_{j=2}^{q+1} \rho(|b_{kj}|; \lambda_2, s), \qquad (2)$$

where $\|\cdot\|$ is the ℓ_2 norm, $\rho(t; \lambda, s) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda s}\right)_+ dx$ is the MCP (minimax concave penalty, (Zhang, 2010)), and b_{kj} is the jth element of b_k. λ_1 and λ_2 are data-dependent tuning parameters, and s is the regularization parameter, per the terminologies in Zhang (2010). The robust estimator is defined as the maximizer of $L_{\lambda_1,\lambda_2,\theta}(\zeta)$. An interaction term (or main effect) is concluded as important if its estimate is nonzero.

In recent genetic interaction analysis, it has been stressed that the "main effects, interactions" hierarchy should be respected. That is, if an interaction term is identified as important, its corresponding main effect(s) should be automatically identified. G-E interaction analysis has its uniqueness. The E variables usually have a low dimensionality and are manually chosen. As such, selection is usually not conducted on the E variables (if desirable, this can be easily achieved). Thus for G-E interaction analysis, the hierarchy postulates that if an G-E interaction is identified as important, its corresponding main G effect is automatically identified. In the adopted sparse group penalty, the first penalty, which is a group MCP, determines which groups are selected. Here one group corresponds to one genetic variable and its interactions. As the group MCP does not have within-group sparsity, the second penalty is imposed, where we penalize the interaction terms and determine which are nonzero. With the special design that the second penalty is only imposed on interactions, important interactions correspond to important groups, automatically leading the estimates of the corresponding main G effects nonzero. As such, the combination of the two penalties guarantees the hierarchy. We note that although sparse group penalization has been studied in the literature (Liu et al., 2013), it has been very rarely coupled with robust loss functions. It is also noted that MCP can be potentially replaced by other penalties.

2.3 Computation

In this section, we develop an efficient algorithm to compute the maximizer of $L_{\lambda_1,\lambda_2,\theta}(\zeta)$. The basic strategy is to iteratively approximate the objective function by its quadratic minorization. Then a coordinate-wise updating procedure is used to find the maximizer of each approximated objective function. The maximizer then serves as the starting point for the next minorization. Overall, this is a coordinate-descent (CD) algorithm nested in a Minorize-Maximization (MM) algorithm.

Let $\mathbf{W}(\zeta)$ be a diagonal matrix with the ith diagonal element $\mathbf{W}_{i,i} = 2\omega_i \exp(-(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta)/\theta$. Also let $\mathbf{v}(\zeta) = (v_1, \cdots, v_n)^\top$ with $v_i = y_i - \mathbf{u}_{i,}^\top \zeta$. Define $\mathbf{U}_{,-j}$ as the sub-matrix of $\mathbf{U}$ with the jth column excluded. Define $\mathbf{u}_{,j}$ as the jth column of matrix $\mathbf{U}$, and $u_{i,j}$ as the jth component of vector $\mathbf{u}_{i,}$. Similarly, define $\boldsymbol{\zeta}_{-j}$ as the sub-vector of ζ with the jth element excluded. For the exponential squared objective function in (1), its first- and second-order derivatives with respect to ζ are

$$\frac{\partial Q_\theta(\zeta)}{\partial \zeta_j} = 2\sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta) u_{i,j}(y_i - \mathbf{u}_{i,}^\top \zeta)/\theta = \mathbf{u}_{,j}^\top \mathbf{W}(\zeta)\mathbf{v}(\zeta),$$

$$\frac{\partial^2 Q_\theta(\zeta)}{\partial \zeta_j \partial \zeta_k} = 2\sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta) u_{i,j} u_{i,k}[2(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta - 1]/\theta.$$

If $(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta > 0.5$, $\frac{\partial^2 Q_\theta(\zeta)}{\partial \zeta_j \partial \zeta_k} \geq 0$. On the other hand if $(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta \leq 0.5$, $\frac{\partial^2 Q_\theta(\zeta)}{\partial \zeta_j \partial \zeta_k} \leq 0$. Hence, to find the maximizer of $Q_\theta(\zeta)$, the simple Newton-Raphson approach may lead to infinity if the starting value is too far from the true value. To tackle this problem, a minorization of $Q_\theta(\zeta)$ is used to approximate $Q_\theta(\zeta)$. Note that $\frac{\partial^2 Q_\theta(\zeta)}{\partial \zeta_j \partial \zeta_k} \geq -2\sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,}^\top \zeta)^2/\theta) u_{i,j} u_{i,k}/\theta$. Hence a minorized approximation to $Q_\theta(\zeta)$ at ζ^m is

$$Q_\theta(\zeta^m) + \mathbf{v}^\top(\zeta^m)\mathbf{W}(\zeta^m)\mathbf{U}(\zeta - \zeta^m) - \frac{1}{2}(\zeta - \zeta^m)^\top \mathbf{U}^\top \mathbf{W}(\zeta^m)\mathbf{U}(\zeta - \zeta^m).$$

Note that $\zeta^m = (\alpha_0^m, \ldots, \alpha_q^m, b_1^{m\top}, \ldots, b_p^{m\top})^\top$ with $b_k^m = (\beta_k^m, \gamma_{1,k}^m, \ldots, \gamma_{q,k}^m)^\top$. For the penalty, we apply a local linear approximation at ζ^m, which is given by

$$-\sum_{k=1}^{p} \dot{\rho}(\|b_k^m\|; \lambda_1, s)\frac{|\beta_k^m|}{\|b_k^m\|}|\beta_k| - \sum_{k=1}^{p}\sum_{j=1}^{q}\left\{\dot{\rho}(\|b_k^m\|; \lambda_1, s)\frac{|\gamma_{j,k}^m|}{\|b_k^m\|} + \dot{\rho}(|\gamma_{j,k}^m|; \lambda_2, s)\right\}|\gamma_{j,k}|$$

if the terms that do not depend on ζ are ignored, where $\dot{\rho}(t; \lambda, s) = sgn(t)\left(\lambda - \frac{|t|}{s}\right)_+$. If we replace $Q_\theta(\zeta)$ in (2) with its minorized approximation and plug in the approximation of the penalty, the penalized objective function then has the form

$$L_{\lambda_1,\lambda_2,\theta}(\zeta|\zeta^m) = Q(\zeta^m) + \mathbf{v}^\top(\zeta^m)\mathbf{W}(\zeta^m)\mathbf{U}(\zeta - \zeta^m)$$
$$- \frac{1}{2}(\zeta - \zeta^m)^\top \mathbf{U}^\top \mathbf{W}(\zeta^m)\mathbf{U}(\zeta - \zeta^m) - \sum_{k=1}^{p} \dot{\rho}(\|b_k^m\|; \lambda_1, s)\frac{|\beta_k^m|}{\|b_k^m\|}|\beta_k|$$
$$- \sum_{k=1}^{p}\sum_{j=1}^{q}\left\{\dot{\rho}(\|b_k^m\|; \lambda_1, s)\frac{|\gamma_{j,k}^m|}{\|b_k^m\|} + \dot{\rho}(|\gamma_{j,k}^m|; \lambda_2, s)\right\}|\gamma_{j,k}|. \quad (3)$$

This function has a "weighted quadratic + penalty" form and can be optimized using the coordinate-descent approach.

The algorithm starts with $m = 0$ and $\zeta^m = \mathbf{0}$, where m is the index of the MM iteration. At iteration m, the objective function is approximated by its minorization $L_{\lambda_1,\lambda_2,\theta}(\zeta|\zeta^m)$ given in (3). Then the penalized weighted quadratic function is maximized using the coordinate-descent algorithm. Denote $\bar{\zeta}^{old}$ as the estimate of ζ before updating. We update each element of the estimate and denote the new estimate as $\bar{\zeta}^{new}$. This is repeated until the distance between $\bar{\zeta}^{old}$ and $\bar{\zeta}^{new}$ is smaller than a prefixed constant. Then $\zeta^{m+1} = \bar{\zeta}^{new}$ serves as the new expansion base point for the next minorization. The overall procedure is repeated until convergence. Convergence properties of the MM and CD techniques have been well studied in the literature. With our problem, the objective function increases at each step and is bounded above, which leads to convergence. In numerical study, we conclude convergence if the difference between two estimates after two consecutive MM steps is small enough. We observe convergence in all numerical examples after a small to moderate number of MM iterations.

The proposed method involves tuning parameters. For s in MCP, we follow (Zhang, 2010) and other published studies, which suggest examining a small number of values or fixing it. In our numerical study, we fix $s = 6$, which has been adopted in published studies (Shi et al., 2014; Xu et al., 2018). We have also examined s values near 6 and observed similar performance (details omitted). In practice, for settings significantly different from ours, other s values may need to be considered. Under low-dimensional settings, (Wang et al., 2013) proposed an iterative approach to select the robust tuning parameter θ. However, their approach is computationally infeasible for high-dimensional data. Under the present setting, for each combination of $(\lambda_1, \lambda_2, \theta)$, we compute the solution. This way, we can obtain a solution surface over a three-dimensional tuning parameter grid. This is feasible as the proposed computational algorithm only involves simple updates and incurs low cost. Then the tuning parameters can be selected using a prediction-based method which proceeds as follows: (a) compute the cross-validated sum of prediction errors for each $(\lambda_1, \lambda_2, \theta)$ combination; (b) for each fixed θ, average the sum of prediction errors over λ_1, λ_2. Select θ that has the smallest average sum of prediction errors; (c) with the selected θ, select λ_1, λ_2 that has the smallest sum of prediction errors. This procedure first groups all (λ_1, λ_2) values together and selects the best θ value. Then with the optimal θ value, the optimal (λ_1, λ_2) values are selected. Our numerical experiments suggest that this procedure generates more stable estimates than directly searching over the three-dimensional $(\lambda_1, \lambda_2, \theta)$ grid.

With a complex robust goodness-of-fit and a penalty that respects the hierarchy, the proposed method is inevitably computationally more expensive than some simpler alternatives. However, as the proposed computational algorithm is composed of relatively simple calculations, the overall computational cost is affordable. With fixed tunings, the analysis of one simulated dataset (described in detail below) takes about nine minutes on a regular laptop. Tuning parameter selection can be conducted in a highly parallel manner to save computer time.

2.4 Consistency Properties

In this section, we rigorously prove that the proposed method can consistently identify the important interactions (and main effects) under ultrahigh-dimensional settings. In the literature, theoretical development for robust methods under high-dimensional settings has been limited. It is especially rare for methods other than the quantile based. With the consistency properties, the proposed method can be preferred over the alternatives whose statistical properties have not been well established. Our theoretical development not only provides a solid ground for the proposed method but also sheds insights for other robust methods under high-dimensional settings.

For any two subsets S_1 and S_2 of $\{1, \cdots, p+q+pq+1\}$ and a matrix H, we denote by $H_{S_1S_2}$ the sub-matrix of H with rows and columns indexed by S_1 and S_2, respectively. Let $\zeta^* = (\alpha_0^*, \ldots, \alpha_q^*, b_1^{*\top}, \ldots, b_p^{*\top})^\top$, where $b_k^* = (\beta_k^*, \gamma_{1,k}^*, \ldots, \gamma_{q,k}^*)^\top$ is the true value of ζ. Here we make the sparsity assumption, under which only a subset of the components of ζ^* is nonzero. Define the three groups of parameters:

$$A_1 = \{\alpha_0^*, \ldots, \alpha_q^*\}, \quad A_2 = \{\gamma_{j,k}^* : \gamma_{j,k}^* \neq 0,\ j = 1, \ldots, q; k = 1, \ldots, p\},$$
$$A_3 = \{\beta_k^* : \beta_k^* \neq 0 \text{ or there exsits some } 1 \leq j \leq q \text{ such that } \gamma_{j,k}^* \neq 0, k = 1, \ldots, p\}.$$

Denote $\mathscr{A}$ as the set of indices of $A_1 \cup A_2 \cup A_3$ in the vector ζ^*. Let $\mathscr{A}^c$ and $|\mathscr{A}|$ denote the complement and cardinality of set $\mathscr{A}$, respectively. We then divide $\mathscr{A}^c$ into there sets of indices $\mathscr{B}_1$, $\mathscr{B}_2$, and $\mathscr{B}_3$, which correspond to the following three sets

$$B_1 = \{\beta_k^* : \beta_k^* = 0, k = 1, \ldots, p\},$$
$$B_2 = \{\gamma_{j,k}^* : \gamma_{j,k}^* = 0 \text{ but } \beta_k^* \neq 0,\ j = 1, \ldots, q; k = 1, \ldots, p\},$$
$$B_3 = \{\gamma_{j,k}^* : \gamma_{j,k}^* = 0 \text{ and } \beta_k^* = 0,\ j = 1, \ldots, q; k = 1, \ldots, p\},$$

respectively. Define

$$D_n(\zeta) = \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_i^\top \zeta)^2/\theta) \frac{2(y_i - \mathbf{u}_i^\top \zeta)}{\theta} \mathbf{u}_i,$$

and

$$I_n(\zeta) = \frac{2}{\theta}\sum_{i=1}^{n}\omega_i \exp(-(y_i - \mathbf{u}_{i,}^{\top}\zeta)^2/\theta)\left(\frac{2(y_i - \mathbf{u}_{i,}^{\top}\zeta)^2}{\theta} - 1\right)\mathbf{u}_{i,}\mathbf{u}_{i,}^{\top}.$$

The following conditions are needed to establish the consistency properties.

C1. T and C are independent, and $P(T \le C|T, X, Z) = P(T \le C|T)$.
C2. The support of T is dominated by that of C. For example, $\tau_T < \tau_C$ or $\tau_T = \tau_C = \infty$, where τ_T and τ_C are the right end points of the support of T and C, respectively.
C3. $E[D_n(\zeta^*)] = 0$.
C4. The distributions of $D_{n,j}(\zeta^*)$'s are subgaussian, that is, $\Pr(|D_{n,j}(\zeta^*)| > t) \le 2\exp\left(-nt^2/\sigma^2\right)$. Moreover, $I_{n,jk}(\zeta) - I_{jk}(\zeta)$'s are subgaussian for all $\zeta \in \Theta = \{\zeta : \|\zeta - \zeta^*\|_2 < \delta\}$, where δ is a positive constant, $I(\zeta) = E[I_n(\zeta)]$, and $I_{jk}(\zeta)$ is the (j,k)th component of matrix $I(\zeta)$. Moreover, there exists a bounded constant κ such that $\boldsymbol{v}^{\top}[I(\zeta^1) - I(\zeta^2)]\boldsymbol{v} \le \kappa\|\zeta^1 - \zeta^2\|_2$ for any $\zeta^1, \zeta^2 \in \Theta$ and $\|\boldsymbol{v}\|_2 = 1$.
C5. $I_{\mathscr{A}\mathscr{A}}(\zeta^*)$ is a $|\mathscr{A}| \times |\mathscr{A}|$ negative-definite matrix. The eigenvalues of $I_{\mathscr{A}\mathscr{A}}(\zeta^*)$ are bounded away from zero and infinity.
C6. $\min_{j,k}\{|\gamma^*_{j,k}| : \gamma^*_{j,k} \ne 0\} \gg \lambda_1 \vee \lambda_2$. $\lambda_1 \wedge \lambda_2 \gg \sqrt{|\mathscr{A}|/n}$.

C1 and C2 have been commonly assumed in the literature. See, for example, (Stute, 1993, 1996; Huang et al., 2007). We note that the independent censoring assumption usually holds in practice, although from a theoretical perspective, quite a few studies have made the weaker conditional independence assumption. We have explored relaxing this assumption and found that alternative and less intuitive assumptions would have to be made. The zero expectations in C3 and C5 ensure the consistency of estimation. C4 is required for Theorem 1, and a similar assumption has been made in (Ma & Du, 2012). C6 requires that the smallest signal does not decay too fast, which is common in studies on high-dimensional inference. The following theorem establishes consistency of the proposed estimator $\widehat{\zeta}$.

Theorem 1 *Suppose that conditions C1-C6 hold.*
Let $\varpi_n = (\lambda_1 \wedge \lambda_2)/\{\max(\Phi_1, \Phi_2, \Phi_3)\}$, *where* $\Phi_t = \|I_{\mathscr{B}_t\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}\|_\infty$, $t = 1, 2, 3$. *If* $|\mathscr{A}| = o(n)$, $\lambda_1 \vee \lambda_2 \to 0$, $n\varpi_n^2 \to \infty$, *and* $\log p = o(n\varpi_n^2)$, *with probability tending to one, we have*

$$(a)\quad \|\widehat{\zeta}_{\mathscr{A}} - \zeta^*_{\mathscr{A}}\|_2 = O_p(\sqrt{|\mathscr{A}|/n}); \qquad (b)\quad \widehat{\zeta}_{\mathscr{A}^c} = \boldsymbol{0}.$$

Proof For the proof, see Appendix. □

This theorem establishes that the proposed method is able to accommodate p with $\log p = o(n\varpi_n^2)$. The penalized robust estimator enjoys the same asymptotic properties as the oracle estimator with probability approaching one. This property holds under high dimensions without restrictive conditions on the errors. To the best of our knowledge, properties of the robust exponential loss, even without censoring, have not been studied under high-dimensional settings. Thus our theoretical investigation can have independent value. Proof of the theorem is presented in Appendix.

3 Simulations

In simulation, we set $n = 300$, $q = 5$, and $p = 1000$. The underlying true model contains a total of 35 nonzero effects, including 5 main E effects, 10 main G effects, and 20 interactions. The "positions" of nonzero main G effects are randomly placed. The nonzero interactions are generated to respect the "main effects, interactions" hierarchy. The nonzero regression coefficients are randomly generated from uniform (0.7, 1.3). We consider both continuous and categorical distributions to mimic, for example, gene expression and SNP data. Specifically, under the continuous scenario, the E and G factors are generated from multivariate normal distributions with marginal means zero, marginal variances one, and the following variance matrix structures: Independent, AR(0.3), AR(0.8), Band(0.3), Band(0.6), and CS(0.2). Under the independent scenario, all factors have zero correlations. Under the AR(ρ) correlation structure, for the ith and jth factors, $corr = \rho^{|i-j|}$. Under the Band(ρ) correlation structure, for the ith and jth factors, $corr = \rho \cdot I(|i - j| = 2) + 0.3 \cdot I(|i - j| = 1) + I(|i - j| = 0)$. Under the CS($\rho$) correlation structure, for the ith and jth factors, the correlation coefficient $corr = \rho^{I(i \neq j)}$. Under the categorical scenario, we first apply the same data generating approach as described above to obtain $\mathbf{U}$. Then for each $u_{i,j}$, the categorical measurement is generated as $I(u_{i,j} > -0.7)$. The threshold value -0.7 is chosen such that the proportion of 1's for each factor is roughly 75%. Under each of the above simulation settings, consider the random error distribution $(1 - \xi)N(0, 1) + \xi Cauchy$, with the contamination probability $\xi = 0$, 0.1, and 0.3. When $\xi = 0$, the error distribution has no contamination and favors the nonrobust approaches, while the latter two values lead to different levels of contamination. The log event times are generated from the AFT model. The censoring times are generated independently from Weibull distributions. The censoring parameters are adjusted so that the censoring rates are about 25%. Beyond the above scenarios, we also consider a set of parallel scenarios, under which there are 10 main E effects, 20 main G effects, and 40 interactions (that is, the number of important effects is doubled), and the nonzero coefficients are generated from uniform (0.4, 0.6) (that is, the signal levels are reduced by about 50%). Other settings remain the same.

The simulated data are analyzed using the proposed method. In addition, we also consider two alternatives: (a) the nonrobust method that adopts the weighted least squared loss and the same penalty as the proposed, and (b) the quantile regression-based method that adopts an L_1 robust loss and the same penalty as the proposed. We note that multiple other methods are potentially applicable. Comparing with the nonrobust method can directly establish the merit of being robust. The quantile regression-based approach is the most popular for high-dimensional data (Wu & Ma, 2015). Thus these two alternatives are the most sensible to compare with.

All three methods involve tuning parameters. To eliminate the (possibly different) effects of tuning parameter selection on identification accuracy, we consider a sequence of tuning parameter values, evaluate identification accuracy at each value, and calculate the AUC (area under the ROC curve) as the overall measure. This approach has been adopted extensively in published studies (Zhu et al., 2014).

Summary statistics are computed based on 500 replicates. The AUC results for interactions and main effects combined are presented in Tables 1 and 2, respectively, for the scenarios with 35 and 70 important effects. To be thorough, we have also evaluated identification accuracy for interactions and main effects separately and present the AUC results in Tables 4, 5, 6 and 7 in Appendix. For all three methods, the AUC value decreases as the contamination proportion increases, as expected. In Table 1, the proposed method outperforms the two alternatives under all except one scenario. In Table 2, it dominates the alternatives. Under some scenarios, the proposed method leads to a significant improvement in identification accuracy. For example in Table 1, with the continuous G distribution, 30% contamination, and Band(0.3) correlation, the proposed method has a mean AUC of 0.901, while the alternatives have mean AUCs of 0.761 and 0.789. Compared to the nonrobust alternative, the proposed method also has smaller standard errors (Table 3).

We have also experimented with a few other scenarios and made similar observations. In particular, we have examined the scenarios where the event and censoring times have weak to moderate correlations and observed similar satisfactory performance (details omitted). The proposed method and two alternatives respect the hierarchy. We have also looked into simpler alternatives, including MCP and Lasso, which may violate the hierarchy, and observed inferior performance.

4 Analysis of the TCGA Lung Adenocarcinoma Data

Adenocarcinoma of the lung is the leading cause of cancer death worldwide. Profiling studies have been extensively conducted searching for its prognostic factors. Here we analyze the TCGA (The Cancer Genome Atlas Research Network, 2014) data on the prognosis of lung adenocarcinoma. The TCGA data were recently collected and published by NCI and have high quality. The prognosis outcome of interest is overall survival. The dataset contains measurements on 43 clinical/environmental variables and 18,897 gene expressions. There are a total of 468 patients, among whom 117 died during follow-up. The median follow-up time is 8 months. We select four E factors for downstream analysis, namely, age, gender, smoking pack years, and smoking history. These factors have a relatively low missing rate in the TCGA dataset and have been previously suggested as potentially related to lung cancer prognosis. There are a total of 436 samples with both E and G measurements available. Among them, 110 died during follow-up, and the median follow-up time is 23 months. For the 326 censored subjects, the median follow-up time is 6 months. In principle, the proposed method can directly analyze all of the available gene expressions. To improve stability and reduce the computational cost, we conduct marginal prescreening. Specifically, genes are screened based on their univariate regression significance (p-value less than or equal to 0.1) and interquartile range (above the median of all interquartile ranges). Similar prescreenings have been adopted in the literature. A total of 819 gene expressions are included in the downstream model fitting. Note that with the

Table 1 Simulation: identification of both G-E interactions and main G effects. In each cell, mean AUC (se). There are a total of 35 nonzero effects, with coefficients ∼uniform (0.7, 1.3)

ξ	Cor	Proposed	Nonrobust	Quantile
Continuous				
0	AR(0)	**0.891(0.065)**	0.842(0.095)	0.806(0.043)
	AR(0.3)	**0.971(0.050)**	0.917(0.096)	0.832(0.025)
	AR(0.8)	**0.981(0.041)**	0.923(0.066)	0.881(0.024)
	Band(0.3)	**0.972(0.057)**	0.908(0.106)	0.828(0.024)
	Band(0.6)	**0.978(0.044)**	0.930(0.078)	0.725(0.024)
	CS(0.2)	**0.920(0.069)**	0.827(0.096)	0.854(0.024)
0.1	AR(0)	**0.824(0.077)**	0.733(0.114)	0.782(0.042)
	AR(0.3)	**0.951(0.057)**	0.858(0.130)	0.815(0.031)
	AR(0.8)	**0.970(0.061)**	0.841(0.119)	0.873(0.034)
	Band(0.3)	**0.945(0.093)**	0.850(0.143)	0.802(0.021)
	Band(0.6)	**0.959(0.058)**	0.865(0.131)	0.704(0.037)
	CS(0.2)	**0.898(0.087)**	0.779(0.109)	0.846(0.043)
0.3	AR(0)	0.769(0.086)	0.646(0.097)	**0.775(0.045)**
	AR(0.3)	**0.889(0.101)**	0.742(0.147)	0.788(0.021)
	AR(0.8)	**0.942(0.077)**	0.754(0.124)	0.856(0.025)
	Band(0.3)	**0.901(0.093)**	0.761(0.140)	0.789(0.022)
	Band(0.6)	**0.924(0.075)**	0.785(0.131)	0.691(0.041)
	CS(0.2)	**0.845(0.092)**	0.661(0.117)	0.831(0.045)
Categorical				
0	AR(0)	**0.890(0.062)**	0.838(0.092)	0.778(0.045)
	AR(0.3)	**0.963(0.054)**	0.913(0.093)	0.802(0.028)
	AR(0.8)	**0.975(0.041)**	0.918(0.068)	0.843(0.021)
	Band(0.3)	**0.971(0.041)**	0.932(0.080)	0.787(0.033)
	Band(0.6)	**0.972(0.039)**	0.925(0.079)	0.702(0.042)
	CS(0.2)	**0.917(0.082)**	0.818(0.097)	0.822(0.047)
0.1	AR(0)	**0.835(0.085)**	0.756(0.115)	0.749(0.043)
	AR(0.3)	**0.944(0.055)**	0.856(0.130)	0.785(0.033)
	AR(0.8)	**0.970(0.037)**	0.867(0.102)	0.831(0.041)
	Band(0.3)	**0.953(0.052)**	0.862(0.119)	0.764(0.025)
	Band(0.6)	**0.965(0.044)**	0.861(0.128)	0.678(0.032)
	CS(0.2)	**0.895(0.086)**	0.752(0.115)	0.803(0.035)
0.3	AR(0)	**0.771(0.090)**	0.635(0.118)	0.738(0.043)
	AR(0.3)	**0.895(0.087)**	0.722(0.131)	0.771(0.024)
	AR(0.8)	**0.946(0.057)**	0.785(0.119)	0.817(0.028)
	Band(0.3)	**0.897(0.115)**	0.748(0.153)	0.741(0.027)
	Band(0.6)	**0.921(0.083)**	0.751(0.140)	0.649(0.047)
	CS(0.2)	**0.822(0.110)**	0.660(0.113)	0.787(0.031)

Table 2 Simulation: identification of both G-E interactions and main G effects. There are a total of 70 nonzero effects, with coefficients ∼uniform (0.4, 0.6)

ξ	Cor	Proposed	Nonrobust	Quantile
Continuous				
0	AR(0)	**0.678(0.043)**	0.649(0.042)	0.645(0.051)
	AR(0.3)	**0.800(0.045)**	0.768(0.057)	0.771(0.046)
	AR(0.8)	**0.916(0.058)**	0.828(0.054)	0.812(0.044)
	Band(0.3)	**0.827(0.057)**	0.781(0.067)	0.787(0.039)
	Band(0.6)	**0.865(0.052)**	0.816(0.061)	0.719(0.025)
	CS(0.2)	**0.717(0.065)**	0.645(0.047)	0.668(0.037)
0.1	AR(0)	**0.651(0.040)**	0.623(0.047)	0.619(0.045)
	AR(0.3)	**0.737(0.069)**	0.668(0.085)	0.672(0.038)
	AR(0.8)	**0.892(0.050)**	0.779(0.081)	0.795(0.047)
	Band(0.3)	**0.790(0.061)**	0.710(0.100)	0.759(0.055)
	Band(0.6)	**0.827(0.060)**	0.767(0.080)	0.754(0.041)
	CS(0.2)	**0.691(0.061)**	0.613(0.053)	0.672(0.042)
0.3	AR(0)	**0.605(0.052)**	0.551(0.042)	0.561(0.048)
	AR(0.3)	**0.697(0.064)**	0.601(0.058)	0.633(0.037)
	AR(0.8)	**0.838(0.081)**	0.679(0.093)	0.719(0.042)
	Band(0.3)	**0.713(0.079)**	0.608(0.080)	0.668(0.039)
	Band(0.6)	**0.754(0.085)**	0.648(0.102)	0.651(0.035)
	CS(0.2)	**0.668(0.059)**	0.568(0.065)	0.611(0.041)
Categorical				
0	AR(0)	**0.675(0.045)**	0.645(0.040)	0.643(0.038)
	AR(0.3)	**0.784(0.057)**	0.769(0.067)	0.758(0.042)
	AR(0.8)	**0.909(0.058)**	0.826(0.052)	0.799(0.051)
	Band(0.3)	**0.799(0.058)**	0.776(0.065)	0.774(0.034)
	Band(0.6)	**0.847(0.063)**	0.827(0.062)	0.688(0.039)
	CS(0.2)	**0.719(0.064)**	0.634(0.041)	0.677(0.049)
0.1	AR(0)	**0.654(0.052)**	0.596(0.060)	0.604(0.037)
	AR(0.3)	**0.748(0.063)**	0.683(0.093)	0.695(0.052)
	AR(0.8)	**0.869(0.085)**	0.764(0.086)	0.772(0.041)
	Band(0.3)	**0.772(0.071)**	0.712(0.093)	0.733(0.039)
	Band(0.6)	**0.806(0.067)**	0.736(0.099)	0.729(0.048)
	CS(0.2)	**0.684(0.058)**	0.595(0.051)	0.638(0.034)
0.3	AR(0)	**0.614(0.056)**	0.557(0.049)	0.571(0.052)
	AR(0.3)	**0.697(0.065)**	0.614(0.068)	0.632(0.047)
	AR(0.8)	**0.824(0.092)**	0.694(0.103)	0.727(0.045)
	Band(0.3)	**0.720(0.071)**	0.639(0.085)	0.655(0.036)
	Band(0.6)	**0.749(0.087)**	0.644(0.090)	0.648(0.045)
	CS(0.2)	**0.666(0.056)**	0.574(0.050)	0.629(0.038)

main G effects as well as interactions, the number of unknown parameters is much larger than the sample size.

Detailed estimation results are presented in Table 3 for the proposed method and Tables 8 and 9 in Appendix for the two alternatives. It is observed that the three methods lead to quite different findings. Specifically, the proposed and quantile methods share four common main G effects and four interactions. Otherwise, there is no overlap in identification. The "signals" in practical data can be weaker than those in simulated data, leading to the significant differences across methods.

With the proposed method, sixteen genes are identified to have interactions with either age or smoking status. As for many other cancer types, age has been identified as a critical factor in lung cancer prognosis. Smoking has been confirmed as the most important E factor for lung cancer risk and prognosis. In the literature, G-E interaction analysis for lung cancer prognosis is still very limited. However, there have been many studies on the functionalities of genes. Searching such studies can provide a partial support to the validity of our analysis results. Among the identified genes, many have been implicated in cancer in the literature. Specifically, the AGPAT family, which includes AGPAT6 as a member, has been found to play a role in multiple cancer types. For example, AGPAT2 and AGPAT11 have been found to be upregulated in ovarian, breast, cervical, and colorectal cancers (Agarwal and Garg, 2010). Another gene that is worth attention is ATF6, which acts both as a sensor and a transcription factor during endoplasmic reticulum stress. ATF6α has been found to promote hepatocarcinogenesis and cancer cell proliferation through activating downstream target gene BIP. Its efficiency of stress recognition and signaling has been found to decrease with age (Naidoo, 2009). We find that gene COLCA2 (colorectal cancer associated 2) interacts with smoking pack years. Studies have shown that COLCA2 may have critical functions in suppressing tumor formation in epithelial cells (Peltekova et al., 2014). We also identify an interaction between NOS1AP and age. It has been found that the protein complex of SCRIB, NOS1AP, and VANGL1 regulates cell polarity and migration, and this complex can be associated with cancer progression (Anastas et al., 2012). An interaction between PPP1R15B and smoking pack years has also been identified. It has been suggested that PPP1R15B is likely to be regulated by Nrf2, which has a protective response to smoking induced oxidative stress in the lung (Taylor et al., 2008). Also, PPP1R15B may promote cancer cell proliferation.

To complement the identification and estimation analysis, we also evaluate stability. Specifically, we randomly select 3/4 of the subjects and apply the proposed method and alternatives. This procedure is repeated 200 times. We then compute the probability that an interaction is identified. Similar procedures have been extensively adopted in published studies. The stability results are also provided in Tables 3, 8, and 9. We see that most of the identified interactions are relatively stable, with many having the probabilities of being identified close to one.

Table 3 Analysis of the TCGA lung adenocarcinoma data using the proposed method. The identified interactions are denoted as "gene * environmental variable". For the interactions, values in "()" are the stability results

Effect	Estimate ×100
Age	8.817
Smoking pack years	–0.358
AGPAT6	55.343
ANKRD46	4.646
ATF6	40.350
C1ORF27	7.708
COLCA2	–1.808
CAND1	32.138
DNAJC21	6.652
DYRK2	–24.595
HERPUD2	–40.358
LCMT2	40.151
NOS1AP	–28.707
PIGZ	–19.058
PPP1R15B	–2.411
TROVE2	–5.979
WIPI2	–18.739
YTHDF3	21.524
AGPAT6 * age	0.202(0.995)
ANKRD46 * age	0.546(0.537)
ATF6 * age	0.493(0.193)
C1ORF27 * age	–0.072(0.989)
COLCA2 * smoking pack years	0.315(0.993)
CAND1 * smoking pack years	–0.716(0.649)
DNAJC21 * age	–0.280(1.000)
DYRK2 * age	0.496(0.330)
HERPUD2 * age	–0.393(0.975)
LCMT2 * age	–0.222(0.927)
NOS1AP * age	0.140(0.734)
PIGZ * age	0.046(0.397)
PPP1R15B * smoking pack years	0.711(0.998)
TROVE2 * age	–0.416(0.890)
WIPI2 * age	–0.205(1.000)
YTHDF3 * age	–0.201(0.839)

Table 4 Simulation: identification of main G effects. In each cell, mean AUC (se). There are a total of 10 nonzero main effects, with coefficients ∼uniform (0.7, 1.3)

ξ	Cor	Robust	Nonrobust	Quantile
Continuous				
0	AR(0)	0.94(0.055)	**0.964(0.052)**	0.867(0.041)
	AR(0.3)	0.985(0.019)	**0.998(0.002)**	0.882(0.052)
	AR(0.8)	0.987(0.019)	**0.994(0.019)**	0.912(0.032)
	BAND(0.3)	0.985(0.020)	**0.999(0.003)**	0.841(0.046)
	BAND(0.6)	0.985(0.022)	**0.998(0.007)**	0.792(0.047)
	CS(0.2)	**0.938(0.047)**	0.923(0.061)	0.863(0.039)
0.1	AR(0)	**0.883(0.079)**	0.834(0.135)	0.852(0.044)
	AR(0.3)	**0.975(0.028)**	0.956(0.108)	0.841(0.053)
	AR(0.8)	**0.981(0.052)**	0.922(0.127)	0.891(0.034)
	BAND(0.3)	**0.967(0.075)**	0.942(0.136)	0.836(0.036)
	BAND(0.6)	**0.975(0.036)**	0.944(0.128)	0.789(0.044)
	CS(0.2)	**0.921(0.079)**	0.851(0.126)	0.855(0.031)
0.3	AR(0)	**0.837(0.102)**	0.716(0.127)	0.792(0.028)
	AR(0.3)	**0.942(0.082)**	0.829(0.168)	0.814(0.031)
	AR(0.8)	**0.970(0.070)**	0.841(0.135)	0.855(0.042)
	BAND(0.3)	**0.943(0.079)**	0.836(0.158)	0.821(0.028)
	BAND(0.6)	**0.954(0.067)**	0.871(0.144)	0.811(0.053)
	CS(0.2)	**0.886(0.095)**	0.704(0.142)	0.824(0.058)
Categorical				
0	AR(0)	0.931(0.047)	**0.956(0.050)**	0.857(0.044)
	AR(0.3)	0.970(0.030)	**0.998(0.010)**	0.872(0.058)
	AR(0.8)	0.976(0.028)	**0.992(0.023)**	0.883(0.044)
	BAND(0.3)	0.974(0.027)	**0.999(0.002)**	0.832(0.041)
	BAND(0.6)	0.974(0.029)	**0.999(0.010)**	0.824(0.036)
	CS(0.2)	**0.937(0.051)**	0.918(0.065)	0.844(0.051)
0.1	AR(0)	**0.894(0.078)**	0.868(0.137)	0.853(0.039)
	AR(0.3)	**0.963(0.036)**	0.959(0.104)	0.875(0.056)
	AR(0.8)	**0.975(0.029)**	0.939(0.090)	0.897(0.064)
	BAND(0.3)	**0.968(0.036)**	0.960(0.090)	0.841(0.042)
	BAND(0.6)	**0.971(0.033)**	0.934(0.117)	0.829(0.048)
	CS(0.2)	**0.923(0.075)**	0.824(0.138)	0.846(0.057)
0.3	AR(0)	**0.838(0.098)**	0.707(0.153)	0.788(0.029)
	AR(0.3)	**0.932(0.075)**	0.808(0.159)	0.818(0.035)
	AR(0.8)	**0.967(0.041)**	0.863(0.143)	0.854(0.048)
	BAND(0.3)	**0.931(0.101)**	0.842(0.173)	0.819(0.035)
	BAND(0.6)	**0.946(0.065)**	0.821(0.155)	0.813(0.044)
	CS(0.2)	**0.854(0.106)**	0.704(0.136)	0.828(0.052)

Table 5 Simulation: identification of G-E interactions. In each cell, mean AUC (se). There are a total of 20 nonzero interactions, with coefficients ∼uniform (0.7, 1.3)

ξ	Cor	Robust	Nonrobust	Quantile
Continuous				
0	AR(0)	**0.866(0.088)**	0.784(0.137)	0.761(0.047)
	AR(0.3)	**0.963(0.075)**	0.873(0.143)	0.776(0.062)
	AR(0.8)	**0.976(0.061)**	0.888(0.098)	0.862(0.066)
	BAND(0.3)	**0.964(0.085)**	0.861(0.159)	0.803(0.058)
	BAND(0.6)	**0.973(0.066)**	0.895(0.117)	0.659(0.043)
	CS(0.2)	**0.904(0.092)**	0.775(0.134)	0.848(0.039)
0.1	AR(0)	**0.792(0.090)**	0.684(0.124)	0.734(0.053)
	AR(0.3)	**0.938(0.082)**	0.811(0.158)	0.786(0.048)
	AR(0.8)	**0.962(0.075)**	0.801(0.128)	0.864(0.039)
	BAND(0.3)	**0.932(0.112)**	0.806(0.165)	0.775(0.041)
	BAND(0.6)	**0.948(0.080)**	0.826(0.150)	0.633(0.052)
	CS(0.2)	**0.881(0.105)**	0.740(0.130)	0.829(0.058)
0.3	AR(0)	0.733(0.094)	0.612(0.095)	**0.753(0.046)**
	AR(0.3)	**0.862(0.122)**	0.701(0.152)	0.749(0.052)
	AR(0.8)	**0.927(0.089)**	0.710(0.128)	0.861(0.057)
	BAND(0.3)	**0.879(0.110)**	0.725(0.146)	0.748(0.033)
	BAND(0.6)	**0.907(0.089)**	0.744(0.139)	0.598(0.062)
	CS(0.2)	0.820(0.105)	0.637(0.114)	**0.841(0.045)**
Categorical				
0	AR(0)	**0.866(0.086)**	0.782(0.130)	0.733(0.064)
	AR(0.3)	**0.955(0.079)**	0.869(0.140)	0.728(0.051)
	AR(0.8)	**0.971(0.061)**	0.881(0.100)	0.802(0.039)
	BAND(0.3)	**0.967(0.061)**	0.898(0.119)	0.749(0.048)
	BAND(0.6)	**0.969(0.058)**	0.888(0.119)	0.609(0.051)
	CS(0.2)	**0.900(0.109)**	0.763(0.134)	0.801(0.039)
0.1	AR(0)	**0.801(0.102)**	0.702(0.125)	0.667(0.057)
	AR(0.3)	**0.932(0.077)**	0.806(0.159)	0.702(0.048)
	AR(0.8)	**0.964(0.056)**	0.830(0.123)	0.764(0.055)
	BAND(0.3)	**0.942(0.074)**	0.814(0.151)	0.689(0.053)
	BAND(0.6)	**0.959(0.064)**	0.826(0.148)	0.604(0.045)
	CS(0.2)	**0.875(0.104)**	0.713(0.123)	0.751(0.048)
0.3	AR(0)	**0.734(0.099)**	0.601(0.114)	0.681(0.039)
	AR(0.3)	**0.873(0.104)**	0.681(0.131)	0.687(0.058)
	AR(0.8)	**0.931(0.075)**	0.746(0.120)	0.768(0.048)
	BAND(0.3)	**0.877(0.132)**	0.704(0.157)	0.699(0.059)
	BAND(0.6)	**0.905(0.101)**	0.718(0.146)	0.547(0.067)
	CS(0.2)	**0.800(0.124)**	0.635(0.115)	0.724(0.055)

Table 6 Simulation: identification of main G effects. In each cell, mean AUC (se). There are a total of 20 nonzero main effects, with coefficients ∼uniform (0.4, 0.6)

ξ	Cor	Robust	Nonrobust	Quantile
Continuous				
0	AR(0)	0.735(0.065)	**0.738(0.057)**	0.684(0.042)
	AR(0.3)	0.885(0.049)	**0.894(0.046)**	0.798(0.038)
	AR(0.8)	**0.961(0.045)**	0.925(0.036)	0.811(0.048)
	BAND(0.3)	**0.896(0.046)**	0.894(0.047)	0.809(0.044)
	BAND(0.6)	0.916(0.050)	**0.921(0.047)**	0.792(0.039)
	CS(0.2)	**0.769(0.065)**	0.710(0.059)	0.753(0.049)
0.1	AR(0)	**0.701(0.064)**	0.682(0.071)	0.678(0.033)
	AR(0.3)	**0.806(0.079)**	0.754(0.112)	0.794(0.049)
	AR(0.8)	**0.947(0.046)**	0.871(0.088)	0.806(0.053)
	BAND(0.3)	**0.865(0.071)**	0.809(0.132)	0.801(0.036)
	BAND(0.6)	**0.886(0.064)**	0.860(0.086)	0.789(0.055)
	CS(0.2)	0.738(0.080)	0.659(0.084)	**0.784(0.042)**
0.3	AR(0)	**0.646(0.073)**	0.577(0.067)	0.632(0.044)
	AR(0.3)	**0.774(0.090)**	0.664(0.088)	0.672(0.052)
	AR(0.8)	**0.896(0.081)**	0.743(0.127)	0.755(0.041)
	BAND(0.3)	**0.782(0.106)**	0.664(0.112)	0.711(0.065)
	BAND(0.6)	**0.823(0.104)**	0.719(0.141)	0.705(0.053)
	CS(0.2)	**0.708(0.078)**	0.601(0.093)	0.645(0.051)
Categorical				
0	AR(0)	**0.736(0.068)**	0.729(0.054)	0.679(0.045)
	AR(0.3)	0.858(0.066)	**0.901(0.059)**	0.782(0.052)
	AR(0.8)	**0.944(0.055)**	0.922(0.046)	0.797(0.041)
	BAND(0.3)	0.869(0.062)	**0.900(0.057)**	0.787(0.058)
	BAND(0.6)	0.894(0.062)	**0.920(0.045)**	0.762(0.048)
	CS(0.2)	**0.768(0.067)**	0.696(0.055)	0.763(0.048)
0.1	AR(0)	**0.716(0.074)**	0.659(0.096)	0.669(0.051)
	AR(0.3)	**0.826(0.077)**	0.777(0.132)	0.752(0.039)
	AR(0.8)	**0.914(0.089)**	0.837(0.103)	0.786(0.054)
	BAND(0.3)	**0.838(0.079)**	0.805(0.120)	0.743(0.062)
	BAND(0.6)	**0.867(0.069)**	0.828(0.118)	0.721(0.039)
	CS(0.2)	**0.723(0.070)**	0.642(0.080)	0.678(0.042)
0.3	AR(0)	**0.639(0.074)**	0.588(0.074)	0.613(0.045)
	AR(0.3)	**0.758(0.083)**	0.684(0.097)	0.658(0.047)
	AR(0.8)	**0.877(0.109)**	0.764(0.138)	0.743(0.055)
	BAND(0.3)	**0.789(0.089)**	0.703(0.115)	0.688(0.044)
	BAND(0.6)	**0.806(0.104)**	0.702(0.121)	0.671(0.049)
	CS(0.2)	**0.694(0.074)**	0.599(0.070)	0.648(0.037)

Table 7 Simulation: identification of G-E interactions. In each cell, mean AUC (se). There are a total of 40 nonzero interactions, with coefficients ∼uniform (0.4, 0.6)

ξ	Cor	Robust	Nonrobust	Quantile
Continuous				
0	AR(0)	**0.647(0.048)**	0.605(0.053)	0.606(0.049)
	AR(0.3)	**0.755(0.059)**	0.707(0.079)	0.752(0.058)
	AR(0.8)	**0.892(0.074)**	0.781(0.073)	0.814(0.039)
	BAND(0.3)	**0.791(0.074)**	0.726(0.093)	0.769(0.052)
	BAND(0.6)	**0.838(0.066)**	0.765(0.084)	0.629(0.048)
	CS(0.2)	**0.686(0.077)**	0.609(0.055)	0.602(0.048)
0.1	AR(0)	**0.623(0.041)**	0.593(0.049)	0.563(0.059)
	AR(0.3)	**0.701(0.074)**	0.625(0.085)	0.602(0.055)
	AR(0.8)	**0.863(0.060)**	0.734(0.088)	0.778(0.041)
	BAND(0.3)	**0.750(0.070)**	0.661(0.099)	0.711(0.058)
	BAND(0.6)	**0.795(0.069)**	0.722(0.089)	0.725(0.061)
	CS(0.2)	**0.662(0.059)**	0.586(0.052)	0.596(0.062)
0.3	AR(0)	**0.581(0.050)**	0.537(0.037)	0.503(0.055)
	AR(0.3)	**0.656(0.064)**	0.570(0.051)	0.596(0.049)
	AR(0.8)	**0.807(0.086)**	0.647(0.085)	0.679(0.058)
	BAND(0.3)	**0.677(0.077)**	0.580(0.072)	0.618(0.061)
	BAND(0.6)	**0.718(0.082)**	0.612(0.088)	0.604(0.042)
	CS(0.2)	**0.642(0.060)**	0.550(0.055)	0.571(0.046)
Categorical				
0	AR(0)	**0.640(0.051)**	0.604(0.052)	0.611(0.052)
	AR(0.3)	**0.743(0.067)**	0.706(0.089)	0.733(0.041)
	AR(0.8)	**0.887(0.070)**	0.779(0.070)	0.806(0.034)
	BAND(0.3)	**0.761(0.069)**	0.716(0.089)	0.751(0.039)
	BAND(0.6)	**0.820(0.076)**	0.781(0.087)	0.623(0.059)
	CS(0.2)	**0.688(0.071)**	0.598(0.051)	0.601(0.033)
0.1	AR(0)	**0.619(0.051)**	0.565(0.050)	0.548(0.047)
	AR(0.3)	**0.706(0.068)**	0.637(0.091)	0.647(0.054)
	AR(0.8)	**0.842(0.092)**	0.728(0.089)	0.751(0.061)
	BAND(0.3)	**0.735(0.077)**	0.667(0.092)	0.726(0.041)
	BAND(0.6)	**0.771(0.078)**	0.691(0.102)	0.736(0.034)
	CS(0.2)	**0.658(0.061)**	0.568(0.046)	0.589(0.052)
0.3	AR(0)	**0.597(0.053)**	0.540(0.043)	0.540(0.062)
	AR(0.3)	**0.663(0.070)**	0.579(0.066)	0.604(0.038)
	AR(0.8)	**0.794(0.091)**	0.658(0.094)	0.704(0.058)
	BAND(0.3)	**0.682(0.073)**	0.606(0.080)	0.626(0.046)
	BAND(0.6)	**0.717(0.086)**	0.615(0.084)	0.614(0.047)
	CS(0.2)	**0.646(0.060)**	0.558(0.048)	0.604(0.055)

Table 8 Analysis of the TCGA lung adenocarcinoma data using the nonrobust method. The identified interactions are denoted as "gene * environmental variable". For the interactions, values in "()" are the stability results

Effect	Estimate ×100
Age	0.868
Gender	6.683
Smoking pack years	0.041
Smoking history	–20.163
SPATA33	–7.060
DNAJC21	5.237
EIF4EBP1	8.736
FAM160B1	–0.030
KIAA1586	4.018
LRRC37A4P	3.040
ST6GALNAC1	5.989
TM2D2	10.110
TMEM192	–5.785
TROVE2	–3.019
WIPI2	5.296
SPATA33 * smoking pack years	–0.084(0.812)
DNAJC21 * smoking history	5.245(0.986)
EIF4EBP1 * smoking pack years	–0.087(0.977)
FAM160B1 * gender	11.844(0.982)
KIAA1586 * age	0.107(0.954)
LRRC37A4P * smoking pack years	–0.205(0.998)
LRRC37A4P * smoking history	–6.799(0.998)
ST6GALNAC1 * smoking pack years	–0.149(0.989)
TM2D2 * smoking pack years	–0.176(0.998)
TMEM192 * gender	9.853(0.995)
TROVE2 * gender	7.349(0.929)
WIPI2 * smoking history	13.420(0.995)

5 Discussions

To understand the prognosis of complex diseases, it is essential to study G-E interactions. In "classic" low-dimensional biomedical studies, data contamination is found to be not rare, and it has been suggested that robust methods are needed to accommodate contamination. This study has developed a robust method for high-dimensional genetic interaction analysis, which is still limited in the literature. The proposed method consists of a novel robust loss function and a penalized identification strategy that respects the "main effects, interactions" hierarchy, both of which have novel

Table 9 Analysis of the TCGA lung adenocarcinoma data using the quantile method. The identified interactions are denoted as "gene * environmental variable". For the interactions, values in "()" are the stability results

Effect	Estimate $\times 100$
Age	0.891
ATP6V1C1	0.252
C1ORF27	7.321
SDE2	0.337
CD46	0.584
DNAJC21	1.272
KLHL7	0.932
PTK2	9.426
PVT1	1.148
RAB3GAP2	0.845
TSPAN3	8.557
TWISTNB	0.872
WDR26	1.265
WIPI2	7.227
YWHAZ	1.883
ATP6V1C1 * age	0.0172(0.724)
C1ORF27 * age	0.295(0.899)
SDE2 * age	0.0153(0.758)
CD46 * age	0.0344(0.862)
DNAJC21 * age	0.327(0.791)
KLHL7 * age	0.372(0.514)
PTK2 * age	1.074(0.927)
PVT1 * age	0.876(0.711)
RAB3GAP2 * age	0.923(0.757)
TSPAN3 * age	1.388(0.942)
TWISTNB * age	0.915(0.812)
WDR26 * age	1.279(0.798)
WIPI2 * age	1.891(0.906)
YWHAZ * age	1.596(0.796)

advancements. Also significantly advancing from the literature, we have rigorously established the consistency properties. The theoretical results may seem "familiar", which is "comforting" in that the consistency properties are not sacrificed with the additional robustness, high dimensionality, and interactions. It is worth noting that the consistency results do not demand excessive assumptions on the error distribution, which are usually needed in the existing literature. In simulation, the proposed method outperforms the nonrobust alternative. It is interesting to note that it has

superior performance when there is no contamination. Another important finding is that it also outperforms the quantile-based robust method. Most of the existing high-dimensional robust studies have adopted the quantile regression technique. Our simulation suggests that it is prudent to develop alternative robust methods. In the analysis of TCGA lung cancer data, the proposed method generates results with some overlappings with the quantile regression method, however, none with the nonrobust method. The identified genes have important implications, and the identified interactions are stable.

The proposed study can be potentially extended in multiple directions. In survival analysis, there are many other models beyond the AFT. It can be of interest to develop robust methods based on other models. We have studied G-E interactions. It can be of interest to extend to G-G interactions. In theoretical analysis, one problem left is the breakdown point. Because of the extremely high complexity, this problem has been left uninvestigated in many other robust studies too. In our simulation, we have experimented with contamination rate as high as 30%, which is much higher than many of the existing studies. The superiority of the proposed method over the quantile regression method is observed. The relative efficiency of different robust methods, although of interest, will be postponed to future studies. In data analysis, the proposed method identifies a different set of main effects and interactions. Mining the literature and the stability evaluation can support the validity of findings to a certain extent. More validations need to be pursued in the future.

Appendix

Proof of Theorem 1

Proof Define the oracle estimator $\widehat{\zeta}$ with $\widehat{\zeta}_{\mathscr{A}^c} = 0$ and

$$\widehat{\zeta}_{\mathscr{A}} = \arg\max \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,\mathscr{A}}^{\top} \zeta_{\mathscr{A}})^2/\theta). \tag{4}$$

Recall that the proposed objective function is

$$L_{\lambda_1,\lambda_2,\theta}(\zeta) = Q_\theta(\zeta) - \sum_{k=1}^{p} \rho(\|b_k\|; \lambda_1, s) - \sum_{k=1}^{p} \sum_{j=2}^{q+1} \rho(|b_{kj}|; \lambda_2, s). \tag{5}$$

In what follows, we first establish the estimation consistency of $\widehat{\zeta}$ in Step 1, and then show that $\widehat{\zeta}$ is a local maximizer of $L_{\lambda_1,\lambda_2,\theta}(\zeta)$ is Step 2.

Step 1. Define the objective function

$$R_n(\zeta_{\mathscr{A}}) = \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,\mathscr{A}}^{\top} \zeta_{\mathscr{A}})^2/\theta).$$

Then $\widehat{\zeta}_{\mathscr{A}} = \arg\max R_n(\zeta_{\mathscr{A}})$. Let $r_n = \sqrt{|\mathscr{A}|/n}$. To prove $\|\widehat{\zeta}_{\mathscr{A}} - \zeta^*_{\mathscr{A}}\|_2 = O_p(r_n)$, it suffices to show that for any given $\eta > 0$, there exists a sufficiently large constant $C > 0$,

$$\Pr\Big(\sup_{\zeta_{\mathscr{A}} \in \mathscr{J}} R_n(\zeta_{\mathscr{A}}) < R_n(\zeta^*_{\mathscr{A}})\Big) \geq 1 - \eta, \tag{6}$$

where $\mathscr{J} = \big\{\zeta_{\mathscr{A}} : \|\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}\|_2 = Cr_n\big\}$. This implies that $R_n(\zeta_{\mathscr{A}})$ has a local maximizer $\widehat{\zeta}_{\mathscr{A}}$ that satisfies $\|\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}\|_2 = O_p(r_n)$.

Recall the definitions of $D_n(\zeta)$ and $I_n(\zeta)$. By Taylor's expansion, we have

$$\begin{aligned} R_n(\zeta_{\mathscr{A}}) - R_n(\zeta^*_{\mathscr{A}}) =& \sum_{i=1}^{n} \omega_i \Big\{\exp(-(y_i - \mathbf{u}^\top_{i,\mathscr{A}}\zeta_{\mathscr{A}})^2/\theta) - \exp(-(y_i - \mathbf{u}^\top_{i,\mathscr{A}}\zeta^*_{\mathscr{A}})^2/\theta)\Big\} \\ =& D_{n,\mathscr{A}}(\zeta^*)^\top(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}) \\ &+ \frac{1}{2}(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}})^\top I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta})(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}) \\ \doteq& Q_1 + Q_2, \end{aligned} \tag{7}$$

where $\bar{\zeta}$ lies between ζ^* and ζ. By C3 and C4, we have that for all $j \in \{1, \cdots, p + q + pq + 1\}$ and any given t, $\Pr(|D_{n,j}(\zeta^*)| > t) \leq 2\exp\big(-nt^2/\sigma^2\big)$. Then $E(|\sqrt{n}D_{n,j}(\zeta^*)|) < K < \infty$ for all j. With Markov's inequality,

$$\Pr(\|D_{n,\mathscr{A}}(\zeta^*)\|_2 > t) \leq E[\|\sqrt{n}D_{n,\mathscr{A}}(\zeta^*)\|_2^2]/(nt^2) \leq |\mathscr{A}|K/(nt^2).$$

By the Cauchy-Schwarz inequality, $Q_1 \leq C\|D_{n,\mathscr{A}}(\zeta^*)\|_2 r_n$. Let $t = C\rho_* r_n/3$, where ρ_* is the smallest eigenvalue of $-I_{AA}(\zeta^*_A)$. From C5, we have that ρ_* is bounded away from zero and infinity. Then we have

$$\Pr(Q_1 \leq \frac{1}{3}\rho_* C^2 r_n^2) \leq 1 - \frac{9K}{C^2\rho_*^2}. \tag{8}$$

For Q_2, we have

$$\begin{aligned} 2Q_2 =& (\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}})^\top I_{\mathscr{A}\mathscr{A}}(\zeta^*)(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}) \\ &+ (\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}})^\top \big\{I_{\mathscr{A}\mathscr{A}}(\bar{\zeta}) - I_{\mathscr{A}\mathscr{A}}(\zeta^*)\big\}(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}) \\ &+ (\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}})^\top \big\{I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta}) - I_{\mathscr{A}\mathscr{A}}(\bar{\zeta})\big\}(\zeta_{\mathscr{A}} - \zeta^*_{\mathscr{A}}) \\ \doteq& Q_{21} + Q_{22} + Q_{23}. \end{aligned} \tag{9}$$

Since $\lambda_{\max}(I_{\mathscr{A}\mathscr{A}}(\zeta^*)) \leq -\rho_*$ by C5, we have

$$Q_{21} \leq -\rho_* C^2 r_n^2. \tag{10}$$

Under C4, we have

$$Q_{22} \leq \kappa \|\bar{\zeta} - \zeta^*\|_2 C^2 r_n^2 < \kappa C^3 r_n^3 < \frac{1}{6} C^2 \rho_* r_n^2. \tag{11}$$

The second inequality holds since $\bar{\zeta}$ lies between ζ^* and ζ, which yields $\|\bar{\zeta} - \zeta^*\|_2 < Cr_n$. When n is sufficiently large, the last inequality holds. With C4 and Bonferroni's inequality,

$$\Pr(\|I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta}) - I_{\mathscr{A}\mathscr{A}}(\bar{\zeta})\|_F^2 \geq \rho_*^2/9) \leq 2|\mathscr{A}|^2 \exp\left(-n\rho_*^2/\sigma^2\right),$$

where $\|\cdot\|_F$ denotes the Frobenius norm. By the inequality $\lambda_{\max}(I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta}) - I_{\mathscr{A}\mathscr{A}}(\bar{\zeta})) \leq \|I_{n,\mathscr{A}\mathscr{A}}(\zeta^*) - I_{\mathscr{A}\mathscr{A}}(\zeta^*)\|_F$, we have

$$Q_{23} \leq \frac{1}{3}\rho_* C^2 r_n^2 \text{ with probability at least } 1 - 2|\mathscr{A}|^2 \exp\left(-n\rho_*^2/\sigma^2\right). \tag{12}$$

Combining (9), (10), (11), and (12), we have

$$\Pr(Q_2 < -\frac{1}{2}\rho_* C^2 r_n^2) \geq 1 - 2|\mathscr{A}|^2 \exp\left(-n\rho_*^2/\sigma^2\right). \tag{13}$$

With (7), (8), and (13), we have

$$R_n(\zeta_{\mathscr{A}}) - R_n(\zeta_{\mathscr{A}}^*) < -\frac{1}{6}\rho_* C^2 r_n^2 < 0 \tag{14}$$

with probability at least

$$1 - \frac{9K}{C^2\rho_*^2} - 2|\mathscr{A}|^2 \exp\left(-n\rho_*^2/\sigma^2\right).$$

Note that ρ_* is bounded away from zero and infinity in C5. As $n \to \infty$, the above probability is bigger than $1 - \frac{16K}{C^2\rho_*^2}$. Let $C = 4\rho_*^{-1}\sqrt{K/\eta}$, then we can conclude (6).

Step 2. Next we show that the oracle estimator $\widehat{\zeta}$ studied in Step 1 satisfies the Karush-Kuhn-Tucher (KKT) condition, and then $\widehat{\zeta}$ is a local maximizer of $L_{\lambda_1,\lambda_2,\theta}(\zeta)$. Based on the results in Step 1 and C6, we only need to check the following conditions

$$\left\|D_{n,\mathscr{B}_1}(\widehat{\zeta})\right\|_\infty < \lambda_1, \quad \left\|D_{n,\mathscr{B}_2}(\widehat{\zeta})\right\|_\infty < \lambda_2, \quad \left\|D_{n,\mathscr{B}_3}(\widehat{\zeta})\right\|_\infty < \lambda_1 + \lambda_2 \tag{15}$$

hold with asymptotic probability one, where $\|v\|_\infty = \max_i |v_i|$ for any vector $v = (v_1, \cdots, v_{|\mathscr{A}^c|})$. Applying Taylor's expansion,

$$D_{n,\mathscr{B}_1}(\widehat{\zeta}) = D_{n,\mathscr{B}_1}(\zeta^*) + I_{n,\mathscr{B}_1\mathscr{A}}(\widetilde{\zeta})(\widehat{\zeta}_{\mathscr{A}} - \zeta_{\mathscr{A}}^*), \tag{16}$$

where $\widetilde{\zeta}$ lies between ζ^* and $\widehat{\zeta}$. From (4) and the proof of Theorem 1(a), we have

$$\widehat{\zeta}_{\mathscr{A}} - \zeta_{\mathscr{A}}^* = -I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta})^{-1} D_{n,\mathscr{A}}(\zeta^*), \tag{17}$$

where $\bar{\zeta}$ lies between ζ^* and $\widehat{\zeta}$, which is defined in Step 1. By substituting (17) into (16),

$$D_{n,\mathscr{B}_1}(\widehat{\zeta}) = D_{n,\mathscr{B}_1}(\zeta^*) - I_{n,\mathscr{B}_1\mathscr{A}}(\widetilde{\zeta})I_{n,\mathscr{A}\mathscr{A}}(\bar{\zeta})^{-1}D_{n,\mathscr{A}}(\zeta^*). \tag{18}$$

Here we define

$$\Delta^*_{n,\mathscr{B}_1} = D_{n,\mathscr{B}_1}(\zeta^*) - I_{\mathscr{B}_1\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}D_{n,\mathscr{A}}(\zeta^*).$$

Inspired by the deduction of Q_2 in Step 1, we can establish that

$$\Pr(\|D_{n,\mathscr{B}_1}(\widehat{\zeta})\|_\infty > \lambda_1) \asymp \Pr(\|\Delta^*_{n,\mathscr{B}_1}\|_\infty > \lambda_1).$$

That is, we only need to focus on $\|\Delta^*_{n,\mathscr{B}_1}\|_\infty$ in order to evaluate the probability of $\{\|D_{n,\mathscr{B}_1}(\widehat{\zeta})\|_\infty < \lambda_1\}$ in (15). Note that,

$$\begin{aligned}\|\Delta^*_{n,\mathscr{B}_1}\|_\infty &\le \|D_{n,\mathscr{B}_1}(\zeta^*)\|_\infty + \|I_{\mathscr{B}_1\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}D_{n,\mathscr{A}}(\zeta^*)\|_\infty \\ &\le \|D_n(\zeta^*)\|_\infty + \|I_{\mathscr{B}_1\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}\|_\infty\|D_n(\zeta^*)\|_\infty. \end{aligned}\tag{19}$$

Recall that $\Phi_1 = \|I_{\mathscr{B}_1\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}\|_\infty$. If

$$\|D_n(\zeta^*)\|_\infty < \frac{\lambda_1}{1+\Phi_1},$$

along with (19), we have $\|\Delta^*_{n,\mathscr{B}_1}\|_\infty < \lambda_1$. Similarly, we also need

$$\|D_n(\zeta^*)\|_\infty < \frac{\lambda_2}{1+\Phi_2}, \text{ and } \|D_n(\zeta^*)\|_\infty < \frac{\lambda_1+\lambda_2}{1+\Phi_3}$$

to satisfy the other two conditions in (15), where $\Phi_2 = \|I_{\mathscr{B}_2\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}\|_\infty$ and $\Phi_3 = \|I_{\mathscr{B}_3\mathscr{A}}(\zeta^*)I_{\mathscr{A}\mathscr{A}}(\zeta^*)^{-1}\|_\infty$. Based on the above discussions, we have

$$\|D_n(\zeta^*)\|_\infty < \frac{\lambda_1 \wedge \lambda_2}{1+\max_{t=1}^3 \Phi_t} < \min\{\frac{\lambda_1}{1+\Phi_1}, \frac{\lambda_2}{1+\Phi_2}, \frac{\lambda_1+\lambda_2}{1+\Phi_3}\}.$$

We now derive the probability bound for the above event. By Bonferroni's inequality and C4, we can obtain

$$\Pr\left\{\|D_n(\zeta^*)\|_\infty < \frac{\lambda_1 \wedge \lambda_2}{1+\max_{t=1}^3 \Phi_t}\right\} \ge 1 - 2(pq+p+q+1)\exp\left(-\frac{n(\lambda_1\wedge\lambda_2)^2}{(1+\max_{t=1}^3\Phi_t)^2\sigma^2}\right).$$

Combining the results in Steps 1 and 2, we conclude that $\widehat{\zeta}$ is a local maximizer of $L_{\lambda_1,\lambda_2,\theta}(\zeta)$ with probability at least

$$1 - O\left(p\exp\left(-\frac{n(\lambda_1 \wedge \lambda_2)^2}{(1+\max_{t=1}^{3}\Phi_t)^2\sigma^2}\right)\right),$$

and satisfies $\|\widehat{\zeta}_{\mathscr{A}} - \zeta^*_{\mathscr{A}}\|_2 = O_p(\sqrt{|\mathscr{A}|/n})$, $\widehat{\zeta}_{\mathscr{A}^c} = 0$. With C6, $\log p = O(n\varpi_n^2)$, and $\varpi_n = (\lambda_1 \wedge \lambda_2)/\{\max(\Phi_1, \Phi_2, \Phi_3)\}$, this tail probability is exponentially small. The theorem is thus proved.

References

Agarwal, A. K., & Garg, A. (2010). Enzymatic activity of the human 1-acylglycerol-3-phosphate-o-acyltransferase isoform 11: upregulated in breast and cervical cancers. *Journal of Lipid Research, 51*, 2143–2152.

Anastas, J., Biechele, T., Robitaille, M., Muster, J., Allison, K., Angers, S., & Moon, R. (2012). A protein complex of SCRIB, NOS1AP and VANGL1 regulates cell polarity and migration, and is associated with breast cancer progression. *Oncogene, 31*, 3696.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics, 41*, 1111–1141.

Bowman, L. (2011). Doctors, researchers worry about accuracy of social security "death file". http://projects.scrippsnews.com/story/doctors-researchers-worry/. Accessed 30 Apr. 2015

Comprehensive molecular profiling of lung adenocarcinoma. (2014). The cancer genome atlas research network. *Nature, 511*, 543–550.

Fall, K., Stromberg, F., Rosell, J., Andren, O., & Varenhorst, E. (2008). Reliability of death certificates in prostate cancer patients. *Scandinavian Journal of Urology, 42*, 352–357.

He, S., Chen, H., Zhu, Z., Ward, D., Cooper, H., Viant, M., Heath, J., & Yao, X. (2015). Robust twin boosting for feature selection from high-dimensional omics data with label noise. *Information Sciences, 291*, 1–18.

Huang, J., Ma, S., & Xie, H. (2007). Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica, 17*, 1533–1548.

Huber, P., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: Wiley.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews Genetics, 6*, 287–298.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., & Ma, S. (2013). Identification of gene-environment interactions in cancer studies using penalization. *Genomics, 102*, 189–194.

Ma, S., & Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statistica Sinica, 22*, 1003–1020.

Naidoo, N. (2009). ER and aging-protein folding and the ER stress response. *Ageing Research Reviews, 8*, 150–159.

Peltekova, V., Lemire, M., Qazi, A., Zaidi, S., Trinh, Q., Bielecki, R., Rogers, M., Hodgson, L., Wang, M., D'souza, D., et al. (2014). Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer-associated variants. *International Journal of Cancer, 134*, 2330–2341.

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., & Ma, S. (2014). A penalized robust method for identifying gene-environment interactions. *Genetic Epidemiology, 38*, 220–230.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *The Journal of Multivariate Analysis, 45*, 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables ae present. *Scandinavian Journal of Statistics, 23*, 461–471.

Taylor, R., Acquaah-Mensah, G., Singhal, M., Malhotra, D., & Biswal, S. (2008). Network inference algorithms elucidate Nrf2 regulation of mouse lung oxidative stress. *PLOS Computational Biology, 4*, e1000166.

Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health, 31*, 21–36.

Wang, X., Jiang, Y., Huang, M., & Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association, 108*, 632–643.

Wu, C., Cui, Y., & Ma, S. (2014). Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Statistics in Medicine, 33*, 4988–4998.

Wu, C., & Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in Bioinformatics, 16*, 873–883.

Xu, Y., Wu, M., Ma, S., & Ejaz Ahmed, S. (2018). Robust gene-environment interaction analysis using penalized trimmed regression. *Journal of Statistical Computation and Simulation, 88*, 3502–3528.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*, 894–942.

Zhu, R., Zhao, H., & Ma, S. (2014). Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genetic Epidemiology, 38*, 353–368.

A Novel Approach for Improving Accuracy for Distributed Storage Networks

Liu Lu, Ke Yuanyuan, and Yuan Yong

Abstract With the development of storage technology and Internet technology, cloud storage continues to make its impact. Scalability, reliability, and lowered costs have made cloud storage widely used with success in businesses and individuals. The advent of the blockchain has brought some changes. As the incentive layer for IPFS, Filecoin allows storage resources to become tradable, greatly extending storage capacity. However, the process of testing the integrity of data still needs constant improvement. In this chapter, we propose a new data audit proof, in which nodes continuously upload hashed data that has been added to random numbers, and the smart contract will compare the result to verify the integrity of the data. Meanwhile, data owner could calculate and then challenge to verify the data integrity. There are audit miners responsible for regulating the behavior of miners and the protection of users' data, and audit miners in a state of semi-participation. It is demonstrated later in the chapter that this proof is accurate enough and resistant to attacks.

Keywords Distributed storage networks · Cloud storage · Blockchain

1 Introduction

Storage technology has evolved rapidly over the last few decades, with continuously decreasing hard disk prices and ever faster data speeds. However, the rapid growth of the online economy and big data technology has caused the need for data storage to expand exponentially, leading to the idea of cloud storage, in which data will be stored on cloud servers provided by third parties, and thus users can access data in a timely

L. Lu · K. Yuanyuan · Y. Yong (✉)
School of Mathematics, Renmin University of China, Beijing, China
e-mail: yong.yuan@ruc.edu.cn

L. Lu
e-mail: liulu0309@ruc.edu.cn

K. Yuanyuan
e-mail: ke_yy@163.com

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_4

and more convenient manner. Typically, cloud storage providers use technologies such as distributed storage (Mattson et al., 1970) to significantly reduce storage costs. However, the centralized storage makes cloud storage providers vulnerable to single points of failure and can create risks such as overstepping provider privileges and causing information leakage. Efficient centralized storage provisioning will remain mainstream in the future, but there is also an emerging and urgent need to meet users' needs for information security.

Traditional cloud storage providers, such as Amazon and Google, build cloud storage architectures with vast resources, using distributed storage technology to serve billions of users. Distributed storage means that data are spread across multiple storage servers and these scattered storage resources form a virtual storage device, effectively storing data in various places across the provider. The benefits of distributed storage are increased system reliability, availability, and access efficiency, as well as improved scalability. But the disadvantages are also obvious. We store our data on Google Cloud Drive on the basis that we trust Google to protect our data from being tampered with or lost, which can also lead to other disadvantages. The central server is vulnerable to attacks from adversaries, and internal failures and malpractice can also lead to data loss. As such, the security of cloud storage has also been a focus of attention in recent years. Traditional symmetric encryption algorithms put the keys on a central server, which makes it easier for attackers to get these keys and thus reduces the security of information. Moreover, data integrity verification whether data are stored efficiently and without deletion is also a crucial part of cloud storage services. Literature (Priyadharshini, 2012) summarizes the data integrity verification of traditional cloud storage, which is performed by TPA (Third Party Auditor) between the user and the CSP (Cloud Service Provider) to validate the data. The user poses a challenge to verify the integrity of their cloud data, and the TPA responds by comparing the original data, or the hash value of it according to literature (Zikratov et al., 2017), to verify the integrity. However, inefficiencies and tripartite or joint evil behavior can make opaque audit proofs unreliable. The convenience of centralized services brings with it the corresponding pitfalls. With the emergence of Bitcoin, decentralized technology continues to be improved, and decentralized storage brings an important addition to the traditional storage market.

The idea of providing decentralized storage has become popular with the rise of blockchain technology, and their combination could be considered a perfect fit. Blockchain enables reaching the consensus among decentralized, untrusted nodes. Its development has facilitated intensive research in several technologies such as cryptography, data structures, and consensus algorithms. When data are stored in multiple copies on the hard drives of different nodes, we cannot guarantee that all nodes are trustworthy. How to ensure the security and integrity of the data is a very crucial issue. After ensuring the stability of the storage, we also need to consider how to motivate people to become nodes and provide their own storage capacity, which requires a reasonable incentive mechanism.

Much of the current research is focused on issues such as access control, integrity verification, data retrieval, and traceability. Many platforms that offer distributed storage have already been launched. For example, the Sia

(Vorick & Champine, 2023) storage system, which was online earlier, has been unable to be developed effectively due to its less than optimal incentive design. IPFS (Benet, 2014), as a relatively complete platform, is a distributed storage system protocol for distributing and storing resources of various data types. Filecoin (Protocol Labs, 2023), as its incentive layer, incentivizes storage miners and retrieval miners to complete their own work by issuing tokens. Taking Filecoin as an example, there are three roles in Filecoin: client, storage miner, and retrieval miner. Clients pay for the service of storing and retrieving data. They can choose from a selection of available service providers. If they want to store private data, they need to encrypt it before submitting it to the service provider. Storage miners store clients' data for a reward. They decide for themselves how much space to provide for storage. After the client and the storage miner have reached an agreement, the miner is obliged to provide proof of their stored data on an ongoing basis. Everyone can view this proof and make sure that the storage miner is reliable. Retrieval miners give data to customers upon their request. They can retrieve data from clients or storage miners. Retrieval miners and clients use small payments to exchange data and tokens. The data are fragmented and the client pays a small amount of tokens for each fragment. Retrieval miners can also act as storage miners at the same time.

We will now show how a decentralized storage network stores and audits. As shown in Fig. 1, we demonstrate a cloud service with blockchain participation in two aspects: storage and audit. Data owners upload their data to miners on the server, who store the data and record the transactions on the blockchain. The blockchain also verifies data owner's information and protects the user's privacy. In order to ensure that their data are stored intact on the server, data owner challenges the TPA, which

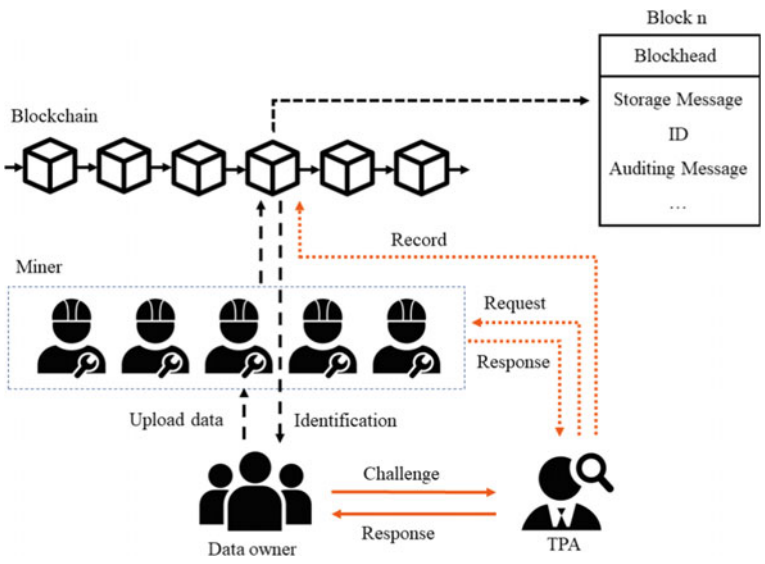

Fig. 1 Decentralized storage network framework

sends a request to the system and verifies the data provided by the miner in response to data owner's challenge. The verified result is then recorded on the blockchain. So each block in the blockchain stores information such as the height of the block, the block header, information about the previous block, a timestamp, storage message, ID, and Auditing message.

Data integrity verification in distributed storage, i.e., the red lines in Fig. 1, is our concern. Data integrity verification is the verification that data is stored intact in the storage space of each untrusted node. This affects the security of the data and is key to the availability of the storage service. Current data integrity validation can be divided into two kinds, one for traditional cloud-based data integrity validation and the other for blockchain-based data integrity validation. In turn, audit solutions using blockchain technology can be divided into whether or not TPA is involved. Most of these audit schemes verify against raw data and avoid dishonest behaviors such as delayed audits, sybil attack, and generation attack through consensus and incentive mechanisms. Blockchain-based data integrity verification can be used not only for auditing cloud data in cloud storage networks but also for different data scenarios to improve the security of the system. However, efficiency and accuracy cannot be achieved together in the process of decentralized data auditing. This will be described in Sect. 3. Most of the verification schemes that have worked better in current research do not run in public blockchains or require the participation of trusted central nodes. Instead, in fully decentralized blockchains, most are more efficient in order to ensure availability. However, the accuracy of verification cannot be fully guaranteed and the system is vulnerable to dishonest attacks. Our algorithm will improve long-term efficiency and stability in a fully decentralized blockchain with guaranteed accuracy.

Our work is based on a modification of Filecoin for verifying the integrity of data in distributed storage. The audit miner in this algorithm is semi-involved and determines whether the data are kept intact by comparing the hash values of the data shards. If the result does not meet the desired goal, the audit miner will first ensure the integrity of the data and then find the storage miner that created the problem, acting as a reasonable supervisor. In Sect. 2, we will summarize the past work on distributed audit algorithms and describe the characteristics of each platform. In Sect. 3, we will present our audit algorithm and analyze its advantages and the problems it solves. Sect. 4 will analyze the fault tolerance of this audit proof.

2 Related Works

2.1 Audit Research

Ensuring data integrity in cloud computing has always been an important issue, and it is a guarantee that cloud computing can be widely used. Traditional data integrity verification can be divided into deterministic and probabilistic types. The dishonest behavior of TPA is also an important issue for audit algorithms when they are entrusted to perform audit integrity verification work. Blockchain technology

with a decentralized architecture no longer relies excessively on the honesty of third parties and reaching an overall consensus based on a reasonable consensus and incentive mechanism and then a mutual benefit for all parties is the core element to be explored at this stage.

Literature (Zikratov et al., 2017) proposes a private blockchain called Zeppar, which determines the integrity of data by comparing the hash values of files. The use of cryptographic techniques to verify data integrity by comparing the original data is a common and applicable method. Such an approach is also used in literature (Wei et al., 2020), where smart contracts monitor data changes based on the unique hash value corresponding to the file generated by the Merkle Hash Tree (MHT). Verifying data integrity by constructing MHT is a relatively convenient method, e.g., in literature (Bai et al., 2018; Li et al., 2020). In literature (Li et al., 2020), data owner (DO) stores the verification tag of the data on the blockchain and verifies the data integrity by constructing MHT. After the blockchain network receives a request from the DO, it calculates the MHT root of the specified data, the CSP receives the DO's challenge and also calculates the corresponding MHT root, and the DO verifies the integrity of the data by comparing the two. We can find that neither the method of comparing file hash values nor the construction of MHT requires the involvement of TPA. Such an approach can be very efficient for verification but will compromise on the degree of centralization or be less fault-tolerant. It is relatively suitable for distributed storage systems where efficiency is required.

In order to ensure the activity of the data, some auditing schemes use the provision of random numbers to avoid users falsifying the results of data validation in advance. Literature (Pinheiro et al., 2020) uses the user's data information to generate random challenges and uses the smart contract to audit the challenge-response information sent by the CSP. The audit scheme also assesses the trustworthiness of each CSP.

2.2 Distributed Storage Project

- Sia: A relatively early decentralized storage platform, Sia in literature Vorick and Champine (2023) enables storage contracts to be formed between peer-to-peer nodes. The contracts are stored in the blockchain, making them publicly auditable. Sia divides files into 30 parts, encrypts each part using the Threefish algorithm, and distributes them to different nodes. Reed-Solomon erasure coding makes it possible to fully recover a file by requiring only any 10 of the 30 parts. With Merkle Tree (Ralph, 1988), nodes are required to upload storage proofs (Maxwell, 2023) within a certain time frame or be penalized.
- Filecoin: Literature (Benet, 2014) proposes a distributed peer-to-peer web protocol: IPFS (InterPlanetary File System). Based on a content addressing protocol, it makes network transmission faster, content storage easier and nodes protection safer. Filecoin can be considered the incentive layer of the IPFS system, providing decentralized cloud storage in the form of tokens distributed in a rational way. Its audit algorithm Proof-of-Replication shown in literature (Protocol Labs, 2017)

deferred encoding of data to get a copy of the data and then generates a zero-knowledge proof to guarantee the correctness of the encoding process. Its other consensus algorithm, Proof-of-Spacetime, requires miners to periodically generate Merkle proofs for the replicas and submit them to the blockchain compressed with zero-knowledge proofs for tokens reward. Such an incentive encourages miners to store data correctly and to prove data liveness to obtain proof of work as a reward.

- Arewave: Arweave cloud storage platform is similar to Filecoin in that it features a service that provides permanent storage. It has designed a new consensus algorithm, Proof of Access, which is based on the concept that new blocks require random validation of previous blocks. This turns the original blockchain into a network of blocks, where nodes no longer need to store exponentially growing amounts of data, but only certain data, allowing the data to be distributed evenly across the system to achieve distributed storage.
- Storj: Storj Labs (2018) built at Kademlia is not a fully decentralized cloud storage system and it is dedicated to data storage durability and storage quality. Satellite nodes act as fully trusted nodes in storj for data management and data integrity review. The data are sliced after encryption and the data integrity is guaranteed by Proof of Retrievability (Juels et al., 2007) consensus algorithm. The satellite nodes are responsible for communication between the user and the storage node, for storing metadata for the user, as well as auditing and enforcing Proof of Retrievability. The presence of the satellite nodes makes storj resistant to Byzantine attacks, but at the expense of the network's performance, resulting in poor scalability.

Table 1 has given the difference among these four platforms. We can find that their audit proofs are different and lead to other differences in other natures.

However, there are still some flaws. The current work almost verifies the integrity of distributed storage data under specific conditions, but none of it has a systematic analysis of the limitations of auditing. We will analyze the compromise factors that

Table 1 Distributed storage networks comparison

	Degree of decentralization	Storage location	Consensus algorithm	Audit proof
Sia	Fully	Off chain	Proof of work	Proof of storage (Maxwell, 2023)
Filecoin	Fully	Off chain	Expected consensus (Protocol Labs, 2017)	Proof of replication, proof of spacetime
Arweave	Fully	On chain	Proof of work, proof of access	Proof of access
Storj	Satellite nodes exist	Off chain	Proof of work, proof of stake	Proof of retrievability (Juels et al., 2007)

can arise from audit algorithms in distributed storage in the next section. We also analyze what requirements the Filecoin platform should have for auditing and what constraints it should have on storage miners. We design an audit proof for distributed storage and prove that it is sufficiently accurate and fault-tolerant.

3 Audit Algorithm

3.1 An Audit Framework

In this chapter, we will reformulate the audit proof of Filecoin to address the current problems of Filecoin platform. Our goal is to retain the decentralized nature and allow the distributed storage network to complete the audit process on its own. Audit miners will only appear when necessary. This will ensure the accuracy of the audit and improve the efficiency of all nodes in reaching consensus on the audit results. We propose the audit impossibility proposition regarding the distributed storage networks as follows:

Proposition 1 (Audit impossibility): The degree of decentralization, the accuracy of audit results, and audit efficiency cannot be reached at the same time.

When integrity checks are performed on an absolutely centralized storage server, CSP can invest significant resources in a way that increases the efficiency and accuracy of the audit, as many cloud storage providers do nowadays. This is the approach that currently dominates the cloud storage market. However, with decentralization, we cannot perform fast and efficient integrity checks on untrustworthy storage nodes based on today's computing power and the sheer volume of data. How to balance accuracy and efficiency is currently the key issue for auditing in all distributed storage. For Filecoin, decentralization is its biggest advantage. However, too frequent data auditing not only affects the accuracy of the data audit results, but also causes the system to be less stable when the nodes are offline. Therefore, to improve the efficiency of auditing while ensuring the accuracy of the audit results is the issue considered in this chapter.

Our design starts by slicing and numbering the data owners encrypted data using the shard technique and then generates multiple copies (k copies) by replication, which will be stored randomly on storage miners. When auditing these files, we will take the last 16 bits of the hash of the previous block as the new random number $\mathcal{N}$, which all miners will add to each of their stored shards for hashing. The result will need to be uploaded to the hash pool in a certain order with the miners' signatures. All the hash values are automatically matched by the smart contract. By determining whether the corresponding hash value is equal to k, it is concluded that the data are stored intact in the distributed storage network. This allows a simple comparison of

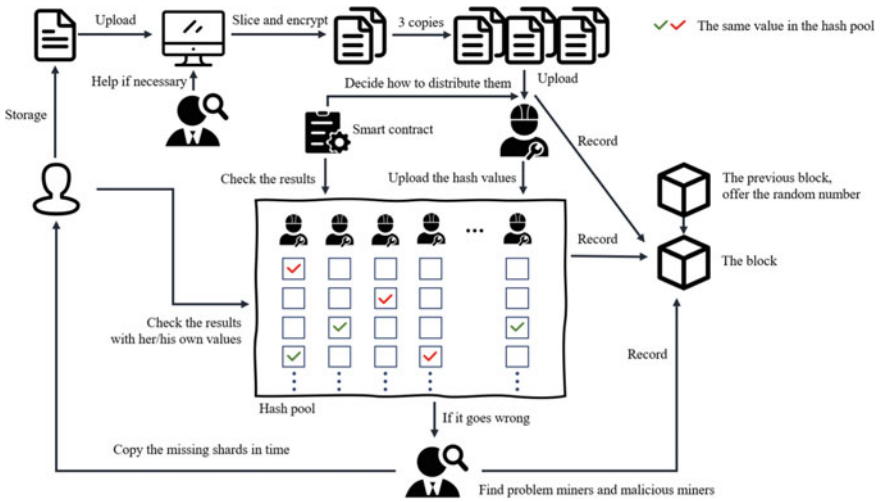

Fig. 2 Algorithm overview

the results to determine whether the data owners' data are completely stored across the network. The data owner can also, but not necessarily, add his/her own shard data to the random number $\mathcal{N}$ and hash them. The result is then compared across the hash pool to determine if the data are stored correctly on the miners by finding the same k values in the network result. If the storage miner is not validly stored, the audit miner needs to find the problem miner quickly and back up the data in time (Fig. 2).

There are three roles in our platform, data owners U, storage miners M, and audit miners A. Data owners upload encrypted data according to their needs and can challenge the integrity of the data. The storage miner stores the data sequentially as assigned by the smart contract as well as uploads proof of data integrity every once in a while. The audit miner is responsible for handling the distribution of data, as well as reviewing and supervising miners, protecting data integrity, regulating content, and assuming legal responsibility. The number of audit miners is limited and storage miners can be audit miners at the same time. Audit miners only appear if there are problems with the audit.

3.2 Data Uploading

From the moment the user uploads data, the user U_i should divide his/her data $D(i)$ into several shards by using slicing and encryption technology in order to keep the data secure. If he/she does not have enough computing power to handle too large data, he/she can upload them to audit miner A for slicing and encrypting and then pay some tokens. All the shards are then distributed by the audit miner to storage

miner M and back to user U_i. Data slicing is a common technique used in distributed storage to protect the data. We use $D(i, j)$ to denote the jth shard of U_i's data. The number of shards U_i has, J_i, will be determined by the size of U_i's data. Replication is also used to replicate k copies of $D(i, j)$: $D(i, j, k)$. Before uploading, the U_i can add a random number N known only to him/her to calculate the result for all $D(i, j)$ and encrypt the result as a validation audit option later. Next, the miner M_a is randomly sent a request to store the corresponding shard or not, with a specific request (Definition 1). M_a that receives the request has to choose whether to store the data or not, depending on its storage capacity. The miner who confirms storage will store the corresponding $D(i, j)$ on his local hard disk. The label (i, j) of the data $D(i, j)$ will only be stored in the smart contract and will not be transmitted to the miner who stored it. The miner will not know the exact label (i, j) of the data he/she stores, but will only number them sequentially according to the order in which he/she stores $D(i, j)$. If $D(i, j)$ is the sixth data storage of M_a, then the corresponding $D(i, j)$ is $M_a(6)$. We would use $M_a()$ to express the set of shards stored by M_a. This allows for better protection of the user's information and data, and prevents the exchange of content between miners as much as possible.

We can effectively prevent malicious miners from sybil attacks or other attacks by slicing and replicating the data and storing them in a decentralized manner. We also require an appropriate specific request for sending shards to avoid joint attacks by miners.

Definition 1 *(Request of distributing shards):* The distribution of the set $\{D(i, j, k), \forall i, j\}$ to miners is subject to the following principles:

1. The number of M_a storing the data of $D(i)$ cannot be less than half of J_i.
2. No miner M_a will receive two or more storage requests for a single copy of data $D(i, j)$.
3. No miner M_a will receive storage requests for $D(i, j)$ and $D(i, j + 1)$.
4. Miners M_a and $M_{a'}$ will not receive storage requests for $D(i, j)$ and $D(i, j^{'})$ together.
5. No miner will store more than y copies of $D(i)$.
6. Miners M_a and $M_{a'}$ will store no more than z identical shards in the shards pool.

This ensures that the data are stored in a sufficiently decentralized manner, with enough miners storing the data owner's data together, so that a single point of failure does not have a major impact on the overall storage. It also ensures that the user's data are not stolen in its entirety, guaranteeing the security of the data. Definition 1 also makes the data stored by the two nodes different, avoiding outsourcing attack. We will specifically analyze the effectiveness of our algorithm in Sect. 4.

3.3 Self-integrity Verification

Now all data owner's data has been uploaded to each storage miner. We then need to continuously interact with all miners to ensure that data liveness is guaranteed and

the data are being stored intact. This is the core of the work in this chapter. We have described in Sect. 2, the current auditing methods, both fully decentralized and not fully decentralized, that are able to do the job but not well in the accuracy of audit results or audit efficiency. This chapter proposes a solution that does not require the data owner's data to be compared and achieves self-auditing through self-comparison in the blockchain network, which substantially improves the long-term stability of the system. At the same time, our proof is more efficient and can quickly reach a consensus on the integrity of all data in a short period of time with responses from all nodes. We also allow data owners to initiate challenges and quickly check the integrity of their own data through the hash algorithm.

The blockchain network audits whether the storage miners have correctly stored the corresponding data within a period T. To ensure timeliness, we use the last 16 bits of the hash value of the previous block as a random number $\mathcal{N}$. After getting $\mathcal{N}$, the miner has to upload the result of the hash operation of all his shards and $\mathcal{N}$ together with his signature M_a_sign within a specified time T. Now we obtain a new set: $\{hash(M_a(), \mathcal{N}), M_a\_sign\}$ to express the result of the hash of all M_a's shards and its signature. It is important to note that the set is ordered, again according to the order in which M_a stores the shards. The advantage of this design is that even when faced with a pile of results, the smart contract can determine the corresponding label (i, j) based on its position. We will use $H(a, b)$ to denote the hash value corresponding to the bth shard of the miner a with $\mathcal{N}$, $D(i, j)\_hash$ to denote the hash value of $D(i, j)$ with $\mathcal{N}$ (Table 2).

After the storage miner M_a has uploaded his/her $\{hash(M_a(), \mathcal{N}), M_a\_sign\}$, the smart contract will quickly determine if the number of $H(a, b)$ is equal to the number of shards already stored by M_a, and if it does not match, invalidate this result and demand M_a to recalculate and upload the new result. If the result matches, the result is accepted and moves on to the hash pool. Next, the smart contract will compare the number of occurrences of all the results in the hash pool. If there are exactly k identical results, i.e., if there are k sets (a, b) s.t. all the results of $H(a, b)$ are equal, then it will be decided that all the copies of the shard have been stored correctly. This would be the best result that can be achieved. All the storage miners need to store their data correctly for their own benefits. If all miners store correctly, all the nodes can quickly and accurately obtain the result that the data are stored intact. We will now determine whether the data are stored correctly based on the occurrences of each hash value.

Definition 2 *(strong integrity)*: The number of occurrences of $D(i, j)\_hash$ is exactly equal to k.

Definition 3 *(weak integrity)*: The number of occurrences of $D(i, j)\_hash$ is greater than or equal to 2 and less than k.

Table 2 Notations for operations/implications

Symbol	Notations for operations/implications
U	Data owners
M	Storage miners
A	Audit miners
U_i	The jth data owner
$D(i)$	Data owner i's data
$D(i, j)$	The jth shard of U_i's data
k	The number of copies
J_i	The number of U_i's shards
$D(i, j, k)$	All of the U_i's shards
M_a	The ath storage miner
(i, j)	The label of $D(i, j)$
$M_a(6)$	The sixth shard stored by M_a
$M_a()$	The set of M_a's storage
T	Cycle time for storage miners uploads
N	The random number set by the user
$\mathcal{N}$	The random number from the previous block
M_a_sign	M_a's digital signature
$H(a, b)$	$M_a(b)$'s hash value with $\mathcal{N}$
$D(i, j)\_hash$	$D(i, j)$'s hash value with $\mathcal{N}$

If all shards achieve strong integrity, we can assume that the storage network has stored all data correctly and that all nodes would agree on this. If all shards achieve weak integrity, we can assume that all data are stored securely on the storage network. Weak integrity is a lower requirement for data availability in storage networks. During auditing, it is more of a constraint on the miners, so strong integrity is what is required by distributed storage networks.

We will now discuss what to do if strong integrity is not achieved. If the number of occurrences of a hash value is greater than k, the possible scenario is that the miners are jointly misbehaving with each other and copying the same result for output. This is because when the storage miner receives the shard corresponding to that result, no other shards are received, and only if the miner has stored other miners' shards. In this case, the $k + \alpha$ results are assigned a number (i, j) based on their location, and the numbers are then compared to find the miner with the incorrect result by audit miners A. The first step is to find the set of $\{(i', j')\}$ corresponding to the wrong hash value, and then check whether the number of occurrences of hash value is k. If it is k, the shard has been completely stored in the storage network. Otherwise, this

number can only be less than k, if so, A needs to find which miners did not upload the right value and ask them to upload in time. If they upload the wrong results, ask them to re-store correctly to solve the problem.

In fact, it is more often the case that the number is less than k. In case when the hash values whose number is less than k, we need the corresponding miners to upload proofs of the correct storage of the corresponding $D(i, j)$. The following results may occur:

1. The miner correctly stores $D(i, j)$ and uploads the correct hash result.
2. The miner correctly stores $D(i, j)$ but uploads the wrong hash result.
3. The miner incorrectly stored $D(i, j)$ but uploaded the correct result.
4. The miner incorrectly stored $D(i, j)$ and uploaded the wrong result.

Audit miners A need to immediately copy $D(i, j)$ to ensure that they couldn't be lost. After that, A will handle errant storage miners as above. Such handling effectively avoids errors caused by miners offline. We will also judge storage miners who make frequent errors as malicious miners. If for the same $D(i, j)$, all the results of the hash operation are different or it is not possible to distinguish the correctness of the result, then A can ask all miners storing the $D(i, j)$ to recalculate it with the random number N and compare it with the result calculated by U_i. In time, copy the data of the miners that output the correct result and ask U_i to re-add another random number N to the calculation and keep the result for future use (Fig. 3).

The above is the process by which a blockchain storage network audits of its own. This process allows for quick consensus to be reached under the condition that all the data are stored correctly, as well as finding malicious nodes if consensus is not reached.

3.4 Data Owner's Integrity Verification

After the data owner gets $\mathcal{N}$, he/she can also get a set of hash values $H(i, j)$ generated by U_i by performing a hash operation on his/her own data shards $D(i, j)$. Smart contract will look for k occurrences of these values in the hash pool to determine whether his/her data have been stored completely. If exactly each result occurs k times, then it is almost certain that U_i's data has been stored correctly. If not, then the storage miner in problem can be found quickly and the data copied by the audit miner in his/her storage in time. Such an audit approach improves the shortcomings of self-integrity verification and increases the accuracy of data integrity verification.

3.5 The Game of Miners Versus Storage Networks

Storage miners can only earn if they store the user's data correctly and upload $H(a, b)$ correctly. If the miner wants to earn without storing correctly, he needs to join with

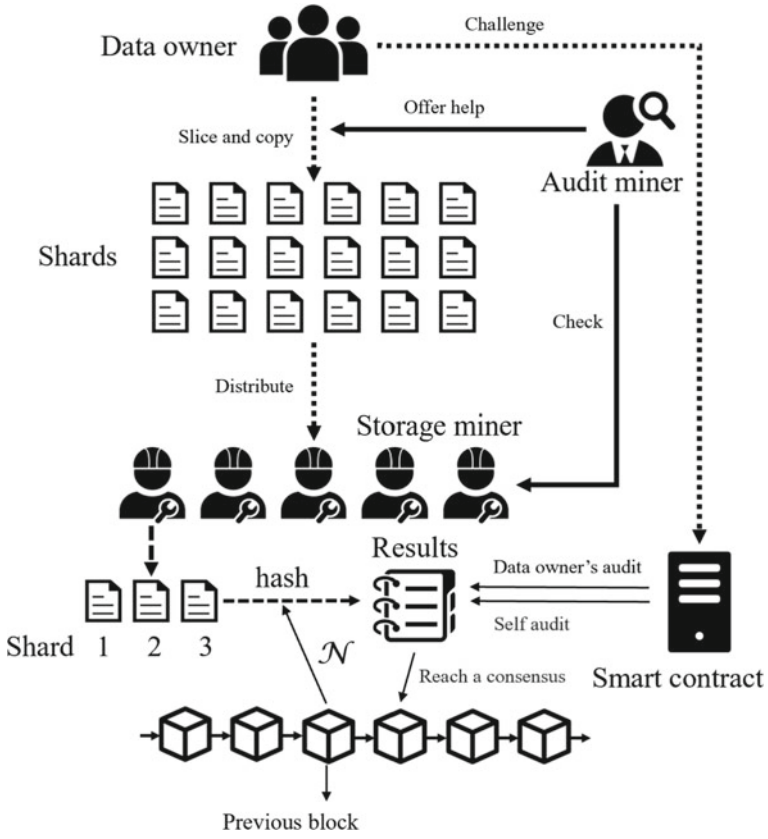

Fig. 3 Audit algorithm

other miners. The miner does not know the number (i, j) of the data he is storing, so he needs to send a request to all miners in the network. And other miners can be rewarded by reporting those malicious nodes. The union of storage miners does not earn a reward, only the individual fulfillment of the storage function makes the storage network maximize its benefits. For audit miners, audit miners are only given the appropriate audit access if there is a problem with the storage miner. Audit miners are only able to earn more rewards by continuously completing audit tasks and tasks delegated by data owners. These ensure that all miners are driven by profit to achieve stability.

Thus our system satisfies incentive-compatible property and also data integrity, recoverability, publicly verifiable, and auditability. The satisfaction of the five properties is obvious. These are the same properties that filecoin satisfies. We can say that our audit proof is reasonable.

4 Fault-Tolerance Verification

We now describe three attacks that are common in distributed storage networks.

- Sybil Attack (Douceur, 2002): Sybil attack is a type of attack in peer-to-peer networks in which a node in the network operates multiple identities actively at the same time and undermines the authority/power in reputation systems. In a distributed storage network, a malicious miner can create multiple sybil identities pretending to store many copies in order to be rewarded, but only one copy is stored in his local.
  In our proof, a miner cannot claim to have stored multiple shards, as the number of shards per share is limited to k. Meanwhile, there is a little additional gain for a malicious miner to pretend to store multiple copies by creating multiple identities. Since each miner stores different content and for two storage miners, they have the number of the same shards less than z. We control the revenue in such a way that storage miners will not receive enough benefit in creating a witch identity, making them less likely to take risks for it. Subsequently, we can limit such a situation even further by monitoring IP address, generating $M_a()$ proofs, etc. Such a scenario makes sybil attacks much less profitable.
- Outsourcing Attack: By relying on fast access to data from other storage providers, malicious miners promise to store more data than they can actually store.
  If a malicious miner wants to launch an outsourcing attack, the miner cannot know the shard number and can only determine if there is an overlapping shard by sharing the miner's $H(a, b)$ set with each other; if there is an overlapping shard, the hash result can be quickly retrieved later in the audit. But the benefit to the provider is weak, and the inclusion of an exposing mechanism keeps miners from going to extremes for the weak benefit. So we can conclude that the benefits of a small number of miners cooperating are much less than the risks associated with incomplete storage.
- Generation Attack: Malicious miners claim to have more storage than they actually have through a small program to gain a greater advantage in the mining competition.
  With slicing and cryptography, miners cannot effectively generate data with small program. The generated proof results need to be computed by hash function, and a small change can lead to a huge difference in results. There are strict penalties for generation attack in Filecoin, so this attack can be substantially avoided from an incentive point of view.

5 Concluding Remarks

In this chapter, we focus on current research on auditing and point out the imperfections of current auditing. We also analyze the audit requirements for Filecoin and redesign an audit algorithm for it. The algorithm determines whether the data have been stored intact in the storage network by comparing the results in the hash pool by means of storage miners uploading the hash results. The audit miner is set to a semi-participating state and will only join in time to gain access if a problem arises. Such an auditing algorithm is relatively accurate and secure for decentralized storage networks. Besides, it is obtained that the algorithm is highly fault-tolerant.

Our algorithm is not yet well designed in terms of incentives and needs to prove that the algorithm can be put into widespread use. Incentives are a key part of getting the algorithm used, and it is important to play the game between miners and the storage network so that both sides can get the optimal solution for their interests. The regulation of the data is also something that needs to be considered in the next phase. Our algorithm needs to be more complete in the future.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (72171230), and the Science and Technology Development Fund, Macau SAR (File No. 0050/2020/A1).

References

Bai, L. H., Xue, J. T., Xu, C. X., et al. (2018). DStore: A distributed cloud storage system based on smart contracts and blockchain. In *International Conference on Algorithms and Architectures for Parallel Processing*. Springer.

Benet, J. (2014). IPFS—Content addressed, versioned. *P2P File System*.

Douceur, J. R. (2002). The sybil attack. Springer.

Juels, A., Kaliski, B. S., PORs, J. (2007). Proofs of retrievability for large files. In *Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS* (Vol. 07, pp. 584–597). ACM.

Li, J., Wu, J., Jiang, G., et al. (2020). Blockchain-based public auditing for big data in cloud storage. *Information Processing & Management, 57*(6), 102382.

Mattson, R. L., Gecsei, et al. (1970). Evaluation techniques for storage hierarchies. *IBM Systems Journal*.

Maxwell, G. Proof of storage to make distributed resource consumption costly. https://bitcointalk.org/index.php?topic=310323

Pinheiro, A., Canedo, E. D., Sousa, R., et al. (2020). Monitoring file integrity using blockchain and smart contracts. *IEEE Access, 8,* 198548–198579.

Priyadharshini, B. (2012). Data integrity in cloud storage. IEEE.

Protocol Labs. Filecoin: A decentralized storage network. https://filecoin.io/filecoin.pdf

Protocol Labs. Technical report: Expected consensus.

Protocol Labs. Technical report: Proof-of-replication.

Ralph, C. (1988). Merkle: A digital signature based on a conventional encryption function. In C. Pomerance (Ed.), *Advances in cryptology, CRYPTO* (Vol. 87, pp. 369–378). Springer.
Storj Labs. Inc. Storj: A decentralized cloud storage network framework.
Vorick, D., & Champine, L. Sia: Simple decentralized storage. https://blockchainlab.com/pdf/whitepaper3.pdf
Wei, P. C., Wang, D., Zhao, Y., et al. (2020). Blockchain data-based cloud data integrity protection mechanism. *Future Generation Computer Systems, 102,* 902–911.
Zikratov, I., Kuzmin, A., Akimenko, V., et al. (2017). Ensuring data integrity using blockchain technology. In *2017 20th Conference of Open Innovations Association (FRUCT)*.

Iterative Learning Control Based on Random Variance Reduction Gradient Method

Yihua Gao, Dong Shen, and Jiaxi Qian

Abstract Traditional iterative learning control (ILC) algorithms usually assume that full system information and operation data can be utilized. However, due to the uncertainty and complexity of actual systems, it is difficult to access full system information and operation data accurately and completely. In this chapter, a novel ILC scheme based on stochastic variance reduced gradient (SVRG) is proposed. This scheme is not only suitable for resolving the incomplete information problem, but also converges efficiently under both strongly convex and non-strongly convex control objectives. To demonstrate the advantages, this chapter studied two scenarios, i.e., random error data dropout and model-free data-driven approach, and proposed two SVRG-based ILC algorithms for these two scenarios, respectively. It is theoretically demonstrated and experimentally verified that the proposed SVRG-based ILC scheme converges faster than both the full gradient and stochastic gradient methods for the two involved scenarios.

Keywords Iterative learning control · Variance reduction gradient method · ILC algorithms

Y. Gao
Renmin University of China, Beijing, China
e-mail: gaoyihua@ruc.edu.cn

D. Shen (✉)
Distributed Artificial Intelligence Laboratory, Renmin University of China, Beijing, China
e-mail: dshen@ruc.edu.cn

J. Qian
School of Mathematics, Renmin University of China, Beijing, China
e-mail: qianjx1917@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_5

1 Introduction

1.1 Background

Iterative learning control (ILC) is a control method applicable to systems doing repeated operations. The basic idea is to use the input and error signals from previous iteration to improve the input of the next iteration. Arimoto et al. first proposed iterative learning control for robotic arms in 1984 and clarified the basic idea of iterative learning control (Arimoto et al., 1984). Subsequently, academia has published numerous chapters around ILC. It has gradually become one of the important branches in the field of control and is widely used in robotics, industrial production and hard disk manufacturing, and other controlled systems with repeated operation.

To achieve excellent control performance, most ILC assume that full operational data and system information can be obtained and utilized. However, in real systems, data delays and dropouts often occur due to various uncertainties. On the other hand, when the system structure is complex or unstable, it is difficult to obtain the system information accurately. To solve the incomplete information problems, it is of great theoretical and practical significance to design ILC algorithms with high performance.

Information incompleteness can be classified into two categories, objective and subjective incompleteness. Information incompleteness caused by objective factors is often related to the uncertainty of the system itself. For example, during the transmission of the signal, the instability of the channel can cause data packet loss. Three main random packet dropout models have been developed for this problem: the random sequence model, the Bernoulli distribution model, and the Markov chain model. Shen (2018) designed an iterative learning control algorithm based on the stochastic approximation algorithm corresponding to the three models and proved that the algorithms satisfy mean-square convergence and probabilistic strong convergence. Information incompleteness due to subjective factors usually artificially assumes that the system information is unknown, thus avoiding the complexity of system modeling and system instability. For example, Oomen et al. (2014) designed a model-free data-driven iterative learning control algorithm for H_∞-parametric estimation of multi-input multi-output (MIMO) systems, which obtains the full gradient by conducting $n_0 \times n_1$ experiments on $n_0 \times n_1$-dimensional MIMO systems. However, this algorithm is difficult to be applied to large MIMO systems due to the excessive number of experiments. Subsequently, Aarnoudse et al. (Owens et al., 2009) designed an iterative learning control algorithm based on the stochastic approximation method by constructing a random matrix to estimate the gradient, which effectively reduces the number of experiments.

It is important to note that the effect of information incompleteness on ILC tracking performance is essentially the robustness of ILC. However, this robustness differs for objective and subjective-type information incompleteness problems. The former is usually model-based and emphasizes modeling to analyze the causes of information deficiency. While the latter is data-based and is generally not concerned with

the causes of information deficiency but with the inherent limitations of information deficiency on control performance. Model-based and data-based control methods are not opposed. To achieve the best control effect, the two control methods can also be used in combination. Existing studies on ILC for solving the information incompleteness problems are usually based on stochastic approximation method or other gradient methods. In this chapter, we will use a stochastic variance reduction gradient (SVRG) method to give a general framework for solving the system information incompleteness problems.

1.2 Design and Analysis of SVRG-Based ILC

Modeling the control objective as an optimization function, for a deterministic discrete-time linear system, Owens et al. (Aarnoudse & Oomen, 2020) proposed a gradient-type ILC algorithm based on optimization ideas and analyzed the stability, monotonicity, and robustness of the algorithm. For noisy discrete-time linear systems, Yang and Ruan (2017) proposed an enhanced gradient-based ILC algorithm that can effectively converge in the presence of perturbations in the system. However, the above gradient-based ILC algorithm requires full error and system information for each iteration, and when this information is not fully available, the traditional gradient-based ILC algorithm is no longer applicable.

Notice that in Machine Learning, Stochastic Gradient Descent (SGD) method replaces the total gradient by randomly selecting a partial gradient each time. Corresponding to the control problems, the partial gradient can also be obtained when there is insufficient information about the error or the system. This correlation inspires us to find suitable stochastic gradient methods to solve errors or system information insufficient problems.

In order to improve the convergence speed and apply to non-smooth and non-strongly convex objective functions, recent research in Machine Learning has produced a large number of improved versions of stochastic gradient descent algorithms, including momentum method, variance reduction method, and incremental aggregated gradients. Allen-Zhu (2018) divided these algorithms to three types according to their complexity under strongly convex conditions. The first generation is the momentum-based gradient algorithm, the second generation includes the variance reduction-based gradient algorithm and the proximal stochastic variance reduction gradient algorithm, the third generation includes the Katyusha algorithm and incremental aggregated gradient algorithms. In most cases, the complexity of the algorithms decreases with the growth of generation. Considering specific control problems, the algorithms in first generation are slow to converge and often fail to meet the practical needs, while the algorithms in third generation require accurate system modeling to achieve faster convergence and are difficult to apply to data-driven ILC. Therefore, the research in this chapter is mainly based on the algorithm in second generation—Stochastic Variance Reduction Gradient (SVRG) algorithm.

1.3 Main Work and Organization

The purpose of this chapter is to construct SVRG-based ILC and use this framework to solve specific information incompleteness problems. As representatives of objective and subjective information incompleteness, two scenarios, error data random dropouts and model-free ILC, are selected in this chapter to give the corresponding SVRG-based ILC algorithms, respectively. The contribution is threefold.

1. Propose a SVRG-based ILC framework for single-input single-output (SISO) systems. The algorithm is shown to converge linearly under smooth and strongly convex conditions.
2. Apply the SVRG-based ILC framework to error data random data dropouts and give the convergence proof of the algorithm.
3. Extend the SVRG-based ILC framework to multi-input multi-output (MIMO) systems in model-free data-driven scenario and prove the convergence of the algorithm under smooth and non-strongly convex conditions.

Section 2 serves as the basis of the chapter, giving the SVRG-based ILC framework for SISO systems. Section 3 applies the framework to error data random dropouts problem. Section 4 extends the framework to MIMO systems in model-free scenario. Since Sect. 2 only gives the algorithm framework and does not cover the specific scenario, Sect. 2 does not give numerical simulations and contains only three parts: system description, algorithm design, and convergence analysis. Both Sects. 3 and 4 include four parts: system description, algorithm design, convergence analysis, and numerical simulation.

2 SVRG-Based ILC Framework

As the basis of the following sections, this section uses SISO systems to give the basic framework of SVRG-based ILC algorithm. This section includes three parts: system description, algorithm design, and convergence analysis.

2.1 System Description

Consider the following single-input single-output (SISO) discrete-time linear system

$$\begin{cases} x_k(t+1) = Ax_k(t) + Bu_k(t), \\ y_k(t) = Cx_k(t), \end{cases} \tag{1}$$

where $t = 0, 1, \ldots, N-1$ is time index, and $k = 1, 2, 3, \ldots$ denotes the iteration index. $x_k(t) \in \mathbb{R}^n$, $u_k(t) \in \mathbb{R}$, and $y_k(t) \in \mathbb{R}$ represent the system state, input and

output, respectively. $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n}$, and $C \in \mathbb{R}^{1\times n}$ are the system matrices. The initial condition is the same for each iteration, i.e., $x_k(0) = x_0, \forall k \in \mathbb{N}^*$.

Taking $t = 0, 1, \ldots, N-1$ in (1) yields

$$\begin{aligned}
y_k(1) &= CAx_k(0) + CBu_k(0) = CBu_k(0) + CAx_0,\\
y_k(2) &= CAx_k(1) + CBu_k(1) = CABu_k(0) + CBu_k(1) + CA^2x_0,\\
&\vdots\\
y_k(N) &= CAx_k(N-1) + CBu_k(N-1)\\
&= CA^{N-1}Bu_k(0) + CA^{N-2}Bu_k(1) + \cdots\\
&\quad + CABu_k(N-2) + CBu_k(N-1) + CA^Nx_0.
\end{aligned}$$

Combining the above equations, system (1) can be rewritten in the following equivalent form

$$y_k = Hu_k + Kx_0, \tag{2}$$

where $u_k = [u_k(0), u_k(1), \ldots, u_k(N-1)]^T \in \mathbb{R}^n$, $y_k = [y_k(1), y_k(2), \ldots, y_k(N)]^T \in \mathbb{R}^n$,

$$H = \begin{bmatrix} h_{11} & 0 & \cdots & 0\\ h_{21} & h_{22} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ h_{N1} & h_{N2} & \cdots & h_{NN} \end{bmatrix} = \begin{bmatrix} CB & 0 & \cdots & 0\\ CAB & CB & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ CA^{N-1}B & CA^{N-2}B & \cdots & CB \end{bmatrix}, K = \begin{bmatrix} CA\\ CA^2\\ \vdots\\ CA^N \end{bmatrix}.$$

For further analysis, the following assumptions are required.

Assumption 1 The input/output coupling matrix $CB \neq 0$.

Assumption 2 For desired trajectory $y_d(t)$, there exists a unique desired input $u_d(t)$ and initial state $x_d(0)$ such that

$$\begin{cases} x_d(t+1) = Ax_d(t) + Bu_d(t),\\ y_d(t) = Cx_d(t). \end{cases} \tag{3}$$

Also written in the form of (2), we have

$$y_d = Hu_d + Kx_d(0). \tag{4}$$

Remark 1 Assumptions 1 and 2 describe the realizability of the system for desired trajectory y_d. To be specific, Assumption 1 means that the relative degree of the system is 1. Assumption 2 describes the existence of an input signal u_d that can precisely trace y_d. If the system does not satisfy the Assumption 2, the system output can only be as close as possible to the desired trajectory y_d.

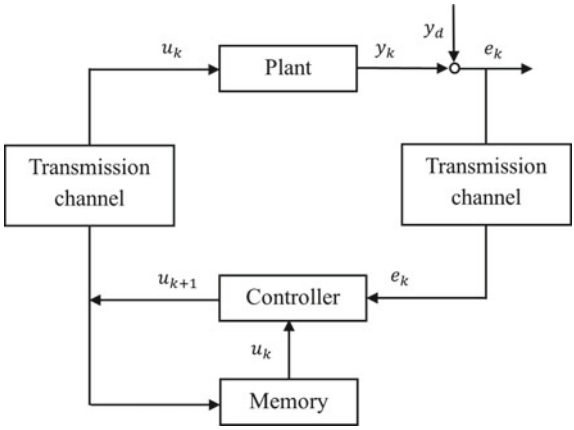

Fig. 1 Block diagram of ILC

Assumption 3 The initial states of (1) and (3) are identical, i.e., $x_d(0) = x_k(0) = x_0, \forall k$. Assume that $x_0 = 0$.

Remark 2 Assumption 3 is based on the requirement for system repeatability in ILC. In order to simplify the algorithm, without loss of generality, take $x_0 = 0$. It is easy to verify that the result of this chapter is also valid when $x_0 \neq 0$.

In this chapter, the above three assumptions will be followed, but in fact, the SVRG-based ILC can also be established when Assumptions 1 and 2 are appropriately relaxed. Section 4 will give specific explanations on how to relax these assumptions.

Figure 1 illustrates the basic framework of ILC. The plant takes input u_k and generates output y_k and gets the error $e_k = y_d - y_k$ between the output and the desired trajectory y_d, which is transmitted to the controller. The controller uses error e_k and input u_k to calculate the input signal u_{k+1} for the next batch and transmits it to the plant. Our goal is to find a sequence of input $\{u_k\}$, s.t.

$$\lim_{k\to\infty} \|e_k\| = \lim_{k\to\infty} \|y_d - y_k\| = 0, \tag{5}$$

where $\|\cdot\|$ is the vector 2-norm and its induced matrix norm, and henceforth refers to this norm if not otherwise specified.

By Assumptions 2 and 3, (5) is equivalent to the optimization problem of function F:

$$F(u_k) \triangleq \frac{1}{2N}\|e_k\|^2 = \frac{1}{2N}\|y_d - Hu_k\|^2, \lim_{k\to\infty} F(u_k) = 0. \tag{6}$$

2.2 Algorithm Design

The traditional gradient-based ILC updating law (Gu et al., 2019) is

$$u_{k+1} = u_k - \eta_k \nabla_k, \tag{7}$$

where η_k denotes step length, and ∇_k is the gradient of the objective function. From (6), we have

$$\nabla_k = \nabla F(u_k) = -\frac{1}{N} H^T e_k. \tag{8}$$

In (8), calculating the full gradient requires all the information of conjugate matrix H^T and error e_k. To give the gradient under partial information, consider decomposing the error $e_k = [e_k(1), e_k(2), \ldots, e_k(N)]^T$ according to the time index. Let $f_i(u_k) = \frac{1}{2}\|e_k(i)\|^2 = \frac{1}{2}\left(y_d(i) - h_i^T u_k\right)^2$, where $h_i = [h_{i1}, \ldots, h_{ii}, 0, \ldots, 0]^T$ denotes the i-th row of the matrix H. Then, equation (6) can be rewritten as

$$F(u_k) = \frac{1}{N} \sum_{i=1}^{n} f_i(u_k). \tag{9}$$

Take the gradient of both sides, we have

$$\nabla F(u_k) = \frac{1}{N} \sum_{i=1}^{n} \nabla f_i(u_k) = \frac{1}{N} \sum_{i=1}^{n} -h_i e_k(i). \tag{10}$$

Define random gradient $\tilde{\nabla}_k$ as a discrete random variable that takes value uniformly over $\{\nabla f_i(u_k)\}_{i=1}^N$, satisfying $P\left(\tilde{\nabla}_k = \nabla f_i(u_k)\right) = \frac{1}{N}$. Therefore $\mathbb{E}\left[\tilde{\nabla}_k\right] = \nabla_k$, i.e., $\tilde{\nabla}_k$ is unbiased estimation of ∇_k.

Note that by decomposing (6)–(9), calculating the specific value $\nabla f_i(u_k)$ of random vector $\tilde{\nabla}_k$ only requires one row of the system matrix H and one-dimensional information of the error e_k. Therefore the decomposition can effectively reduce the information required for each iteration. This technique of gradient decomposition is the basis for solving the ILC of information incompleteness using SVRG method in this chapter. In Sects. 3 and 4, two specific decomposition methods are presented for incompleteness of error and system information, respectively.

Consider the stochastic gradient descent (SGD) method used in Machine Learning. Replacing the full gradient ∇_k with the stochastic gradient $\tilde{\nabla}_k$ in the ILC updating law (7), we can obtain the SGD-based ILC algorithm. However, the convergence rate of SGD algorithm is $O(1/\sqrt{k})$ even under strongly convex condition (Allen-Zhu, 2018), which cannot meet the practical requirements. This is because although the stochastic gradient $\tilde{\nabla}_k$ is unbiased estimate of the full gradient ∇_k, the variance accumulates as the iteration increases. To reduce the variance, Johnson and Zhang (2013) proposed a general stochastic variance reducted gradient (SVRG) descent method. By recording a "snapshot" $\tilde{u}^s$ every few updates to construct an converging upper bound of the gradient, the rate of convergence of SVRG method is $O\left(\rho^k\right)$ under strongly convex condition and $O(1/k)$ under non-strongly convex

condition. Based on this method, the input updated with "snapshot" $\tilde{u}^s$ is denoted as $u_{s,k}$, and the SVRG-based ILC updating law is

$$u_{s,k+1} = u_{s,k} - \eta \left(\nabla f_i \left(u_{s,k} \right) - \nabla f_i \left(\tilde{u}^s \right) + \nabla F \left(\tilde{u}^s \right) \right). \tag{11}$$

For system (2), the SVRG-based ILC algorithm with updating law (11) is shown in Algorithm 1.

Algorithm 1 SISO SVRG-based ILC framework for SISO systems

Input: $\eta, u_{0,0}$;
$m \leftarrow 2N$; $\tilde{u}^0 \leftarrow u_{0,0}$;
for $s \leftarrow 0$ to $S-1$ **do**
  $u_{s,0} \leftarrow \tilde{u}^s$, $\mu_s \leftarrow \nabla F(\tilde{u}^s)$;
  for $k \leftarrow 0$ to $m-1$ **do**
    $w_k \leftarrow \nabla f_i \left(u_{s,k} \right) - \nabla f_i (\tilde{u}^s) + \mu_s$; where i from $\{1, 2, \ldots, N\}$ randomly
    $u_{s,k+1} \leftarrow u_{s,k} - \eta w_k$;
  end for
  Option I: $\tilde{u}^{s+1} \leftarrow \frac{1}{m} \sum_{k=0}^{m-1} u_{s,k}$;
  Option II: $\tilde{u}^{s+1} \leftarrow u_{s,m}$;
end for

Algorithm 1 has two loops. The outer loop updates the "snapshot" $\tilde{u}^s$ once when the inner loop iterates m times. The iteration length m is taken as an integer multiple of N, which is empirically set to $2N$. Line 9 and 10 of Algorithm 1 shows two ways of updating the "snapshot", Option I and Option II. Option I takes the average of the first $m-1$ inputs as the "snapshot", without using $u_{s,m}$, so actually the inner loop only requires $m-1$ iterations. The corresponding Option II takes the mth-iteration and uses $u_{s,m}$ as the "snapshot". The two "snapshot" updating methods do not change the convergence of Algorithm 1 (Bottou et al., 2018). Due to the limitation of space, we only prove the convergence of Option I under strongly convex conditions and Option II under non-strongly convex conditions in this section and Sect. 4, respectively.

2.3 *Convergence Analysis*

This subsection is divided into two parts, first giving the convex optimization knowledge required for the proof of this chapter and then analyzing the convergence of the system (2) when the "snapshot" update method of Algorithm 1 is set for Option I.

2.3.1 Preliminaries of Convex Optimization

The basics of convex optimization required for this chapter are given below (Lyubashevsky, 2005).

Definition 1 *(Smoothness)* Suppose S is an nonempty convex subset of $\mathbb{R}^d$, $f : S \to \mathbb{R} \in C^1$. If $\exists L > 0$, s.t. $\forall x, y \in S$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

then we say that f is L-smooth or $\nabla f(x)$ is L-Lipschitz continuous on S, where L is the Lipschitz constant.

Definition 2 *(Strong convexity)* Suppose S is an nonempty convex subset of $\mathbb{R}^d$, $f : S \to \mathbb{R} \in C^1$. If $\exists \sigma > 0$, s.t. $\forall x, y \in S$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\sigma}{2}\|x - y\|^2,$$

then we say that f is σ-strongly convex on S. When $\sigma = 0$, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$, f is convex.

Definition 3 *(Conditional number)* If f is L-smooth and σ-strongly convex, $\kappa = L/\sigma$ is the conditional number of f.

Theorem 1 *For convex function f, the followings are equivalent:*

a. $\nabla f(x)$ *is* L*-Lipschitz continuous,*
b. $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$,
c. $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$,
d. $\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$.

Proof $a \to b$: Denote $g(t) = f(t(y - x) + x)$, then $f(x) = g(0)$, $f(y) = g(1)$, and $g'(t) = \langle \nabla f(t(y - x) + x), y - x\rangle$. Therefore,

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &= g(1) - g(0) - \langle \nabla f(x), y - x\rangle \\ &= \int_0^1 g'(t)dt - \langle \nabla f(x), y - x\rangle \\ &= \int_0^1 \langle \nabla f(t(y - x) + x) - \nabla f(x), y - x\rangle dt \\ &\leq \int_0^1 \|\nabla f(t(y - x) + x) - \nabla f(x)\| \cdot \|y - x\| dt \\ &\leq \int_0^1 L\|t(y - x)\| \cdot \|y - x\| dt = \frac{L}{2}\|y - x\|^2. \end{aligned}$$

$b \to c$: Denote $f_x(z) = f(z) - \langle \nabla f(x), z\rangle$, for $\forall z, z' \in \mathbb{R}^d$,

$$f\left(z'\right) - f(z) \leq \left\langle \nabla f(z), z' - z \right\rangle + \frac{L}{2}\left\|z' - z\right\|^2,$$
$$f\left(z'\right) - f(z) - \left\langle \nabla f(x), z' - z \right\rangle \leq \left\langle \nabla f(z) - \nabla f(x), z' - z \right\rangle + \frac{L}{2}\left\|z' - z\right\|^2,$$
$$f_x\left(z'\right) - f_x(z) \leq \left\langle \nabla f_x(z), z' - z \right\rangle + \frac{L}{2}\left\|z' - z\right\|^2. \tag{12}$$

By using the convexity of f, we have

$$\begin{aligned} f_x\left(z'\right) - f_x(z) &= f\left(z'\right) - f(z) - \left\langle \nabla f(x), z' - z \right\rangle \\ &\geq \left\langle \nabla f(z), z - z' \right\rangle - \left\langle \nabla f(x), z - z' \right\rangle = \left\langle \nabla f_x(z), z - z' \right\rangle. \end{aligned}$$

Therefore $f_x(z)$ is also convex, since $\nabla f_x(z) = \nabla f(z) - \nabla f(x)$, $f_x(z)$ achieves its minimum at $z = x$. By (12),

$$\begin{aligned} f_x(x) = \min_{z'} f_x\left(z'\right) &\leq \min_{z'}\left\{ f_x(z) + \left\langle \nabla f_x(z), z' - z \right\rangle + \frac{L}{2}\left\|z' - z\right\|^2 \right\} \\ &= f_x(z) + \min_{\|y\|=1} \min_{t \geq 0}\left\{ t\left\langle \nabla f_x(z), y \right\rangle + \frac{L}{2}t^2 \right\} \\ &= f_x(z) + \min_{\|y\|=1}\left\{ -\frac{\left\langle \nabla f_x(z), y \right\rangle^2}{2L} \right\} \\ &= f_x(z) - \frac{1}{2L}\left\|\nabla f_x(z)\right\|^2. \end{aligned}$$

Therefore $f_x(z) - f_x(x) \geq \frac{1}{2L}\left\|\nabla f_x(z)\right\|^2$, which implies

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L}\left\|\nabla f_x(y)\right\|^2 = \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

$c \to d$: Swapping x and y in $c.$, we have

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Summing the two equations, we have

$$\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

$d \to a$: $\|\nabla f(y) - \nabla f(x)\|^2 \leq L\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|\nabla f(x) - \nabla f(y)\| \cdot \|x - y\|$ by Cauchy inequality, thus $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. □

Theorem 2 *Let $f(x) = \frac{1}{2}x^T Qx + q^T x + c$, where Q is positive definite. Then $f(x)$ is L-smooth and σ-strongly convex, where $L = \lambda_M$, and $\sigma = \lambda_m$. λ_M, and λ_m are the maximum and minimum eigenvalues of Q, respectively.*

Proof Since $\nabla f(x) = Qx + q$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq \|Q(x-y)\| \leq \|Q\| \cdot \|x-y\| = \lambda_M \|x-y\|.$$

Hence $f(x)$ is λ_M-smooth. It is easy to verify that

$$f(x) - f(y) - \langle \nabla f(y), x-y \rangle = \frac{1}{2}(x-y)^T Q(x-y).$$

Since Q is positive definite, the orthogonal similarity can be diagonalized as $Q = P^T \Lambda P$, where P is the orthogonal matrix, Λ is the diagonal matrix of eigenvalues, and $\lambda_m > 0$. Thus

$$\frac{1}{2}(x-y)^T Q(x-y) = \frac{1}{2}z^T \Lambda z = \frac{1}{2}\sum_i \lambda_i z_i^2 \geq \frac{1}{2}\lambda_m \|z\|^2 = \frac{1}{2}\lambda_m \|x-y\|^2.$$

Thus $f(x)$ is λ_m-strongly convex. □

For system (2) and objective function (6), we have:

Proposition 1 *Each f_i is convex and L-smooth.*

Proof For $f(x) = \frac{1}{2}\left(q^T x + c\right)^2$, where $q = [q_1, q_2, \ldots, q_n]^T \in \mathbb{R}^n, x \in \mathbb{R}^n, c \in \mathbb{R}$. Obviously, f is convex, and $\nabla f(x) = qq^T x + cq$, $\|\nabla f(x) - \nabla f(y)\| = \left\|qq^T(x-y)\right\| \leq \left\|qq^T\right\| \cdot \|x-y\| \leq \|q\|^2 \cdot \|x-y\|$. Therefore, for each f_i in (9), let $L = \max_i\{\|h_i\|^2\} > 0$, $h_i = [h_{i1}, \ldots, h_{ii}, 0, \ldots, 0]^T$, then f_i is L-smooth. □

Proposition 2 *F is L-smooth and σ-strongly convex.*

Proof By (6), we have

$$\begin{aligned} F(u_k) =& \frac{1}{2N}(y_d - Hu_k)^T (y_d - Hu_k) \\ =& \frac{1}{2N}\left(u_k^T H^T H u_k - y_d^T H u_k - u_k^T H^T y_d + y_d^T y_d\right), \end{aligned}$$

where $H^T H$ is positive definite. Then by Theorem 2, $F(u_k)$ is L-smooth and σ-strongly convex, where L and σ are the $\frac{1}{2N}$ of the maximal and minimal eigenvalues of $H^T H$, respectively. □

2.3.2 Proof of Convergence

In Algorithm 1, set the "snapshot" updating as Option I. Then, under the assumptions of system (2), the convergence of Algorithm 1 is given by the following theorem.

Theorem 3 *If each f_i is convex and L-smooth, and F is σ-strongly convex. We denote the optimal point $u^* = \operatorname{argmin}_u F(u)$, and assume that m is large enough such that*

$$\alpha = \frac{1}{\sigma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1.$$

Then the convergence of Algorithm 1 satisfies

$$\mathbb{E}\left[F\left(\tilde{u}^S\right) - F\left(u^*\right)\right] \leq \alpha^S \left(F\left(\tilde{u}^0\right) - F\left(u^*\right)\right).$$

Proof Since f_i is convex and L-smooth, for any i, by Theorem 1,

$$\left\|\nabla f_i(u) - \nabla f_i\left(u^*\right)\right\|^2 \leq 2L\left[f_i(u) - f_i\left(u^*\right) - \left\langle \nabla f_i\left(u^*\right), u - u^*\right\rangle\right]. \tag{13}$$

Since $\frac{1}{N}\sum_{i=1}^n \nabla f_i(u) = \nabla F(u)$, and $\nabla F\left(u^*\right) = 0$, we regard ∇f_i as random vectors which take values from $\{\nabla f_i\}_{i=1}^N$, then

$$\mathbb{E}\left[\left\|\nabla f_i(u) - \nabla f_i\left(u^*\right)\right\|^2\right] = \frac{1}{N}\sum_{i=1}^n \left\|\nabla f_i(u) - \nabla f_i\left(u^*\right)\right\|^2 \leq 2L\left[F(u) - F\left(u^*\right)\right]. \tag{14}$$

For any fixed s, we set $w_k = \nabla f_i\left(u_{s,k}\right) - \nabla f_i\left(\tilde{u}^s\right) + \nabla F\left(\tilde{u}^s\right)$, then

$$\begin{aligned}
\mathbb{E}\left[\|w_k\|^2\right] &\leq 2\mathbb{E}\left[\left\|\nabla f_i\left(u_{s,k}\right) - \nabla f_i\left(u^*\right)\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\nabla f_i\left(\tilde{u}^s\right) - \nabla f_i\left(u^*\right) - \nabla F\left(\tilde{u}^s\right)\right\|^2\right] \\
&= 2\mathbb{E}\left[\left\|\nabla f_i\left(u_{s,k}\right) - \nabla f_i\left(u^*\right)\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\left[\nabla f_i\left(\tilde{u}^s\right) - \nabla f_i\left(u^*\right)\right] - \mathbb{E}\left[\nabla f_i\left(\tilde{u}^s\right) - \nabla f_i\left(u^*\right)\right]\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla f_i\left(u_{s,k}\right) - \nabla f_i\left(u^*\right)\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla f_i\left(\tilde{u}^s\right) - \nabla f_i\left(u^*\right)\right\|^2\right] \\
&\leq 4L\left[F\left(u_{s,k}\right) - F\left(u^*\right) + F\left(\tilde{u}^s\right) - F\left(u^*\right)\right].
\end{aligned} \tag{15}$$

In the above, we have used the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and the property $\mathbb{E}\left[\|\zeta - \mathbb{E}\zeta\|^2\right] = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}\zeta\|^2 \leq \mathbb{E}\|\zeta\|^2$, as well as (14). Notice that $\mathbb{E}\left[w_k\right] = \nabla F\left(u_{s,k}\right)$, thus

$$
\begin{aligned}
&\mathbb{E}\left[\left\|u_{s,k+1}-u^*\right\|^2\right] \\
&=\left\|u_{s,k}-u^*\right\|^2-2\eta\mathbb{E}\left[\left\langle w_k,u_{s,k}-u^*\right\rangle\right]+\eta^2\mathbb{E}\left[\left\|w_k\right\|^2\right] \\
&\leq\left\|u_{s,k}-u^*\right\|^2-2\eta\left[\left\langle\nabla F\left(u_{s,k}\right),u_{s,k}-u^*\right\rangle\right] \\
&\quad+4L\eta^2\left[F\left(u_{s,k}\right)-F\left(u^*\right)+F\left(\tilde{u}^s\right)-F\left(u^*\right)\right] \\
&\leq\left\|u_{s,k}-u^*\right\|^2-2\eta\left[F\left(u_{s,k}\right)-F\left(u^*\right)\right] \\
&\quad+4L\eta^2\left[F\left(u_{s,k}\right)-F\left(u^*\right)+F\left(\tilde{u}^s\right)-F\left(u^*\right)\right] \\
&=\left\|u_{s,k}-u^*\right\|^2-2\eta(1-2L\eta)\left[F\left(u_{s,k}\right)-F\left(u^*\right)\right] \\
&\quad+4L\eta^2\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right].
\end{aligned}
$$

In the above, we have used (15) and the convexity of F, i.e., $\left\langle\nabla F\left(u_{s,k}\right),u_{s,k}-u^*\right\rangle\geq F\left(u_{s,k}\right)-F\left(u^*\right)$.

We sum up the expectations of the above equation for $k=0,1,\ldots,m-1$. Using the convexity of F and the selection of $\tilde{u}^{s+1}$ under Option I, we have $F\left(\tilde{u}^{s+1}\right)=F\left(\frac{1}{m}\sum_{k=0}^{m-1}u_{s,k}\right)\leq\frac{1}{m}\sum_{k=0}^{m-1}F\left(u_{s,k}\right)$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|u_{s,m}-u^*\right\|^2\right]+2\eta(1-2L\eta)m\mathbb{E}\left[F\left(\tilde{u}^{s+1}\right)-F\left(u^*\right)\right] \\
&\leq\mathbb{E}\left[\left\|u_{s,0}-u^*\right\|^2\right]+4L\eta^2m\mathbb{E}\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right] \\
&\leq\frac{2}{\sigma}\mathbb{E}\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right]+4L\eta^2m\mathbb{E}\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right].
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
\mathbb{E}\left[F\left(\tilde{u}^{s+1}\right)-F\left(u^*\right)\right]&\leq\left(\frac{1}{\sigma\eta(1-2L\eta)m}-\frac{2L\eta}{1-2L\eta}\right)\mathbb{E}\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right] \\
&=\alpha\mathbb{E}\left[F\left(\tilde{u}^s\right)-F\left(u^*\right)\right].
\end{aligned}
$$

Summing up the above equation for $s=0,1,\ldots,S-1$, we have

$$
\mathbb{E}\left[F\left(\tilde{u}^S\right)-F\left(u^*\right)\right]\leq\alpha^S\mathbb{E}\left[F\left(\tilde{u}^0\right)-F\left(u^*\right)\right]. \qquad \square
$$

Remark 3 Theorem 3 indicates that Algorithm 1 has the rate of convergence $O\left(\alpha^S\right)$. And this convergence rate is related to the value of α. If the information of the system H is known, to make α as small as possible, we generally take $\eta=\frac{0.1}{L}$, $m=\Theta(n)$, so that the value of α is close to $\frac{1}{2}$. If the system information is unknown, we need to find the appropriate η and m by experiment.

3 SVRG-Based ILC Under Random Data Dropouts

This section follows the SISO system in Sect. 2, but assumes that random data dropouts occur in the error signal transmission. This section consists of four parts: system description, algorithm design, performance analysis, and numerical simulation.

3.1 System Description

In SISO system (2), we still hold Assumptions 1–3, but assuming that data dropouts occur in the transmission of the error signal, as shown in Fig. 2. We further assume that the dropouts satisfy the Bernoulli distribution model (Shen, 2018). Therefore, the ILC updating law (7) becomes

$$u_{k+1} = u_k + \eta \frac{1}{N} H^T \Gamma_k e_k, \tag{16}$$

where $\Gamma_k = \text{diag}\{\gamma_k(1), \gamma_k(2), \ldots, \gamma_k(N)\}$. $\{\gamma_k(i)\}_{i=1}^N$ is i.i.d following Bernoulli distribution. Let $\gamma \triangleq \mathbb{E}\left[\gamma_k(i)\right]$ be the successful transmission rate, where $\gamma_k(i) = 0$ means data dropout occurs in the i-th time of k-th batch, and otherwise, data dropout does not occur.

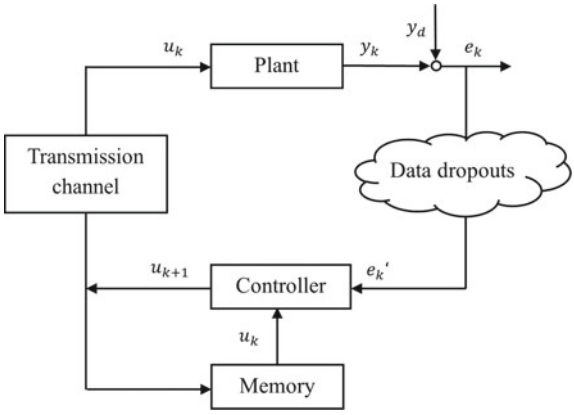

Fig. 2 ILC with random data dropouts

3.2 Algorithm Design

Based on gradient descent method, there are two main approaches to solve the data dropouts problem:

(1) Obtain the full gradient by retransmissing. For each transmission, the controller stores the successfully transmitted data and asks the lost data to be retransmitted until all data are received. This method eliminates the effect of data dropouts by retransmissing.
(2) Use successfully transmitted data to construct random gradient. The data is updated directly using successfully transmitted data each iteration, as shown in the update law (16).

The first method requires a lot of wasted time when data retransmission is slow. Although the second method saves the time of data retransmission, the actual running time may be larger than first method when data retransmission.

Based on the framework of Algorithm 1, the SVRG-based ILC under error data dropouts can be constructed by utilizing the second method for each iteration, but calculating the full gradient every several iterations using the first method. This algorithm does not require data retransmission in most cases compared to the first method and has a significant improvement in convergence speed compared to the second method. Thus it can achieve a good balance between convergence rate and data retransmission speed, and it is more suitable for general data dropout cases.

The formal construction of the algorithm is given below.

Firstly, we take the random gradient $\tilde{\nabla}_k \triangleq -\frac{1}{\gamma N} H^T \Gamma_k e_k$, and we note that $\mathbb{E}\left[\tilde{\nabla}_k\right] = -\frac{1}{N} H^T e_k = \nabla_k$. For the convenience of proof, we present $\tilde{\nabla}_k$ as

$$\nabla \tilde{F}_k (u_k) = \frac{1}{\gamma N} \sum_{\{i|\gamma_k(i)=1\}} \nabla f_i (u_k), \tag{17}$$

where ∇f_i is defined in the same way as (10). Notice that if there is no dropout, $\gamma = 1$, and therefore (17) is equivalent to (10).

Remark 4 Equation (17) is similar to the Batch Gradient Descent (BGD) method in Machine Learning, but they are fundamentally different. In (17), the number of ∇f_i in each summation $\sum_i \nabla f_i$ varies according to the value of the random vector $\{\gamma_k(i)\}_{i=1}^N$. But in BGD, the number of ∇f_i is fixed. Therefore, the algorithm based on gradient $\nabla \tilde{F}_k$ cannot be directly applied to BGD.

Secondly, similar to (11), ILC updating law under random data dropouts is constructed:

$$u_{s,k+1} = u_{s,k} - \eta \left(\nabla \tilde{F}_k \left(u_{s,k}\right) - \nabla \tilde{F}_k \left(\tilde{u}^s\right) + \nabla F \left(\tilde{u}^s\right) \right). \tag{18}$$

Finally, change the iteration length m in Algorithm 1 from $2N$ to $\lceil 2\gamma N \rceil$. Because the number of summation in each batch is $\mathbb{E}\left[\sum_{i=1}^{N} \gamma_k(i)\right] = \gamma N$ in the desired sense, $\nabla \tilde{F}_k$ is equivalent to γN sum of ∇f_i.

In conclusion, SVRG-based ILC under random data dropouts is shown in Algorithm 2.

Algorithm 2 Data dropout SISO SVRG-based ILC

Input: $\eta, u_{0,0}$;
$m \leftarrow 2\gamma N$; $\tilde{u}^0 \leftarrow u_{0,0}$;
for $s \leftarrow 0$ to $S-1$ **do**
  $u_{s,0} \leftarrow \tilde{u}^s$, $\mu_s \leftarrow \nabla F(\tilde{u}^s)$;
  for $k \leftarrow 0$ to $m-1$ **do**
    $w_k \leftarrow \nabla f_i\left(u_{s,k}\right) - \nabla f_i(\tilde{u}^s) + \mu_s$;
    $u_{s,k+1} \leftarrow u_{s,k} - \eta w_k$;
  end for
  $\tilde{u}^{s+1} \leftarrow \frac{1}{m}\sum_{k=0}^{m-1} u_{s,k}$;
end for

For the "snapshot" of Algorithm 2, the update method is taken as Option I in Algorithm 1, and the recommended iteration length is set to $\lceil 2\gamma N \rceil$. When γ is unknown, we need to find the appropriate m by experiments.

3.3 Convergence Analysis

By Proposition 1, every f_i is convex and L-smooth, and the following proposition holds:

Proposition 3 *Each value of $\nabla \tilde{F}_k(u_k)$ is convex and L'-smooth, where $L' = L/\gamma$, and L is the Lipschitz constant corresponding to the smoothness of f_i in Proposition 1.*

Proof Since every f_i is convex and L-smooth, $\frac{1}{\gamma N}\sum_{i=1}^{N} \nabla f_i(u_k)$ is L'-Lipschitz continuous, where $L' = L/\gamma$. Because each value of $\nabla \tilde{F}_k(u_k)$ is a linear combination of ∇f_i, the summation number does not exceed $\frac{1}{\gamma N}\sum_{i=1}^{N} \nabla f_i(u_k)$. Therefore $\nabla f_i(u_k)$; as a result, $\nabla \tilde{F}_k(u_k)$ is convex and L'-smooth. □

For the convergence of Algorithm 2, we have the following theorem.

Theorem 4 *If each value of $\nabla \tilde{F}_k(u_k)$ is convex and L'-smooth, F is L-smooth and σ-strongly convex, and for the optimal point $u^* = \operatorname{argmin}_u F(u)$, assuming that m is large enough s.t.*

$$\alpha = \frac{1}{\sigma\eta(1-2L'\eta)m} + \frac{2L'\eta}{1-2L'\eta} < 1.$$

Then the convergence of Algorithm 2 satisfies

$$\mathbb{E}\left[F\left(\tilde{u}^S\right) - F\left(u^*\right)\right] \leq \alpha^S \left(F\left(\tilde{u}^0\right) - F\left(u^*\right)\right).$$

Proof By Proposition 3, replacing f_i in the proof of Theorem 3 with $\nabla \tilde{F}_k$, equation (13) is rewritten as:

$$\left\|\nabla \tilde{F}_k(u) - \nabla \tilde{F}_k\left(u^*\right)\right\|^2 \leq \frac{1}{\gamma N} \sum_{\{i|\gamma_k(i)=1\}} 2L'\left[f_i(u) - f_i\left(u^*\right) - \left\langle \nabla f_i\left(u^*\right), u - u^*\right\rangle\right].$$

The corresponding Eq. (14) is

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla \tilde{F}_k(u) - \nabla \tilde{F}_k\left(u^*\right)\right\|^2\right] \\
&\leq \frac{1}{\gamma N}\mathbb{E}\left[\sum_{\{i|\gamma_k(i)=1\}} 2L'\left[f_i(u) - f_i\left(u^*\right) - \left\langle \nabla f_i\left(u^*\right), u - u^*\right\rangle\right]\right] \\
&= \frac{1}{\gamma N}\gamma \mathbb{E}\left[2L'\left[F(u) - F\left(u^*\right) - \left\langle \nabla F\left(u^*\right), u - u^*\right\rangle\right]\right] \\
&= 2L'\mathbb{E}\left[F(u) - F\left(u^*\right)\right].
\end{aligned}$$

The rest of the proof repeats the proof of Theorem 3. □

Remark 5 Theorem 4 shows that Algorithm 2 also converges linearly, and the speed of convergence is related to α. For the choice of m, note that Theorem 4 differs from Theorem 3 in the Lipschitz constant corresponding to the smoothness of the condition. By Proposition 3, with $L' = L/\gamma$, for

$$\alpha = \frac{1}{\sigma\eta\left(1 - 2L'\eta\right)m} + \frac{2L'\eta}{1 - 2L'\eta}.$$

We can consider multipling m by γ times, i.e., changing m from $2N$ to $\lceil 2\gamma N \rceil$, to approximately keep the convergence of Algorithm 2.

3.4 Numerical Simulation

In SISO system (1), take the system matrix (A, B, C) as

$$A = \begin{bmatrix} 0.50 & -0.25 & 1.00 \\ 0.15 & 0.30 & -0.50 \\ -0.75 & 0.25 & -0.25 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0\ 0\ 1.0 \end{bmatrix}.$$

Take desired trajectory $y_d(t) = \sin(2\pi t/50)$, time length $N = 50$, initial state $x_0 = 0$, and initial input $u_0 = 0$. When $\gamma = 0.9$ and $\gamma = 0.6$, the ILC based on full gradient (GD), stochastic gradient (SGD) and stochastic variance reduced gradient (SVRG) is shown in Fig. 3.

When calculating the full gradient, each data retransmission increases 1 to the iteration number, which means that the controller skips one round of computation until all error information is completely transmitted. Take the optimal step that the three methods can converge.

When $\gamma = 0.9$, Fig. 3a shows that the SVRG-based ILC converges slightly faster than the GD- and SGD-based ILC. When $\gamma = 0.6$, Fig. 3b illustrates a significant difference in the convergence speed of the three types ILC, from fast to slow for SVRG-, SGD-, and GD-based ILC. In summary, the SVRG-based ILC under error data dropouts, i.e., Algorithm 2, outperforms the GD- and SGD-based ILC under different successful transmission rates, and the difference becomes more significant as the γ decreases.

4 Model-Free SVRG-Based ILC for MIMO Systems

This section extends the Algorithm 1 in Sect. 2 from SISO systems to MIMO systems. Firstly, a system description of the discrete linear MIMO system is given. Secondly, the existing model-free data-driven methods are introduced, and a new model-free data-driven ILC based on SVRG method is constructed. Thirdly, the convergence of the algorithm under non-strongly convex conditions is proved. Finally, numerical simulations are established to verify the convergence performance of SVRG-based ILC in deterministic and noisy systems.

4.1 System Description

Consider the following discrete linear multi-input multi-ouput (MIMO) system $\mathscr{J}$, which has q inputs $u_k^1, u_k^2, \ldots, u_k^q$, and p outputs $y_k^1, y_k^2, \ldots, y_k^p$. Rewrite the system in the form of (2),

$$\boldsymbol{y}_k = \mathscr{J}\boldsymbol{u}_k, \tag{19}$$

where

$$\mathscr{J} = \begin{bmatrix} J_{11} & J_{12} & \cdots & J_{1q} \\ J_{21} & J_{22} & \cdots & J_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ J_{p1} & J_{p2} & \cdots & J_{pq} \end{bmatrix}, \boldsymbol{u}_k = \begin{bmatrix} u_k^1 \\ u_k^2 \\ \vdots \\ u_k^q \end{bmatrix}, \boldsymbol{y}_k = \begin{bmatrix} y_k^1 \\ y_k^2 \\ \vdots \\ y_k^p \end{bmatrix}.$$

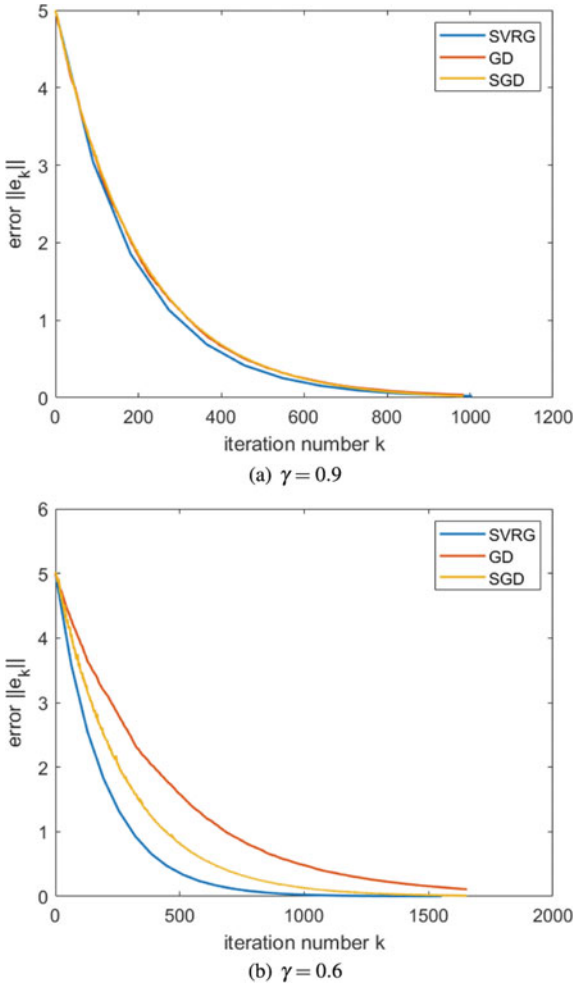

Fig. 3 Comparison of three gradient-based ILC under error data dropouts

Each $J_{ij} \in \mathbb{R}^{N\times N}$ has the same properties as matrix H in (2). $y_k^i = \left[y_k^i(1), \ldots, y_k^i(N)\right]^T$, $u_k^j = \left[u_k^j(0), \ldots, u_k^j(N-1)\right]^T$, N is the length of time, and the desired trajectory is

$$\boldsymbol{y}_d = \left[\left(y_d^1\right)^T, \left(y_d^2\right)^T, \ldots, \left(y_d^p\right)^T\right]^T.$$

For this system, consider the following assumptions:

Assumption 4 System matrix $\mathscr{J} \neq 0$.

Assumption 5 The dimension of input signal does not exceed the dimension of output signal, i.e., $p \geq q$.

Remark 6 If $p = q = 1$, the system (19) degenerates to SISO system (2). Unlike Assumptions 1 and 4 can no longer guarantee that the system matrix $\mathscr{J}$ is of full rank. Assumption 4 is the most fundamental, since if $\mathscr{J} = 0$, any input signal cannot track y_d. Assumption 2 is also relaxed from the system, the reasons will be given in the proof of the convergence. In addition, Assumption 5 is added for the MIMO system, because if the input dimension is larger than the output dimension, it means that there is a redundant information.

For desired trajectory $\boldsymbol{y}_d$, consider control objective similar to (6), i.e., to find a sequence $\{\boldsymbol{u}_k\}$, s.t.

$$G(\boldsymbol{u}_k) = \frac{1}{2p}\|\boldsymbol{e}_k\|^2 = \frac{1}{2p}\left\|\boldsymbol{y}_d - \mathscr{J}\boldsymbol{u}_k\right\|^2, \lim_{k\to\infty} G(\boldsymbol{u}_k) = G\left(\boldsymbol{u}^*\right), \tag{20}$$

where $\boldsymbol{u}^*$ is the input when G takes the minimum value, and the error signal $\boldsymbol{e}_k$ is

$$\boldsymbol{e}_k = \boldsymbol{y}_d - \boldsymbol{y}_k = \begin{bmatrix} y_k^1 - y_d^1 \\ y_k^2 - y_d^2 \\ \vdots \\ y_k^p - y_d^p \end{bmatrix} = \begin{bmatrix} e_k^1 \\ e_k^2 \\ \vdots \\ e_k^p \end{bmatrix}.$$

4.2 Algorithm Design

The full gradient of G in (20) is $\nabla_k = \nabla G(\boldsymbol{u}_k) = -\frac{1}{p}\mathscr{J}^T\left(\boldsymbol{y}_d - \mathscr{J}\boldsymbol{u}_k\right)$. We need the information of $\mathscr{J}^T$ to calculate the full gradient. However, in model-free learning, we want to obtain the gradient by conducting experiments on system $\mathscr{J}$ only. For this purpose, Oomen et al. (2014) gives the following method to estimate $\mathscr{J}^T$.

Lemma 1 *For SISO system $\mathscr{J} = J_{11}$, its transpose $\mathscr{J}^T$ can be obtained by matrix multiplication*

$$(J_{11})^T = \mathcal{T}_N J_{11} \mathcal{T}_N,$$

where $\mathcal{T}$ is the N-order permutation matrix whose anti-diagonal is 1, i.e.,

$$\mathcal{T}_N = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix}.$$

Therefore the full gradient of SISO system $-\frac{1}{p}(J_{11})^T \boldsymbol{e}_k = -\frac{1}{p}\mathcal{T}_N J_{11}\mathcal{T}_N \boldsymbol{e}_k$ can be obtained by a single experiment.

Lemma 2 *For MIMO system $\mathscr{J}$, whose transpose $\mathscr{J}^T$ is*

$$\mathscr{J}^T = \begin{bmatrix} (J_{11})^T & \cdots & (J_{p1})^T \\ \vdots & \ddots & \vdots \\ (J_{1q})^T & \cdots & (J_{pq})^T \end{bmatrix} = \underbrace{\begin{bmatrix} \mathcal{T}_N & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathcal{T}_N \end{bmatrix}}_{\mathcal{T}^{qN}} \underbrace{\begin{bmatrix} J_{11} & \cdots & J_{p1} \\ \vdots & \ddots & \vdots \\ J_{1q} & \cdots & J_{pq} \end{bmatrix}}_{\tilde{\mathscr{J}}} \underbrace{\begin{bmatrix} \mathcal{T}_N & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathcal{T}_N \end{bmatrix}}_{\mathcal{T}^{pN}}.$$

For symmetric MIMO systems, $-\tilde{\mathscr{J}} \neq \mathcal{J}$, so the full gradient $-\frac{1}{p}\mathscr{J}^T \boldsymbol{e}_k$ of MIMO system cannot be obtained from a single experiment on system $\mathscr{J}$. The method proposed by Oomen et al. (2014) estimates $\mathscr{J}^T$ from pq experiments:

$$\mathscr{J}^T = \mathscr{J}^{qN} \left(\sum_{i=1}^{q} \sum_{j=1}^{p} \mathcal{L}^{ij} \mathscr{J} \mathcal{L}^{ij} \right) \mathscr{J}^{pN}, \tag{21}$$

where $\mathcal{L}^{ij}$ is a matrix consisting of $q \times p$ blocks. In $\mathcal{L}^{ij}$, the (i, j) block is unit matrix of order N, and the remaining blocks are all 0:

$$\mathcal{L}^{ij} = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & I_N & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{qN \times pN}.$$

In (21), the left multiplication matrix $\mathcal{L}^{ij}$ takes the i-th row of $\mathscr{J}$, and the right multiplication matrix $\mathcal{L}^{ij}$ takes out the j-th column of $\mathscr{J}$. The two multiplications lead to a great loss of system information. We would like to improve the above method by extracting as much system information as possible. Therefore, consider the following decomposition as (9).

Set $g_i(\boldsymbol{u}_k) = \frac{1}{2}\left\|e_k^i\right\|^2 = \frac{1}{2}\left\|y_d^i - \sum_{j=1}^{q} J_{ij} u_k^j\right\|^2$, then (20) can be written as

$$G(\boldsymbol{u}_k) = \frac{1}{p} \sum_{i=1}^{n} g_i(\boldsymbol{u}_k). \tag{22}$$

Taking gradient on the both sides, we have

$$\nabla G(\boldsymbol{u}_k) = \frac{1}{p} \sum_{i=1}^{n} \nabla g_i(\boldsymbol{u}_k), \tag{23}$$

where

$$\nabla g_i\left(\boldsymbol{u}_k\right) = -\begin{bmatrix} J_{i1}^T e_k^i \\ J_{i2}^T e_k^i \\ \vdots \\ J_{iq}^T e_k^i \end{bmatrix} \in \mathbb{R}^{qN}.$$

Note that the $\nabla g_i\left(\boldsymbol{u}_k\right)$ can be calculate by one line of the system matrix. The following lemma can help us design a controller to take out this information.

Lemma 3 *Calculating $\nabla g_i\left(\boldsymbol{u}_k\right)$ only needs a single experiment.*

Proof First, we note that

$$\underbrace{[0,\ldots,0,\mathscr{J}_N,0,\ldots,0]}_{\mathcal{T}_i^{qN}} \underbrace{\begin{bmatrix} J_{11} & J_{12} & \cdots & J_{1q} \\ J_{21} & J_{22} & \cdots & J_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ J_{p1} & J_{p2} & \cdots & J_{pq} \end{bmatrix}}_{\mathscr{J}} \underbrace{\begin{bmatrix} \mathcal{T}_N & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathcal{T}_N \end{bmatrix}}_{\mathcal{T}^{qN}} = \left[J_{i1}^T,\ldots,J_{iq}^T\right] \in \mathbb{R}^{N\times qN},$$

where $\mathcal{T}_i^{pN} \in \mathbb{R}^{N\times pN}$ is the matrix whose i-th block is $\mathcal{T}_N$ and the rest blocks are 0.

For $e_k^i \in \mathbb{R}^N$,

$$e_k^i = \underbrace{[0,\ldots,0,I_N,0,\ldots,0]}_{\mathscr{L}_i} \boldsymbol{e}_k,$$

where $\mathscr{L}_i \in \mathbb{R}^{N\times qN}$ is the matrix whose i-th block is an identity matrix of order N and the rest blocks are 0.

The matrix multiplication method can retrieve a row of information of the system, but it cannot directly obtain the matrix for further computation. Therefore, simple and easy-to-implement linear mappings are considered to change the matrix to suitable dimension.

Set $E_k^i \in \mathbb{R}^{qN\times q}$ and define a linear mapping Ψ:

$$\Psi e_k^i = E_k^i = \begin{bmatrix} e_k^i & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e_k^i \end{bmatrix},$$

where E_k^i is the matrix blocked by $N \times 1$, with e_k^i on the diagonal and 0 in the rest of the blocks.

Since

$$\mathcal{T}_i^{pN}\mathcal{J}\mathcal{T}^{qN}E_k^i = \left[J_{i1}^T e_k^i,\ldots,J_{iq}^T e_k^i\right] \in \mathbb{R}^{N\times q},$$

we can define the linear map Φ: $\mathbb{R}^{N\times q} \to \mathbb{R}^{qN}$. It maps matrix in $\mathbb{R}^{N\times q}$ to $\mathbb{R}^{qN}$ by arranging each column of the matrix in order to a vector, i.e.,

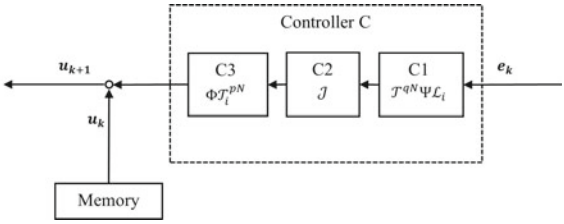

Fig. 4 Controller for model-free MIMO systems

$$\Phi\left[J_{i1}^T e_k^i, \ldots, J_{iq}^T e_k^i\right] = \begin{bmatrix} J_{i1}^T e_k^i \\ J_{i2}^T e_k^i \\ \vdots \\ J_{iq}^T e_k^i \end{bmatrix}.$$

Combining above, we have

$$\nabla g_i\left(\boldsymbol{u}_k\right) = -\Phi \mathscr{J}_i^{pN} \mathscr{J} \mathcal{T}^{qN} L_i \boldsymbol{e}_k.$$

Thus, $\nabla g_i\left(\boldsymbol{u}_k\right)$ can be calculated in a single experiment. □

Based on Lemma 3, a controller can be designed as shown in Fig. 4.

This controller can reduce the calculation of the full gradient in (21) from pq experiments to p experiments. However, when the system is noisy, there is no guarantee that the partial gradient estimated for each experiment $\nabla g_i\left(\boldsymbol{u}_k\right)$ all correspond to the same full gradient $\nabla G\left(\boldsymbol{u}_k\right)$. In nosiy systems, we can use the random gradient $\tilde{\nabla}_k$, which takes values uniformly $\{\nabla g_i\left(\boldsymbol{u}_k\right)\}_{i=1}^N$ s.t. $P\left(\tilde{\nabla}_k = \nabla g_i\left(\boldsymbol{u}_k\right)\right) = \frac{1}{N}$. However, the SGD method converges slowly. Combining the convergence speed and the effect of noise in the system, we consider to design a SVRG-based ILC algorithm similar to Algorithm 1, as shown in Algorithm 3.

Algorithm 3 Data-driven MIMO SVRG-based ILC

Input: $\eta, \boldsymbol{u}_{\boldsymbol{0,0}}$;
$m \leftarrow 2p$; $\tilde{\boldsymbol{u}}^{\boldsymbol{0}} \leftarrow \boldsymbol{u}_{\boldsymbol{0,0}}$;
for $s \leftarrow 0$ to $S-1$ **do**
  $\boldsymbol{u}_{\boldsymbol{s,0}} \leftarrow \tilde{\boldsymbol{u}}^{\boldsymbol{s}}$, $\boldsymbol{\mu}_s \leftarrow \nabla G\left(\tilde{\boldsymbol{u}}^{\boldsymbol{s}}\right)$;
  for $k \leftarrow 0$ to $m-1$ **do**
    $\boldsymbol{w_k} \leftarrow \nabla g_i\left(\boldsymbol{u}_{\boldsymbol{s,k}} - \tilde{\boldsymbol{u}}^{\boldsymbol{s}}\right) + \boldsymbol{\mu}_s$; where i from $\{1, 2, \ldots, p\}$ randomly
    $\boldsymbol{u}_{\boldsymbol{s,k+1}} \leftarrow \boldsymbol{u}_{\boldsymbol{s,k}} - \eta \boldsymbol{w_k}$;
  end for
  $\tilde{\boldsymbol{u}}^{\boldsymbol{s+1}} \leftarrow \boldsymbol{u}_{\boldsymbol{s,m}}$;
end for

Algorithm 3 uses Option II in Algorithm 1 to update the "snapshot". If m is twice as many as p, a total of $2p$ experiments are required for each internal iteration.

p experiments are required to compute the full gradient, so a total of $3p$ system experiments are required for each iteration. Compared with the SGD-based ILC, Algorithm 3 requires p more systematic experiments per m iterations to compute the full gradient in order to accelerate the convergence.

4.3 Convergence Analysis

From the following two propositions, we will see that G is not necessarily strongly convex.

Proposition 4 *G and g_i are convex and L-smooth.*

Proof We note that

$$\nabla g_i(\boldsymbol{x}) = -\begin{bmatrix} J_{i1}^T \\ J_{i2}^T \\ \vdots \\ J_{iq}^T \end{bmatrix} \left(y_d^i - \sum_{j=1}^{q} J_{ij} x^j \right) = -J_i^T \left(y_d^i - J_i \boldsymbol{x} \right),$$

where $J_i = \left[J_{i1} J_{i2} \dots J_{iq} \right], \boldsymbol{x} \in \mathbb{R}^{qN}. g_i$ is convex, and

$$\begin{aligned} \|\nabla g_i(\boldsymbol{x}) - \nabla g_i(\boldsymbol{y})\| &= \left\| J_i J_i^T (\boldsymbol{x} - \boldsymbol{y}) \right\| \\ &\le \left\| J_i J_i^T \right\| \cdot \|x - y\| \le \sum_{j=1}^{q} \left\| J_{ij} \right\|^2 \cdot \|x - y\|. \end{aligned}$$

Let $L = \max_i \left\{ \sum_{j=1}^{q} \lambda_{ij} \right\}$, where λ_{ij} is the maximum eigenvalue of $J_{ij}^T J_{ij}$. Since $J_{ij}^T J_{ij}$ is always semipositive definite, $\lambda_{ij} = 0$ if and only if $J_{ij} = 0$. Hence by Assumption 4, $L > 0$, g_i is convex and L-smooth.

Since G is a convex combination of g_i, G is convex and L-smooth. □

When $p = q = 1$, $\mathscr{J} = J_{11} = \text{diag}\{1, 0, \dots, 0\}$, G is not strong convex. The following proposition gives a sufficient condition for G to be strongly convex.

Proposition 5 *If the system matrix $\mathscr{J}$ is of full column rank, then G is σ-strongly convex.*

Proof By Assumption 5, $p \ge q$, and

$$G(\boldsymbol{u}_k) = \frac{1}{2p} \left(\boldsymbol{u}_k^T \mathscr{J}^T \mathscr{J} \boldsymbol{u}_k - \boldsymbol{y}_d^T \mathscr{J} \boldsymbol{u}_k - \boldsymbol{u}_k^T \mathscr{J}^T \boldsymbol{y}_d + \boldsymbol{y}_d^T \boldsymbol{y}_d \right).$$

Since $\mathscr{J}^T\mathscr{J}$ is positive definite if and only if $\mathscr{J}$ has full column rank. Therefore by Theorem 2.2, $G(u_k)$ is σ-strongly convex when $\mathscr{J}$ has full column rank, and the strong convexity factor is $\frac{1}{2p}$ of the minimum eigenvalue of matrix $\mathscr{J}^T\mathscr{J}$. □

When G is strongly convex, we can prove that Algorithm 3 converges linearly similar to Algorithm 1 (Bottou et al., 2018). The convergence proof of Algorithm 3 under the strongly convex condition is omitted because of limited space. We only give the convergence proof under non-strongly convex.

First we have the following lemma (Reddi et al., 2016).

Lemma 4 *Assume that* $c_k, c_{k+1}, \beta > 0$,

$$c_k = c_{k+1}\left(1 + \eta\beta + 2\eta^2 L^2\right) + \eta^2 L^3.$$

If η, β *and* c_{k+1} *are chosen such that*

$$\mathrm{T}_k = \left(\eta - \frac{c_{k+1}\eta}{\beta} - \eta^2 L - 2c_{k+1}\eta^2\right) > 0,$$

then each iteration of Algorithm 3 has an upper bound

$$\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] \leq \frac{R_{s,k} - R_{s,k+1}}{\mathrm{T}_k},$$

where $R_{s,k} \triangleq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right) + c_t\left\|\boldsymbol{u}_{s,k} - \tilde{\boldsymbol{u}}^s\right\|^2\right]$.

Proof Since g_i is L-smooth,

$$g_i\left(\boldsymbol{u}_{s,k+1}\right) \leq g_i\left(\boldsymbol{u}_{s,k}\right) + \left\langle \nabla g_i\left(\boldsymbol{u}_{s,k}\right), \boldsymbol{u}_{s,k+1} - \boldsymbol{u}_{s,k}\right\rangle + \frac{L}{2}\left\|\boldsymbol{u}_{s,k+1} - \boldsymbol{u}_{s,k}\right\|^2.$$

For fixed s, let $\boldsymbol{w}_k = \nabla g_i\left(\boldsymbol{u}_{s,k}\right) - \nabla g_i\left(\tilde{\boldsymbol{u}}^s\right) + \boldsymbol{\mu}_s$, then $\mathbb{E}[\boldsymbol{w}_k] = \nabla G\left(\boldsymbol{u}_{s,k}\right)$. Since $\boldsymbol{u}_{s,k+1} - \boldsymbol{u}_{s,k} = -\eta\boldsymbol{w}_k$, we use the above equation and take the expectation on both sides to obtain

$$\mathbb{E}\left[G\left(\boldsymbol{u}_{s,k+1}\right)\right] \leq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right) - \eta\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] + \frac{L\eta^2}{2}\mathbb{E}\left[\|\boldsymbol{w}_k\|^2\right]. \tag{24}$$

In addition, for $\left\|\boldsymbol{u}_{s,k+1} - \tilde{\boldsymbol{u}}^s\right\|^2$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{u}_{s,k+1}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right] \\
&=\mathbb{E}\left[\left\|\boldsymbol{u}_{s,k+1}-\boldsymbol{u}_{s,k}\right\|^{2}+\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}+2\left\langle\boldsymbol{u}_{s,k+1}-\boldsymbol{u}_{s,k}, \boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\rangle\right] \\
&=\mathbb{E}\left[\eta^{2}\left\|\boldsymbol{w}_{k}\right\|^{2}+\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right]-2\eta\mathbb{E}\left[\left\langle\nabla G\left(\boldsymbol{u}_{s,k}\right), \boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\rangle\right] \\
&\leq \mathbb{E}\left[\eta^{2}\left\|\boldsymbol{w}_{k}\right\|^{2}+\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right] \\
&\quad-2\eta\mathbb{E}\left[\frac{1}{2\beta}\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}+\frac{\beta}{2}\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right].
\end{aligned}
\tag{25}
$$

In the above, we have used Young's inequality $\langle x, y\rangle \leq \frac{1}{2\beta}\|x\|^{2}+\frac{\beta}{2}\|y\|^{2}$.

For $\mathbb{E}\left[\|\boldsymbol{w}_{k}\|^{2}\right]$, we have the following estimation:

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{w}_{k}\|^{2}\right] &= \mathbb{E}\left[\left\|\nabla g_{i}\left(\boldsymbol{u}_{s,k}\right)-\nabla g_{i}\left(\tilde{\boldsymbol{u}}^{s}\right)+\nabla G\left(\tilde{\boldsymbol{u}}^{s}\right)\right\|^{2}\right] \\
&= \mathbb{E}\left[\left\|\nabla g_{i}\left(\boldsymbol{u}_{s,k}\right)-\nabla g_{i}\left(\tilde{\boldsymbol{u}}^{s}\right)+\nabla G\left(\tilde{\boldsymbol{u}}^{s}\right)-\nabla G\left(\boldsymbol{u}_{s,k}\right)+\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla g_{i}\left(\boldsymbol{u}_{s,k}\right)-\nabla g_{i}\left(\tilde{\boldsymbol{u}}^{s}\right)-\left(\nabla G\left(\boldsymbol{u}_{s,k}\right)-\nabla G\left(\tilde{\boldsymbol{u}}^{s}\right)\right)\right\|^{2}\right] \\
&\quad+2\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla g_{i}\left(\boldsymbol{u}_{s,k}\right)-\nabla g_{i}\left(\tilde{\boldsymbol{u}}^{s}\right)\right\|^{2}\right]+2\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right] \\
&\leq 2L^{2}\mathbb{E}\left[\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right]+2\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right],
\end{aligned}
\tag{26}
$$

where the inequality is given by $\|a+b\|^{2} \leq 2\|a\|^{2}+2\|b\|^{2}$, $\mathbb{E}\left[\|\zeta-\mathbb{E}\zeta\|^{2}\right]=\mathbb{E}\|\zeta\|^{2}-\|\mathbb{E}\zeta\|^{2} \leq \mathbb{E}\|\zeta\|^{2}$ and the smoothness of g_{i}, i.e., $\|\nabla g_{i}\left(\boldsymbol{u}_{s,k}\right)-\nabla g_{i}\left(\tilde{\boldsymbol{u}}^{s}\right)\| \leq \|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\|^{2}$.

Denoting $R_{s,k} \triangleq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right)+c_{k}\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right]$, by (24) and (25), we have

$$
\begin{aligned}
R_{s,k+1} &\leq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right)-\eta\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right]+\frac{L\eta^{2}}{2}\mathbb{E}\left[\|\boldsymbol{w}_{k}\|^{2}\right] \\
&\quad+c_{k+1}\mathbb{E}\left[\eta^{2}\left\|\boldsymbol{w}_{k}\right\|^{2}+\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right] \\
&\quad-2c_{k+1}\eta\mathbb{E}\left[\frac{1}{2\beta}\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}+\frac{\beta}{2}\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right] \\
&\leq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right)-\eta\left(1-\frac{c_{k+1}}{\beta}\right)\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^{2}\right] \\
&\quad+\eta^{2}\left(\frac{L}{2}+c_{k+1}\right)\mathbb{E}\left[\|\boldsymbol{w}_{k}\|^{2}\right] \\
&\quad+c_{t+1}(1+\eta\beta)\mathbb{E}\left[\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^{s}\right\|^{2}\right]
\end{aligned}
$$

From (26), we have

$$\begin{aligned} R_{s,k+1} \leq \mathbb{E}\left[G\left(\boldsymbol{u}_{s,k}\right)\right] + \left(c_{k+1}\left(1+\eta\beta+2\eta^2L^3\right)+\eta^2L^3\right)\mathbb{E}\left[\left\|\boldsymbol{u}_{s,k}-\tilde{\boldsymbol{u}}^s\right\|^2\right] \\ -\left(\eta-\frac{c_{k+1}\eta}{\beta}-L\eta^2-2c_{k+1}\eta^2\right)\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] \\ = R_{s,k}-\left(\eta-\frac{c_{t+1}\eta}{\beta}-L\eta^2-2c_{k+1}\eta^2\right)\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right]. \end{aligned}$$

Let $\mathrm{T}_k \triangleq \left(\eta-\frac{c_{k+1}\eta}{\beta}-\eta^2L-2c_{k+1}\eta^2\right)$, $\mathrm{T}_k>0$, then

$$\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] \leq \frac{R_{s,k}-R_{s,k+1}}{\mathrm{T}_k}. \qquad \square$$

Because of the complexity of non-strongly convex problem, we do not consider convergence criteria in Theorems 3 and 4 such as $\mathbb{E}\left[G(\boldsymbol{u})-G\left(\boldsymbol{u}^*\right)\right] \leq \varepsilon$, but instead proving $\mathbb{E}\left[\|\nabla G(\boldsymbol{u})\|^2\right] \leq \varepsilon$ for Algorithm 3. Note that if G is σ-strongly convex, it is easy to verify that

$$G(\boldsymbol{u})-G\left(\boldsymbol{u}^*\right) \leq \frac{1}{2\sigma}\|\nabla G(\boldsymbol{u})\|^2.$$

Thus by $\mathbb{E}\left[\|\nabla G(\boldsymbol{u})\|^2\right] \leq \varepsilon$, we have $\mathbb{E}\left[G(\boldsymbol{u})-G\left(\boldsymbol{u}^*\right)\right] \leq \varepsilon$. However, this relationship does not always hold under non-strongly convex case (Ghadimi & Lan, 2013). The following theorem gives the proof of the convergence $\mathbb{E}\left[\|\nabla G(\boldsymbol{u})\|^2\right] \leq \varepsilon$ of Algorithm 3.

Theorem 5 *Suppose each g_i is convex and L-smooth, and G is convex. For $0 \leq k \leq m-1$, $c_k, c_{k+1}, \beta > 0$, $c_m = 0$ satisfying*

$$c_k = c_{k+1}\left(1+\eta\beta+2\eta^2L^2\right)+\eta^2L^3.$$

and η, β, c_{k+1} are chosen such that

$$\mathrm{T}_k = \left(\eta-\frac{c_{k+1}\eta}{\beta}-\eta^2L-2c_{k+1}\eta^2\right) > 0.$$

Let $\tau_m = \min_k \mathrm{T}_k$, $\boldsymbol{u}_a$ be a uniformly distributed random vector with values $\{\boldsymbol{u}_{s,k} \mid 0 \leq s \leq S-1, 0 \leq k \leq m-1\}$. Denote that $\boldsymbol{u}^* = \mathrm{argmin}_{\boldsymbol{u}} G(\boldsymbol{u})$, then Algorithm 3 satisfies

$$\mathbb{E}\left[\|\nabla G\left(\boldsymbol{u}_a\right)\|^2\right] \leq \frac{\mathbb{E}\left[G\left(\tilde{\boldsymbol{u}}^0\right)-G\left(\boldsymbol{u}^*\right)\right]}{Sm\tau_m}.$$

Proof We take $k = 0, 1, \ldots, m-1$ in Lemma 1 and sum up to obtain

$$\sum_{k=0}^{m-1}\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] \leq \frac{R_{s,0}-R_{s,m}}{\tau_m} = \frac{\mathbb{E}\left[G\left(\tilde{\boldsymbol{u}}^s\right)-G\left(\tilde{\boldsymbol{u}}^{s+1}\right)\right]}{\tau_m}.$$

By the definition of $R_{s,k}$, we choose $\boldsymbol{u}_{s,0}=\tilde{\boldsymbol{u}}^s$, $\boldsymbol{u}_{s,m}=\tilde{\boldsymbol{u}}^{s+1}$ in $\tilde{\boldsymbol{u}}^s$. We take $s=0,1,\ldots,S-1$ in the above and sum up to get

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_a\right)\right\|^2\right] &= \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{k=0}^{m-1}\mathbb{E}\left[\left\|\nabla G\left(\boldsymbol{u}_{s,k}\right)\right\|^2\right] \\ &\leq \frac{\mathbb{E}\left[G\left(\tilde{\boldsymbol{u}}^0\right)-G\left(\tilde{\boldsymbol{u}}^s\right)\right]}{Sm\tau_m} \leq \frac{\mathbb{E}\left[G\left(\tilde{\boldsymbol{u}}^0\right)-G\left(\boldsymbol{u}^*\right)\right]}{Sm\tau_m}.\end{aligned}$$

In the above, use the definition of $\boldsymbol{u}_a$ and $G\left(\tilde{\boldsymbol{u}}^s\right)\geq G\left(\boldsymbol{u}^*\right)$. □

Remark 7 Theorem 5 shows that the convergence of Algorithm 3 is $O(1/Sm)$ under non-strongly convex conditions. The convergence of the corresponding SGD-based ILC under non-strongly convex conditions is $O(1/\sqrt{Sm})$ ((Johnson & Zhang, 2013)). Moreover, the theorem states that the convergence of Algorithm 3 is only related to the step size η but not to the choice of the number of iterations m. Since the system information is unknown, η needs to be estimated by experiment.

Remark 8 Theorem 5 indicates that $G\left(\boldsymbol{u}_k\right)$ can approach the optimal value $G\left(\boldsymbol{u}^*\right)=(1/2p)\|\boldsymbol{y}_d-\mathscr{J}\boldsymbol{u}^*\|^2$, i.e., $\boldsymbol{u}_k$ can converge to the optimal input $\boldsymbol{u}^*$. Similarly, the Assumption 2 can also be changed to $\lim_{k\to\infty}F\left(u_k\right)=F\left(u^*\right)$ without affecting the convergence of Algorithm 1 and Algorithm 2. If $F\left(u^*\right)=0$, then $\lim_{k\to\infty}F\left(u_k\right)=0$.

Remark 9 System (21) degenerates to SISO system when both input and output are one dimension. At this point, the theorem indicates that when system (2) satisfies Assumption 4 (Assumption 1 need not to be satisfied), Algorithm 1 and Algorithm 2 still converge with updating the "snapshot" as Option II. But the convergence rate becomes $O(1/Sm)$ when the objective function is not strongly convex.

4.4 Numerical Simulation

We take MIMO systems with 21×21 input-output dimensions randomly generated by the Matlab drss function (Aarnoudse & Oomen, 2021) and set the time length to $N=42$. The desired trajectory $\boldsymbol{y}_d$ is 0.025 in each dimension. Model-free ILC based on full gradient (GD), stochastic gradient (SGD), and stochastic variance reduced gradient (SVRG) are performed in Fig. 5. We take the optimal step of each algorithm after multiple experiments.

As shown in Fig. 5a, the GD- and SVRG-based ILC converge similarly for deterministic systems, and both are faster than the SGD-based ILC algorithm.

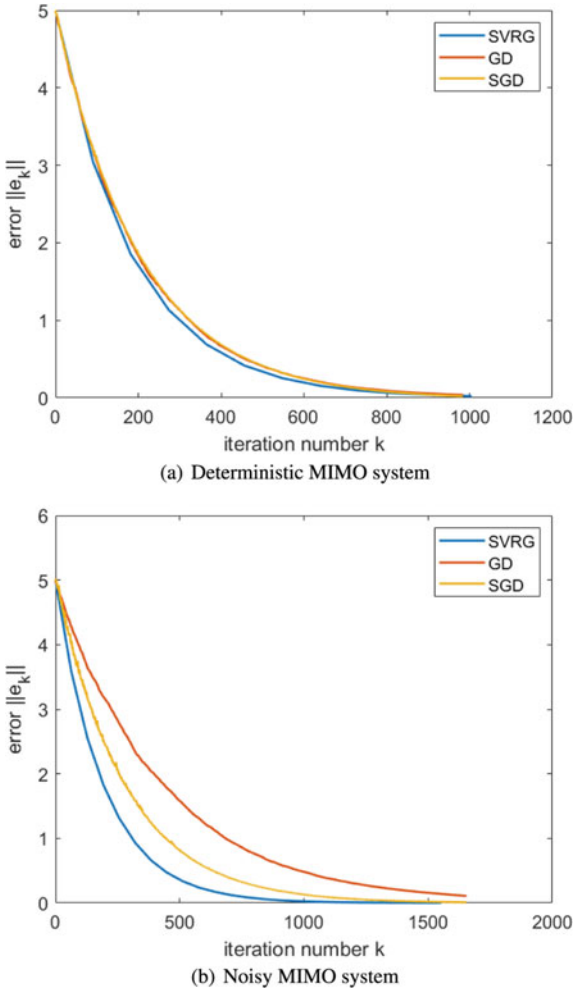

Fig. 5 Comparison of three data-driven gradient-based ILC under MIMO systems

For randomly generated MIMO systems with input-output dimensions of 21×21 and $N = 42$, we add Gaussian white noise to the system.

From Fig. 5b, we can see that both GD- and SGD-based ILC converge worse than SVRG type ILC when the system is noisy, and SVRG-based ILC can still maintain excellent convergence when the system is noisy.

5 Conclusions

This chapter focuses on exploring ILC based on the SVRG method. Firstly, Sect. 2 gives the basic framework of SVRG-based ILC and proves that the algorithm converges at a rate of $O(\alpha^k)$ under smooth and strongly convex condition. Secondly, Sect. 3 designs a SVRG-based ILC algorithm to solve random error data dropouts and proves that the algorithm converges linearly. Finally, Sect. 4 constructs a model-free SVRG-based ILC by improving the existing model-free algorithm for MIMO systems and proves that the convergence rate is $O(1/k)$ under smooth and convex condition. Compared to the GD- and SGD-based ILC, two numerical simulations in Sects. 3 and 4 verify that the SVRG-based ILC has superior convergence rate in both the random error dropouts and model-free contexts, respectively.

It should be noted that the SVRG-based ILC framework given in this chapter is not only applicable to the random error dropouts and model-free problems but can also be utilized to solve other error or system information deficient problems by properly decomposing the control objectives. Future research includes comparing its advantages and disadvantages with the stochastic approximation (SA) method, extending the framework to other information deficient problems, and attempting to develop algorithms with faster convergence based on this framework.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62173333), Beijing Natural Science Foundation (Z210002), and Research Fund of Renmin University of China (2021030187).

References

Aarnoudse, L., & Oomen, T. (2020). Model-free learning for massive MIMO systems: Stochastic approximation Adjoint iterative learning control. *IEEE Control Systems Letters,5*(6), 1946–1951.

Aarnoudse, L., & Oomen,T. (2021). Conjugate gradient MIMO iterative learning control using data-driven stochastic gradients. In *2021 60th IEEE Conference on Decision and Control (CDC)* (pp. 3749–3754).

Allen-Zhu, Z. (2018). Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research, 18,* 1–51.

Arimoto, S., Kawamura, S., & Miyazaki, F. (1984). Bettering operation of robots by learning. *The Journal of Intelligent and Robotic Systems, 1*(2), 123–140.

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review,60*(2), 223–311.

Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization,23*, 2341–2368.

Gu, P., Tian, S., & Chen, Y. (2019). Iterative learning control based on Nesterov accelerated gradient method. *IEEE Access,7*, 115836–115842.

Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems,1*, 315–323.

Nesterov, Y. (2005). *Introductory lectures on convex programming volume: A basic course* (Vol. I). Kluwer Academic Publishers.

Oomen, T., van der Maas, R., Rojas, C. R., & Hjalmarsson, H. (2014). Iterative data-driven H-infinity norm estimation of multivariable systems with application to robust active vibration isolation. *IEEE Transactions on Control Systems Technology,22*(6), 2247–2260.

Owens, D. H., Hatonen, J. J., & Daley, S. (2009). Robust monotone Radient-based discrete-time iterative learning control. *The International Journal of Robust and Nonlinear Control,19*(6), 634–661.

Reddi, J. S., Hefny, A., Sra, S., Poczos, B., & Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (Vol. 48, pp. 314–323).

Shen, D. (2018). Iterative learning control with incomplete information: A survey. *IEEE/CAA Journal of Automatica Sinica, 5*(5), 885–901.

Yang, X., & Ruan, X. (2017). Reinforced gradient-type iterative learning control for discrete linear time-invariant systems with parameters uncertainties and external noises. *IMA Journal of Mathematical Control and Information,34*(4), 1117–1133.

A Generalization of NTRUEncrypt

Zheng Zhiyong, Liu Fengxia, Huang Wenlin, Xu Jie, and Tian Kun

Abstract The main purpose of this chapter is to give a more general construction of NTRU based on ideal matrices and q-ary lattice theory. To understand our construction, first we discuss a more general form of the ordinary cyclic code, namely ϕ-cyclic code, which firstly appeared in (Lopez-Permouth et al., 2009; Shi et al., 2020); thus, we give a more generalized NTRUEncrypt from replacing finite field by real number field $\mathbb{R}$.

Keywords ϕ-cyclic code · Ideal matrices · Convolutional modular Lattice · NTRU

1 ϕ-Cyclic Code

Let F_q be a finite field with q elements and q be a power of a prime number, $F_q[x]$ be the polynomial ring of F_q with variable x. Let F_q^n be the n-dimensional linear space over F_q, and $a = (a_0, a_1, ..., a_{n-1}) \in F_q^n$ be a fixed vector in F_q^n with $a_0 \neq 0$, the associated polynomial of a given by

$$\phi(x) = \phi_a(x) = x^n - a_{n-1}x^{n-1} - \cdots - a_1 x - a_0 \in F_q[x], a_0 \neq 0. \quad (1)$$

Z. Zhiyong · L. Fengxia · H. Wenlin · X. Jie · T. Kun (✉)
Engineering Research Center of Ministry of Education for Financial Computing and Digital Engineering, Renmin University of China, Beijing 100872, China
e-mail: tkun19891208@ruc.edu.cn

Z. Zhiyong
e-mail: zhengzy@ruc.edu.cn

L. Fengxia
e-mail: liu_fx@ruc.edu.cn

H. Wenlin
e-mail: wenlin@ruc.edu.cn

X. Jie
e-mail: xujie0665@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_6

Let $< \phi(x) >$ be the principal ideal generated by $\phi(x)$ in $F_q[x]$. There is a one to one correspondence between F_q^n and the quotient ring $R = F_q[x]/ < \phi(x) >$, given by

$$c = (c_0, c_1, ..., c_{n-1}) \in F_q^n \rightleftarrows c(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1} \in R. \quad (2)$$

In fact, this correspondence is also an isomorphism of Abel groups. One may extend this correspondence to subsets of F_q^n and R by

$$C \subset F_q^n \rightleftarrows C(x) = \{c(x)|c \in C\} \subset R. \quad (3)$$

If $C \subset F_q^n$ is a linear subspace of F_q^n of dimension k, then C is called a linear code in coding theory and written by $C = [n, k]$ as usual. Each vector $c = (c_0, c_1, ..., c_{n-1}) \in C$ is called a codeword of length n. Obviously, $C = [n, 0]$ and $C = [n, n]$ are two trivial codes. Another one is called constant codes, of which is almost trivial given by

$$C = \{(b, b, ..., b)|b \in F_q\}, \text{ and } C = [n, 1].$$

According to the given polynomial $\phi(x) = \phi_a(x)$, we may define a linear transformation τ_ϕ in F_q^n,

$$\begin{aligned} \tau_\phi(c) &= \tau_\phi((c_0, c_1, ..., c_{n-1})) \\ &= (a_0 c_{n-1}, c_0 + a_1 c_{n-1}, c_1 + a_2 c_{n-1}, ..., c_{n-2} + a_{n-1} c_{n-1}). \end{aligned} \quad (4)$$

It is easily seen that $\tau_\phi : F_q^n \to F_q^n$ is a linear transformation.

Definition 1 Let $C \subset F_q^n$ be a linear code. It is called a ϕ-cyclic code, if

$$\forall c \in C \Rightarrow \tau_\phi(c) \in C. \quad (5)$$

In other words, a linear code C is a ϕ-cyclic code, if and only if C is closed under linear transformation τ_ϕ. Clearly, if $a = (1, 0, ..., 0)$, and $\phi_a(x) = x^n - 1$, then the ϕ-cyclic code is precisely the ordinary cyclic code (see this chapter of Lopez-Permouth et al. (2009)).

Remark 1 The ϕ-cyclic code we give here is polycyclic code in fact, which firstly appeared in (Lopez-Permouth et al., 2009; Shi et al., 2020), but we mainly concern for its application to McEliece and Niederriter's cryptosystems. We first show that there is a one to one correspondence between ϕ-cyclic codes in F_q^n and ideals in $R = F_q[x]/ < \phi(x) >$.

Theorem 1 *Let $C \subset F_q^n$ be a subset, then C is a ϕ-cyclic code, if and only if $C(x)$ is an ideal of R.*

Proof We use column notation for vector in F_q^n, then linear transformation τ_ϕ may be written as

$$\tau_\phi \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} a_0 c_{n-1} \\ c_0 + a_1 c_{n-1} \\ \vdots \\ c_{n-2} + a_{n-1} c_{n-1} \end{pmatrix}, \forall c = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} \in F_q^n.$$

Let T_ϕ be a $n \times n$ square matrix over F_q,

$$T_\phi = \left(\begin{array}{ccc|c} 0 & \cdots & 0 & a_0 \\ \hline & & & a_1 \\ & I_{n-1} & & \vdots \\ & & & a_{n-1} \end{array}\right) \in F_q^{n \times n}, \tag{6}$$

where I_{n-1} is the $(n-1) \times (n-1)$ unit matrix. The matrix expression of τ_ϕ is as follows:

$$\tau_\phi \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = T_\phi \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} a_0 c_{n-1} \\ c_0 + a_1 c_{n-1} \\ \vdots \\ c_{n-2} + a_{n-1} c_{n-1}. \end{pmatrix}. \tag{7}$$

Suppose $C \subset F_q^n$ and $C(x)$ is an ideal of R, it is clear that C is a linear code of F_q^n. To prove C is a ϕ-cyclic code, we note that for any polynomial $c(x) \in C(x)$, then $xc(x) \in C(x)$ if and only if $\tau_\phi(c) \in C$, namely, if $c(x) \in C(x)$, then

$$xc(x) \in C(x) \Leftrightarrow \tau_\phi(c) \in C \Leftrightarrow T_\phi c \in C. \tag{8}$$

Therefore, if $C(x)$ is an ideal of R, then we have immediately that C is a ϕ-cyclic code of F_q^n.

Conversely, if $C \subset F_q^n$ is a ϕ-cyclic code, then for all $k \geqslant 1$, we have

$$\forall c \in C \Rightarrow T_\phi^k c \in C, k \geqslant 1.$$

It follows that

$$\forall c(x) \in C(x) \Rightarrow x^k c(x) \in C(x), 0 \leqslant k \leqslant n-1,$$

which implies $C(x)$ is an ideal of R. This is the proof of Theorem 1. □

By Theorem 1, to find a ϕ-cyclic code, it is enough to find an ideal of R. There are two trivial ideals $C(x) = 0$ and $C(x) = R$, the corresponding ϕ-cyclic codes are $C = [n, 0]$ and $C = F_q^n$, respectively, which are called trivial ϕ-cyclic code. To find non-trivial ϕ-cyclic codes, we make use of homomorphic theorems, which is a

standard technique in Algebra. Let π be the natural homomorphism from $F_q[x]$ to its quotient ring $R = F_q[x]/ < \phi(x) >$, $\ker\pi =< \phi(x) >$,

$$< \phi(x) > \subset N \subset F_q[x] \xrightarrow{\pi} R = F_q[x]/ < \phi(x) >, \tag{9}$$

where N is an ideal of $F_q[x]$, of which is containing $\ker\pi =< \phi(x) >$. Since $F_q[x]$ is a principal ideal domain, then $N =< g(x) >$ is a principal ideal generated by a monic polynomial $g(x) \in F_q[x]$. It is easy to see that

$$< \phi(x) > \subset < g(x) > \Leftrightarrow g(x)|\phi(x).$$

It follows that all ideals N satisfying (1.9) are given by

$$\{< g(x) > \mid g(x) \in F_q[x] \text{ is monic and } g(x)|\phi(x)\}.$$

We write by $< g(x) > \bmod \phi(x)$, the image of $< g(x) >$ under π, it is easy to check

$$< g(x) > \bmod \phi(x) = \{h(x)g(x) \mid h(x) \in F_q[x] \text{ and } \deg h(x) + \deg g(x) < n\}, \tag{10}$$

more precisely, which is a representative elements set of $< g(x) > \bmod \phi(x)$, by homomorphism theorem in ring theory, all ideals of R given by

$$\{< g(x) > \bmod \phi(x) \mid g(x) \in F_q[x] \text{ is monic and } g(x)|\phi(x)\}. \tag{11}$$

Let d be the number of monic divisors of $\phi(x)$ in $F_q[x]$, it follows immediately that

Corollary 1 *The number of ϕ-cyclic code in F_q^n is d.*

To compare the ϕ-cyclic code and ordinary cyclic code, we see a simple example.

Example 1 Constant code C is always a cyclic code for $1 + x + \cdots + x^{n-1}|x^n - 1$, and its generated polynomial is just $1 + x + \cdots + x^{n-1}$. But constant code C in F_q^n is not always a ϕ-cyclic code, it is a ϕ-cyclic code if and only if $1 + x + \cdots + x^{n-1}|\phi(x)$, an equivalent condition for $1 + x + \cdots + x^{n-1}|\phi(x)$ is

$$a_{n-1} = a_{n-2} = \cdots = a_1 = b, \text{ and } a_0 = 1 + b.$$

Definition 2 Let C be a ϕ-cyclic code and $C(x) = g(x) \bmod \phi(x)$. We call $g(x)$ as the generated polynomial of C, where $g(x)$ is monic and $g(x)|\phi(x)$.

Lemma 1 *Let $g(x) = g_0 + g_1x + \cdots + g_{n-k-1}x^{n-k-1} + x^{n-k}$ be the generated polynomial of a ϕ-cyclic code C, where $1 \leqslant k \leqslant n-1$, and $g(x)|\phi(x)$, then $C = [n, k]$, and a generated matrix for C is the following block matrix:*

$$G = \begin{pmatrix} g \\ \tau_\phi(g) \\ \tau_\phi^2(g) \\ \vdots \\ \tau_\phi^{k-1}(g) \end{pmatrix}_{k\times n}, \tag{12}$$

where $g = (g_0, g_1, ..., g_{n-k-1}, 1, 0, ..., 0) \in C$ *is the corresponding codeword of* $g(x)$*, and* $\tau_\phi^i(g) = \tau_\phi^{i-1}(\tau_\phi(g))$ *for* $1 \leqslant i \leqslant n-1$.

Proof By assumption, $C(x) =< g(x) > \bmod \phi(x)$, then $\{g, \tau_\phi(g), ..., \tau_\phi^{k-1}(g)\} \subset C$, we are to prove it is a basis of C. First, these vectors are linearly independent. Otherwise, we have

$$\sum_{i=0}^{k-1} b_i \tau_\phi^i(g) = 0, \text{ for some } b_i \in F_q, \tag{13}$$

and the corresponding polynomial is zero, namely

$$\left(\sum_{i=0}^{k-1} b_i x^i\right) g(x) = 0.$$

It follows that

$$\sum_{i=0}^{k-1} b_i x^i = 0 \Rightarrow b_i = 0 \text{ for all } i, 0 \leqslant i \leqslant k-1.$$

Next, if $c \in C$, and $c(x) \in C(x)$, by (1.10), there is a polynomial $b(x) = b_0 + b_1 x + \cdots + b_{k-2}x^{k-2} + x^{k-1}$ such that

$$c(x) = b(x)g(x) = \left(\sum_{i=0}^{k-1} b_i x^i\right) g(x), \text{ where } b_{k-1} = 1.$$

Thus, we have the corresponding codeword of $C(x)$

$$c = \sum_{i=0}^{k-1} b_i \tau_\phi^i(g).$$

This shows that $\{g, \tau_\phi(g), ..., \tau_\phi^{k-1}(g)\}$ is a basis of C, and a generated matrix for C is

$$G = \begin{pmatrix} g \\ \tau_\phi(g) \\ \tau_\phi^2(g) \\ \vdots \\ \tau_\phi^{k-1}(g) \end{pmatrix}_{k\times n}.$$

We have Lemma 1 at once. □

To describe a parity check matrix for a ϕ-cyclic code, for any $c = (c_0, c_1, ..., c_{n-1}) \in F_q^n$, we write

$$\overline{c} = (c_{n-1}, c_{n-2}, ..., c_1, c_0) \in F_q^n.$$

Lemma 2 *Suppose C is a ϕ-cyclic code with generated polynomial $g(x)$, where $g(x)|\phi(x)$ and $degg(x) = n - k$. Let $h(x)g(x) = \phi(x)$, where $h(x) = h_0 + h_1x + \cdots + h_{k-1}x^{k-1} + x^k$. Then a parity check matrix for C is*

$$H = \begin{pmatrix} \overline{h} \\ \tau_\phi(\overline{h}) \\ \vdots \\ \tau_\phi^{n-k-1}(\overline{h}) \end{pmatrix}_{(n-k)\times n}. \tag{14}$$

Proof Since $h(x)g(x) = \phi(x)$, it means that $h(x)g(x) = 0$ in $R = F_q[x]/ < \phi(x) >$, thus we have

$$g_0h_i + g_1h_{i-1} + \cdots + g_{n-k}h_{i-n+k} = 0, \forall 0 \leqslant i \leqslant n-1.$$

It follows that $GH' = 0$, where G is a generated matrix for C given by (1.12). Therefore, H is a parity check matrix for C. □

A separable polynomial in Algebra means that it has no multiple roots in its splitting field. The following lemma shows that there is an unit element in any non-zero ideal of R, when $\phi(x)$ is a separable polynomial.

Lemma 3 *Suppose $\phi(x)$ is a separable polynomial of F_q, and $C(x) = g(x)$ mod $\phi(x)$ is an ideal of R with $degg(x) \leqslant n - 1$, then there exists an element $d(x) \in C(x)$ such that*

$$c(x)d(x) = c(x), \text{ for all } c(x) \in C(x).$$

Proof Let $h(x)g(x) = \phi(x)$. Since $\phi(x)$ is a separable polynomial, then $\gcd(g(x), h(x)) = 1$, and there are two polynomials $a(x)$ and $b(x)$ in $F_q[x]$ such that

$$a(x)g(x) + b(x)h(x) = 1.$$

Let

$$d(x) = a(x)g(x) = 1 - b(x)h(x) \in C(x).$$

If $c(x) \in C(x)$, by (1.10), we write $c(x) = g(x)g_1(x)$, it follows that

$$c(x)d(x) \equiv a(x)g(x)g(x)g_1(x) \equiv (1 - b(x)h(x))g(x)g_1(x)$$

$$\equiv g(x)g_1(x) \equiv c(x)(\text{mod } \phi(x)).$$

Thus, we have $c(x)d(x) = c(x)$ in R. □

Next, we discuss maximal ϕ-cyclic code. Let $C(x) = g(x) \bmod \phi(x)$, and $g(x)$ be an irreducible polynomial in $F_q[x]$, we call the corresponding ϕ-cyclic code C a maximal ϕ-cyclic code, because $< g(x) >$ is a maximal ideal in $F_q[x]$.

Lemma 4 *Let C be a maximal ϕ-cyclic code with generated polynomial $g(x)$, β be a root of $g(x)$ in some extensions of F_q, then*

$$C(x) = \{a(x) \mid a(x) \in R \text{ and } a(\beta) = 0\}. \tag{15}$$

Proof If $a(x) \in C(x)$, by (1.10) we have $a(\beta) = 0$ immediately. Conversely, if $a(x) \in F_q[x]$ and $a(\beta) = 0$, since $g(x)$ is irreducible, thus we have $g(x)|a(x)$, and (1.15) follows at once. □

An important application of maximal ϕ-cyclic code is to construct an error-correcting code, so that we may obtain modified McEliece-Niederriter's cryptosystem. To do this, let $1 \leqslant m < \sqrt{n}$, and F_{q^m} be an extension field of F_q of degree m. Suppose $F_{q^m} = F_q(\theta)$, where θ is a primitive element of F_{q^m} and $F_q(\theta)$ is the simple extension containing F_q and θ. Let $g(x) \in F_q[x]$ be the minimum polynomial of θ, then $g(x)$ is an irreducible polynomial of degree m of $F_q[x]$. It is well-known that F_{q^m} is a Galois extension of F_q, so that all roots of $g(x)$ are in F_{q^m}. Let $\beta_1, \beta_2, ..., \beta_m$ be all roots of $g(x)$, the Vandermonde matrix $V(\beta_1, \beta_2, ..., \beta_m)$ defined by

$$H = V(\beta_1, \beta_2, ..., \beta_m) = \begin{pmatrix} 1 & \beta_1 & \beta_1^2 & \cdots & \beta_1^{n-1} \\ 1 & \beta_2 & \beta_2^2 & \cdots & \beta_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \beta_m & \beta_m^2 & \cdots & \beta_m^{n-1} \end{pmatrix}_{m \times n}, \tag{16}$$

where $\beta_1 = \theta$ and each β_i is a vector of $(F_q)^m$. For arbitrary monic polynomial $h(x) \in F_q[x]$, $\deg h(x) = n - m$, let $\phi(x) = h(x)g(x)$ and C be a maximal ϕ-cyclic code generated by $g(x)$. It is easy to verify that

$$c \in C \Leftrightarrow cH' = 0.$$

Therefore, H is a parity check matrix for C. If we choose the primitive element θ, so that any $d - 1$ columns in H are linearly independent, then the minimum distance of C is greater than d, and C is a t-error-correcting code, where $t = [\frac{d}{2}]$.

The public key cryptosystems based on algebraic coding theory were created by Lyubashevsky and Micciancio (2006) and Micciancio and Regev (2009), a suitable t-error-correcting code plays a key role in their construction. The error-correcting code C should satisfy the following requirements:

(i) C should have a relatively large error-correcting capability so that a reasonable number of message vectors can be used;
(ii) C should allow an efficient decoding algorithm so that the decryption can be carried out within a short time.

Our results supply a different way to choose an error-correcting code by selecting arbitrary irreducible polynomials $g(x) \in F_q[x]$ of degree m and roots of $g(x)$ rather than an irreducible factor of $x^n - 1$ and the roots of unit such as ordinary BCH code and Gappa code.

In fact, for any positive integer m, there is at least an irreducible polynomial $g(x) \in F_q[x]$ with degree m. Let $N_q(m)$ be the number of irreducible polynomials of degree m in $F_q[x]$, then we have (see Theorem 3.25 of Lidl and Niederreiter (1983))

$$N_q(m) = \frac{1}{m}\sum_{d|m} u\left(\frac{m}{d}\right) q^d = \frac{1}{m}\sum_{d|m} u(d) q^{\frac{m}{d}},$$

where $u(d)$ is Mobiüs function.

Assuming one has selected two monic and irreducible polynomials $g(x)$ and $h(x)$ with $\deg g(x) = m$ and $\deg h(x) = n - m$, let $\phi(x) = g(x)h(x)$, then one may obtain ϕ-cyclic code C generated by $g(x)$ or $h(x)$, which is more convenient and more flexible than the ordinary methods.

2 A Generalization of NTRUEncrypt

The public key cryptosystem NTRU proposed in 1996 by Hoffstein, Pipher, and Silverman is the fastest known lattice-based encryption scheme, although its description relies on arithmetic over polynomial quotient ring $Z[x]/ < x^n - 1 >$, it was easily observed that it could be expressed as a lattice-based cryptosystem (see IEEE Computer Society (2000)). For the background materials, we refer to (Coppersmith & Shamir, 1997; Hoffstein et al., 1998, 2017; Lint, 1999; McEliece, 1978; Micciancio, 2001). Our strategy in this section is to replace $Z[x]/ < x^n - 1 >$ by a more general polynomial ring $Z[x]/ < \phi(x) >$ and obtain a generalization of NTRUEncrypt, where $\phi(x)$ is a monic polynomial of degree n with integer coefficients.

In this section, we denote $\phi(x)$ and R by

$$\begin{aligned} \phi(x) =& x^n - a_{n-1}x^{n-1} - \cdots - a_1 x - a_0 \in Z[x], \\ R =& Z[x]/ < \phi(x) >, a_0 \neq 0. \end{aligned} \tag{17}$$

Let $H_\phi \in Z^{n\times n}$ be a square matrix given by

$$H = H_\phi = \left(\begin{array}{c|c} 0 \ \cdots \ 0 & a_0 \\ \hline & a_1 \\ I_{n-1} & \vdots \\ & a_{n-1} \end{array}\right)_{n\times n}, \tag{18}$$

where I_{n-1} is $(n-1)\times(n-1)$ unit matrix. Obviously, $\phi(x)$ is the characteristic polynomial of H, and H defines a linear transformation of $\mathbb{R}^n \to \mathbb{R}^n$ by $x \to Hx$, where $\mathbb{R}$ is real number field and x is a column vector of $\mathbb{R}^n$. We may extend this transformation to $\mathbb{R}^{2n}$ and denote σ by

$$\sigma\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}H\alpha\\ H\beta\end{pmatrix}, \text{ where } \begin{pmatrix}\alpha\\ \beta\end{pmatrix} \in \mathbb{R}^{2n}. \tag{19}$$

Of course, σ is again a linear transformation of $\mathbb{R}^{2n} \to \mathbb{R}^{2n}$.

According to Micciancio (2001), a q-ary lattice is a lattice L such that $qZ^n \subset L \subset Z^n$, where q is a positive integer.

Definition 3 A q-ary lattice L is called convolutional modular lattice, if L is in even dimension $2n$ satisfying

$$\forall \begin{pmatrix}\alpha\\ \beta\end{pmatrix} \in L \Rightarrow \sigma\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}H\alpha\\ H\beta\end{pmatrix} \in L. \tag{20}$$

In other words, a convolutional modular lattice is a q-ary lattice in even dimension and is closed under the linear transformation σ.

Recalling the secret key $\begin{pmatrix}f\\ g\end{pmatrix}$ of NTRU is a pair of polynomials of degree $n-1$, we may regard f and g as column vectors in Z^n. To obtain a convolutional modular lattice containing $\begin{pmatrix}f\\ g\end{pmatrix}$, we need some help of ideal matrices. An ideal matrix generated by a vector f is defined by

$$H^*(f) = H^*_\phi(f) = [f, Hf, H^2 f, ..., H^{n-1} f]_{n\times n}, \tag{21}$$

which is a block matrix in terms of each column $H^k f$ $(0 \leqslant k \leqslant n-1)$. It is easily seen that $H^*(f)$ is a generalization of the classical circulant matrices (see Davis (1994)), in fact, let $\phi(x) = x^n - 1$, and $f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1} \in Z[x]$, the ideal matrix $H^*_\phi(f)$ generated by f is given by

$$H^*(f) = H^*_\phi(f) = \begin{pmatrix} f_0 & f_{n-1} & \cdots & f_1 \\ f_1 & f_0 & \cdots & f_2 \\ \vdots & \vdots & & \vdots \\ f_{n-1} & f_{n-2} & \cdots & f_0 \end{pmatrix}, \phi(x) = x^n - 1,$$

which is known as a circulant matrix. On the other hand, ideal matrix and ideal lattice play an important role in Ajtai's construction of a collision resistant Hash function, the related materials we refer to (Ajtai, 1996; Ajtai & Dwork, 1997; Lint, 1999; Niederreiter, 1986; Plantard & Schneider, 2013; Pradhan et al., 2019).

First, we have to establish some basic properties for an ideal matrix $H^*(f)$, most of them are known when $H^*(f)$ is a circulant matrix.

Lemma 5 *Suppose H and $H^*(f)$ are given by (2.2) and (2.5), respectively, then for any $f \in \mathbb{R}^n$, we have*

$$H \cdot H^*(f) = H^*(f) \cdot H, \quad \forall f \in \mathbb{R}^n.$$

Proof Since $\phi(x) = x^n - a_{n-1}x^{n-1} - \cdots - a_1 x - a_0$ is the characteristic polynomial of H, by the Hamilton-Cayley theorem, we have

$$H^n = a_0 I_n + a_1 H + \cdots + a_{n-1} H^{n-1}. \tag{22}$$

Let

$$b = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \end{pmatrix}, \text{ and } H = \begin{pmatrix} 0 & a_0 \\ I_{n-1} & b \end{pmatrix}.$$

By (2.5) we have

$$H^*(f)H = [f, Hf, ..., H^{n-1}f] \begin{pmatrix} 0 & a_0 \\ I_{n-1} & b \end{pmatrix}$$

$$= [Hf, H^2 f, ..., H^{n-1}f, a_0 f + a_1 Hf + \cdots + a_{n-1} H^{n-1} f]$$

$$= [Hf, H^2 f, ..., H^{n-1}f, H^n f]$$

$$= H[f, Hf, ..., H^{n-1}f] = H \cdot H^*(f).$$

The lemma follows. □

Lemma 6 *For any $f = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{R}^n$ we have*

$$H^*(f) = f_0 I_n + f_1 H + \cdots + f_{n-1} H^{n-1}. \tag{23}$$

Proof We use induction on n to show this conclusion. If $n = 1$, it is trivial. Suppose it is true for n, we consider the case of $n + 1$. For this purpose, we write $H = H_n$, $e_1, e_2, ..., e_n$ the n column vectors of unit in $\mathbb{R}^n$, namely

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \cdots e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

and

$$H_{n+1} = \begin{pmatrix} 0 & A_0 \\ e_1 & Hn \end{pmatrix},$$

where $A_0 = (0, 0, ..., a_0) \in \mathbb{R}^n$ is a row vector. For any k, $1 \leqslant k \leqslant n - 1$, it is easy to check that

$$H_n e_k = e_{k+1}, H_n^k e_1 = e_{k+1} \text{ and } H_{n+1}^k = \begin{pmatrix} 0 & A_0 H_n^{k-1} \\ e_k & H_n^k \end{pmatrix}.$$

Let $f = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} \in \mathbb{R}^{n+1}$, we denote f' by

$$f' = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} \in \mathbb{R}^n, \text{ and } f = \begin{pmatrix} f_0 \\ f' \end{pmatrix}.$$

By the assumption of induction, we have

$$H_n^*(f') = [f', H_n f', ..., H_n^{n-1} f'] = f_1 I_n + f_2 H_n + \cdots + f_n H_n^{n-1}.$$

It follows that

$$H_{n+1}^*(f) = \left[\begin{pmatrix} f_0 \\ f' \end{pmatrix}, H_{n+1} \begin{pmatrix} f_0 \\ f' \end{pmatrix}, \cdots, H_{n+1}^n \begin{pmatrix} f_0 \\ f' \end{pmatrix} \right]$$

$$= f_0 I_n + f_1 H_{n+1} + \cdots + f_n H_{n+1}^n.$$

We complete the proof of Lemma 2. □

We always suppose that $\phi(x) \in Z[x]$ is a separable polynomial and $w_1, w_2, ..., w_n$ are complex number roots of $\phi(x)$, of which are different from each other. The Vandermonde matrix V_ϕ generated by $\{w_1, w_2, ..., w_n\}$ is

$$V_\phi = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ w_1 & w_2 & \cdots & w_n \\ \vdots & \vdots & & \vdots \\ w_1^{n-1} & w_2^{n-1} & \cdots & w_n^{n-1} \end{pmatrix}, \quad \text{and} \quad \det(V_\phi) \neq 0.$$

Lemma 7 *Let $f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1} \in \mathbb{R}[x]$, then we have*

$$H^*(f) = V_\phi^{-1} \, diag \, \{f(w_1), f(w_2), ..., f(w_n)\} V_\phi, \tag{24}$$

where diag $\{f(w_1), f(w_2), ..., f(w_n)\}$ is the diagonal matrix.

Proof By Theorem 3.2.5 of Davis (1994), for H, we have

$$H = V_\phi^{-1} \, \text{diag} \, \{w_1, w_2, ..., w_n\} V_\phi. \tag{25}$$

By Lemma 2, it follows that

$$H^*(f) = V_\phi^{-1} \, \text{diag} \, \{f(w_1), f(w_2), ..., f(w_n)\} V_\phi.$$

□

Now, we summarize some basic properties for ideal matrix as follows.

Theorem 2 *Let $f \in \mathbb{R}^n$, $g \in \mathbb{R}^n$ be two column vectors and $H^*(f)$ be the ideal matrix generated by f, then we have the following:*

(i) $H^*(f)H^*(g) = H^*(g)H^*(f)$.

(ii) $H^*(f)H^*(g) = H^*(H^*(f)g)$.

(iii) $det\,(H^*(f)) = \Pi_{i=1}^n f(w_i)$.

(iv) $H^*(f)$ is an invertible matrix if and only if $\phi(x)$ and $f(x)$ are coprime, i.e. $gcd\,(\phi(x), f(x)) = 1$.

Proof (i) and (ii) follow from Lemma 2 immediately, (iii) and (iv) follow from Lemma 3. Here we only give an equivalent form of (ii). Let

$$f * g = H^*(f)g. \tag{26}$$

Then by (ii) we have

$$H^*(f * g) = H^*(f)H^*(g). \tag{27}$$

□

To construct a convolutional modular lattice containing vector $\begin{pmatrix} f \\ g \end{pmatrix}$, let $\begin{pmatrix} f \\ g \end{pmatrix} \in Z^{2n}$, $(H^*(f))'$ be the transpose of $H^*(f)$, and

$$A = [(H^*(f))', (H^*(g))'] = \begin{pmatrix} f' & g' \\ f'H' & g'H' \\ f'(H')^2 & g'(H')^2 \\ \vdots & \vdots \\ f'(H')^{n-1} & g'(H')^{n-1} \end{pmatrix}_{n \times 2n}, \tag{28}$$

$$A' = \begin{pmatrix} H^*(f) \\ H^*(g) \end{pmatrix} = \begin{pmatrix} f & Hf & \cdots & H^{n-1}f \\ g & Hg & \cdots & H^{n-1}g \end{pmatrix}_{2n \times n}. \tag{29}$$

We consider A and A' as matrices over Z_q, i.e. $A \in Z_q^{n \times 2n}$, $A' \in Z_q^{2n \times n}$, a q-ary lattice $\wedge_q(A)$ is defined by (see Micciancio (2001))

$$\wedge_q(A) = \{y \in Z^{2n} \mid \text{there exists } x \in Z^n \Rightarrow y \equiv A'x (\text{mod } q)\}. \tag{30}$$

Under the above notations, we have the following.

Theorem 3 *For any column vectors $f \in Z^n$ and $g \in Z^n$, then $\wedge_q(A)$ is a convolutional modular lattice, and $\begin{pmatrix} f \\ g \end{pmatrix} \in \wedge_q(A)$.*

Proof It is known that $\wedge_q(A)$ is a q-ary lattice, i.e.

$$qZ^{2n} \subset \wedge_q(A) \subset Z^{2n}.$$

We only prove that $\wedge_q(A)$ is fixed under the linear transformation σ given by (2.4). If $y \in \wedge_q(A)$, then $y \equiv A'x (\text{mod } q)$ for some $x \in Z^n$, by Lemma 2, we have

$$\sigma(y) \equiv \begin{pmatrix} HH^*(f)x \\ HH^*(g)x \end{pmatrix} = \begin{pmatrix} H^*(f)Hx \\ H^*(g)Hx \end{pmatrix}$$

$$\equiv A'Hx (\text{mod } q).$$

It means that $\sigma(y) \in \wedge_q(A)$ whenever $y \in \wedge_q(A)$. Let

$$e = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in Z^n \Rightarrow H^*(f)e = f, \text{ and } H^*(g)e = g.$$

We have $\begin{pmatrix} f \\ g \end{pmatrix} \in \wedge_q(A)$, and Theorem 3 follows. □

Since $\wedge_q(A) \subset Z^{2n}$, then there is a unique Hermite Normal Form of basis N, which is a upper triangular matrix given by

$$N = \begin{pmatrix} I_n & H^*(h) \\ 0 & qI_n \end{pmatrix}, \text{ where } h \equiv (H^*(f))^{-1} g (\text{mod } q). \tag{31}$$

Next, we consider parameters system of NTRU. To choose the parameters of NTRU, let d_f be a positive integer and $\{p, 0, -p\}^n \subset Z^n$ be a subset of Z^n, of which has exactly $d_f + 1$ positive entries and d_f negative ones, the remaining $n - 2d_f - 1$ entries will be zero. We take some assumption conditions for the choice of parameters as follows:

(i) $\phi(x) = x^n - a_{n-1}x^{n-1} - \cdots - a_1 x - a_0 \in Z[x]$ with $a_0 \neq 0$, and $\phi(x)$ is separable polynomial, n, p, q, d_f are positive integers with n prime, $1 < p < q$ and gcd $(p, q) = 1$.
(ii) $f(x)$ and $g(x)$ are two polynomials in $Z[x]$ of degree $n - 1$, the constant term of $f(x)$ is 1, and

$$f(x) - 1 \in \{p, 0, -p\}^n, \quad g \in \{p, 0, -p\}^n.$$

(iii) $H^*(f)$ is invertible modulo q.
(iv) $d_f < (\frac{q}{2} - 1)/4p - \frac{1}{2}$.

Under the above conditions, by Lemma 2, we have

$$H^*(f) \equiv I_n (\text{mod } p), \text{ and } H^*(g) \equiv 0 (\text{mod } p). \tag{32}$$

Now, we state a generalization of NTRU as follows.

- Private key. The private key in generalized NTRU is a short vector $\begin{pmatrix} f \\ g \end{pmatrix} \in Z^{2n}$. The lattice associated with a private key is $\wedge_q(A)$, which is a convolutional modular lattice containing a private key.
- Public key. The public key of the generalized NTRU is the HNF basis N of $\wedge_q(A)$, which is given by (2.15).
- Encryption. An input message is encoded as a vector $m \in \{1, 0, -1\}^n$ with exactly $d_f + 1$ positive entries and d_f negative ones. The vector m is concatenated with a randomly chosen vector $r \in \{1, 0, -1\}^n$ also with exactly $d_f + 1$ positive entries and d_f negative ones, to obtain a short error vector $\begin{pmatrix} m \\ r \end{pmatrix} \in \{1, 0, -1\}^{2n}$. Let

$$\begin{pmatrix} c \\ 0 \end{pmatrix} = N \begin{pmatrix} m \\ r \end{pmatrix} \equiv \begin{pmatrix} m + H^*(h)r \\ 0 \end{pmatrix} (\text{mod } q), \tag{33}$$

where h is given by (2.15). Then, the n-dimensional vector c

$$c \equiv m + H^*(h)r (\text{mod } q)$$

is the ciphertext.

- Decryption. Suppose the entries of n-dimensional vector c belong to interval $[-\frac{q}{2}, \frac{q}{2}]$, then ciphertext c is decrypted by multiplying it by the secret matrix $H^*(f)$ mod q, it follows that

$$H^*(f)c \equiv H^*(f)m + H^*(f)H^*(h)r \equiv H^*(f)m + H^*(g)r (\text{mod } q). \quad (34)$$

Here, we use the identity (ii) of Theorem 2, namely

$$H^*(f)H^*(g) = H^*(H^*(f)g).$$

If the above conditions (iv) are satisfied, it is easily seen that the coordinates of vector $H^*(f)m + H^*(g)r$ are all bounded by $\frac{q}{2}$ in absolute value, or, with high probability, even for larger value of d_f. The decryption process is completed by reducing (2.18) modulo p, to obtain

$$H^*(f)m + H^*(g)r \equiv mI_n (\text{mod } p).$$

Thus, one gets plaintext m from ciphertext c.

References

Ajtai, M. (1996). Generating hard instances of lattice problems. In *Proceedings of the 28th STOC* (pp. 99–108).

Ajtai, M., & Dwork, C. (1997). A public-key cryptosystem with worst-case/average-case equivalence. In *Proceedings of the 29th STOC* (pp. 284–293).

Coppersmith D., & Shamir, A. (1997). Lattice attacks on NTRU. In W. Fumy (Eds.), *Advances in cryptology, lecture notes in computer science* (Vol. 1233, pp. 52–61). Springer.

Davis, P. J. (1994). *Circulant matrices* (2nd ed.). Chelsea Publishing.

Hoffstein, J., Pipher, J., & Schanck, J. M., et al. (2017) Choosing parameters for NTRUEncrypt. In H. Handschuh (Eds.), *Topics in cryptology, lecture notes in computer science* (Vol. 10159, pp. 3–18). Springer.

Hoffstein, J., Pipher, J., & Silverman, J. H. (1998). NTRU: A ring based public key cryptosystem. In J. P. Buhler (Eds.), *Algorithmic number theory, lecture notes in computer science* (Vol. 1423, pp. 267–288). Springer.

IEEE Computer Society. (2000). IEEE standard specifications for public-key cryptography. *IEEE Std, 1363–2000*, 1–228.

Lidl, R., & Niederreiter, H. (1983). Finite fields. *Encyclopedia of Mathematics and Its Applications, 20*.

Lopez-Permouth, S. R., Parra-Avila, B. R., & Szabo, S. (2009). Dual generalizations of the concept of cyclicity of codes. *Advances in Mathematics of Communications, 3*(3), 227–234.

Lint, J. H. V. (1999) *Introduction to coding theory* (Vol. 86). Springer. GTM.
Lyubashevsky, V., & Micciancio, D. (2006) Generalized compact knapsacks are collision resistant. In M. Bugliesi, B. Preneel, V. Sassone & I. Wegener (Eds.), *Automata, languages and programming, lecture notes in computer science* (Vol. 4052, pp. 144–155). Springer.
McEliece, R. J. (1978). *A public-key cryptosystem based on algebraic coding theory* (pp. 42–44). DSN Progress Report: Jet Propulsion Laboratory.
Micciancio, D. (2001). Improving lattice based cryptosystems using the Hermite normal form. In J. H. Silverman (Eds.), *Cryptography and lattices, lecture notes in computer science* (Vol. 2146, pp. 126–145). Springer.
Micciancio, D., & Regev, O. (2009). Lattice-based cryptography. In D. J. Bernstein, J. Buchmann & E. Dahmen (Eds.), *Post-quantum cryptography* (pp. 147–191). Springer.
Niederreiter, H. (1986). Knapsack-type cryptosystems and algebraic coding theory. *Problems of Control and Information Theory, 15*, 159–166.
Plantard T., & Schneider, M. (2013). Creating a challenge for ideal lattices (pp. 1–17).
Pradhan, P. K., Rakshit, S., & Datta, S. (2019) Lattice based cryptography: Its applications, areas of interest and future scope. In *Proceedings of the Third International Conference on Computing Methodologies and Communication* (pp. 988–993).
Stehle, D., & Steinfeld, R. (2011) Making NTRU as secure as worst-case problems over ideal lattices. In K. G. Paterson (Eds.), *Advances in cryptology, lecture notes in computer science* (Vol. 6632, pp. 27–47). Springer.
Shi, M., Li, X., Sepasdar, Z., et al. (2020). Polycyclic codes as invariant subspaces. *Finite Fields and Their Applications, 68*.

Cyclic Lattices, Ideal Lattices, and Bounds for the Smoothing Parameter

Zheng Zhiyong, Liu Fengxia, Lu Yunfan, and Tian Kun

Abstract Cyclic lattices and ideal lattices were introduced by Micciancio (2002), Lyubashevsky and Micciancio (2006), respectively, which play an efficient role in Ajtai's construction of a collision resistant Hash function (see Ajtai (1996), Ajtai and Dwork (1997)) and in Gentry's construction of fully homomorphic encryption (see Gentry (2009)). Let $R = Z[x]/\langle \phi(x) \rangle$ be a quotient ring of the integer coefficients polynomials ring, Lyubashevsky and Micciancio regarded an ideal lattice as the correspondence of an ideal of R, but they neither explain how to extend this definition to whole Euclidean space $\mathbb{R}^n$, nor exhibit the relationship of cyclic lattices and ideal lattices. In this chapter, we regard the cyclic lattices and ideal lattices as the correspondences of finitely generated R-modules, so that we may show that ideal lattices are actually a special subclass of cyclic lattices, namely, cyclic integer lattices. In fact, there is a one to one correspondence between cyclic lattices in $\mathbb{R}^n$ and finitely generated R-modules (see Theorem 4). On the other hand, since R is a Noether ring, each ideal of R is a finitely generated R-module, so it is natural and reasonable to regard ideal lattices as a special subclass of cyclic lattices (see Corollary 7). It is worth noting that we use a more general rotation matrix here, so our definition and results on cyclic lattices and ideal lattices are more general forms. As an application, we provide a cyclic lattice with an explicit and countable upper bound for the smoothing parameter (see Theorem 5). It is an open problem that is the shortest vector problem on cyclic lattice NP-hard (see Micciancio (2002)). Our results may be viewed as a substantial progress in this direction.

Z. Zhiyong · L. Fengxia (✉) · L. Yunfan · T. Kun
Engineering Research Center of Ministry of Education for Financial Computing and Digital Engineering, Renmin University of China,Beijing 100872, China
e-mail: liu_fx@ruc.edu.cn

Z. Zhiyong
e-mail: zhengzy@ruc.edu.cn

L. Yunfan
e-mail: luyunfan@ruc.edu.cn

T. Kun
e-mail: tkun19891208@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_7

Keywords Cyclic lattice · Ideal lattice · Finitely generated R-module · Smoothing parameter

1 Discrete Subgroup in $\mathbb{R}^n$

Let $\mathbb{R}$ be the real numbers field, $\mathbb{Z}$ be the integers ring, and $\mathbb{R}^n$ be Euclidean space of which is an n-dimensional linear space over $\mathbb{R}$ with the Euclidean norm $|x|$ given by

$$|x| = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}, \quad \text{where } x' = (x_1, x_1, \cdots, x_n) \in \mathbb{R}^n.$$

We use column vector notation for $\mathbb{R}^n$ through out this chapter, and $x' = (x_1, x_2, \ldots, x_n)$ is transpose of x, which is called row vector of $\mathbb{R}^n$.

Definition 1 Let $L \subset \mathbb{R}^n$ be a non-trivial additive subgroup, it is called a discrete subgroup if there is a positive real number $\lambda > 0$ such that

$$\min_{x \in L, x \neq 0} |x| \geqslant \lambda > 0. \tag{1}$$

As usual, a ball of center x_0 with radius δ is defined by

$$b(x_0, \delta) = \{x \in \mathbb{R}^n \;\Big|\; |x - x_0| \leqslant \delta\}.$$

If L is a discrete subgroup of $\mathbb{R}^n$, then there are only finitely many vectors of L lie in every ball $b(0, \delta)$, thus we always find a vector $\alpha \in L$ such that

$$|\alpha| = \min_{x \in L, x \neq 0} |x| = \lambda > 0, \quad \alpha \in L. \tag{2}$$

α is called one of shortest vector of L and λ is called the minimum distance of L.

Let $B = [\beta_1, \beta_2, \ldots, \beta_m] \in \mathbb{R}^{n \times m}$ be a $n \times m$ dimensional matrix with $\text{rank}(B) = m \leqslant n$, it means that $\beta_1, \beta_2, \ldots, \beta_m$ are m linearly independent vectors in $\mathbb{R}^n$. The lattice $L(B)$ generated by B is defined by

$$L(B) = \sum_{i=1}^{m} x_i \beta_i = \{Bx \mid x \in \mathbb{Z}^m\}, \quad \forall x_i \in \mathbb{Z}, \tag{3}$$

which is all linear combinations of $\beta_1, \beta_2, \ldots, \beta_m$ over $\mathbb{Z}$. If $m = n$, $L(B)$ is called a full-rank lattice.

It is a well-known conclusion that a discrete subgroup L in $\mathbb{R}^n$ is just a lattice $L(B)$. Firstly, we give a detailed proof here by making use of the simultaneous Diophantine approximation theory in real number field $\mathbb{R}$ (see Cassels (1971) and Cassels (1963)).

Lemma 1 *Let $L \subset \mathbb{R}^n$ be a discrete subgroup, $\alpha_1, \alpha_2, \ldots, \alpha_m \in L$ be m vectors of L. Then $\alpha_1, \alpha_2, \ldots, \alpha_m$ are linearly independent over $\mathbb{R}$, if and only if which are linearly independent over $\mathbb{Z}$.*

Proof If $\alpha_1, \alpha_2, \ldots, \alpha_m$ are linearly independent over $\mathbb{R}$, trivially which are linearly independent over $\mathbb{Z}$. Suppose that $\alpha_1, \alpha_2, \ldots, \alpha_m$ are linearly independent over $\mathbb{Z}$, we consider arbitrary linear combination over $\mathbb{R}$. Let

$$a_1\alpha_1 + a_2\alpha_2 + \cdots + a_m\alpha_m = 0, \quad \forall a_i \in \mathbb{R}. \tag{4}$$

We should prove (1.4) is equivalent to $a_1 = a_2 = \cdots = a_m = 0$, which implies that $\alpha_1, \alpha_2, \ldots, \alpha_m$ are linearly independent over $\mathbb{R}$.

By Minkowski's Third Theorem (see Theorem VII of Cassels (1963)), for any sufficiently large $N > 1$, there are a positive integer $q \geqslant 1$ and integers $p_1, p_2, \ldots, p_m \in \mathbb{Z}$ such that

$$\max_{1 \leqslant i \leqslant m} |qa_i - p_i| < N^{-\frac{1}{m}}, \text{ and } 1 \leqslant q \leqslant N. \tag{5}$$

By (1.4), we have

$$|p_1\alpha_1 + p_2\alpha_2 + \cdots + p_m\alpha_m| = |(qa_1 - p_1)\alpha_1 + (qa_2 - p_2)\alpha_2 + \cdots + (qa_m - p_m)\alpha_m|$$

$$\leqslant mN^{-\frac{1}{m}} \max_{1 \leqslant i \leqslant m} |\alpha_i|. \tag{6}$$

Let λ be the minimum distance of L, $\varepsilon > 0$ be any positive real number. We select N such that

$$N > \max\{(\frac{m}{\varepsilon})^m, \ (\frac{m}{\lambda})^m \max_{1 \leqslant i \leqslant m} |\alpha_i|^m\}.$$

It follows that $mN^{-\frac{1}{m}} < \varepsilon$ and

$$mN^{-\frac{1}{m}} \max_{1 \leqslant i \leqslant m} |\alpha_i| < \lambda.$$

By (1.6) we have

$$|p_1\alpha_1 + p_2\alpha_2 + \cdots + p_m\alpha_m| < \lambda.$$

Since $p_1\alpha_1 + p_2\alpha_2 + \cdots + p_m\alpha_m \in L$, thus we have $p_1\alpha_1 + p_2\alpha_2 + \cdots + p_m\alpha_m = 0$, and $p_1 = p_2 = \cdots = p_m = 0$. By (1.5) we have $q|a_i| < \frac{1}{m}\varepsilon$ for all i, $1 \leqslant i \leqslant m$. Since ε is a sufficiently small positive number, we must have $a_1 = a_2 = \cdots = a_m = 0$. We complete the proof of lemma.

Suppose that $B \in \mathbb{R}^{n\times m}$ is an $n \times m$-dimensional matrix and $\text{rank}(B) = m$, B' is the transpose of B. It is easy to verify

$$\text{rank}(B'B) = \text{rank}(B) = m \Rightarrow \det(B'B) \neq 0,$$

which implies that $B'B$ is an invertible square matrix of $m \times m$ dimension. Since $B'B$ is a positive defined symmetric matrix, then there is an orthogonal matrix $P \in \mathbb{R}^{m\times m}$ such that

$$P'B'BP = \text{diag}\{\delta_1, \delta_2, \ldots, \delta_m\}, \tag{7}$$

where $\delta_i > 0$ are the characteristic value of $B'B$, and $\text{diag}\{\delta_1, \delta_2, \ldots, \delta_m\}$ is the diagonal matrix of $m \times m$ dimension.

Lemma 2 *Suppose that $B \in \mathbb{R}^{n\times m}$ with $\text{rank}(B) = m$, $\delta_1, \delta_2, \ldots, \delta_m$ are m characteristic values of $B'B$, and $\lambda(L(B))$ is the minimum distance of lattice $L(B)$, then we have*

$$\lambda(L(B)) = \min_{x\in\mathbb{Z}^m,\ x\neq 0} |Bx| \geqslant \sqrt{\delta}, \tag{8}$$

where $\delta = \min\{\delta_1, \delta_2, \ldots, \delta_m\}$.

Proof Let $A = B'B$, by (1.7), there exists an orthogonal matrix $P \in \mathbb{R}^{m\times m}$ such that

$$P'AP = \text{diag}\{\delta_1, \delta_2, \ldots, \delta_m\}.$$

If $x \in \mathbb{Z}^m$, $x \neq 0$, we have

$$\begin{aligned} |Bx|^2 = x'Ax &= x'P(P'AP)P'x \\ &= (P'x)'\,\text{diag}\{\delta_1, \delta_2, \ldots, \delta_m\}P'x \\ &\geqslant \delta|P'x|^2 = \delta|x|^2. \end{aligned}$$

Since $x \in \mathbb{Z}^m$ and $x \neq 0$, we have $|x|^2 \geqslant 1$, it follows that

$$\min_{x\in\mathbb{Z}^m,\ x\neq 0} |Bx| \geqslant \sqrt{\delta}|x| \geqslant \sqrt{\delta}.$$

We have Lemma 2 immediately.

Another application of Lemma 2 is to give a countable upper bound for smoothing parameter (see Theorem 5). Combining Lemmas 1 and 2, we show the following assertion.

Theorem 1 *Let $L \subset \mathbb{R}^n$ be a subset, then L is a discrete subgroup if and only if there is an $n \times m$ dimensional matrix $B \in \mathbb{R}^{n\times m}$ with $\text{rank}(B) = m$ such that*

$$L = L(B) = \{Bx \mid x \in \mathbb{Z}^m\}. \tag{9}$$

Proof If $L \subset \mathbb{R}^n$ is a discrete subgroup, then L is a free $\mathbb{Z}$-module. By Lemma 1, we have $\text{rank}_{\mathbb{Z}}(L) = m \leqslant n$. Let $\beta_1, \beta_2, \ldots, \beta_m$ be a $\mathbb{Z}$-basis of L, then

$$L = \left\{ \sum_{i=1}^{m} a_i \beta_i \mid a_i \in \mathbb{Z} \right\}.$$

Writing $B = [\beta_1, \beta_2, \ldots, \beta_m]_{n \times m}$, then the rank of matrix B is m, and

$$L = \{Bx \mid x \in \mathbb{Z}^m\} = L(B).$$

Conversely, let $L(B)$ be arbitrary lattice generated by B, obviously, $L(B)$ is an additive subgroup of $\mathbb{R}^n$, by Lemma 2, $L(B)$ is also a discrete subgroup, we have Theorem 1 at once.

Corollary 1 *Let $L \subset \mathbb{R}^n$ be a lattice and $G \subset L$ be an additive subgroup of L, then G is a lattice of $\mathbb{R}^n$.*

Corollary 2 *Let $L \subset \mathbb{Z}^n$ be an additive subgroup, then L is a lattice of $\mathbb{R}^n$. These lattices are called integer lattices.*

According to above Theorem 1, a lattice $L(B)$ is equivalent to a discrete subgroup of $\mathbb{R}^n$. Suppose $L = L(B)$ is a lattice with generated matrix $B \in \mathbb{R}^{n \times m}$, and $\text{rank}(B) = m$, we write $\text{rank}(L) = \text{rank}(B)$, and

$$d(L) = \sqrt{\det(B'B)}. \tag{10}$$

In particular, if $\text{rank}(L) = n$ is a full-rank lattice, then $d(L) = |\det(B)|$ as usual. A sublattice N of L means a discrete additive subgroup of L, the quotient group is written by L/N, and the cardinality of L/N is denoted by $|L/N|$.

Lemma 3 *Let $L \subset \mathbb{R}^n$ be a lattice and $N \subset L$ be a sublattice. If $\text{rank}(N) = \text{rank}(L)$, then the quotient group L/N is a finite group.*

Proof Let $\text{rank}(L) = m$, and $L = L(B)$, where $B \in \mathbb{R}^{n \times m}$ with $\text{rank}(B) = m$. We define a mapping σ from L to $\mathbb{Z}^m$ by $\sigma(Bx) = x$. Clearly, σ is an additive group isomorphism, $\sigma(N) \subset \mathbb{Z}^m$ is a full-rank lattice of $\mathbb{Z}^m$, and $L/N \cong \mathbb{Z}^m/\sigma(N)$. It is a well-known result that

$$|\mathbb{Z}^m/\sigma(N)| = d(\sigma(N)).$$

It follows that

$$|L/N| = |\mathbb{Z}^m/\sigma(N)| = d(\sigma(N)).$$

Lemma 3 follows.

Suppose that $L_1 \subset \mathbb{R}^n$, $L_2 \subset \mathbb{R}^n$ are two lattices of $\mathbb{R}^n$, we define $L_1 + L_2 = \{a + b | a \in L_1, b \in L_2\}$. Obviously, $L_1 + L_2$ is an additive subgroup of $\mathbb{R}^n$, but generally speaking, $L_1 + L_2$ is not a lattice of $\mathbb{R}^n$ again.

Lemma 4 *Let $L_1 \subset \mathbb{R}^n$, $L_2 \subset \mathbb{R}^n$ be two lattices of $\mathbb{R}^n$. If rank$(L_1 \cap L_2)$ =rank(L_1) or rank$(L_1 \cap L_2)$ =rank(L_2), then $L_1 + L_2$ is again a lattice of $\mathbb{R}^n$.*

Proof To prove $L_1 + L_2$ is a lattice of $\mathbb{R}^n$, by Theorem 1, it is sufficient to prove $L_1 + L_2$ is a discrete subgroup of $\mathbb{R}^n$. Suppose that rank$(L_1 \cap L_2)$ =rank(L_1), for any $x \in L_1$, we define a distance function $\rho(x)$ by

$$\rho(x) = \inf\{|x - y| \Big| y \neq x, \ y \in L_2\}.$$

Since there are only finitely many vectors in $L_2 \cap b(x, \delta)$, where $b(x, \delta)$ is any a ball of center x with radius δ. Therefore, we have

$$\rho(x) = \min\{|x - y| \Big| y \neq x, \ y \in L_2\} = \lambda_x > 0. \tag{11}$$

On the other hand, if $x_1 \in L_1$, $x_2 \in L_1$, and $x_1 - x_2 \in L_2$, then there is $y_0 \in L_2$ such that $x_1 = x_2 + y_0$, and we have $\rho(x_1) = \rho(x_2)$. It means that $\rho(x)$ is defined over the quotient group $L_1 + L_2/L_2$. Because we have the following group isomorphic theorem

$$L_1 + L_2/L_2 \cong L_1/L_1 \cap L_2.$$

By Lemma 3, it follows that

$$|L_1 + L_2/L_2| = |L_1/L_1 \cap L_2| < \infty.$$

In other words, $L_1 + L_2/L_2$ is also a finite group. Let $x_1, x_2, \ldots, x_k$ be the representative elements of $L_1 + L_2/L_2$, we have

$$\min_{x \in L_1, y \in L_2, x \neq y} |x - y| = \min_{1 \leqslant i \leqslant k} \rho(x_i) \geqslant \min\{\lambda_{x_1}, \lambda_{x_2}, \ldots, \lambda_{x_k}\} > 0.$$

Therefore, $L_1 + L_2$ is a discrete subgroup of $\mathbb{R}^n$, thus it is a lattice of $\mathbb{R}^n$ by Theorem 1.

Remark 1 The condition rank$(L_1 \cap L_2)$ =rank(L_1) or rank$(L_1 \cap L_2)$ =rank(L_2) in Lemma 4 seems to be necessary. As a counterexample, we see the real line $\mathbb{R}$, let $L_1 = \mathbb{Z}$ and $L_2 = \sqrt{2}\mathbb{Z}$, then $L_1 + L_2$ is not a discrete subgroup of $\mathbb{R}$, thus $L_1 + L_2$ is not a lattice in $\mathbb{R}$. Because $L_1 + L_2 = \{n + \sqrt{2}m \big| n \in \mathbb{Z}, m \in \mathbb{Z}\}$ is dense in $\mathbb{R}$ by Dirichlet's Theorem (see Theorem I of Cassels (1963)).

As a direct consequence, we have the following generalized form of Lemma 4.

Corollary 3 *Let $L_1, L_2, \ldots, L_m$ be m lattices of $\mathbb{R}^n$ and*

$$rank(L_1 \cap L_2 \cap \cdots \cap L_m) = rank(L_j) \text{ for some } 1 \leqslant j \leqslant m.$$

Then $L_1 + L_2 + \cdots + L_m$ is a lattice of $\mathbb{R}^n$.

Proof Without loss of generality, we assume that

$$\text{rank}(L_1 \cap L_2 \cap \cdots \cap L_m) = \text{rank}(L_m).$$

Let $L_1 + L_2 + \cdots + L_{m-1} = L'$, then

$$L' + L_m / L' \cong L_m / L' \cap L_m.$$

Since rank$(L' \cap L_m)$ =rank(L_m), by Lemma 4, we have $L' + L_m = L_1 + L_2 + \cdots + L_m$ is a lattice of $\mathbb{R}^n$ and the corollary follows.

2 Ideal Matrices

Let $\mathbb{R}[x]$ and $\mathbb{Z}[x]$ be the polynomials rings over $\mathbb{R}$ and $\mathbb{Z}$ with variable x, respectively. Suppose that

$$\phi(x) = x^n - \phi_{n-1}x^{n-1} - \cdots - \phi_1 x - \phi_0 \in \mathbb{Z}[x],\ \phi_0 \neq 0, \tag{12}$$

is a polynomial with integer coefficients of which has no multiple roots in complex numbers field $\mathbb{C}$. Let $w_1, w_2, \ldots, w_n$ be the n different roots of $\phi(x)$ in $\mathbb{C}$, the Vandermonde matrix V_ϕ is defined by

$$V_\phi = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ w_1 & w_2 & \cdots & w_n \\ \vdots & \vdots & & \vdots \\ w_1^{n-1} & w_2^{n-1} & \cdots & w_n^{n-1} \end{pmatrix}, \quad \text{and} \quad \det(V_\phi) \neq 0. \tag{13}$$

According to the given polynomial $\phi(x)$, we define a rotation matrix $H = H_\phi$ by

$$H = H_\phi = \left(\begin{array}{ccc|c} 0 & \cdots & 0 & \phi_0 \\ \hline & & & \phi_1 \\ & I_{n-1} & & \vdots \\ & & & \phi_{n-1} \end{array}\right)_{n\times n} \in \mathbb{Z}^{n\times n}, \tag{14}$$

where I_{n-1} is the $(n-1) \times (n-1)$ unit matrix. Obviously, the characteristic polynomial of H is just $\phi(x)$.

We use column notation for vectors in $\mathbb{R}^n$, for any $f = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{R}^n$, the ideal matrix generated by vector f is defined by

$$H^*(f) = [f, Hf, H^2 f, \ldots, H^{n-1} f]_{n\times n} \in \mathbb{R}^{n\times n}, \tag{15}$$

which is a block matrix in terms of each column $H^k f$ $(0 \leqslant k \leqslant n-1)$. Sometimes, f is called an input vector. It is easily seen that $H^*(f)$ is a more general form of the classical circulant matrix (see Davis (1994)) and r-circulant matrix (see Shi (2018), Yasin and Taskara (2013)). In fact, if $\phi(x) = x^n - 1$, then $H^*(f)$ is the ordinary circulant matrix generated by f. If $\phi(x) = x^n - r$, then $H^*(f)$ is the r-circulant matrix.

By (2.4), it follows immediately that

$$H^*(f+g) = H^*(f) + H^*(g), \text{ and } H^*(\lambda f) = \lambda H^*(f), \ \forall \lambda \in \mathbb{R}. \tag{16}$$

Moreover, $H^*(f) = 0$ is a zero matrix if and only if $f = 0$ is a zero vector, thus one has $H^*(f) = H^*(g)$ if and only if $f = g$. Let M^* be the set of all ideal matrices, namely

$$M^* = \{H^*(f) \mid f \in \mathbb{R}^n\}. \tag{17}$$

We may regard H^* as a mapping from $\mathbb{R}^n$ to M^* of which is a one to one correspondence.

In Zheng et al. (2023), we have shown some basic properties of ideal matrix, most of them may be summarized as the following theorem.

Theorem 2 *Suppose that $\phi(x) \in \mathbb{Z}[x]$ is a fixed polynomial with no multiple roots in $\mathbb{C}$, then for any two column vectors f and g in $\mathbb{R}^n$, we have*

(i) $H^*(f) = f_0 I_n + f_1 H + \cdots + f_{n-1} H^{n-1}$;
(ii) $H^*(f)H^*(g) = H^*(H^*(f)g)$ *and* $H^*(f)H^*(g) = H^*(g)H^*(f)$;
(iii) $H^*(f) = V_\phi^{-1}\, diag\{f(w_1), f(w_2), \ldots, f(w_n)\}V_\phi$;
(iv) $det\,(H^*(f)) = \Pi_{i=1}^n f(w_i)$;
(v) $H^*(f)$ *is an invertible matrix if and only if* $(f(x), \phi(x)) = 1$ *in* $\mathbb{R}[x]$,

where V_ϕ is the Vandermonde matrix given by (2.2), w_i $(1 \leqslant i \leqslant n)$ are all roots of $\phi(x)$ in $\mathbb{C}$, and $diag\{f(w_1), f(w_2), \ldots, f(w_n)\}$ is the diagonal matrix.

Proof See Theorem 2 of Zheng et al. (2023).

Let $e_1, e_2, \ldots, e_n$ be unit vectors of $\mathbb{R}^n$, that is

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \cdots, e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

It is easy to verify that

$$H^*(e_1) = I_n, \text{ and } H^*(e_k) = H^{k-1}, \ 1 \leqslant k \leqslant n. \tag{18}$$

This means that the unit matrix I_n and rotation matrices H^k $(1 \leqslant k \leqslant n-1)$ are all the ideal matrices.

Let $\phi(x)\mathbb{R}[x]$ and $\phi(x)\mathbb{Z}[x]$ be the principal ideals generated by $\phi(x)$ in $\mathbb{R}[x]$ and $\mathbb{Z}[x]$, respectively, we denote the quotient rings R and $\overline{R}$ by

$$R = \mathbb{Z}[x]/\phi(x)\mathbb{Z}[x], \text{ and } \overline{R} = \mathbb{R}[x]/\phi(x)\mathbb{R}[x]. \tag{19}$$

There is a one to one correspondence between $\overline{R}$ and $\mathbb{R}^n$ given by

$$f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1} \in \overline{R} \xrightarrow{t} f = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{R}^n.$$

We denote this correspondence by t, that is

$$t(f(x)) = f \text{ and } t^{-1}(f) = f(x), \ \forall f(x) \in \overline{R}, \text{ and } f \in \mathbb{R}^n. \tag{20}$$

If we restrict t in the quotient ring R, then which gives a one to one correspondence between R and $\mathbb{Z}^n$. First, we show that t is also a ring isomorphism.

Definition 2 For any two column vectors f and g in $\mathbb{R}^n$, we define the ϕ-convolutional product $f * g$ by $f * g = H^*(f)g$.

By Theorem 2, it is easy to see that

$$f * g = g * f, \text{ and } H^*(f * g) = H^*(f)H^*(g). \tag{21}$$

Lemma 5 *For any two polynomials $f(x)$ and $g(x)$ in $\overline{R}$, we have*

$$t(f(x)g(x)) = H^*(f)g = f * g.$$

Proof Let $g(x) = g_0 + g_1 x + \cdots + g_{n-1}x^{n-1} \in \overline{R}$, then

$$xg(x) = \phi_0 g_{n-1} + (g_0 + \phi_1 g_{n-1})x + \cdots + (g_{n-2} + \phi_{n-1}g_{n-1})x^{n-1}.$$

It follows that

$$t(xg(x)) = Ht(g(x)) = Hg. \tag{22}$$

Hence, for any $0 \leqslant k \leqslant n-1$, we have

$$t(x^k g(x)) = H^k t(g(x)) = H^k g,\ 0 \leqslant k \leqslant n-1. \tag{23}$$

Let $f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1} \in \overline{R}$, by (i) of Theorem 2, we have

$$t(f(x)g(x)) = \sum_{i=0}^{n-1} f_i t(x^i g(x)) = \sum_{i=0}^{n-1} f_i H^i g = H^*(f)g.$$

The lemma follows.

Theorem 3 *Under ϕ-convolutional product, $\mathbb{R}^n$ is a commutative ring with identity element e_1 and $\mathbb{Z}^n \subset \mathbb{R}^n$ is its subring. Moreover, we have the following ring isomorphisms:*

$$\overline{R} \cong \mathbb{R}^n \cong M^*, \text{ and } R \cong \mathbb{Z}^n \cong M_{\mathbb{Z}}^*,$$

where M^* is the set of all ideal matrices given by (2.6), and $M_{\mathbb{Z}}^*$ is the set of all integer ideal matrices.

Proof Let $f(x) \in \overline{R}$ and $g(x) \in \overline{R}$, then

$$t(f(x) + g(x)) = f + g = t(f(x)) + t(g(x)),$$

and

$$t(f(x)g(x)) = H^*(f)g = f * g = t(f(x)) * t(g(x)).$$

This means that t is a ring isomorphism. Since $f * g = g * f$ and $e_1 * g = H^*(e_1)g = I_n g = g$, then $\mathbb{R}^n$ is a commutative ring with e_1 as the identity elements. Noting $H^*(f)$ is an integer matrix if and only if $f \in \mathbb{Z}^n$ is an integer vector, the isomorphism of subrings follows immediately.

According to property (v) of Theorem 2, $H^*(f)$ is an invertible matrix whenever $(f(x), \phi(x)) = 1$ in $\mathbb{R}[x]$, we show that the inverse of an ideal matrix is again an ideal matrix.

Lemma 6 *Let $f(x) \in \overline{R}$ and $(f(x), \phi(x)) = 1$ in $\mathbb{R}[x]$, then*

$$(H^*(f))^{-1} = H^*(u),$$

where $u(x) \in \overline{R}$ is the unique polynomial such that $u(x)f(x) \equiv 1 \ (mod\ \phi(x))$.

Proof By Lemma 5, we have $u * f = e_1$, it follows that

$$H^*(u)H^*(f) = H^*(e_1) = I_n.$$

Thus we have $(H^*(f))^{-1} = H^*(u)$. It is worth to note that if $H^*(f)$ is an invertible integer matrix, then $(H^*(f))^{-1}$ is not an integer matrix in general.

Sometimes, the following lemma may be useful, especially, when we consider an integer matrix.

Lemma 7 *Let* $f(x) \in \mathbb{Z}[x]$ *and* $(f(x), \phi(x)) = 1$ *in* $\mathbb{Z}[x]$, *then we have* $(f(x), \phi(x)) = 1$ *in* $\mathbb{R}[x]$.

Proof Let Q be the rational number field. Since $(f(x), \phi(x)) = 1$ in $\mathbb{Z}[x]$, then $(f(x), \phi(x)) = 1$ in $\mathbb{Q}[x]$. We know that $\mathbb{Q}[x]$ is a principal ideal domain, thus there are two polynomials $a(x)$ and $b(x)$ in $\mathbb{Q}[x]$ such that

$$a(x)f(x) + b(x)\phi(x) = 1.$$

This means that $(f(x), \phi(x)) = 1$ in $\mathbb{R}[x]$.

3 Cyclic Lattices and Ideal Lattices

As we know that cyclic code plays a central role in the algebraic coding theorem (see Chap. 6 of Lint (1999)). In Zheng et al. (2023), we extended ordinary cyclic code to more general forms, namely ϕ-cyclic codes. To obtain an analogous concept of ϕ-cyclic code in $\mathbb{R}^n$, we note that every rotation matrix H defines a linear transformation of $\mathbb{R}^n$ by $x \to Hx$.

Definition 3 A linear subspace $C \subset \mathbb{R}^n$ is called a ϕ-cyclic subspace if $\forall \alpha \in C \Rightarrow H\alpha \in C$. A lattice $L \subset \mathbb{R}^n$ is called a ϕ-cyclic lattice if $\forall \alpha \in L \Rightarrow H\alpha \in L$.

In other words, a ϕ-cyclic subspace C is a linear subspace of $\mathbb{R}^n$, of which is closed under linear transformation H. A ϕ-cyclic lattice L is a lattice of $\mathbb{R}^n$ of which is closed under H. If $\phi(x) = x^n - 1$, then H is the classical circulant matrix and the corresponding cyclic lattice first appeared in Micciancio (2002), but he does not discuss the further property for these lattices. To obtain the explicit algebraic construction of ϕ-cyclic lattice, we first show that there is a one to one correspondence between ϕ-cyclic subspaces of $\mathbb{R}^n$ and the ideals of $\overline{R}$.

Lemma 8 *Let* t *be the correspondence between* $\overline{R}$ *and* $\mathbb{R}^n$ *given by (2.9), then a subset* $C \subset \mathbb{R}^n$ *is a* ϕ*-cyclic subspace of* $\mathbb{R}^n$, *if and only if* $t^{-1}(C) \subset \overline{R}$ *is an ideal.*

Proof We extend the correspondence t to subsets of $\overline{R}$ and $\mathbb{R}^n$ by

$$C(x) \subset \overline{R} \xrightarrow{t} C = \{c | c(x) \in C(x)\} \subset \mathbb{R}^n. \tag{24}$$

Let $C(x) \subset \overline{R}$ be an ideal, it is clear that $C \subset t(C(x))$ is a linear subspace of $\mathbb{R}^n$. To prove C is a ϕ-cyclic subspace, we note that if $c(x) \in C(x)$, then by (2.11)

$$xc(x) \in C(x) \Leftrightarrow Ht(c(x)) = Hc \in C.$$

Therefore, if $C(x)$ is an ideal of $\overline{R}$, then $t(C(x)) = C$ is a ϕ-cyclic subspace of $\mathbb{R}^n$. Conversely, if $C \subset \mathbb{R}^n$ is a ϕ-cyclic subspace, then for any $k \geqslant 1$, we have $H^k c \in C$ whenever $c \in C$, it implies

$$\forall c(x) \in C(x) \Rightarrow x^k c(x) \in C(x),\ 0 \leqslant k \leqslant n-1,$$

which means that $C(x)$ is an ideal of $\overline{R}$. We complete the proof.

By the above lemma, to find a ϕ-cyclic subspace in $\mathbb{R}^n$, it is enough to find an ideal of $\overline{R}$. There are two trivial ideals $C(x) = 0$ and $C(x) = \overline{R}$, the corresponding ϕ-cyclic subspace are $C = 0$ and $C = \mathbb{R}^n$. To find non-trivial ϕ-cyclic subspaces, we make use of the homomorphism theorems, which is a standard technique in algebra. Let π be the natural homomorphism from $\mathbb{R}[x]$ to $\overline{R}$, $\ker\pi = \phi(x)\mathbb{R}[x]$. We write $\phi(x)\mathbb{R}[x]$ by $< \phi(x) >$. Let N be an ideal of $\mathbb{R}[x]$ satisfying

$$< \phi(x) > \subset N \subset \mathbb{R}[x] \xrightarrow{\pi} \overline{R} = \mathbb{R}[x]/ < \phi(x) > . \tag{25}$$

Since $\mathbb{R}[x]$ is a principal ideal domain, then $N =< g(x) >$ is a principal ideal generated by a monic polynomial $g(x) \in \mathbb{R}[x]$. It is easy to see that

$$< \phi(x) > \subset < g(x) > \Leftrightarrow g(x) | \phi(x) \text{ in } \mathbb{R}[x].$$

It follows that all ideals N satisfying (2) are given by

$$\{< g(x) > \Big| \, g(x) \in \mathbb{R}[x] \text{ is monic and } g(x) | \phi(x)\}.$$

We write by $< g(x) > \bmod \phi(x)$, the image of $< g(x) >$ under π, i.e.

$$< g(x) > \ \bmod \phi(x) = \pi(< g(x) >).$$

It is easy to check

$$< g(x) > \ \bmod \phi(x) = \{a(x)g(x) \mid a(x) \in \mathbb{R}[x] \text{ and } \deg a(x) + \deg g(x) < n\}, \tag{26}$$

more precisely, which is a representative elements set of $< g(x) >$ mod $\phi(x)$. By homomorphism theorem in ring theory, all ideals of $\overline{R}$ are given by

$$\{< g(x) > \text{ mod } \phi(x) \mid g(x) \in \mathbb{R}[x] \text{ is monic and } g(x)|\phi(x)\}. \tag{27}$$

Let d be the number of monic divisors of $\phi(x)$ in $\mathbb{R}[x]$, we have the following.

Corollary 4 *The number of ϕ-cyclic subspace of $\mathbb{R}^n$ is d.*

Next, we discuss ϕ-cyclic lattice, which is the geometric analogy of cyclic code. The ϕ-cyclic subspace of $\mathbb{R}^n$ may be regarded as the algebraic analogy of cyclic code. Let the quotient rings R and $\overline{R}$ be given by (2.8). A R-module is an Abel group $\wedge$ such that there is an operator $\lambda\alpha \in \wedge$ for all $\lambda \in R$ and $\alpha \in \wedge$, satisfying $1 \cdot \alpha = \alpha$ and $(\lambda_1\lambda_2)\alpha = \lambda_1(\lambda_2\alpha)$. It is easy to see that $\overline{R}$ is a R-module, if $\wedge \subset \overline{R}$ and $\wedge$ is a R-module, then $\wedge$ is called a R-submodule of $\overline{R}$. All R-modules we discuss here are R-submodule of $\overline{R}$. On the other hand, if $I \subset R$, then I is an ideal of R, if and only if I is a R-module. Let $\alpha \in \overline{R}$, the cyclic R-module generated by α be defined by

$$R\alpha = \{\lambda\alpha \mid \lambda \in R\}. \tag{28}$$

If there are finitely many polynomials $\alpha_1, \alpha_2, \ldots, \alpha_k$ in $\overline{R}$ such that $\wedge = R\alpha_1 + R\alpha_2 + \cdots + R\alpha_k$, then $\wedge$ is called a finitely generated R-module, which is a R-submodule of $\overline{R}$.

Now, if $L \subset \mathbb{R}^n$ is a ϕ-cyclic lattice, $g \in \mathbb{R}^n$, $H^*(g)$ is the ideal matrix generated by vector g, and $L(H^*(g))$ is the lattice generated by $H^*(g)$. It is easy to show that any $L(H^*(g))$ is a ϕ-cyclic lattice and

$$L(H^*(g)) \subset L, \text{ whenever } g \in L, \tag{29}$$

which implies that $L(H^*(g))$ is the smallest ϕ-cyclic lattice of which contains vector g. Therefore, we call $L(H^*(g))$ is a minimal ϕ-cyclic lattice in $\mathbb{R}^n$.

Lemma 9 *There is a one to one correspondence between the minimal ϕ-cyclic lattice in $\mathbb{R}^n$ and the cyclic R-submodule in $\overline{R}$, namely,*

$$t(Rg(x)) = L(H^*(g)), \text{ for all } g(x) \in \overline{R}$$

and

$$t^{-1}(L(H^*(g))) = Rg(x), \text{ for all } g \in \mathbb{R}^n.$$

Proof Let $b(x) \in R$, by Lemma 5, we have

$$t(b(x)g(x)) = H^*(b)g = H^*(g)b \in L(H^*(g)),$$

and $t(Rg(x)) \subset L(H^*(g))$. Conversely, if $\alpha \in L(H^*(g))$, and $\alpha = H^*(g)b$ for some integer vector b, by Lemma 5 again, we have $b(x)g(x) \in Rg(x)$, and $t(b(x)g(x)) = \alpha$. This implies that $L(H^*(g)) \subset t(Rg(x))$, and

$$t(Rg(x)) = L(H^*(g)).$$

The lemma follows immediately.

Suppose $L = L(\beta_1, \beta_2, \ldots, \beta_m)$ is arbitrary ϕ-cyclic lattice, where $B = [\beta_1, \beta_2, \ldots, \beta_m]_{n\times m}$ is the generated matrix of L. L may be expressed as the sum of finitely many minimal ϕ-cyclic lattices, in fact, we have

$$L = L(H^*(\beta_1)) + L(H^*(\beta_2)) + \cdots + L(H^*(\beta_m)). \tag{30}$$

To state and prove our main results, first, we give a definition of prime spot in $\mathbb{R}^n$.

Definition 4 Let $g \in \mathbb{R}^n$, and $g(x) = t^{-1}(g) \in \overline{R}$. If $(g(x), \phi(x)) = 1$ in $\mathbb{R}[x]$, we call g is a prime spot of $\mathbb{R}^n$.

By (v) of Theorem 2, $g \in \mathbb{R}^n$ is a prime spot if and only if $H^*(g)$ is an invertible matrix, thus the minimal ϕ-cyclic lattice $L(H^*(g))$ generated by a prime spot is a full-rank lattice.

Lemma 10 *Let g and f be two prime spots of $\mathbb{R}^n$, then $L(H^*(g)) + L(H^*(f))$ is a full-rank ϕ-cyclic lattice.*

Proof According to Lemma 4, it is sufficient to show that

$$\operatorname{rank}\big(L(H^*(g)) \cap L(H^*(f))\big) = \operatorname{rank}\big(L(H^*(g))\big) = n. \tag{31}$$

In fact, we should prove in general

$$L(H^*(g) \cdot H^*(f)) \subset L(H^*(g)) \cap L(H^*(f)). \tag{32}$$

Since $H^*(g) \cdot H^*(f)$ is an invertible matrix, then $\operatorname{rank}\big(L(H^*(g) \cdot H^*(f))\big) = n$, and (8) follows immediately.

To prove (9), we note that

$$L(H^*(g) \cdot H^*(f)) = L(H^*(g * f)).$$

It follows that

$$t^{-1}\big(L(H^*(g) \cdot H^*(f))\big) = Rg(x)f(x).$$

It is easy to see that

$$Rg(x)f(x) \subset Rg(x) \cap Rf(x).$$

Therefore, we have

$$L(H^*(g) \cdot H^*(f)) = t(Rg(x)f(x)) \subset L(H^*(g)) \cap L(H^*(f)).$$

This is the proof of Lemma 10.

It is worth to note that (9) is true for the more general case and does not need the condition of prime spot.

Corollary 5 *Let $\beta_1, \beta_2, \ldots, \beta_m$ be arbitrary m vectors in $\mathbb{R}^n$, then we have*

$$L(H^*(\beta_1)H^*(\beta_2)\cdots H^*(\beta_m)) \subset L(H^*(\beta_1)) \cap L(H^*(\beta_2)) \cap \cdots \cap L(H^*(\beta_m)). \tag{33}$$

Proof If $\beta_1, \beta_2, \ldots, \beta_m$ are integer vectors, then (10) is trivial. For the general case, we write

$$L(H^*(\beta_1) \cdot H^*(\beta_2) \cdots H^*(\beta_m)) = L(H^*(\beta_1 * \beta_2 * \cdots * \beta_m)),$$

where $\beta_1 * \beta_2 * \cdots * \beta_m$ is the ϕ-convolutional product, then

$$t^{-1}\big(L(H^*(\beta_1) \cdots H^*(\beta_m))\big) = R\beta_1(x)\beta_2(x) \cdots \beta_m(x).$$

Since

$$R\beta_1(x)\beta_2(x) \cdots \beta_m(x) \subset R\beta_1(x) \cap R\beta_2(x) \cap \cdots \cap R\beta_m(x).$$

It follows that

$$L(H^*(\beta_1)H^*(\beta_2)\cdots H^*(\beta_m)) \subset L(H^*(\beta_1)) \cap L(H^*(\beta_2)) \cap \cdots \cap L(H^*(\beta_m)).$$

We have this corollary.

By Lemma 10, we also have the following assertion.

Corollary 6 *Let $\beta_1, \beta_2, \ldots, \beta_m$ be m prime spots of $\mathbb{R}^n$, then $L(H^*(\beta_1)) + L(H^*(\beta_2)) + \cdots + L(H^*(\beta_m))$ is a full-rank ϕ-cyclic lattice.*

Proof It follows immediately from Corollary 3.

Our main result in this chapter is to establish the following one to one correspondence between ϕ-cyclic lattices in $\mathbb{R}^n$ and finitely generated R-modules in $\overline{R}$.

Theorem 4 *Let $\wedge = R\alpha_1(x) + R\alpha_2(x) + \cdots + R\alpha_m(x)$ be a finitely generated R-module in $\overline{R}$, then $t(\wedge)$ is a ϕ-cyclic lattice in $\mathbb{R}^n$. Conversely, if $L \subset \mathbb{R}^n$ is a ϕ-cyclic lattice in $\mathbb{R}^n$, then $t^{-1}(L)$ is a finitely generated R-module in $\overline{R}$, that is a one to one correspondence.*

Proof If $\wedge$ is a finitely generated R-module, by Lemma 9, we have

$$t(\wedge) = t(R\alpha_1(x) + \cdots + R\alpha_m(x)) = L(H^*(\alpha_1)) + L(H^*(\alpha_2)) + \cdots + L(H^*(\alpha_m)).$$

The main difficulty is to show that $t(\wedge)$ is a lattice of $\mathbb{R}^n$, we require a surgery to embed $t(\wedge)$ into a full-rank lattice. To do this, let $(\alpha_i(x), \phi(x)) = d_i(x)$, $d_i(x) \in \mathbb{Z}[x]$, and $\beta_i(x) = \alpha_i(x)/d_i(x)$, $1 \leqslant i \leqslant m$. Since $\phi(x)$ has no multiple roots by assumption, then $(\beta_i(x), \phi(x)) = 1$ in $\mathbb{R}[x]$. In other words, each $t(\beta_i(x)) = \beta_i$ is a prime spot. It is easy to verify $R\alpha_i(x) \subset R\beta_i(x)$ $(1 \leqslant i \leqslant m)$, thus we have

$$t(\wedge) \subset L(H^*(\beta_1)) + L(H^*(\beta_2)) + \cdots + L(H^*(\beta_m)).$$

By Corollaries 6 and 1, we have $t(\wedge)$ is ϕ-cyclic lattice. Conversely, if $L \subset \mathbb{R}^n$ is a ϕ-cyclic lattice of $\mathbb{R}^n$, and $L = L(\beta_1, \beta_2, \ldots, \beta_m)$, by (7), we have

$$t^{-1}(L) = R\beta_1(x) + R\beta_2(x) + \cdots + R\beta_m(x),$$

which is a finitely generated R-module in $\overline{R}$. We complete the proof of Theorem 4.

As we introduced in abstract, since R is a Noether ring, then $I \subset R$ is an ideal if and only if I is a finitely generated R-module. On the other hand, if $I \subset R$ is an ideal, then $t(I) \subset \mathbb{Z}^n$ is a discrete subgroup of $\mathbb{Z}^n$, thus $t(I)$ is a lattice, we define the following.

Definition 5 Let $I \subset R$ be an ideal, $t(I)$ is called the ϕ-ideal lattice.

Ideal lattice first appeared in Lyubashevsky and Micciancio (2006) (see Definition 3.1 of Lyubashevsky and Micciancio (2006)). As a direct consequence of Theorem 4, we have the following.

Corollary 7 *Let $L \subset \mathbb{R}^n$ be a subset, then L is a ϕ-cyclic lattice if and only if*

$$L = L(H^*(\beta_1)) + L(H^*(\beta_2)) + \cdots + L(H^*(\beta_m)),$$

where $\beta_i \in \mathbb{R}^n$ and $m \leqslant n$. Furthermore, L is a ϕ-ideal lattice if and only if every $\beta_i \in \mathbb{Z}^n$, $1 \leqslant i \leqslant m$.

Corollary 8 *Suppose that $\phi(x)$ is an irreducible polynomial in $\mathbb{Z}[x]$, then any non-zero ideal I of R defines a full-rank ϕ-ideal lattice $t(I) \subset \mathbb{Z}^n$.*

Proof Let $I \subset R$ be a non-zero ideal, then we have $I = R\alpha_1(x) + R\alpha_2(x) + \cdots + R\alpha_m(x)$, where $\alpha_i(x) \in R$ and $(\alpha_i(x), \phi(x)) = 1$. It follows that

$$t(I) = L(H^*(\alpha_1)) + L(H^*(\alpha_2)) + \cdots + L(H^*(\alpha_m)).$$

Since each α_i is a prime spot, we have $\text{rank}(t(I)) = n$ by Corollary 6, and the corollary follows at once.

According to Definition 3.1 of Lyubashevsky and Micciancio (2006), we have proved that any an ideal of R corresponding to a ϕ-ideal lattice, which just is a ϕ-cyclic integer lattice under the more general rotation matrix $H = H_\phi$. Cyclic lattice and ideal lattice were introduced in Lyubashevsky and Micciancio (2006), Micciancio (2002), respectively, to improve the space complexity of lattice-based cryptosystems. Ideal lattices allow to represent a lattice using only two polynomials. Using such lattices, class lattice-based cryptosystems can diminish their space complexity from $O(n^2)$ to $O(n)$. Ideal lattices also allow to accelerate computations using the polynomial structure. The original structure of Micciancio's matrices uses the ordinary circulant matrices and allows for an interpretation in terms of arithmetic in polynomial ring $\mathbb{Z}[x]/ < x^n - 1 >$. Lyubashevsky and Micciancio (2006) later suggested to change the ring to $\mathbb{Z}[x]/ < \phi(x) >$ with an irreducible $\phi(x)$ over $\mathbb{Z}[x]$. Our results here suggest to change the ring to $\mathbb{Z}[x]/ < \phi(x) >$ with any polynomial $\phi(x)$. There are many works subsequent to Micciancio (2002, Lyubashevsky and Micciancio (2006), such as (Feige & Micciancio, 2004; Micciancio & Regev, 2009; Peikert, 2016; Plantard & Schneider, 2013; Pradhan et al., 2019; Stehle & Steinfeld, 2011).

Example 1 It is interesting to find some examples of ϕ-cyclic lattices in an algebraic number field K. Let Q be a rational number field, without loss of generality, an algebraic number field K of degree n is just $K = Q(w)$, where $w = w_i$ is a root of $\phi(x)$. If all $Q(w_i) \subset \mathbb{R}$ $(1 \leqslant i \leqslant n)$, then K is called a totally real algebraic number field. Let O_K be the ring of algebraic integers of K, and $I \subset O_K$ be an ideal, $I \neq 0$. Since there is an integral basis $\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subset I$ such that

$$I = \mathbb{Z}\alpha_1 + \mathbb{Z}\alpha_2 + \cdots + \mathbb{Z}\alpha_n.$$

We may regard every ideal of O_K as a lattice in Q^n, and our assertion is that every non-zero ideal of O_K is corresponding to a full-rank ϕ-cyclic lattice of Q^n. To see this example, let

$$Q[w] = \left\{ \sum_{i=0}^{n-1} a_i w^i \mid a_i \in Q \right\}.$$

It is known that $K = Q[w]$, thus every $\alpha \in K$ corresponds to a vector $\overline{\alpha} \in Q^n$ by

$$\alpha = \sum_{i=0}^{n-1} a_i w^i \xrightarrow{\tau} \overline{\alpha} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \in \mathbb{Q}^n.$$

If $I \subset O_K$ is an ideal of O_K and $I = \mathbb{Z}\alpha_1 + \mathbb{Z}\alpha_2 + \cdots + \mathbb{Z}\alpha_n$, let $B = [\overline{\alpha_1}, \overline{\alpha_2}, \ldots, \overline{\alpha_n}] \in Q^{n\times n}$, which is full-rank matrix. We have $\tau(I) = L(B)$ as a full-rank lattice. It remains to show that $\tau(I)$ is a ϕ-cyclic lattice, we only prove that if $\alpha \in I \Rightarrow H\overline{\alpha} \in \tau(I)$. Suppose that $\alpha \in I$, then $w\alpha \in I$. It is easy to verify that $\tau(w) = e_2$ (see (2.7)) and

$$\tau(w\alpha) = \tau(w) * \tau(\alpha) = H\overline{\alpha} \in \tau(I).$$

This means that $\tau(I)$ is a ϕ-cyclic lattice of Q^n, which is a full-rank lattice.

4 Smoothing Parameter

As an application of the algebraic structure of ϕ-cyclic lattice, we show an explicit upper bound of the smoothing parameter for the ϕ-cyclic lattices. Firstly, we introduce some basic notations.

A Gauss function $\rho_{s,c}(x)$ in $\mathbb{R}^n$ is given by

$$\rho_{s,c}(x) = e^{-\pi|x-c|^2/s^2}, \tag{34}$$

where $x \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, and $s > 0$ is a positive real number. $\rho_{s,c}(x)$ is called the Gauss function around original point c with parameter s. It is easy to see that

$$\int_{\mathbb{R}^n} \rho_{s,c}(x)\mathrm{d}x = s^n.$$

Thus, we may define a probability density function $D_{s,c}(x)$ by

$$D_{s,c}(x) = \rho_{s,c}(x) / \int_{\mathbb{R}^n} \rho_{s,c}(x)\mathrm{d}x = \rho_{s,c}(x)/s^n. \tag{35}$$

Suppose $L \subset \mathbb{R}^n$ is a lattice, let

$$D_{s,c}(L) = \sum_{x\in L} D_{s,c}(x),\ \rho_{s,c}(L) = \sum_{x\in L} \rho_{s,c}(x). \tag{36}$$

The discrete Gauss distribution over L is a probability distribution $D_{L,s,c}$ over L given by

$$D_{L,s,c}(x) = \frac{D_{s,c}(x)}{D_{s,c}(L)} = \frac{\rho_{s,c}(x)}{\rho_{s,c}(L)}. \tag{37}$$

If $c = 0$ is the zero vector of $\mathbb{R}^n$, we write $\rho_{s,0}(x) = \rho_s(x)$, $\rho_{s,0}(L) = \rho_s(L)$, $D_{s,0}(x) = D_s(x)$, and $D_{s,0}(L) = D_s(L)$. Suppose that L is a full-rank lattice and L^* is its dual lattice, we define the smoothing parameter $\eta_\varepsilon(L)$ of L to be the smallest s such that $\rho_{1/s}(L^*) \leqslant 1 + \varepsilon$, more precisely,

$$\eta_\varepsilon(L) = \min\{s : s > 0 \text{ and } \rho_{1/s}(L^*) \leqslant 1 + \varepsilon\}, \tag{38}$$

where $\varepsilon > 0$ is a positive number. Notice that $\rho_{1/s}(L^*)$ is a continuous and strictly decreasing function of s, thus the smoothing parameter $\eta_\varepsilon(L)$ is a continuous and strictly decreasing function of ε.

Let $L = L(\beta_1, \beta_2, \ldots, \beta_n) \subset \mathbb{R}^n$ be a full-rank lattice with a basis $\beta_1, \beta_2, \ldots, \beta_n$, the fundamental region $P(L)$ is given by

$$P(L) = \left\{ \sum_{i=1}^{n} a_i \beta_i | 0 \leqslant a_i < 1,\ 1 \leqslant i \leqslant n \right\}. \tag{39}$$

Suppose that X and Y are two discrete random variables on $\mathbb{R}^n$, the statistical distance between X and Y over L is defined by

$$\triangle(X, Y) = \frac{1}{2} \sum_{a \in L} |P\{X = a\} - P\{Y = a\}|. \tag{40}$$

If X and Y are continuous random variables with probability density function T_1 and T_2, respectively, then $\triangle(X, Y)$ is defined by

$$\triangle(X, Y) = \frac{1}{2} \int_{\mathbb{R}^n} |T_1(z) - T_2(z)| \mathrm{d}z. \tag{41}$$

The smoothing parameter was introduced by Micciancio and Regev (2007), which plays an important role in the statistical information of lattices. An important property of smoothing parameter is for any lattice $L = L(B)$ and any $\varepsilon > 0$, the statistical distance between $D_s \bmod L$ and the uniform distribution over the fundamental region $P(L)$ is at most $\frac{1}{2}(\rho_{1/s}(L(B)^*))$. More precisely, for any $\varepsilon > 0$ and any $s \geqslant \eta_\varepsilon(L(B))$, the statistical distance is at most $\frac{1}{2}\varepsilon$, namely

$$\triangle\big(D_{s,c} \bmod L,\ U(P(L))\big) \leqslant \frac{\varepsilon}{2}. \tag{42}$$

Lemma 11 *Let $L \subset \mathbb{R}^n$ be a full-rank lattice, we have*

$$\eta_{2^{-n}}(L) \leqslant \sqrt{n}/\lambda_1(L^*), \tag{43}$$

where L^* is the dual lattice of L, and $\lambda_1(L^*)$ is the minimum distance of L^*.

Proof See Lemma 3.2 of Micciancio and Regev (2007), or Banaszczyk (1993).

Lemma 12 *Suppose that L_1 and L_2 are two full-rank lattices in $\mathbb{R}^n$, and $L_1 \subset L_2$, then for any $\varepsilon > 0$, we have*

$$\eta_\varepsilon(L_2) \leqslant \eta_\varepsilon(L_1). \tag{44}$$

Proof Let $\eta_\varepsilon(L_1) = s$, we are to show that $\eta_\varepsilon(L_2) \leqslant s$. Since

$$\rho_{1/s}(L_1^*) = 1 + \varepsilon, \text{ and } \sum_{x \in L_1^*} e^{-\pi s^2 |x|^2} = 1 + \varepsilon.$$

It is easy to check that $L_2^* \subset L_1^*$, it follows that

$$1 + \varepsilon = \sum_{x \in L_1^*} e^{-\pi s^2 |x|^2} \geqslant \sum_{x \in L_2^*} e^{-\pi s^2 |x|^2},$$

which implies

$$\rho_{1/s}(L_2^*) \leqslant 1 + \varepsilon,$$

and $\eta_\varepsilon(L_2) \leqslant s = \eta_\varepsilon(L_1)$, thus we have Lemma 12.

According to (2.4), the ideal matrix $H^*(f)$ with input vector $f \in \mathbb{R}^n$ is just the ordinary circulant matrix when $\phi(x) = x^n - 1$. Next lemma shows that the transpose of a circulant matrix is still a circulant matrix. For any $g = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix} \in \mathbb{R}^n$, we denote $\overline{g} = \begin{pmatrix} g_{n-1} \\ g_{n-2} \\ \vdots \\ g_0 \end{pmatrix}$, which is called the conjugation of g.

Lemma 13 *Let $\phi(x) = x^n - 1$, then for any $g = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix} \in \mathbb{R}^n$, we have*

$$(H^*(g))' = H^*(H\overline{g}). \tag{45}$$

Proof Since $\phi(x) = x^n - 1$, then $H = H_\phi$ (see (2.3)) is an orthogonal matrix, and we have $H^{-1} = H^{n-1} = H'$. We write $H_1 = H' = H^{-1}$. The following identity is easy to verify

$$H^*(g) = \begin{pmatrix} \overline{g}' H_1 \\ \overline{g}' H_1^2 \\ \vdots \\ \overline{g}' H_1^n \end{pmatrix}$$

It follows that

$$(H^*(g))' = [H\overline{g}, H(H\overline{g}), \ldots, H^{n-1}(H\overline{g})] = H^*(H\overline{g}),$$

and we have the lemma.

Lemma 14 *Suppose that $g \in \mathbb{R}^n$ and the circulant matrix $H^*(g)$ is invertible. Let $A = (H^*(g))' H^*(g)$, then all characteristic values of A are given by*

$$\{|g(\theta_1)|^2, |g(\theta_2)|^2, \ldots, |g(\theta_n)|^2\},$$

where $\theta_i^n = 1$ $(1 \leqslant i \leqslant n)$ are the n-th roots of unity.

Proof By Lemma 13 and (ii) of Theorem 2, we have

$$A = H^*(H\overline{g})H^*g = H^*(H^*(H\overline{g})g) = H^*(g''),$$

where $g'' = H^*(H\overline{g})g$. Let $g''(x) = t^{-1}(g'')$ be the corresponding polynomial of g''. By (iii) of Theorem 2, all characteristic values of A are given by

$$\{g''(\theta_1), g''(\theta_2), \ldots, g''(\theta_n)\},\ \theta_i^n = 1,\ 1 \leqslant i \leqslant n. \tag{46}$$

Let $g = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix} \in \mathbb{R}^n$. It is easy to see that

$$g''(x) = \sum_{i=0}^{n-1} g_i^2 + \left(\sum_{i=0}^{n-1} g_i g_{1-i}\right) x + \cdots + \left(\sum_{i=0}^{n-1} g_i g_{(n-1)-i}\right) x^{n-1} = |g(x)|^2,$$

where $g_{-i} = g_{n-i}$ for all $1 \leqslant i \leqslant n-1$, then the lemma follows at once.

By definition 4, if $g \in \mathbb{R}^n$ is a prime spot, then there is a unique polynomial $u(x) \in \overline{R}$ such that $u(x)g(x) \equiv 1 \pmod{\phi(x)}$. We define a new vector T_g and its corresponding polynomial $T_g(x)$ by

$$T_g = H\overline{u},\ \text{and}\ T_g(x) = t^{-1}(H\overline{u}). \tag{47}$$

If $g \in \mathbb{Z}^n$ is an integer vector, then $T_g \in \mathbb{Z}^n$ is also an integer vector, and $T_g(x) \in \mathbb{Z}[x]$ is a polynomial with integer coefficients. Our main result on smoothing parameter is the following theorem.

Theorem 5 *Let $\phi(x) = x^n - 1$, $L \subset \mathbb{R}^n$ be a full-rank ϕ-cyclic lattice, then for any prime spots $g \in L$, we have*

$$\eta_{2^{-n}}(L) \leqslant \sqrt{n}(\min\{|T_g(\theta_1)|, |T_g(\theta_2)|, \ldots, |T_g(\theta_n)|\})^{-1}, \tag{48}$$

where $\theta_i^n = 1$, $1 \leqslant i \leqslant n$, and $T_g(x)$ is given by (4.14).

Proof Let $g \in L$ be a prime spot, by Lemma 12, we have

$$L(H^*(g)) \subset L \Rightarrow \eta_\varepsilon(L) \leqslant \eta_\varepsilon(L(H^*(g))), \ \forall \varepsilon > 0. \tag{49}$$

To estimate the smoothing parameter of $L(H^*(g))$, the dual lattice of $L(H^*(g))$ is given by

$$L(H^*(g))^* = L((H^*(u))') = L(H^*(H\overline{u})) = L(H^*(T_g)),$$

where $u(x) \in \overline{R}$ and $u(x)g(x) \equiv 1 \pmod{x^n - 1}$, and T_g is given by (4.14). Let $A = (H^*(T_g))'H^*(T_g)$, by Lemma 14, all characteristic values of A are

$$\{|T_g(\theta_1)|^2, |T_g(\theta_2)|^2, \ldots, |T_g(\theta_n)|^2\}.$$

By Lemma 2, the minimum distance $\lambda_1(L(H^*(g))^*)$ is bounded by

$$\lambda_1(L(H^*(g))^*) \geqslant \min\{|T_g(\theta_1)|, |T_g(\theta_2)|, \ldots, |T_g(\theta_n)|\}. \tag{50}$$

Now, Theorem 5 follows from Lemma 11 immediately.

Let $L = L(B)$ be a full-rank lattice and $B = [\beta_1, \beta_2, \ldots, \beta_n]$. We denote by $B^* = [\beta_1^*, \beta_2^*, \ldots, \beta_n^*]$ the Gram-Schmidt orthogonal vectors $\{\beta_i^*\}$ of the ordered basis $B = \{\beta_i\}$. It is a well-known conclusion that

$$\lambda_1(L) \geqslant |B^*| = \min_{1 \leqslant i \leqslant n} |\beta_i^*|,$$

which yields by Lemma 11 the following upper bound

$$\eta_{2^{-n}}(L) \leqslant \sqrt{n}|B_0^*|^{-1}, \tag{51}$$

where B_0^* is the orthogonal basis of dual lattice L^* of L.

For a ϕ-cyclic lattice L, we observe that the upper bound (4.17) is always better than (4.18) by numerical testing, we give two examples here.

Example 2 Let $n = 3$ and $\phi(x) = x^3 - 1$, the rotation matrix H is

$$H = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

We select a ϕ-cyclic lattice $L = L(B)$, where

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since $L = \mathbb{Z}^3$, thus L is a ϕ-cyclic lattice. It is easy to check

$$|B_0^*| = \min_{1 \leqslant i \leqslant 3} |\beta_i^*| = \frac{\sqrt{3}}{3}.$$

On the other hand, we randomly find a prime spot $g = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \in L$ and $g(x) = x^2$. Since $xg(x) \equiv 1 \pmod{x^3 - 1}$, we have $T_g(x) = x^2$, it follows that $|T_g(\theta_1)| = |T_g(\theta_2)| = |T_g(\theta_3)| = 1$, and

$$\min_{1 \leqslant i \leqslant 3} |T_g(\theta_i)|^{-1} \leqslant |B_0^*|^{-1} = \sqrt{3}.$$

Example 3 Let $n = 4$ and $\phi(x) = x^4 - 1$, the rotation matrix H is

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We select a ϕ-cyclic lattice $L = L(B)$, where

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since $L = \mathbb{Z}^4$, thus L is a ϕ-cyclic lattice. It is easy to check

$$|B_0^*| = \min_{1 \leqslant i \leqslant 4} |\beta_i^*| = \frac{1}{2}.$$

On the other hand, we randomly find a prime spot $g = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \in L$ and $g(x) = x - 2$. Since $(\frac{1}{7}x^3 - \frac{1}{7}x^2 - \frac{2}{7}x - \frac{5}{7})g(x) \equiv 1 \pmod{x^4 - 1}$, we have $T_g(x) = -\frac{2}{7}x^3 - \frac{1}{7}x^2 + \frac{1}{7}x - \frac{5}{7}$, it follows that $|T_g(\theta_1)| = 1$, $|T_g(\theta_2)| = |T_g(\theta_3)| = |T_g(\theta_4)| = \frac{5}{7}$, and

$$\min_{1 \leqslant i \leqslant 4} |T_g(\theta_i)|^{-1} = \frac{7}{5} \leqslant |B_0^*|^{-1} = 2.$$

References

Ajtai, M. (1996). Generating hard instances of the short basis problem. In *Proceedings of 28th STOC* (pp. 99–108).

Ajtai, M., & Dwork, C. (1997). A public-key cryptosystem with worst-case/average-case equivalence. In *Proceedings of 29th STOC* (pp. 284–293).

Banaszczyk, W. (1993). New bounds in some transference theorems in the geometry of numbers. *Mathematische Annalen, 296*(4), 625–635.

Cassels, J. W. S. (1963). *Introduction to diophantine approximation*. Cambridge University Press.

Cassels, J. W. S. (1971). *An introduction to the geometry of numbers*. Springer.

Davis, P. J. (1994). *Circulant matrices* (2nd ed.). Chelsea Publishing.

Feige, U., & Micciancio, D. (2004). The inapproximability of lattice and coding problems with preprocessing. *Journal of Computer and System Sciences, 69*(1), 45–67.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. Stoc.

Lint, J. H. V. (1999) *Introduction to coding theory*. Springer.

Lyubashevsky, V., & Micciancio, D. (2006) Generalized compact knapsacks are collision resistant. In *Proceedings of the 33rd international conference on Automata, Languages and Programming—Proceedings of ICALP 2006* (Vol. 4052, pp. 144–155). Springer LNCS.

Micciancio, D. (2001). The hardness of the closest vector problem with preprocessing. *IEEE Transactions on Information Theory, 47*(3), 1212–1215.

Micciancio, D. (2022). Generalized compact knapsacks, cyclic lattices, and efficient one-way functions from worst-case complexity assumptions: (extended abstract). In *Annual Symposium on Foundations of Computer Science*.

Micciancio, D., & Regev, O. (2007). Worst-case to average-case reductions based on gaussian measures. *SIAM Journal on Computing, 37*(1), 267–302.

Micciancio, D., & Regev, O. (2009). Lattice-based cryptography. In D. J. Bernstein, J. Buchmann & E. Dahmen (Eds.), *Post-quantum cryptography* (pp. 147–191). Springer.

Peikert, C. (2016). A decade of lattice cryptography. *Foundations and trends in theoretical computer science*.

Plantard, T., & Schneider, M. (2013). Creating a challenge for ideal lattices (pp. 1–17).

Pradhan, P. K., Rakshit, S., & Datta, S. (2019). Lattice based cryptography: Its applications, areas of interest and future scope. In *Proceedings of the Third International Conference on Computing Methodologies and Communication* (pp. 988–993).

Regev, O. (2004). Improved inapproximability of lattice and coding problems with preprocessing. *IEEE Transactions on Information Theory, 50*(9), 2031–2037.

Shi, B. J. (2018). The spectral norms of geometric circulant matrices with the generalized k-Horadam numbers. *Journal of Inequalities and Applications, 14*.

Stehle, D., & Steinfeld, R.: Making NTRU as secure as worst-case problems over ideal lattices. In K. G. Paterson(Eds.), *Advances in cryptology, lecture notes in computer sciences* (Vol. 6632, pp. 27–47). Springer.

Yasin, Y., & Taskara, N. (2013). On the inverse of circulant matrix via generalized k-Horadam numbers. *Applied Mathematics and Computation, 223*, 191–196.

Zheng, Z.Y., Huang, W. L., Xu, J., & Tian, K. A generalization of cyclic code and applications to public key cryptosystems.

On the LWE Cryptosystem with More General Disturbance

Zheng Zhiyong and Tian Kun

Abstract The main purpose of this chapter is to give an extension on learning with errors problem (LWE)-based cryptosystem about the probability of decryption error with more general disturbance. In the first section, we introduce the LWE cryptosystem with its application and some previous research results. Then we give a more precise estimation probability of decryption error based on independent identical Gaussian disturbances and any general independent identical disturbances. This upper bound probability could be closed to 0 if we choose applicable parameters. It means that the probability of decryption error for the cryptosystem could be sufficiently small. So we verify our core result that the LWE-based cryptosystem could have high security.

Keywords Learning with errors problem · Decryption error · Probability · General disturbance

1 Introduction

In this section, we describe a cryptosystem based on the learning with errors problem (LWE) (Micciancio & Regev, 2009; Regev, 2005). First, we introduce the LWE problem. Let p be a prime number, m, n be positive integers and consider a list of equations with error as follows:

Z. Zhiyong · T. Kun (✉)
Engineering Research Center of Ministry of Education for Financial Computing and Digital Engineering, Renmin University of China, Beijing 100872, China
e-mail: tkun19891208@ruc.edu.cn

Z. Zhiyong
e-mail: zhengzy@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_8

$$\begin{cases} < s, a_1 > \approx_{\chi} v_1 \pmod{p}, \\ < s, a_2 > \approx_{\chi} v_2 \pmod{p}, \\ \qquad\qquad \vdots \\ < s, a_m > \approx_{\chi} v_m \pmod{p}. \end{cases}$$

Here $s \in \mathbb{Z}_p^n$, $a_1, a_2, \ldots, a_m$ are chosen independently and uniformly from $\mathbb{Z}_p^n$, and $v_1, v_2, \ldots, v_m \in \mathbb{Z}_p$. $< s, a_i >$ is the inner product of two vectors s and a_i. The errors in these equations are generated from a probability distribution $\chi : \mathbb{Z}_p \to \mathbb{R}^+$ on $\mathbb{Z}_p$, i.e. for each equation, we have $v_i =< s, a_i > +e_i$ and $e_i \in \mathbb{Z}_p$ is chosen independently based on the probability distribution χ. The problem of finding $s \in \mathbb{Z}_p^n$ from such equations is called $\mathrm{LWE}_{p,\chi}$. There is an equivalent description for the LWE problem. The input has a pair (A, v) where $A \in \mathbb{Z}_p^{m \times n}$ is chosen uniformly, and the choices of v have two cases. One case for v is chosen uniformly from $\mathbb{Z}_p^m$, the other case is $As + e$ for a uniformly chosen vector $s \in \mathbb{Z}_p^n$ and vector $e \in \mathbb{Z}_p^m$ chosen according to χ^m. The goal is to distinguish between these two cases with non-negligible probability. It is also equivalent with a decoding problem in q-ary lattices (Regev, 2005).

The short integer solution (SIS) problem was first introduced in the seminal work of Ajtai (1996) and has served as the foundation for one-way and collision-resistant hash functions, identification schemes, digital signatures, and other "minicrypt" primitives. A very important work of Regev from 2005 introduced the LWE problem, which is the "encryption-enabling" analogue of the SIS problem (Regev, 2009). In fact, the two problems are very similar and can meaningfully be seen as duals of each other.

The LWE problem is a very robust problem and can be viewed as an extension of a well-known problem in learning theory. It remains hard even if the attacker learns extra information about the secret and errors. Regev gave the worst-case hardness theorem for LWE (Regev, 2009). The complexity of the best-known algorithm is running in exponential time in n (Ajtai et al., 2001; Blum et al., 2003; Kumar & Sivakumar, 2001). This theorem is proved by giving a quantum polynomial-time reduction that uses an oracle for LWE to solve GapSVP_{γ} and SIVP_{γ} in the worst case, thereby transforming any algorithm that solves LWE into a quantum algorithm for lattice problems. The quantum nature of the reduction is meaningful since there are no known quantum algorithms for GapSVP_{γ} and SIVP_{γ} that significantly outperform classical ones, beyond generic quantum speedups. It would be very useful to have a completely classical reduction to give further confidence in the hardness of LWE, which was given in 2009 by Peikert (2009). Regev also gave a public-key cryptosystem whose semantic security can provably be based on the LWE problem, and hence on the conjectured quantum hardness of GapSVP_{γ} and SIVP_{γ} for $\gamma = O(n^{3/2})$ (Regev, 2009). LWE problem has a close relationship with decoding problems in coding theory (Ajtai, 2005; Ajtai & Dwork, 1997; Alekhnovich, 2003; Asokan et al., 2007; Ding, 2004; Kawachi et al., 2007; Peikert, 2007; Peikert et al., 2008; Regev, 2004; Signing et al., 2022). Regev's cryptosystem is secure against passive eavesdroppers since the LWE problem is hard.

Another application of LWE is fully homomorphic encryption (FHE) (Rivest et al., 1978). The earliest FHE constructions were based on average-case assumptions about ideal lattices (Gentry, 2009; Dijk et al., 2010). Later, Brakerski and Vaikuntanathan gave the second generation of FHE constructions, which were based on the LWE problem (Brakerski & Vaikuntanathan, 2011a, b). In 2013, Gentry, Sahai, and Waters proposed an LWE-based FHE scheme that has some unique and advantageous properties, such as homomorphic multiplication does not require any key-switching step, and the scheme can be made identity-based. This yields unbounded FHE based on LWE with just an inverse-polynomial $n^{-O(1)}$ error rate (Gentry et al., 1999).

Now we introduce the efficient lattice-based cryptosystem in the following which has strong theoretical security (Micciancio & Regev, 2009).

- Private key: $S \in \mathbb{Z}_q^{n\times l}$ is uniformly chosen at random.
- Public key: $A \in \mathbb{Z}_q^{m\times n}$ is uniformly chosen at random and $E \in \mathbb{Z}_q^{m\times l}$ is chosen from the distribution $\overline{\psi}_\alpha$. The public key is $(A, P = AS + E)$.
- Encryption: Given $v \in \mathbb{Z}_t^l$ from the message space and a public key (A, P), choose a vector $a \in \{-r, -r+1, \cdots, r\}^m$ uniformly at random, and compute the ciphertext $(u = A^T a, c = P^T a + f(v))$.
- Decryption: Given a ciphertext (u, c) and a private key S, output $f^{-1}(c - S^T u)$.

Here m, n, l, t, q, r are positive integers and $\alpha > 0$. $\overline{\psi}_\alpha$ is defined to be the distribution on $\mathbb{Z}_q$ obtained by sampling a normal variable with mean 0 and standard deviation $\alpha q/\sqrt{2\pi}$, rounding the result to the nearest integer and reduced modulo q. f is defined as the function from $\mathbb{Z}_t^l$ to $\mathbb{Z}_q^l$ by multiplying each coordinate by q/t and rounding to the nearest integer. f^{-1} is defined to be the "inverse" mapping of f by multiplying each coordinate by t/q and rounding to the nearest integer. The definitions of f and f^{-1} are in the next section. The probability of decryption error in one letter for this cryptosystem is approximatively estimated in (Micciancio & Regev, 2009) as

$$\text{error probability per letter} \approx 2\left(1 - \Phi\left(\frac{1}{2t\alpha}\sqrt{\frac{6\pi}{mr(r+1)}}\right)\right), \tag{1}$$

where Φ is the cumulative distribution function of the standard normal distribution, i.e. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t$. We give here a more precise upper bound estimation

$$\text{error probability} \leqslant 2l\left(1 - \Phi\left(\frac{q-t}{2\alpha t q}\sqrt{\frac{6\pi}{mr(r+1)}}\right)\right). \tag{2}$$

This upper bound probability could be closed to 0 if we choose α small enough. It means that the probability of decryption error for the cryptosystem could be sufficiently small. However, the above estimation is based on Gaussian disturbance. In our work, we also give the probability of decryption error for the LWE-based cryptosystem with more general disturbance. By central limit theorem (Riauba, 1975), general

disturbance could be approximated as Gaussian disturbance, then we get the following probability estimation result which is more advanced than that in (Micciancio & Regev, 2009).

$$\text{error probability} \leqslant 2l\left(1-\Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\right)+l\delta. \tag{3}$$

Here β is the standard deviation of disturbance distribution, and δ is a positive real number.

1.1 Innovation and Contribution

Our work gives estimation probability of decryption error based on Gaussian disturbances and proves that the decryption error could be sufficiently small. The most salient innovation and contribution is that for any general disturbances, the decryption error could also be small enough. This indicates high security and reliability of LWE-based cryptosystem. In other words, this cryptosystem is secure enough against passive eavesdroppers and could be applied in many kinds of encryption processes.

2 Methodology

2.1 Preliminary Property

Definition 1 $\forall x \in \mathbb{R}$, let $[x]$ be the closest integer to x, specially, $[x]$ is defined to be $x-\frac{1}{2}$ if the fractional part of x is $\frac{1}{2}$. It is trivial that $-\frac{1}{2} < x-[x] \leqslant \frac{1}{2}$ for all $x \in \mathbb{R}$.

Lemma 1 *t and q are positive integers, $t \leqslant q$. $\forall a \in \mathbb{Z}_t$, let $f(a)=[\frac{q}{t}a] \in \mathbb{Z}_q$. $\forall b \in \mathbb{Z}_q$, let $f^{-1}(b)=[\frac{t}{q}b] \in \mathbb{Z}_t$. Then $f^{-1}(f(a))=a$ for $\forall a \in \mathbb{Z}_t$ holds.*

Remark 1 If $a_1 \equiv a_2 \pmod t$, we have $f(a_1) \equiv f(a_2) \pmod q$, so the definition of f is well defined and reasonable.

Proof of Lemma 1 (1) If $t=q$, then we have $f(a)=[a]=a$ and

$$f^{-1}(f(a))=f^{-1}(a)=[a]=a,\ \forall a \in \mathbb{Z}_t.$$

(2) If $t<q$, then $\frac{q}{2t} > \frac{1}{2}$, we know

$$\frac{q}{t}a - \frac{1}{2} \leqslant \left[\frac{q}{t}a\right] < \frac{q}{t}a + \frac{1}{2}.$$

It follows that

$$\frac{q}{t}a - \frac{q}{2t} < \frac{q}{t}a - \frac{1}{2} \leqslant \left[\frac{q}{t}a\right] < \frac{q}{t}a + \frac{1}{2} < \frac{q}{t}a + \frac{q}{2t}.$$

So we can get

$$\frac{q}{t}a - \frac{q}{2t} < \left[\frac{q}{t}a\right] < \frac{q}{t}a + \frac{q}{2t}.$$

This is equivalent to

$$a - \frac{1}{2} < \frac{t}{q}\left[\frac{q}{t}a\right] < a + \frac{1}{2},$$

and

$$-\frac{1}{2} < \frac{t}{q}\left[\frac{q}{t}a\right] - a < \frac{1}{2}.$$

Thus,

$$[\frac{t}{q}[\frac{q}{t}a] - a] = 0, \text{ and } \left[\frac{t}{q}\left[\frac{q}{t}a\right]\right] = a.$$

This means that

$$f^{-1}(f(a)) = a, \ \forall a \in \mathbb{Z}_t. \qquad \square$$

Lemma 2 *t and q are positive integers, $t > q$. If a is uniformly chosen in $\mathbb{Z}_t$, then*

$$P\{f^{-1}(f(a)) \neq a\} = 1 - \frac{q}{t}.$$

Proof of Lemma 2 $t > q$, from Lemma 1 we have

$$\left[\frac{q}{t}\left[\frac{t}{q}b\right]\right] = b, \ \forall b \in \mathbb{Z}_q.$$

This is equivalent to

$$f\left(\left[\frac{t}{q}b\right]\right) = b, \ \forall b \in \mathbb{Z}_q.$$

So we get

$$f^{-1}\left(f\left(\left[\frac{t}{q}b\right]\right)\right) = f^{-1}(b) = \left[\frac{t}{q}b\right], \ \forall b \in \mathbb{Z}_q.$$

Here $0, \left[\frac{t}{q}\right], \left[\frac{2t}{q}\right], \ldots, \left[\frac{(q-1)t}{q}\right]$ are different from each other in $\mathbb{Z}_t$. Next we prove that the number of a in $\mathbb{Z}_t$ satisfying $f^{-1}(f(a)) = a$ is no more than q. Let A be the set containing all the elements satisfying $f^{-1}(f(a)) = a$ in $\mathbb{Z}_t$. $\forall a_1, a_2 \in A, a_1 \neq a_2$

in $\mathbb{Z}_t$, then we have $f(a_1) \not\equiv f(a_2) \pmod{q}$, i.e. $f(a_1) \neq f(a_2)$ in $\mathbb{Z}_q$. This means the number of A is no more than q.

Above all, it shows that $0, \left[\frac{t}{q}\right], \left[\frac{2t}{q}\right], \ldots, \left[\frac{(q-1)t}{q}\right]$ are just all the numbers in $\mathbb{Z}_t$ such that $f^{-1}(f(a)) = a$. Based on a is uniformly chosen in $\mathbb{Z}_t$, then

$$P\{f^{-1}(f(a)) \neq a\} = 1 - \frac{q}{t}. \qquad \square$$

Corollary 1 *$t, q,$ and l are positive integers. $\forall a = (a_1, a_2, \ldots, a_l) \in \mathbb{Z}_t^l$, let $f(a) = \left(\left[\frac{q}{t}a_1\right], \left[\frac{q}{t}a_2\right], \ldots, \left[\frac{q}{t}a_l\right]\right) \in \mathbb{Z}_q^l$. $\forall b = (b_1, b_2, \ldots, b_l) \in \mathbb{Z}_q^l$, let $f^{-1}(b) = \left(\left[\frac{t}{q}b_1\right], \left[\frac{t}{q}b_2\right], \ldots, \left[\frac{t}{q}b_l\right]\right) \in \mathbb{Z}_t^l$. If a is uniformly chosen in $\mathbb{Z}_t^l$ and $a_1, a_2, \ldots, a_l$ are independent, then*

$$P\{f^{-1}(f(a)) \neq a\} = \max\left\{0, 1 - \left(\frac{q}{t}\right)^l\right\}.$$

Proof of Corollary 1 If $t \leqslant q$, from Lemma 1, we have

$$f^{-1}(f(a_i)) = a_i, \ \forall a_i \in \mathbb{Z}_t, \ \forall 1 \leqslant i \leqslant l.$$

So

$$f^{-1}(f(a)) = a, \ \forall a \in \mathbb{Z}_t^l.$$

$$P\{f^{-1}(f(a)) \neq a\} = 0 = \max\left\{0, 1 - \left(\frac{q}{t}\right)^l\right\}.$$

If $t > q$, from Lemma 2, we have

$$P\{f^{-1}(f(a_i)) = a_i\} = \frac{q}{t}, \ a_i \in \mathbb{Z}_t, \ \forall 1 \leqslant i \leqslant l.$$

Since $a_1, a_2, \ldots, a_l$ are independent, therefore,

$$P\{f^{-1}(f(a)) = a\} = \left(\frac{q}{t}\right)^l, \ a \in \mathbb{Z}_t^l.$$

$$P\{f^{-1}(f(a)) \neq a\} = 1 - \left(\frac{q}{t}\right)^l = \max\left\{0, 1 - \left(\frac{q}{t}\right)^l\right\}. \qquad \square$$

2.2 Probability of Decryption Error Based on Gaussian Disturbance

Now we can calculate the probability of decryption error for the LWE-based cryptosystem. As described in the first section, assume S be the private key, (A, P) be the public key, and we choose $v \in \mathbb{Z}_t^l$ from the message space, encrypt v, and then decrypt it. The ciphertext is $(u = A^T a, c = P^T a + f(v))$. The decryption result is

$$\begin{aligned} f^{-1}(c - S^T u) &= f^{-1}(P^T a + f(v) - S^T u) \\ &= f^{-1}((AS + E)^T a + f(v) - S^T A^T a) \\ &= f^{-1}(E^T a + f(v)). \end{aligned}$$

Here the decryption result $f^{-1}(E^T a + f(v)) \in \mathbb{Z}_t^l$. The decryption error occurs if $f^{-1}(E^T a + f(v)) \neq v$. Since all the parameters are taken to guarantee security and efficiency of the cryptosystem, here we set $q > t$ and obtain the following theorem.

Theorem 1 *t, q, l, m, r are positive integers and $q > t$. $v \in \mathbb{Z}_t^l$, f is defined in the previous section, $E_{m\times l}$ is a Gaussian disturbance matrix with each element chosen independently from the Gaussian distribution with mean 0 and standard deviation $\alpha q/\sqrt{2\pi}$, $a \in \{-r, -r+1, \cdots, r\}^m$ is uniformly chosen at random. Then we have the following inequality of the probability of decryption error.*

$$P\{f^{-1}(E^T a + f(v)) \neq v\} \leqslant 2l\left(1 - \Phi\left(\frac{q-t}{2\alpha t q}\sqrt{\frac{6\pi}{mr(r+1)}}\right)\right).$$

Here Φ is the cumulative distribution function of the standard normal distribution, i.e. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t$.

Proof of Theorem 1 In order to compute the probability of decryption error, we consider one letter first, i.e. the probability of $f^{-1}(E_i^T a + f(v_i)) \neq v_i$, here v_i is the ith coordinate of v, $E_{m\times l} = (E_1, E_2, \ldots, E_l)$, and $f^{-1}(E_i^T a + f(v_i))$ is the ith coordinate of $f^{-1}(E^T a + f(v))$. From Lemma 1, we know that $f^{-1}(f(v_i)) = v_i$ for any $v_i \in \mathbb{Z}_t$ under this condition. We have

$$-\frac{1}{2} < \frac{q}{t}v_i - \left\lceil \frac{q}{t}v_i \right\rceil \leqslant \frac{1}{2}.$$

$$-\frac{t}{2q} \leqslant \frac{t}{q}\left\lceil \frac{q}{t}v_i \right\rceil - v_i < \frac{t}{2q}.$$

So if $|\frac{t}{q}E_i^T a| < \frac{1}{2} - \frac{t}{2q}$, we get

$$\left|\frac{t}{q}E_i^T a + \frac{t}{q}\left[\frac{q}{t}v_i\right] - v_i\right| < \frac{1}{2} - \frac{t}{2q} + \frac{t}{2q} = \frac{1}{2}.$$

$$\left[\frac{t}{q}E_i^T a + \frac{t}{q}\left[\frac{q}{t}v_i\right] - v_i\right] = 0.$$

$$\left[\frac{t}{q}E_i^T a + \frac{t}{q}\left[\frac{q}{t}v_i\right]\right] = v_i.$$

$$f^{-1}(E_i^T a + f(v_i)) = v_i.$$

It means that if $|\frac{t}{q}E_i^T a| < \frac{1}{2} - \frac{t}{2q}$, we can get $f^{-1}(E_i^T a + f(v_i)) = v_i$. Equivalently, if $f^{-1}(E_i^T a + f(v_i)) \neq v_i$, i.e. the decryption error occurs in the ith letter, then $|\frac{t}{q}E_i^T a| \geqslant \frac{1}{2} - \frac{t}{2q}$. So the probability of decryption error in one letter is no more than the probability of $|\frac{t}{q}E_i^T a| \geqslant \frac{1}{2} - \frac{t}{2q}$, i.e.

$$P\{f^{-1}(E_i^T a + f(v_i)) \neq v_i\} \leqslant P\left\{\left|\frac{t}{q}E_i^T a\right| \geqslant \frac{1}{2} - \frac{t}{2q}\right\}.$$

The next step we estimate the probability of $|\frac{t}{q}E_i^T a| \geqslant \frac{1}{2} - \frac{t}{2q}$. Since each coordinate of E_i is chosen independently from the Gaussian distribution with mean 0 and standard deviation $\alpha q/\sqrt{2\pi}$ and the sum of independent Gaussian variables is still a Gaussian variable, $E_i^T a$ is also a Gaussian distribution variable. $a = (a_1, a_2, \ldots, a_m)$ and each a_i is chosen from $\{-r, -r+1, \cdots, r\}$ uniformly at random, then

$$E(a_i) = \frac{-r + (-r+1) + \cdots + r}{2r+1} = 0.$$

$$\mathrm{Var}(a_i) = \frac{(-r)^2 + (-r+1)^2 + \cdots + r^2}{2r+1} = \frac{r(r+1)}{3}.$$

$$E(E_i^T a) = 0.$$

$$\mathrm{Var}(E_i^T a) = \left(\frac{\alpha q}{2\pi}\right)^2 \cdot \frac{r(r+1)}{3}m = \frac{\alpha^2 q^2 mr(r+1)}{6\pi}.$$

Therefore, $E_i^T a$ is treated as a normal distribution with mean 0 and standard deviation $\alpha q\sqrt{mr(r+1)}/\sqrt{6\pi}$. We have

$$P\left\{\left|\frac{t}{q}E_i^T a\right| \geqslant \frac{1}{2} - \frac{t}{2q}\right\} = P\left\{\left|E_i^T a\right| \geqslant \frac{q-t}{2t}\right\}$$

$$= P\left\{|E_i^T a| \Big/ \left(\alpha q\sqrt{\frac{mr(r+1)}{6\pi}}\right) \geqslant \frac{q-t}{2t} \Big/ \left(\alpha q\sqrt{\frac{mr(r+1)}{6\pi}}\right)\right\}$$

$$= P\left\{|E_i^T a|/(\alpha q\sqrt{\frac{mr(r+1)}{6\pi}}) \geqslant \frac{q-t}{2\alpha tq}\sqrt{\frac{6\pi}{mr(r+1)}}\right\}$$

$$= 2\left(1-\Phi\left(\frac{q-t}{2\alpha tq}\sqrt{\frac{6\pi}{mr(r+1)}}\right)\right).$$

So we get the following inequality for the probability of decryption error of the LWE-based cryptosystem

$$\begin{aligned} & P\{f^{-1}(E^T a + f(v)) \neq v\} \\ \leqslant\ & lP\{f^{-1}(E_i^T a + f(v_i)) \neq v_i\} \\ \leqslant\ & lP\left\{\left|\frac{t}{q}E_i^T a\right| \geqslant \frac{1}{2} - \frac{t}{2q}\right\} \\ =\ & 2l\Big(1-\Phi\left(\frac{q-t}{2\alpha tq}\sqrt{\frac{6\pi}{mr(r+1)}}\right)\Big). \end{aligned}$$

□

This upper bound probability estimation is more precise than (1). The upper bound could be as closed as 0 if we choose α small enough. It means that the probability of decryption error for the LWE-based cryptosystem could be made very small with an appropriate setting of parameters.

2.3 Probability of Decryption Error for General Disturbance

In this section, we estimate the probability of decryption error for the LWE-based cryptosystem when the noise matrix $E = (E_{ij})_{m\times l}$ is chosen independently from a general common variable.

Theorem 2 *t, q, l, r are positive integers and $q > t$, m is a undetermined positive integer. $v \in \mathbb{Z}_t^l$, f is defined in the second section, $E_{m\times l}$ is a general disturbance matrix with each element chosen independently from a common random variable of mean 0 and standard deviation β, $a \in \{-r, -r+1, \cdots, r\}^m$ is uniformly chosen at random. For any $\delta > 0$, we can find positive integer m, such that the following inequality of the probability of decryption error holds.*

$$P\{f^{-1}(E^T a + f(v)) \neq v\} \leqslant 2l\Big(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\Big) + l\delta.$$

Here Φ is the cumulative distribution function of the standard normal distribution, i.e. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t$.

Proof of Theorem 2 Similarly as the proof of Theorem 1, we need to estimate the probability of $|\frac{t}{q} E_i^T a| \geqslant \frac{1}{2} - \frac{t}{2q}$. Since the coordinates of E_i^T are independent identically distributed, E_i^T and a are also independent, by central limit theorem (Riauba, 1975), $E_i^T a$ is approximately normal distribution with mean 0 and standard deviation $d = \sqrt{mVar(E_{ij})Var(a_i)} = \beta\sqrt{\frac{mr(r+1)}{3}}$. Thus, for any sufficiently small $\delta > 0$, there is a positive integer m such that

$$P\left\{\left|\frac{t}{q} E_i^T a\right| \geqslant \frac{1}{2} - \frac{t}{2q}\right\} = P\left\{|E_i^T a| \geqslant \frac{q-t}{2t}\right\}$$

$$= P\left\{|E_i^T a| \Big/ \left(\beta\sqrt{\frac{mr(r+1)}{3}}\right) \geqslant \frac{q-t}{2t} \Big/ \left(\beta\sqrt{\frac{mr(r+1)}{3}}\right)\right\}$$

$$= P\left\{|E_i^T a| \Big/ \left(\beta\sqrt{\frac{mr(r+1)}{3}}\right) \geqslant \frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right\}$$

$$= 2\Big(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\Big) + \varepsilon.$$

Here $|\varepsilon| \leqslant \delta$. Then we get the following inequality for the probability of decryption error of the LWE-based cryptosystem for general disturbance

$$P\{f^{-1}(E^T a + f(v)) \neq v\}$$

$$\leqslant lP\{f^{-1}(E_i^T a + f(v_i)) \neq v_i\}$$

$$\leqslant lP\left\{\left|\frac{t}{q} E_i^T a\right| \geqslant \frac{1}{2} - \frac{t}{2q}\right\}$$

$$= 2l\Big(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\Big) + l\varepsilon.$$

$$\leqslant 2l\Big(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\Big) + l\delta.$$

□

This probability could be also closed to 0 if we choose the parameter $\beta\sqrt{m}$ and δ small enough. Therefore, the probability of decryption error of the LWE-based cryptosystem for general disturbance could be made very small, which leads to high security.

Example 1 Let $t = 2, q = 5, l = 1, m = 1, r = 1, \delta = 10^{-3}$, $v \in \mathbb{Z}_2$ is uniformly chosen at random, the disturbance E is a random variable with the distribution ψ_β such that $P\{E = k\} = \frac{\beta^k}{2\cdot k!}e^{-\beta}$ for integer k and $P\{E = 0\} = e^{-\beta}$ with parameter $\beta = 10^{-3}$, $a \in \{-1, 0, 1\}$ is uniformly chosen at random. Then the probability of decryption error

$$P\{f^{-1}(Ea + f(v)) \neq v\} = P\left\{\left[\frac{2}{5}\left(Ea + \left[\frac{5}{2}v\right]\right)\right] \neq v\right\}$$

$$= \frac{1}{2}P\left\{\left[\frac{2}{5}Ea\right] \neq 0\right\} + \frac{1}{2}P\left\{\left[\frac{2}{5}(Ea + 2)\right] \neq 1\right\}$$

$$\leqslant \frac{1}{2}P\{E \neq 0\} + \frac{1}{2}P\{E \neq 0\}$$

$$= 1 - P\{E = 0\} = 1 - e^{-0.001} < 10^{-3}.$$

On the other hand,

$$2l\left(1 - \Phi\left(\frac{q - t}{2\beta t}\sqrt{\frac{3}{mr(r + 1)}}\right)\right) + l\delta > 10^{-3}.$$

So it follows that

$$P\{f^{-1}(Ea + f(v)) \neq v\} < 2l\left(1 - \Phi\left(\frac{q - t}{2\beta t}\sqrt{\frac{3}{mr(r + 1)}}\right)\right) + l\delta.$$

The inequality in Theorem 2 holds.

Example 2 Let $t = 2, q = 5, l = 1, m = 1, r = 1, \delta = 10^{-4}$, $v \in \mathbb{Z}_2$ is uniformly chosen at random, the disturbance E is a Laplace distribution variable with parameter $\lambda = 0.05$ and probability density function $f(x) = \frac{1}{2\lambda}e^{-\frac{|x|}{\lambda}}$ rounding to the nearest integer, $a \in \{-1, 0, 1\}$ is uniformly chosen at random. Similarly as Example 1, the probability of decryption error

$$P\{f^{-1}(Ea + f(v)) \neq v\} = P\left\{\left[\frac{2}{5}\left(Ea + \left[\frac{5}{2}v\right]\right)\right] \neq v\right\}$$

$$\leqslant 1 - P\{E = 0\} = 1 - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}} \mathrm{d}x = e^{-10} < 10^{-4}.$$

On the other hand,

$$2l\left(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\right) + l\delta > 10^{-4}.$$

It follows that

$$P\{f^{-1}(Ea + f(v)) \neq v\} < 2l\left(1 - \Phi\left(\frac{q-t}{2\beta t}\sqrt{\frac{3}{mr(r+1)}}\right)\right) + l\delta.$$

The inequality in Theorem 2 holds.

3 Results and Conclusions

In this work, we first introduce the LWE problem and LWE-based cryptosystem. We give a more precise estimation probability of decryption error based on independent identical Gaussian disturbances. The salient significance of our work is that for any general independent identical disturbances, we also give the estimation probability of decryption error using central limit theorem. The upper bound probability could be closed to 0 if we choose applicable parameters. It means that the probability of decryption error for the cryptosystem could be sufficiently small. Then we confirm that the LWE-based cryptosystem could have high security.

4 Discussions

4.1 Future Work

Although we have reached our objective in this work, there are still many interesting works to study in this research area in the future. We will focus on the fully homomorphic encryption (FHE)-based cryptosystem later, which is an application of LWE (Brakerski & Vaikuntanathan, 2011a, b; Dijk et al., 2010; Gentry, 2009; Gentry et al., 1999). Fully homomorphic encryption was known to have abundant applications in cryptography, but for three decades no plausibly secure scheme was known until

2009. To date, the FHE-based cryptography has more than three generations. The third generation FHE scheme based on LWE problem is proved that has some unique and advantageous properties (Gentry et al., 1999). It also remains some improvable technique which needs to be studied in depth.

References

Ajtai, M. (1996). Generating hard instances of lattice problems. *Quaderni di Matematica,* 1–32.

Ajtai, M. (2005). Representing hard lattices with O(NLOGN) bits. In *Proceedings of the 37th ACM Symposium* (pp. 94–103).

Ajtai, M., & Dwork, C. (1997). A public-key cryptosystem with worst-case/average-case equivalence. In *Proceedings of the 29th ACM Symposium* (pp. 284–293).

Ajtai, M., Kumar, R., & Sivakumar, D. (2001). A sieve algorithm for the shortest lattice vector problem. In *Proceedings of the 33rd ACM Symposium* (pp. 601–610).

Alekhnovich, M. (2003). More on average case versus approximation complexity. In *Proceedings of the 44th Annual IEEE Symposium* (pp. 298–307).

Asokan, N., Kostiainen, K., Ginzboorg, P., Ott, J., Luo, C., Asokan, P., Ginzboorg, P., & Ott, P. (2007). Applicability of identity-based cryptography for disruption-tolerant networking. In *International MobiSys Workshop on Mobile Opportunistic Networking* (pp. 52–56).

Blum, A., Kalai, A., & Wasserman, H. (2003). Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM,* 506–519.

Brakerski, Z., & Vaikuntanathan, V. (2011a). Fully homomorphic encryption from ring-LWE and security for key dependent messages. *Annual Cryptology Conference* (pp. 505–524).

Brakerski, Z., & Vaikuntanathan, V. (2011b). Efficient fully homomorphic encryption from (standard) LWE. In *Proceedings of 52nd IEEE Foundations of Computer Science* (pp. 831–871).

Dijk, M., Gentry, C., Halevi, S., & Vaikuntanathan, V. (2010). Fully homomorphic encryption over the integers. In *International Conference on Theory and Applications of Cryptographic Techniques* (pp. 24–43).

Ding, J. (2004). A new variant of the Matsumoto-Imai cryptosystem through perturbation. *In Public Key Cryptography PKC, 2004,* 305–318.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In *STOC* (pp. 169–178).

Gentry, C., Sahai, A., & Waters, B. (2013). Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. *Annual Cryptology Conference* (pp. 75–92).

Kawachi, A., Tanaka, K., & Xagawa, K. (2007). Multi-bit cryptosystems based on lattice problems. In *Public Key Cryptography PKC 2007* (pp. 315–329).

Kumar, R., & Sivakumar, D. (2001). On polynomial approximation to the shortest lattice vector length. In *Proceedings of the 12th ACM Symposium* (pp. 126–127).

Micciancio, D., & Regev, O. (2009). Lattice-based cryptography. In D. J. Bernstein, J. Buchmann, & E. Dahmen (Eds.), *Post-quantum cryptography* (pp. 147–191). Springer.

Peikert, C. (2007). Limits on the hardness of lattice problems in l_p norms. In *Proceedings of the 22nd IEEE Annual CCC* (pp. 333–346).

Peikert, C. (2009). Public-key cryptosystems from the worst-case shortest vector problem. In *Proceedings of the 41st ACM Symposium* (pp. 333–342).

Peikert, C., Vaikuntanathan, V., & Waters, B. (2008). A framework for efficient and composable oblivious transfer. In *Annual Cryptology Conference* (pp. 1–28).

Regev O. (2004). New lattice based cryptographic constructions. *Journal of the ACM,* 899–942.

Regev, O. (2005). On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the 37th ACM Symposium* (pp. 84–93).
Regev, O. (2009). On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM,* 1–40.
Riauba, B. (1975). A central limit theorem for dependent random variables. *Lithuanian Mathematical Journal,* 185–200.
Rivest, R., Adleman, L., & Dertouzos, M. (1978). On data banks and privacy homomorphisms. In *Foundations of secure computation* (pp. 169–180).
Signing, V., Tegue, G., Kountchou, M., Njitacke, Z., Tsafack, N., Nkapkop, J., Etoundi, C., & Kengne, J. (2022). A cryptosystem based on a chameleon chaotic system and dynamic DNA coding. *Chaos, Solitons and Fractals,111777.*

On the High Dimensional RSA Algorithm—A Public Key Cryptosystem Based on Lattice and Algebraic Number Theory

Zheng Zhiyong, Liu Fengxia, and Chen Man

Abstract The most known public key cryptosystem was introduced in 1978 by Rivest et al. (1978) and is now called the RSA public key cryptosystem in their honor. Later, a few authors gave a simple extension of RSA over algebraic numbers field (see Takagi and Naito (2015), Uematsu et al. (1985, 1986)), but they require that the ring of algebraic integers is Euclidean ring, and this requirement is much more stronger than the class number one condition. In this chapter, we introduce a high dimensional form of RSA by making use of the ring of algebraic integers of an algebraic number field and the lattice theory. We give an attainable algorithm (see Algorithm 1) which is significant both from the theoretical and practical point of view. Our main purpose in this chapter is to show that the high dimensional RSA is a lattice based on public key cryptosystem indeed, of which would be considered as a new number in the family of post-quantum cryptography (see Peikert (2014), Pradhanet al. (2019)). On the other hand, we give a matrix expression for any algebraic number fields (see Theorem 2), which is a new result even in the sense of classical algebraic number theory.

Keywords RSA · The ring of algebraic integers · Ideal matrix · Ideal lattice · HNF basis

Z. Zhiyong · L. Fengxia (✉) · C. Man
Engineering Research Center of Ministry of Education for Financial Computing and Digital Engineering, Renmin University of China, Beijing 100872, China
e-mail: liufengxia91@126.com

Z. Zhiyong
e-mail: zhengzy@ruc.edu.cn

C. Man
e-mail: cm2019000758@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_9

1 Introduction

Let Q, $\mathbb{R}$, $\mathbb{C}$ be the rational numbers field, real numbers field, and complex numbers field, respectively, and $\mathbb{Z}$ be the integers ring. Let $E \subset \mathbb{C}$ be an algebraic numbers field of degree n, and $R \subset E$ be the ring of algebraic integers of E. Suppose that $A \subset R$ is a non-zero ideal(all ideals in this chapter are non-zero), then the factor ring R/A is a finite ring, we denote by $N(A)$ the number of elements of R/A, which is called the norm of A, and denote by $\varphi(A)$ the number of invertible elements of R/A, which is called the Euler totient function of A. For any $\alpha \in R$, the principal ideal generated by α is denoted by αR, then α is an invertible element of R/A if and only if $(\alpha R, A) = 1$. It is known (see Theorem 1.19 of Narkiewicz (2004)) that

$$\varphi(A) = N(A) \prod_{P|A} (1 - \frac{1}{N(P)}) \tag{1}$$

where the product is extended over all prime ideals P dividing A. Moreover, if $\alpha \in R$ and $(\alpha R, A) = 1$, then

$$\alpha^{\varphi(A)} \equiv 1 (\text{mod } A). \tag{2}$$

To generalize that RSA to arbitrary algebraic number fields E, we first show the following assertion.

Theorem 1 *Let P_1 and P_2 be two distinct prime ideals of R and $A = P_1 P_2$, then for any $\alpha \in R$ and integer $k \geq 0$, we have*

$$\alpha^{k\varphi(A)+1} \equiv \alpha (\text{mod } A). \tag{3}$$

Proof Let $\alpha \in R$. If $(\alpha R, A) = 1$, then (3) follows directly from (2). If $(\alpha R, A) = A$, then $\alpha R \subset A$ and $\alpha \in A$, (3) is trivial. Thus, we only consider the cases of $(\alpha R, A) = P_1$ and $(\alpha R, A) = P_2$. If $(\alpha R, A) = P_1$, then $(\alpha R, P_2) = 1$, by (2) we have

$$\alpha^{\varphi(P_2)} \equiv 1 (\text{mod } P_2).$$

It follows that

$$\alpha^{k\varphi(A)} \equiv 1 (\text{mod } P_2), \quad \forall k \in \mathbb{Z}, \quad k \geq 0.$$

Therefore, there exists an element $\beta \in P_2$ such that

$$\alpha^{k\varphi(A)} = 1 + \beta.$$

We thus have

$$\alpha^{k\varphi(A)+1} = \alpha + \alpha\beta, \quad \text{and} \quad \alpha^{k\varphi(A)+1} \equiv \alpha (\text{mod } A),$$

since $\alpha\beta \in A$. The same reason gives (3) when $(\alpha R, A) = P_2$.

Table 1 RSA in the ring of algebraic integers

RSA in the Ring of Algebraic Integers
• **Parameters**: $n \geq 1$ is a positive integer, E/Q is an algebraic numbers field of degree n, $R \subset E$ is the ring of algebraic integers of E. P_1 and P_2 are two prime ideals of R, $A = P_1 P_2$, R/A is the factor ring, S is a set of coset representatives of R/A, $\varphi(A)$ is the Euler function of A, $1 \leq e < \varphi(A)$ and $1 \leq d < \varphi(A)$ are two positive integers such that $ed \equiv 1 (\text{mod } \varphi(A))$
• **Public keys**: The ideal A and positive integer e are the public keys.
• **Private keys**: The prime ideals P_1, P_2 and the positive integer d are the private keys
• **Encryptions**: For any input message $\alpha \in S$, the ciphertext c is $c \equiv \alpha^e (\text{mod } A)$
• **Decryption**: $c^d \equiv \alpha^{ed} \equiv \alpha (\text{ mod } A)$, one can find plaintext α from c in S

According to Theorem 1, one can easily extend the classical RSA over an algebraic number field as follows (also see Takagi and Naito (2015)), but it does not give the proof of (3)).

Obviously, if $n = 1$, the above algorithm is the ordinary RSA. However, it is difficult to find the prime ideals in R and to construct a set of coset representatives of R/A yet. In Takagi and Naito (2015), the author supposed the ring R is a Euclidean ring, so that S can be constructed by Euclidean algorithm in R. The simplest way is to select an prime element α in R, so that the principal ideal αR is a prime ideal. In algorithm I, we would precisely construct a set of coset representatives for the factor ring R/A by the lattice theory. Here we give an approximate construction of the set of coset representatives for factor ring R/A.

If $P \subset R$ is a prime ideal, then $P \cap \mathbb{Z} = p\mathbb{Z}$, where $p \in \mathbb{Z}$ is a rational prime number. Since R/P is a finite field and $\mathbb{Z}/(p\mathbb{Z}) \subset R/P$, thus $N(P) = p^f$, where f $(1 \leq f \leq n)$ is called the degree of P. We write $pR = P_1^{e_1} P_2^{e_2} \cdots P_g^{e_g}$, where $P = P_1$ and P_i are distinct prime ideals, e_i is called the ramification index of P_i. There exists a remarkable relation among ramification indexes and degrees (see Theorem 3 of page 181 of Ireland and Rosen (1990))

$$\sum_{i=1}^{g} e_i f_i = n. \tag{4}$$

Let $\{\alpha_1, \alpha_2, \cdots \alpha_n\} \subset R$ be an integral basis for E/Q, $A = P_1 P_2$. Suppose that $P_1 \cap \mathbb{Z} = p\mathbb{Z}$ and $P_2 \cap \mathbb{Z} = q\mathbb{Z}$, then $A \cap \mathbb{Z} = pq\mathbb{Z}$, where p and q are two distinct rational prime numbers.

Lemma 1 *Let*

$$S_1 = \left\{ \sum_{i=1}^{n} a_i \alpha_i \mid 0 \le a_i < pq,\ a_i \in \mathbb{Z},\ 1 \le i \le n \right\}. \tag{5}$$

Then S_1 covers a set of coset representatives of R/A. Moreover, if the degrees of P_1 and P_2 are n, then S_1 is precisely an set of coset representatives of R/A.

Proof Since $A = P_1 P_2$, $P_1 \cap \mathbb{Z} = p\mathbb{Z}$, and $P_2 \cap \mathbb{Z} = q\mathbb{Z}$, we have $pqR \subset A$, thus R/pqR maps onto R/A. To prove the first assertion, it is enough to show that S_1 is a set of coset representatives of R/pqR. Since $\{\alpha_1, \alpha_2, \ldots \alpha_n\}$ is an integral basis and

$$R = \mathbb{Z}\alpha_1 + \mathbb{Z}\alpha_2 + \cdots + \mathbb{Z}\alpha_n.$$

Suppose that $\alpha = \sum_{i=1}^{n} m_i \alpha_i \in R$, write $m_i = a_i pq + r_i$, where $0 \le r_i < pq$. Clearly

$$\alpha \equiv \sum_{i=1}^{n} r_i \alpha_i (\text{mod } pqR).$$

Thus every coset of pqR contains an element of S_1. If $\sum_{i=1}^{n} r_i \alpha_i = \sum_{i=1}^{n} r_i' \alpha_i$ are in S_1 and in the same coset mod pqR, then

$$\sum_{i=1}^{n} \left(r_i - r_i'\right) \alpha_i \equiv 0 (\text{mod } \ pqR).$$

Since α_i are linearly independent, it follows that

$$r_i \equiv r_i' (\text{ mod } \ pq) \quad \text{and} \quad r_i = r_i', \quad 1 \le i \le n.$$

Next, suppose that the degrees of P_1 and P_2 are n, then $N(P_1) = p^n$ and $N(P_2) = q^n$, by (4) we thus have $P_1 = pR$, $P_2 = qR$, and $A = pqR$. The second assertion follows immediately.

If one replaces S by S_1 in Table 1, then the successful probability of decryption is

$$N(A)/p^n q^n = p^{f_1 - n} q^{f_2 - n}, \tag{6}$$

where f_1 and f_2 are the degrees of P_1 and P_2, respectively.

We note that $f_1 = f_2 = n$ if and only if $P_1 = pR$ and $P_2 = qR$; in this special case, we may give a numerical explanation. It is easy to see that

$$\varphi(A) = \varphi(pR)\varphi(qR) = \left(p^n - 1\right)\left(q^n - 1\right).$$

By Theorem 1, for any $a \in \mathbb{Z}$, we have

$$a^{k(p^n-1)(q^n-1)+1} \equiv a(\text{mod } pq), \quad k \in \mathbb{Z}, \quad k \geq 0. \tag{7}$$

Since S_1 is a set of coset representatives of R/A, $\alpha = \sum_{i=1}^{n} a_i\alpha_i \in S_1$, we may regard α as a vector $(a_1, a_2, \ldots, a_n) \in \mathbb{Z}_{pq}^n$. Let $m = pq$, $1 \leq e < (p^n - 1)(q^n - 1)$ and $1 \leqslant d < (p^n - 1)(q^n - 1)$ such that

$$ed \equiv 1(\text{mod } (p^n - 1)(q^n - 1)).$$

Then for every input message $\alpha = (a_1, a_2, \cdots, a_n)$, we use the public key (m, e) and private key (p, q, d) to encryption and decryption for each a_i in order, obviously, these are the algorithms given by Takagi and Naito (2015), we consider these algorithms are just a simple repeat of RSA.

The main purpose of this chapter is to show that the high dimensional form of RSA algorithm is a lattice based on cryptosystem in general. To do this, we first establish a relationship between an algebraic number field E and the Euclidean space Q^n. Let $\mathbb{R}^n$ be the Euclidean space which is a linear space over $\mathbb{R}$ with the Euclidean norm $|x|$,

$$|x| = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}, \quad \text{where } x' = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n. \tag{8}$$

We use the column notation for vector in $\mathbb{R}^n$, and x' is the transpose of x, which is called a row vector in $\mathbb{R}^n$. $Q^n \subset \mathbb{R}^n$ is a subspace of $\mathbb{R}^n$.

Without loss of generality, an algebraic number field E of degree n may be expressed as $E = Q(\theta)$, where θ is an algebraic integer of degree n and $Q(\theta)$ is the field generated by θ over Q. Let $\phi(x)$ be the minimal polynomial of θ,

$$\phi(x) = x^n - \phi_{n-1}x^{n-1} - \cdots - \phi_1 x - \phi_0 \in \mathbb{Z}[x], \tag{9}$$

where all $\phi_i \in \mathbb{Z}$. It is known that

$$E = Q[\theta] = \left\{\sum_{i=0}^{n-1} a_i\theta^i \mid a_i \in Q\right\}. \tag{10}$$

We define an one to one correspondence between E and Q^n by τ:

$$\alpha = \sum_{i=0}^{n-1} a_i\theta^i \in E \longrightarrow^{\tau} \overline{\alpha} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \in Q^n \tag{11}$$

and write $\tau(\alpha) = \overline{\alpha}$ or $\alpha \overset{\tau}{\to} \overline{\alpha}$. In fact, τ is a homomorphism of additive group from E to Q^n, because of $\tau(a\alpha) = a\tau(\alpha)$ for all $a \in \mathbb{Q}$.

As usual, the trace and norm mappings from E to Q are denoted by

$$\mathrm{tr}(\alpha) = \mathrm{tr}_{E/Q}(\alpha), \quad \text{and} \quad N(\alpha) = N_{E/Q}(\alpha).$$

It is known (see corollary of page 58 of Narkiewicz (2004)) that

$$N(\alpha R) = |N(\alpha)|, \quad \forall \alpha \in R. \tag{12}$$

A full-rank lattice L is a discrete addition subgroup of $\mathbb{R}^n$, the equivalent expression for L is (See Micciancio and Regev (2009), Zheng et al. (2023))

$$L = L(B) = \left\{ Bx \mid x \in \mathbb{Z}^n \right\}, \tag{13}$$

where $B = \left[\overline{\beta}_1, \overline{\beta}_2, \cdots, \overline{\beta}_n\right]_{n\times n} \in \mathbb{R}^{n\times n}$ is an invertible matrix of $n \times n$ dimension, B is called a generated matrix of L. If $L \subset Q^n$, we call L a rational lattice, if $L \subset \mathbb{Z}^n$, we call L an integer lattice. It is not difficult to see that every ideal of R corresponds to an rational lattice, we have the following.

Lemma 2 *Let $A \subset R$ be an ideal and $A \neq 0$, then $\tau(A)$ is a rational lattice.*

Proof Let $\{\beta_1, \beta_2, \cdots, \beta_n\} \subset A$ be an integral basis for E/Q, one has

$$A = \mathbb{Z}\beta_1 + \mathbb{Z}\beta_2 + \cdots + \mathbb{Z}\beta_n.$$

It follows that

$$\tau(A) = \mathbb{Z}\overline{\beta}_1 + \mathbb{Z}\overline{\beta}_2 + \cdots + \mathbb{Z}\overline{\beta}_n,$$

where $\overline{\beta}_i = \tau(\beta_i) \in Q^n$. Let $B = [\overline{\beta}_1, \overline{\beta}_2, \cdots, \overline{\beta}_n]$, since $\{\beta_1, \beta_2, \cdots, \beta_n\}$ is linearly independent over Q, thus B is an invertible matrix, and we have

$$\tau(A) = L(B) = \{Bx \mid x \in \mathbb{Z}^n\}.$$

The lemma follows at once.

Let $L \subset Q^n$ be a rational lattice, of which be corresponded by an ideal A in E for some suitable algebraic number field E, we call L an ideal lattice. Ideal lattice was first introduced by Lyubashevsky and Micciancio (2006) in the case of integer lattice, here we generalize this notation to the case of rational lattices. For more detailed discussion about ideal lattice, we refer to (Zheng et al., 2023).

To give an attainable algorithm for high dimensional RSA, we require the following NC-property for the algebraic number field E.

$$\text{NC- property:} \qquad E = Q(\theta) \quad \text{and} \quad R = \mathbb{Z}[\theta], \tag{14}$$

where

Table 2 Algebraic number fields with NC-property

Algebraic Number Fields with NC-property
• Quadratic Fields(see Proposition 13.1.1 of Ireland and Rosen (1990)) $E = Q(\sqrt{d})$, where $d \in \mathbb{Z}$ is a square-free integer and $d = 2, 3 (\mathrm{mod}\ 4)$
• Cyclotemic Fields (see Theorem 2.6 of Washington (1982)) $E = Q(\xi_n)$, where $\xi_n = e^{2\pi i/n}$ is a primitive n-th root of unity
• Totally Real Algebraic Number Fields (see Proposition 2.16 of Washington (1982)) $E = Q(\xi_n + \xi_n^{-1})$, and $E \subset \mathbb{R}$ is the maximal real subfield of $Q(\xi_n)$

$$\mathbb{Z}[\theta] = \left\{ \sum_{i=0}^{n-1} a_i \theta^i \mid a_i \in \mathbb{Z},\ 1 \le i \le n \right\}. \tag{15}$$

Some of the well-known algebraic number fields satisfy the NC-property, we list a few as follows (Table 2).

2 Ideal Matrices

Suppose that θ is an algebraic integer of degree n, $\phi(x) = x^n - \phi_{n-1}x^{n-1} - \cdots - \phi_1 x - \phi_0 \in \mathbb{Z}[x]$ is the minimal polynomial of θ, thus $\phi(x)$ is irreducible. Let $\theta = \theta_0, \theta_1, \theta_2, \cdots, \theta_{n-1}$ be n different roots of $\phi(x)$, the Vandermonde matrix of $\phi(x)$ is defined by

$$V = V_\phi = \left[\theta_j^i\right]_{0 \le i,j \le n-1}, \quad \text{and} \quad \Delta = \det(V_\phi) \ne 0. \tag{16}$$

According to $\phi(x)$, we denote the rotation matrix or adjoint matrix (see page 116 of Manin and Panchishkin (2005)) by

$$H = H_\phi = \left(\begin{array}{c|c} \begin{matrix} 0 & \cdots & 0 \end{matrix} & \phi_0 \\ \hline \multirow{2}{*}{I_{n-1}} & \phi_1 \\ & \vdots \\ & \phi_{n-1} \end{array}\right) \in \mathbb{Z}^{n\times n}, \tag{17}$$

where I_{n-1} is the unit matrix of $n-1$ dimension.

Definition 1 An ideal matrix $H^*(\overline{f})$ generated by the input vector $\overline{f} \in \mathbb{R}^n$ is defined by

$$H^*(\overline{f}) = \left[\ \overline{f}, H\overline{f}, \cdots, H^{n-1}\overline{f}\ \right]_{n\times n} \in \mathbb{R}^{n\times n} \tag{18}$$

and all ideal matrices are denoted by

$$M_{\mathbb{R}}^{*}=\left\{H^{*}(\overline{f}) \mid \overline{f} \in \mathbb{R}^{n}\right\} \quad \text{and} \quad M_{Q}^{*}=\left\{H^{*}(\overline{f}) \mid \overline{f} \in Q^{n}\right\}. \tag{19}$$

Definition 2 For any two vectors $\overline{f}$ and $\overline{g}$ in $\mathbb{R}^n$, the ϕ-conventional product is defined by

$$\overline{f} \otimes \overline{g}=H^{*}(\overline{f}) \overline{g} \tag{20}$$

and the m-multi product is denoted by

$$\overline{f}^{\otimes m}=\overbrace{\overline{f} \otimes \overline{f} \otimes \cdots \otimes \overline{f}}^{m}, \quad m \in \mathbb{Z}, \quad m \geq 1. \tag{21}$$

Remark 1 If $\phi(x)=x^{n}-1$, then H_{ϕ} is the classical circulant matrix (see Davis (1994)), and conventional product with circulant matrix was first proposed by Hoffstein et al. (1998), which plays a key role in their cryptosystem. In Zheng et al. (2023), we generalized this definition with more general rotation matrices.

By (18), $H^{*}(\overline{f})=0$ is a zero matrix if and only if $\overline{f}=0$ is a zero vector, and $H^{*}(\overline{f}+\overline{g})=H^{*}(\overline{f})+H^{*}(\overline{g})$, then $H^{*}(\overline{f})=H^{*}(\overline{g})$ if and only if $\overline{f}=\overline{g}$. Thus we may regard $H^{*}: \mathbb{R}^{n} \rightarrow \mathrm{M}_{\mathbb{R}}^{*}$ as an one to one correspondence, which is also a homomorphism of Abel group.

The main aim of this subsection is to show the Q^n is a field under the ϕ-conventional product and M_Q^* is also a field under the ordinary additive and product of matrices, both of which are isomorphic to the algebraic number field $E=Q(\theta)$. To do this, we require some basic properties of the ideal matrices.

Let $\overline{e}_1, \overline{e}_2, \cdots, \overline{e}_n$ be the unit vectors of $\mathbb{R}^n$, namely

$$\overline{e}_{1}=\begin{pmatrix}1\\0\\ \vdots\\0\end{pmatrix}, \overline{e}_{2}=\begin{pmatrix}0\\1\\ \vdots\\0\end{pmatrix}, \cdots, \quad \overline{e}_{n}=\begin{pmatrix}0\\0\\ \vdots\\1\end{pmatrix}. \tag{22}$$

Lemma 3 *Let τ be defined by (11), then we have*

$$\begin{cases}\tau\left(\theta^{k}\right)=\overline{e}_{k+1}, & 0 \leq k \leq n-1\\ H^{*}\left(\overline{e}_{k}\right)=H^{k-1}, & 1 \leqslant k \leqslant n.\end{cases} \tag{23}$$

Proof $\tau\left(\theta^{k}\right)=\overline{e}_{k+1}$ follows directly from the definition of τ. We use induction to prove $H^{*}\left(\overline{e}_{k}\right)=H^{k-1}$. It is easy to see that $H^{*}\left(\overline{e}_{1}\right)=I_{n}$, the unit matrix of n dimension. Suppose that $H^{*}\left(\overline{e}_{k-1}\right)=H^{k-2}$, for $k \geq 2$, note that $\overline{e}_{k}=H \overline{e}_{k-1}$, it follows that

$$\begin{aligned} H^*(\overline{e}_k) &= \left[H\overline{e}_{k-1}, H^2\overline{e}_{k-1}, \cdots, H^n\overline{e}_{k-1}\right] \\ &= H\left[\overline{e}_{k-1}, H\overline{e}_{k-1}, \cdots, H^{n-1}\overline{e}_{k-1}\right] \\ &= HH^*(\overline{e}_{k-1}) = HH^{k-2} = H^{k-1}. \end{aligned}$$

The lemma follows immediately.

Since $\phi(x)$ is the characteristic polynomial of H, by the Hamilton-Cayley theorem, we have

$$\phi(H) = 0, \quad \text{or } H^n = \phi_0 + \phi_1 H + \cdots + \phi_{n-1}H^{n-1}. \tag{24}$$

Therefore, all the rotation matrices $H^k (k \geq 0)$ are the ideal matrices, especially, the unit matrix $I_n = H^*(\overline{e}_1)$ is an ideal matrix.

Let $\mathbb{R}[x]$ be the polynomials ring and $\mathbb{R}(x)/\langle\phi(x)\rangle$ be the quotient ring, where $\langle\phi(x)\rangle$ is the principal ideal generated by $\phi(x)$ in $\mathbb{R}[x]$. We establish an one to one correspondence t between $\mathbb{R}^n$ and $\mathbb{R}[x]/\langle\phi(x)\rangle$ by

$$\overline{f} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{R}^n \xrightarrow{t} f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1} \in \mathbb{R}[x]/\langle\phi(x)\rangle \tag{25}$$

and write $t(\overline{f}) = f(x)$, or $t^{-1}(f(x)) = \overline{f}$.

Lemma 4 *For any $\overline{f} \in \mathbb{R}^n$, the ideal matrix $H^*(\overline{f})$ is given by*

$$H^*(\overline{f}) = f(H) = f_0 I_n + f_1 H + \cdots + f_{n-1}H^{n-1}. \tag{26}$$

Moreover, if $F(x) \in \mathbb{R}[x]$ and $F(x) \equiv f(x)(mod\ \phi(x))$, then $f(H) = F(H)$.

Proof Writing $\overline{f} = f_0\overline{e}_1 + f_1\overline{e}_2 + \cdots + f_{n-1}\overline{e}_n$, by Lemma 3, we have

$$\begin{aligned} H^*(\overline{f}) &= f_0 H^*(\overline{e}_1) + f_1 H^*(\overline{e}_2) + \cdots + f_{n-1}H^*(\overline{e}_n) \\ &= f_0 I_n + f_1 H + \cdots + f_{n-1}H^{n-1} = f(H). \end{aligned}$$

Suppose that $F(x) \equiv f(x)(\text{mod } \phi(x))$, by (24), we have $f(H) = F(H)$ immediately.

Lemma 5 *Let $\overline{f}$ and $\overline{g}$ be two vectors in $\mathbb{R}^n$, and $f(x)$, $g(x)$ be the corresponding polynomials, respectively, then we have*

$$t(\overline{f} \otimes \overline{g}) \equiv f(x)g(x)(mod\ \ \phi(x)). \tag{27}$$

Proof Since t is a bijection, it is suffice to show that

$$t^{-1}(f(x)g(x)) = \overline{f} \otimes \overline{g}. \tag{28}$$

Let $g(x) = g_0 + g_1(x) + \cdots + g_{n-1}x^{n-1} \in \mathbb{R}[x]/\langle \phi(x) \rangle$, then

$$\begin{aligned} xg(x) &= g_0 x + \cdots + g_{n-1}x^n \\ &= g_{n-1}\phi_0 + (g_0 + \phi_1 g_{n-1})x + \cdots + (g_{n-2} + \phi_{n-1}g_{n-1})x^{n-1}. \end{aligned}$$

It follows that

$$t^{-1}(xg(x)) = Ht^{-1}(g(x)) = H\overline{g}.$$

More generally, we have

$$t^{-1}\left(x^k g(x)\right) = H^k t^{-1}(g(x)) = H^k \overline{g}, \quad 0 \le k \le n-1. \tag{29}$$

Let $f(x) = f_0 + f_1 x + \cdots + f_{n-1}x^{n-1}$, then

$$t^{-1}(f(x)g(x)) = \sum_{k=0}^{n-1} f_k t^{-1}\left(x^k g(x)\right) = \sum_{k=0}^{n-1} f_k H^k \overline{g} = H^*(\overline{f})\overline{g} = \overline{f} \otimes \overline{g}.$$

The lemma follows immediately.

Lemma 6 *For any two vectors* $\overline{f} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{R}^n$, $\overline{g} = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix} \in \mathbb{R}^n$, *we have the following properties for ideal matrices:*

i $H^*(\overline{f})H^*(\overline{g}) = H^*\left(\overline{g}\right)H^*(\overline{f})$;
ii $H^*(\overline{f})H^*(\overline{g}) = H^*(H^*(\overline{f})\overline{g})$;
iii $H^*(\overline{f}) = V_\phi^{-1}\,\mathrm{diag}\left\{f\left(\theta_0\right), f\left(\theta_1\right), \cdots, f\left(\theta_{n-1}\right)\right\} V_\phi$;
iv $\det\left(H^*(\overline{f})\right) = \prod_{i=0}^{n-1} f\left(\theta_i\right)$;
v *If* $\overline{f} \in Q^n$, $\overline{f} \neq 0$, *then* $H^*(\overline{f})$ *is an invertible matrix and*

$$\left(H^*(\overline{f})\right)^{-1} = H^*(\overline{u}),$$

where $u(x) \in Q[x]$ *is the unique polynomial such that* $u(x)f(x) \equiv 1 (mod\ \phi(x))$ *in* $Q[x]$.

Proof By Lemma 4, we have

$$H^*(\overline{f})H^*(\overline{g}) = f(H)g(H) = g(H)f(H) = H^*(\overline{g})H^*(\overline{f}).$$

To prove (ii), we write $H^*(\overline{f})\overline{g} = \overline{f} \otimes \overline{g}$, it follows that

$$H^*\left(H^*(\overline{f})\overline{g}\right) = H^*(\overline{f} \otimes \overline{g}) = f(H)g(H) = H^*(\overline{f}) \cdot H^*(\overline{g}).$$

By Theorem 3.5 of Davis (1994), we have

$$H = V_\phi^{-1} \operatorname{diag}\{\theta_0, \theta_1, \cdots, \theta_{n-1}\} V_\phi. \tag{30}$$

It follows that

$$H^*(\overline{f}) = f(H) = V_\phi^{-1} \operatorname{diag}\{f(\theta_0), f(\theta_1), \cdots, f(\theta_{n-1})\} V_\phi.$$

Since $\operatorname{diag}\{f(\theta_0), f(\theta_1), \cdots, f(\theta_{n-1})\}$ is a diagonal matrix, we have

$$\det\left(H^*(\overline{f})\right) = \det\left(\operatorname{diag}\{f(\theta_0), f(\theta_1), \cdots, f(\theta_{n-1})\}\right) = \prod_{i=0}^{n-1} f(\theta_i).$$

To show the last assertion, since $\overline{f} \in Q^n$, $\overline{f} \neq 0$, and $\phi(x)$ is an irreducible polynomial, thus we have $(f(x), \phi(x)) = 1$ in $Q[x]$, There are $u(x) \in Q[x]$ and $v(x) \in Q[x]$ such that

$$u(x)f(x) + v(x)\phi(x) = 1.$$

By (29) and noting that $t^{-1}(1) = \overline{e}_1 \in \mathbb{R}^n$, we have $\overline{u} \otimes \overline{f} = \overline{e}_1$. It follows that

$$H^*(\overline{u}) \cdot H^*(\overline{f}) = H^*(\overline{e}_1) = I_n.$$

We complete the proof of Lemma.

Next, we discuss the algebraic number field $E = Q(\theta)$ and recall τ is an one to one correspondence between E and Q^n.

Lemma 7 *For any two elements α and β in E, we have*

$$\tau(\alpha\beta) = \tau(\alpha) \otimes \tau(\beta) = \overline{\alpha} \otimes \overline{\beta}. \tag{31}$$

Proof Let $\beta = \beta_0 + \beta_1\theta + \cdots + \beta_{n-1}\theta^{n-1}$, where $\beta_i \in Q$, it is easily seen that

$$\theta\beta = \phi_0\beta_{n-1} + (\beta_0 + \phi_1\beta_{n-1})\theta + \cdots + (\beta_{n-2} + \phi_{n-1}\beta_{n-1})\theta^{n-1},$$

thus we have $\tau(\theta\beta) = H\tau(\beta) = H\overline{\beta}$, and

$$\tau\left(\theta^k\beta\right) = H^k\tau(\beta) = H^k\overline{\beta}, \quad 0 \leq k \leq n-1. \tag{32}$$

Let $\alpha = \alpha_0 + \alpha_1\theta + \cdots + \alpha_{n-1}\theta^{n-1}$, by Lemma 4, we have

$$\tau(\alpha\beta) = \sum_{k=0}^{n-1} \alpha_k \tau\left(\theta^k \beta\right) = \sum_{k=0}^{n-1} \alpha_k H^k \overline{\beta} = H^*(\overline{\alpha})\overline{\beta} = \overline{\alpha} \otimes \overline{\beta},$$

the lemma follows immediately.

Let $A = \left(a_{ij}\right)_{n\times n}$ be a square matrix, and the trace of A is defined by $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$ as usual. The main result of this subsection is the following theorem.

Theorem 2 *Let $E = Q(\theta)$ be an algebraic number field of degree n, and $\phi(x) \in \mathbb{Z}[x]$ be the minimal polynomial of θ. Then the linear space Q^n is a field under the ϕ-conventional product, and all of the ideal matrices M_Q^* generated by rational vectors is also a field with the ordinary additive and product of matrices. Both of them are isomorphic to E, namely*

$$E \cong Q^n \cong M_Q^*. \tag{33}$$

Moreover, let $\alpha \in E$, $\mathrm{tr}(\alpha)$ and $N(\alpha)$ be the trace and norm of α, then we have

$$\mathrm{tr}(\alpha) = \mathrm{Tr}\left(H^*(\overline{\alpha})\right), \quad \textit{and} \quad N(\alpha) = \det\left(H^*(\overline{\alpha})\right). \tag{34}$$

Proof $\tau : E \to Q^n$ given by (11), it is clearly that

$$\tau(\alpha+\beta) = \tau(\alpha) + \tau(\beta), \quad \text{and} \quad \tau(\alpha\beta) = \tau(\alpha) \otimes \tau(\beta).$$

Thus Q^n is a field under the ϕ-conventional product and $E \cong Q^n$. By Lemma 6, we have

$$H^*(\overline{\alpha}+\overline{\beta}) = H^*(\overline{\alpha}) + H^*(\overline{\beta}) \quad \text{and} \quad H^*\left(\overline{\alpha} \otimes \overline{\beta}\right) = H^*(\overline{\alpha}) H^*(\overline{\beta}),$$

thus M_Q^* is also a field and $E \cong Q^n \cong M_Q^*$.

The main difficulty is to prove (34). We observe that θ induces a linear transformation of E/Q by $\alpha \to \theta\alpha$, and the matrix of this linear transformation under basis $\left\{1, \theta, \theta^2, \cdots, \theta^{n-1}\right\}$ is just H, namely

$$\theta\left(1, \theta, \theta^2, \cdots, \theta^{n-1}\right) = \left(1, \theta, \theta^2, \cdots, \theta^{n-1}\right) H.$$

By the definition of trace, we have

$$\mathrm{tr}(\theta) = \mathrm{Tr}(H), \quad \text{and } \mathrm{tr}(\theta^k) = \mathrm{Tr}(H^k), \quad , 1 \le k \le n-1.$$

Let $\alpha = \alpha_0 + \alpha_1\theta + \cdots + \alpha_{n-1}\theta^{n-1} \in E$, it follows that

$$\mathrm{tr}(\alpha) = \sum_{k=0}^{n-1} \alpha_i \,\mathrm{tr}\left(\theta^k\right) = \sum_{i=0}^{n-1} \alpha_i \,\mathrm{Tr}\left(H^k\right) = \mathrm{Tr}\left(\sum_{k=0}^{n-1} \alpha_i H^k\right) = \mathrm{Tr}\left(H^*(\overline{\alpha})\right).$$

To show that conclusion on the norm, let $\alpha^{(i)}(0 \leq i \leq n-1)$ be the n conjugations of α in the smallest normal extension of Q containing E, where $\alpha^{(0)} = \alpha = \alpha_0 + \alpha_1\theta + \cdots + \alpha_{n-1}\theta^{n-1}$. It is easily seen that

$$\alpha^{(i)} = \sum_{k=0}^{n-1} \alpha_k \theta_i^k, \quad \text{where } \theta_0 = \theta \text{ and } 0 \leq i \leq n-1.$$

By property (iii) of Lemma 6, we have

$$N(\alpha) = \prod_{i=0}^{n-1} \alpha^{(i)} = \prod_{i=0}^{n-1} \alpha(\theta_i) = \det\left(H^*(\overline{\alpha})\right).$$

We complete the proof of Theorem 2.

The cyclic lattice in $\mathbb{R}^n$ was introduced by Micciancio (2007), (also see Zheng et al. (2023)), which plays an important role in Ajtai's construction of collision resistant Hash function (see Ajtai and Dwork (1997)). As an application, we show that every ideal in an algebraic number field corresponds to a cyclic lattice:

Corollary 1 *Let $A \subset R$ be an ideal and $A \neq 0$, then $\tau(A) \subset Q^n$ is a cyclic lattice.*

Proof Suppose that $\alpha \in A$. Since $\theta \in R$, then $\theta\alpha \in A$. By (31), we have

$$\tau(\theta\alpha) = H\overline{\alpha} \in \tau(A).$$

Thus $\tau(A)$ is a cyclic lattice.

3 High Dimensional RSA

In this section, we give an attainable algorithm for the high dimensional RSA by making use of lattice theory, and this algorithm is significant both from the theoretical and practical point of view. Suppose that the algebraic numbers field E satisfying the NC-property, then $R = \mathbb{Z}[\theta]$ is the ring of algebraic integers of E, the restriction of correspondence τ gives a ring isomorphism from R to $\mathbb{Z}^n$. Let $\mathbb{Z}(x)$ be the ring of integer coefficients polynomials and $(\phi(x))$ be the principal ideal generated by $\phi(x)$ in $\mathbb{Z}(x)$, it is easy to see that $R \cong \mathbb{Z}[x]/(\phi(x))$. Let $M_{\mathbb{Z}}^*$ be the set of ideal matrices generated by an integral vector, i.e.

$$M_{\mathbb{Z}}^* = \left\{H^*(\overline{f}) \mid \overline{f} \in \mathbb{Z}^n\right\}. \tag{35}$$

Then the following four rings are isomorphic from each other

$$\mathbb{Z}[x]/(\phi(x)) \cong R \cong \mathbb{Z}^n \cong M_{\mathbb{Z}}^*. \tag{36}$$

For any polynomial $\alpha(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1}x^{n-1} \in \mathbb{Z}[x]/(\phi(x))$, the corresponding algebraic integer is $\alpha = \alpha_0 + \alpha_1\theta + \cdots + \alpha_{n-1}\theta^{n-1} \in R$, we write this isomorphism by

$$\alpha(x) \to \alpha \xrightarrow{\tau} \overline{\alpha} \xrightarrow{H^*} H^*(\alpha). \tag{37}$$

A ϕ-ideal lattice means an integer lattice of which corresponds an ideal of $\mathbb{Z}(x)/(\phi(x))$, it was first introduced by Lyubashevsky and Micciancio in (see also Zheng et al. (2023)), which also plays a key role in Gentry's construction for the full homomorphic cryptosystem (see Gentry (2009)), and Fluckiger and Suarez (2006) extended this definition to total real number field.

Lemma 8 *Let E be an algebraic numbers field with NC- property, $R = \mathbb{Z}[\theta]$ be the ring of algebraic integers of E. Then there is an one to one correspondence between ideals of R and the ϕ-ideal lattices. Moreover, if $\alpha \in R$, then we have*

$$\tau(\alpha R) = L\left(H^*(\overline{\alpha})\right). \tag{38}$$

In general, suppose that $A \subset R$ is an ideal and $A \neq 0$, then there exist two elements α and β in A such that

$$\tau(A) = L\left(H^*(\overline{\alpha})\right) + L\left(H^*(\overline{\beta})\right). \tag{39}$$

Proof Since there is an one to one correspondence between the ϕ-ideal lattices and the ideals of $\mathbb{Z}[x]/(\phi(x))$ (See Corollary of Zheng et al. (2023)), by (36), the first assertion follows immediately. Let $\alpha \in R$, then $\alpha R = \{\alpha x \mid x \in R\}$, by Lemma 7 we have

$$\tau(\alpha x) = H^*(\alpha)\overline{x}, \quad \text{where } \overline{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \in \mathbb{Z}^n.$$

It follows that

$$\tau(\alpha R) = \left\{H^*(\alpha)\overline{x} \mid \overline{x} \in \mathbb{Z}^n\right\} = L\left(H^*(\overline{\alpha})\right).$$

To prove (39), it is known that any ideal of R is generated by at most two elements (see corollary 5 of page 11 of Narkiewicz (2004)), namely, $A = \alpha R + \beta R$, then we have

$$\tau(A) = \tau(\alpha R) + \tau(\beta R) = L\left(H^*(\overline{\alpha})\right) + L\left(H^*(\overline{\beta})\right).$$

To introduce an attainable algorithm for high dimensional RSA, we require some basic results from lattice theory. Let $L = L(B) \subset \mathbb{R}^n$ be a full-rank lattice, and the determinant of L is defined by

$$d(L) = |\det(B)|. \tag{40}$$

Suppose that the generated matrix $B = \left[\overline{b}_1, \overline{b}_2, \cdots, \overline{b}_n\right], \overline{b}_i \in \mathbb{R}^n$ is the column vectors of B. Since $\left\{\overline{b}_1, \overline{b}_2, \cdots, \overline{b}_n\right\}$ is a basis for $\mathbb{R}^n$, let $B^* = \left\{\overline{b}_1^*, \overline{b}_2^*, \cdots, \overline{b}_n^*\right\}$ be the corresponding orthogonal basis, where $\overline{b}_1^* = \overline{b}_1$, and $\overline{b}_i^*$ is obtained by the Gram-Schmidt orthogonal process in order.

A basis B is called in Hermited Normal Form (HNF) if it is upper triangular, all elements on the diagonal are strictly positive, and any other elements b_{ij} satisfies $0 \le b_{ij} < b_{ii}$. It is easy to see that every integer lattice $L = L(B)$ has a unique basis in Hermited Normal Form, denoted by HNF(L) (see Theorem 2.4.3 of Cohen (1993)). Moreover, given any basis B for lattice L, HNF(L) can be efficiently computed from B (see Cohen (1993), Micciancio (2001)).

Proposition 1 *Let $L = L(B)$ and $B = (b_{ij})_{n\times n}$ be the basis in HNF. Then the corresponding orthogonal basis B^* is a diagonal matrix, namely*

$$B^* = \operatorname{diag}\{b_{11}, b_{22}, \cdots, b_{nn}\}. \tag{41}$$

Moreover, we have

$$d(L) = \prod_{i=1}^{n} b_{ii}. \tag{42}$$

Proof See Micciancio (2001).

Definition 3 Let $L = L(B) \subset \mathbb{R}^n$ be a full-rank lattice, and $B^* = \left[\overline{b}_1^*, \overline{b}_2^*, \cdots, \overline{b}_n^*\right]$ be the corresponding orthogonal basis, the orthogonal parallelepiped $F(B^*)$ is defined by

$$F(B^*) = \left\{\sum_{i=1}^{n} x_i \overline{b}_i^* \mid 0 \le x_i < 1 \text{ and } x_i \in \mathbb{R}\right\}. \tag{43}$$

Proposition 2 *Let $L = L(B) \subset \mathbb{Z}^n$ be an integer lattice, $B = \mathrm{HNF}(L)$ be the basis in* HNF *and $B^* = \operatorname{diag}\{b_{11}, b_{22}, \cdots, b_{nn}\}$ be the corresponding orthogonal basis, $F(B^*)$ is the orthogonal parallelepiped given by (43), then S is a set of coset representatives for the quotient group $\mathbb{Z}^n/L$, where*

$$S = F\left(B^*\right) \cap \mathbb{Z}^n = \left\{x' = (x_1, x_2, \cdots, x_n) \mid \forall x_i \in \mathbb{Z} \ \text{ and } \ 0 \le x_1 < b_{ii}\right\}.$$

Proof See Sect. 4.1 of Micciancio (2001).

Now, we return to the algebraic numbers field $E = Q[\theta]$ (with NC-property). Let $\alpha, \beta \in R$ be two algebraic integers, by Lemma 8, the principal ideal αR corresponds to the minimal ϕ-ideal lattice $L(H^*(\overline{\alpha}))$. Thus $A = (\alpha R)(\beta R) = \alpha\beta R$ corresponds to $L\left(H^*(\overline{\alpha} \otimes \beta)\right)$.

Definition 4 For given $\alpha, \beta \in R$, $\tau(\alpha) = \overline{\alpha}$, and $\tau(\beta) = \overline{\beta}$, we denote the lattice $L_{\alpha,\beta}$ by

$$L_{\alpha,\beta} = L\left(H^*(\overline{\alpha} \otimes \overline{\beta})\right). \tag{44}$$

The HNF basis of $L_{\alpha,\beta}$ is denoted by $B_{\alpha,\beta}$ and the corresponding orthogonal basis is denoted by

$$B^*_{\alpha,\beta} = \operatorname{diag}\{b_1, b_2, \cdots, b_n\}, \tag{45}$$

where $b_i \in \mathbb{Z}$ and $b_i \geq 1$. The parallelepiped is given by

$$S_{\alpha,\beta} = \left\{(x_1, x_2, \cdots, x_n) \in \mathbb{Z}^n \mid x_i \in \mathbb{Z} \quad \text{and} \quad 0 \leq x_i < b_i\right\}. \tag{46}$$

Lemma 9 *Let $\alpha \in R$, $\beta \in R$, and $A = \alpha\beta R$. Then $S_{\alpha,\beta}$ given by (46) is corresponding to a set of coset representatives of the factor ring R/A in the algebraic numbers field E with NC-property.*

Proof By Proposition 1, it is easy to see that

$$\left|S_{\alpha,\beta}\right| = \prod_{i=1}^{n} b_i = \left|\det\left(H^*(\overline{\alpha} \otimes \overline{\beta})\right)\right| = \left|\det\left(H^*(\overline{\alpha})\right)\right| \cdot \left|\det\left(H^*(\overline{\beta})\right)\right| = d\left(L_{\alpha,\beta}\right).$$

By Theorems 2 and (12), we have

$$N(A) = |N(\alpha \cdot \beta)| = |N(\alpha)| \cdot |N(\beta)| = \left|\det\left(H^*(\overline{\alpha})\right)\right| \cdot \left|\det\left(H^*(\overline{\beta})\right)\right| = d\left(L_{\alpha,\beta}\right).$$

It follows that $N(A) = \left|S_{\alpha,\beta}\right|$. Since E satisfies NC-property, if $\alpha \in R$, then $\overline{\alpha} = \tau(\alpha) \in \mathbb{Z}^n$, hence $\alpha \equiv \beta (\operatorname{mod} A)$ in R, if and only if

$$\overline{\alpha} \equiv \overline{\beta}\left(\operatorname{mod} L_{\alpha,\beta}\right).$$

The lemma follows from Proposition 2 immediately.

The main result of this subsection is the following theorem.

Theorem 3 *Let E be an algebraic numbers field of degree n with NC-property, $\alpha \in R$, $\beta \in R$ be two distinct prime elements, $A = \alpha\beta R$, and $L_{\alpha,\beta}$ be the lattice given by (44). Then for any $\overline{a} \in \mathbb{Z}^n$, $k \in \mathbb{Z}$, $k \geq 0$, we have*

$$\overline{a}^{\otimes(k\varphi(\alpha,\beta)+1)} \equiv \overline{a}\left(mod \quad L_{\alpha,\beta}\right), \tag{47}$$

where

$$\varphi(\alpha, \beta) = \left(\left|\det\left(H^*(\overline{\alpha})\right)\right| - 1\right)\left(\left|\det\left(H^*(\overline{\beta})\right)\right| - 1\right). \tag{48}$$

Proof Since E satisfies NC-property, $\overline{a} \in \mathbb{Z}^n$, then $a = \tau^{-1}(\overline{a}) \in R$. By Theorem 1, we have

$$a^{k\varphi(A)+1} \equiv a(\operatorname{mod} \quad A).$$

It is easy to see that

Table 3 Algorithm I

Algorithm 9.1: RSA in the Algebraic Numbers Field

$n \geq 1$ is a positive integer, E/Q is an algebraic numbers field with NC-property of degree n, $R \subset E$ is the ring of algebraic integers of E, $\alpha \in R$, $\beta \in R$ are two distinct prime elements of R, $A = \alpha\beta R$ is a principal ideal of R, $H^*(\overline{\alpha} \otimes \overline{\beta})$ is the ideal matrix corresponding to A, $L_{\alpha,\beta} = L\left(H^*(\overline{\alpha} \otimes \overline{\beta})\right)$ is the lattice generated by $H^*(\overline{\alpha} \otimes \overline{\beta})$, $B_{\alpha,\beta} = \mathrm{HNF}\left(L_{\alpha,\beta}\right)$ is the basis of $L_{\alpha,\beta}$ in HNF, $B^*_{\alpha,\beta} = \mathrm{diag}\{b_1, b_2, \cdots, b_n\}$ is the corresponding orthogonal basis

- **Parameters**: $\varphi(\alpha, \beta) = (|\det(H^*(\overline{\alpha}))| - 1)\left(\left|\det\left(H^*(\overline{\beta})\right)\right| - 1\right)$, $S_{\alpha,\beta} = \{x' = (x_1, x_2, \cdots, x_n) \in \mathbb{Z}^n \mid 0 \leq x_i < b_i\}$. $1 \leq e < \varphi(\alpha, \beta)$, $1 \leq d < \varphi(\alpha, \beta)$, such that $ed \equiv 1 (\mathrm{mod}\ \varphi(\alpha, \beta))$
- **Public keys**: The rotation matrix H, the lattice $L(B_{\alpha,\beta}) = L_{\alpha,\beta}$ and the positive integer e are public keys
- **Private keys**: Ideal matrices $H^*(\overline{\alpha})$, $H^*(\overline{\beta})$, the basis $H^*(\overline{\alpha} \otimes \overline{\beta})$ of $L_{\alpha,\beta}$ and positive integer d are private keys.
- **Encryption**: For any input message $\overline{a} \in S_{\alpha,\beta}$, the ciphertext $\overline{c}$ is given by $\overline{c} \equiv \overline{a}^{\otimes e} (\mathrm{mod}\ \ L_{\alpha,\beta})$.
- **Decryption**: $\overline{c}^{\otimes d} \equiv \overline{a}^{\otimes de} \equiv \overline{a}^{\otimes(k\varphi(\alpha,\beta)+1)} \equiv \overline{a} (\mathrm{mod}\ L_{\alpha,\beta})$. One can find the plaintext $\overline{a}$ from $\overline{c}$ in $S_{\alpha,\beta}$

$$\begin{aligned}\varphi(A) = \varphi(\alpha R)\varphi(\beta A) &= (N(\alpha R) - 1)(N(\beta R) - 1)\\ &= (|N(\alpha)| - 1)(|N(\beta)| - 1)\\ &= \left(\left|\det\left(H^*(\overline{\alpha})\right)\right| - 1\right)\left(\left|\det\left(H^*(\overline{\beta})\right)\right| - 1\right)\\ &= \varphi(\alpha, \beta).\end{aligned}$$

By Lemma 8, we have

$$\tau(A) = \tau(\alpha\beta R) = L\left(H^*(\overline{\alpha} \otimes \overline{\beta})\right) = L_{\alpha,\beta} \ \text{ and } \ \tau\left(a^{k\varphi(\alpha,\beta)+1}\right) = \overline{a}^{\otimes(k\varphi(\alpha,\beta)+1)}.$$

Therefore, (47) follows immediately.

According to the above theorem, we may describe an attainable algorithm for high dimensional RSA as follows (Table 3).

Remark 2 If the class number $h_E = 1$, in other words, R is a UFD, then the prime elements are equivalent to irreducible elements in R, and one can find prime elements α from $\alpha(x) \in \mathbb{Z}[x]/(\phi(x))$ and $\alpha(x)$ irreducible.

4 Security and Example

The classical RSA public key cryptosystem is nowadays used in a wide variety of applications ranging from web browsers to smart cords. Since its initial publication in 1978, many researchers have tried to look for vulnerabilities in the system. Some clever attacks have been found (see Bonech (2002), Coppersmith (2001)). However, none of the known attacks is devastating and the ordinary RSA system is still considered secure.

The security of high dimensional RSA depends on virtually factoring of an element of the algebraic integers ring R into product of of distinct prime elements. Factoring on R is much more complicated than factoring of a positive integer, and none of efficient method is known up to day, thus we consider the high dimensional RSA almost absolutely secure.

To see the size of private keys, since $\det(H^*(\overline{\alpha})) = N(\alpha)$, it may be extremely huge, for example, if $\alpha = p \in \mathbb{Z}$, $\beta = q \in \mathbb{Z}$ are prime numbers, then

$$\det\left(H^*(\overline{\alpha})\right) = N(\alpha) = p^n, \quad \det\left(H^*(\overline{\beta})\right) = q^n$$

and

$$\varphi(\alpha, \beta) = \left(p^n - 1\right)\left(q^n - 1\right),$$

which is much larger than pq, the latter is the site of public key of the classical RSA cryptosystem.

The lattice based on cryptography has been intensively studied for the past two decades. The GGH cryptosystem proposed by Goldreich et al. (1997) is perhaps the most intuitive encryption scheme based on lattices. The public key is a "bad" basis for a lattice, and Micciancio proposed in (2001) to use, as the public basis, the Hermite Normal Form $B = \mathrm{HNF}(L)$. The private key of GGH is an exceptionally good basis for L. The security of GGH relies on the assumption that it is difficult to find a special basis for L from a known basis of L. In this sense, we regard the high dimensional RSA as secure as GGH/HNF cryptosystem at least.

Another number theoretic cryptosystem based on the lattice is NTRUEncrypt. The public key cryptosystem NTRU proposed in 1996 by Hoffstein et al. (1998) is the fastest known lattice-based encryption scheme, although its description relies on arithmetic over polynomial quotient ring $Z[x]/\langle x^n - 1\rangle$, it was easily observed that it could be expressed as a lattice based on cryptosystem. NTRU uses a q-ary convolutional modular lattice(see Micciancio and Regev (2009), Zheng (2022)), its public key is also the HNF basis of L, and the private key is a special basis of L containing two secrete polynomials $f(x)$ and $g(x)$. Obviously, our algorithm I is at least as hard as solving NTRUEncrypt.

Unfortunately, neither GGH nor NTRU is supported by a proof of security showing that breaking the cryptosystem is at least as hard as solving some underlying lattice problem; they are primarily practical proposals aimed at offering a concrete alternative to RSA or other number theoretic cryptosystems (see page 166 of Mic-

ciancio and Regev (2009)). However, the significance of this chapter is to show that the real alternative of RSA is the high dimensional RSA we present here rather than GGH and NTRU.

Example 1 Finally, we give an example and see how to work the high dimensional RSA in a quadratic field. Let $E = Q(\sqrt{d}), d \in \mathbb{Z}$ be a square-free integer and $d \equiv 2$, or 3 mod 4, thus E satisfies the NC-property. Let δ_E be the discriminant of E, and it is known that $\delta_E = 4d$ (see Proposition 13.1.2 of Ireland and Rosen (1990)). Let $p \in \mathbb{Z}$ be an odd prime satisfying the following condition:

$$p \nmid 4d, \quad \text{and} \quad x^2 \equiv d(\text{mod } p) \text{ is not solvable in } \mathbb{Z}. \tag{49}$$

By Proposition 13.1.3 of Ireland and Rosen (1990), we know that p is a prime element in E.

According to Algorithm I, we select two large primes p and q of which satisfying (49). Let $\alpha = p$ and $\beta = q$, then

$$\bar{\alpha} = \begin{pmatrix} p \\ 0 \end{pmatrix}, \ \bar{\beta} = \begin{pmatrix} q \\ 0 \end{pmatrix}, \ H^*(\overline{\alpha}) = \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}, \ \text{and} \ H^*(\overline{\beta}) = \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix}.$$

It follows that

$$H^*(\overline{\alpha} \otimes \overline{\beta}) = H^*(\overline{\alpha}) H^*(\overline{\beta}) = \begin{pmatrix} pq & 0 \\ 0 & pq \end{pmatrix}, \ L_{\alpha,\beta} = L\left(H^*(\overline{\alpha} \otimes \overline{\beta})\right) \tag{50}$$

and

$$S_{\alpha,\beta} = \left\{ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{Z}^2 \mid 0 \le x_1, x_2 < pq \right\}. \tag{51}$$

It is easy to see that

$$\varphi(\alpha, \beta) = (p^2 - 1)(q^2 - 1). \tag{52}$$

In this special case, the two-dimensional RSA may be described as follows (Table 4).

We can similarly deal with the cases of Cyclotomic Fields. Let $n = \varphi(m)$ for some positive integers m, $\xi_m = e^{2\pi i/m}$, $E = Q(\xi_m)$, and $R \subset E$ be the ring of algebraic integers of E. Suppose that $p \in \mathbb{Z}$ is a rational prime number, then p is a prime element of R if and only if (see Theorem 2 of page 196 of Ireland and Rosen (1990))

$$p \nmid m \ \text{ and } \ p^{\varphi(m)} \equiv 1(\text{mod } m). \tag{53}$$

Suppose that $p \in \mathbb{Z}$ and $q \in \mathbb{Z}$ are two distinct prime numbers satisfying (53), we obtain the lattice $L(H^*(\overline{p} \otimes \overline{q}))$ and an attainable algorithm in $Q(\xi_m)$.

Table 4 RSA in a quadratic field

RSA in A Quadratic Field
• **Parameters**: $E = Q(\sqrt{d})$, d is a square-free integer and $d \equiv 2$ or $3(\text{mod } 4)$ the rotation matrix $H = \begin{pmatrix} 0 & d \\ 1 & 0 \end{pmatrix}$, p, q are two large and distinct prime numbers of which satisfy (49). $N = pq$ and $\chi(N) = \left(p^2 - 1\right)\left(q^2 - 1\right)$ $L = L(B)$ is a lattice, $B = \begin{pmatrix} \mathrm{N} & 0 \\ 0 & \mathrm{N} \end{pmatrix}$. $1 \le e < \chi(N)$, $1 \le d_1 < \chi(N)$ such that $ed_1 \equiv 1(\text{mod } \chi(N))$
• **Public keys**: H, N and the positive integer e are public keys
• **Private keys**: p, q and the positive integer d_1 are private keys
• **Encryption**: For any $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \in \mathbb{Z}_{pq}^2$, the ciphertext $c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \in \mathbb{Z}^2$ given by $c \equiv a^{\otimes e}(\text{mod } L)$
• **Decryption**: $c^{\otimes d_1} \equiv a^{\otimes d_1 e} \equiv a(\text{mod } L)$. One can find the plaintext a from c in $\mathbb{Z}_{pq}^2$

References

Ajtai, M., & Dwork, C. (1997). A Public-key cryptosystem with worst-case/avarage—Case equivalence. In *29th ACM Symposium on Theory of Computing* (pp. 284–293).

Bonech, D. (2002). Twenty years of attacks on the RSA cryptosystem. *Notices of the AMS, 46*(2), 203–213.

Coppersmith, D. (2001). Finding small solutions to small degree polynomials. *Lecture Notes in Computer Science, 2146*, 20–31.

Cohen, H. (1993). A course in computational algebraic number theory. In *Graduate texts in mathematics*. Springer.

Davis, P. J. (1994). *Circulant matrices* (2nd ed.). New York: Chelseea Publishing.

Fluckiger, E. B., & Suarez, I. (2006). Ideal lattices over totally real number fields and Euclidean minima. *Archiv Der Mathematik, 86*(3), 217–225.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In *STOC* (pp. 169–178).

Goldreich, O., Goldwasser, S., & Halevi, S. (1997). Public-key cryptosystems from lattice reduction problems. In *Advances in cryptology, lecture notes in computer* (Vol. 1294, pp. 112–131)

Hoffstein, J., Pipher, J., & Silverman, J. H. (1998). NTRU: A ring-based public key cryptosystem. In *Proceedings of ANTS-III* (Vol. 1423, pp. 267–288). LNCS.

Ireland, K., & Rosen, M. (1990). *A classical introduction to modern number theory*. Springer.

Lyubashevsky, V., & Micciancio, D. (2006). Generalized compact knapsacks are collision resistant. In *33rd international conference on automata, languages and programming* (Vol. 2, pp. 144–155). Springer.

Manin, Y. I., & Panchishkin, A. A. (2005). *Introduction to modern number theory: Fundamental problems ideas and theories*. Springer.

Micciancio, D., & Regev, O. (2009). Lattice-based cryptography, post quantum cryptography (pp. 147–191). Springer.

Micciancio, D. (2007). Generalized compact knapsacks, cyclic lattices, and efficient one way functions. *Computational Complexity, 16*(4), 365–411.

Micciancio, D. (2001). Improving Lattice based cryptosystems using the Hermite normal form. In *CaLC* (pp. 126–145). Springer.
Narkiewicz, W. (2004). *Elementary and analytic theory of algebraic numbers*. Springer.
Peikert, C. (2014). A decade of lattice cryptography. *Foundations and Trends in Theoretical Computer Science, 10*(4), 2–3.
Pradhan, P. K., Rakshit, S., & Datta, S.: Lattice based cryptography. In *Proceedings of the Third International Conference on Computing Methodologies and Communication, ICCMC*.
Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM, 21*, 120–126.
Takagi, T., & Naito, S. (2015). Construction of RSA cryptosystem over the algebraic field using ideal theory and investigation of its security. *Electronics and Communications in Japan (Part III Fundamental Electronic Science), 83*(8), 19–29.
Uematsu, Y., et al. (1985). On the extension of RSA cryptosystem. *Tech Rep, 1985*, 85–89.
Uematsu, Y., et al. (1986). A note on extension of RSA cryptosystem and consideration of amount of computation. In *Encryption and Information Security Work Shop* (pp. 27–29).
Washington, L. C. (1982). Introduction to cyclotomic fields (graduate texts in mathematics) (Vol. 83). Springer.
Zheng, Z. Y., Liu, F. X., Lu, Y. F., & Tian, K. Cyclic lattices, ideal lattices and bounds for the smoothing parameter. Tech. https://doi.org/10.36227/techrxiv.17626391.v1.
Zheng, Z. Y., Liu, F. X., Xu, J., Huang, W. L., & Tian, K. A generalization of NTRUEncrypt. arXiv:2112.14115 [cs.IT].
Zheng, Z. Y. (2022). Modern cryptography Volume 1—A classical introduction to informational and mathematical principle. Springer.

Central Bank Digital Currency Cross-Border Payment Model Based on Blockchain Technology

Mao Hanyu

Abstract Since the turn of the twenty-first century, the growth of the globalized economy and trade has accelerated, and the cross-border payment system, which is an essential component of the international financial infrastructure, has played a significant role in the global economy and trade. However, traditional cross-border payments present risks and challenges, such as expensive processing fees, limited payment efficiency, information asymmetry in the trade process, and reliance on a highly centralized cross-border payment system. This chapter is based on consortium blockchain technology and utilizes Polkadot's Parachain, Relay chain, and cross-chain technologies as references; a scalable, high-efficiency, high-security, and privacy-protecting central bank digital currency cross-border payment model is designed. Analyzed the usage of hash digest technology and CoinJoin technology to avoid the tracing of transactions in order to protect privacy. The issuance of multi-country central bank digital currency or stablecoin anchored to a basket of fiat currencies is discussed as the currency in circulation in the model. Finally, the central bank digital currency cross-border payment development trend is summarized and forecasted.

Keywords Payment model · Cross-border · Blockchain technology · CBDC

1 Introduction

Since 2020, the global digital transformation has been developing rapidly, and the era of central bank digital currency (CBDC) is accelerating, with China's digital currency—e-CNY leading the world. The launch of the e-CNY not only promotes the healthy development of China's digital economy but also benefits the RMB internationalization plan and speeds up the pace of RMB internationalization. At the same time, CBDC has especially significant advantages in cross-border payments,

M. Hanyu (✉)
School of Mathematics, Renmin University of China, Beijing, China
e-mail: 2020103607@ruc.edu.cn

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_10

which can effectively solve the problems of extended time, high cost, low efficiency, and low transparency faced by current cross-border payments. In addition, building a cross-border payment network system based on CBDC will also be a pivotal key to unlocking the opportunity to break the monopoly position of the US dollar and reshape the global cross-border payment system. Therefore, CBDC cross-border payments will inject new vitality into the rapid growth of our economy and will also play a pivotal role in establishing a fair and equitable international monetary settlement system.

With the rapid development of CBDC, CBDC cross-border payments are becoming a research hotspot in the central bank's digital currency research area. According to a survey by BIS, more than 50% of central banks consider cross-border payments as one of the crucial reasons for accelerating the development of CBDC. Traditional cross-border payments suffer from high fees, low efficiency, information asymmetry in the cross-border trade process, and the potential financial risk of a highly centralized cross-border payment system. The CBDC cross-border payment system, with the characteristics of high payment efficiency, low cost, and high transparency, is not only conducive to solving the current existence of cross-border trade friction and breaking the centralized cross-border payment system, but also conducive to eliminating the use of competitive currency devaluation, currency war, and other vicious behaviors between countries, promoting the peaceful development of financial markets, and laying a moderately centralized cross-border payment system with a healthy market foundation for international trade (Yang, 2020). Therefore, a large number of central banks and international organizations have started to try to explore the application of CBDC in cross-border payments. On February 26, 2022, the United States, together with the European Union, the United Kingdom, and Canada, issued a joint statement announcing that Russia is banned from using the Society for Worldwide Interbank Financial Telecommunications (SWIFT) international settlement system. It undoubtedly accelerated the research of countries investigating the idea of bypassing SWIFT for cross-border transactions.

Currently, the research on cross-border payment of central bank digital currency is still in the initial stage. There is a lack of in-depth research for a scalable and high-efficiency cross-border payment model, which leads to a lack of necessary theoretical research and essential technical support for its development. Therefore, it is significant to design a scalable, high-efficiency, CBDC cross-border payment model based on blockchain.

CBDC cross-border payments issued by central banks have become a significant trend. In this chapter, we use Polkadot's Parachain, Relay Chain, and cross-chain technologies as references for the CBDC cross-border payment model, and we commit to designing a scalable, high-efficiency, highly secure, and privacy-preserving CBDC cross-border payment model based on consortium chain.

2 CBDC Cross-Border Payment Development Current Situation

CBDC cross-border payments can be made in two ways: first, retail central bank digital currencies (CBDCs) in a given jurisdiction are available to people both inside and outside the jurisdiction, with no coordination between central banks; and second, central banks work together to establish access and settlement arrangements between different retail or wholesale CBDCs (Wan & Wu, 2022). CBDC cross-border payments can be divided into four quadrants: "same system and same currency", "same system and different currency", "same currency and different currency", "same currency and different system", and "different currency and different system". Among them, "same system and different currency" and "different currency and different system" are the most typical scenarios for cross-border payments and will be a key research focus in the future.

At this stage, for CBDC cross-border payments research, the following three models are used to achieve cross-border and cross-currency interoperability, enhancing compatibility of CBDCs systems; linking multiple CBDC systems; integrating multiple CBDCs in a single multi-CBDC (mCBDC) system (Auer et al., 2021). Models linking multiple CBDC systems include the Stella project of the European Central Bank and the Bank of Japan (2019); and the Jasper-Ubin project of the Bank of Canada (BOC) and the Monetary Authority of Singapore (MAS) (2019). Jura project for cross-border payment between Banque de France and Swiss National Bank. Integrating multiple CBDCs in a single mCBDC system mainly contains the Aber project of the UAE and the Central Bank of Saudi Arabia (2020); Dunbar, a joint project of the Monetary Authority of Singapore and the BIS (2022); and the Inthanon-LionRock project of the Bank of Thailand and the Hong Kong Monetary Authority (2020). In 2021, with the addition of the Digital Currency Institute of the People's Bank of China and the United Arab Emirates Bank, the project evolved into its third phase. It was renamed the mCBDC Bridge (mBridge) Project (Inthanon-LionRock to mBridge-Building a multi CBDC platform for international payments, 2021).

Recently, the CBDC projects Jasper, Ubin, and Stella have completed their experiments. All these projects continue the line "from wholesale payments to voucher payments to cross-border payments" (Yao, 2021). Thus, enabling cross-border payments is the ultimate goal and an essential part of the CBDC research route. Moreover, the experimental results of these representative wholesale CBDC projects show that current technology and design solutions can support Real-Time Gross Settlement (RTGS) in terms of efficiency and can also realize Liquidity Saving Mechanism (LSM) in terms of functionality (Huang, 2022). Also, these projects show that cross-chain technology is a crucial issue for CBDC cross-border payments. Although CBDC cross-border payments have become a research hotspot both at home and abroad, most existing research scholars focus on the two fields of economics and law for CBDC cross-border payments, and the proposed CBDC cross-border payment model has not been sufficiently investigated on a technical level.

At the technical level, most research by central banks or international organizations has focused on linking multiple CBDC systems and integrating multiple CBDCs in a single mCBDC system. However, most projects are still at the experimental stage with the participation of only a few countries and lack a certain degree of scalability in practice. At the same time, the cross-chain technology used to link multiple CBDC systems, hash time-locked contract (HTLC), is limited in its application scenarios, where the two sides of a transaction need to establish N^2 magnitude of transaction channels between them, and the number of transaction channels grows in power as N increases. Therefore, the scalability of hash time-locked contract (HTLC) is deficient and may not be suitable for application to large-scale economies. While integrating multiple CBDCs in a single mCBDC system can avoid complex hash time-lock contract (HTLC) and improve payment efficiency, the establishment of privacy groups inevitably introduces multi-ledger-style behaviors and constraints that hinder the realization of transaction atomicity. Therefore, research on cross-chain technology and the introduction of effective privacy protection mechanisms to achieve transaction atomicity and improve transaction efficiency while ensuring transaction privacy are the focus of future research.

3 Polkadot Technology Overview

Polkadot is a scalable heterogeneous multi-chain technology that provides a more general cross-chain protocol. Any blockchain system compatible with Polkadot's cross-chain protocol will be able to complete cross-chain interconnection (Polkadot, 2016). Polkadot is envisioned as a new form of blockchain "blockchain network" and one of the critical infrastructures of the future web 3.0. As shown in Fig. 1, Polkadot is completed with Parachain, relay chain, and bridge. It uses various Parachain technologies to satisfy the needs of different applications. It uses Relay chain technology to unify the management of consensus security and data interaction, which can solve the scalability and isolability problems of current blockchain technology.

3.1 Relay Chain and Parachain Technology

The Parachain is a member blockchain of Polkadot that collects and processes transactions and transmits them to a Relay chain. Each participating Parachain has a high degree of autonomy and flexibility. Each Parachain can be designed and focused on a specific scenario as long as it follows the protocols set by Polkadot. The Relay chain is the core of the Polkadot network, responsible for maintaining the whole network's security, coordinating consensus among different Parachains, and forwarding cross-chain transactions between each Parachain. The consensus mechanism of the Relay chain uses an asynchronous Byzantine fault-tolerant algorithm to reach consensus.

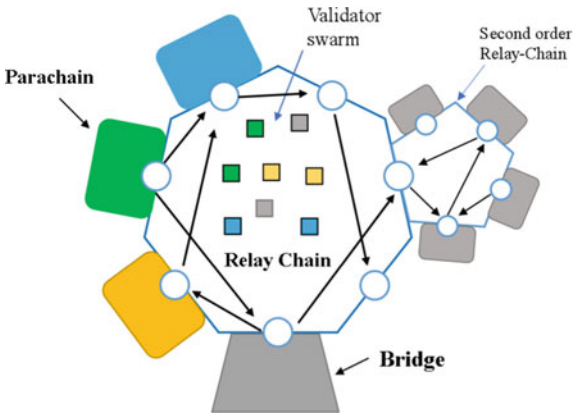

Fig. 1 Polkadot architecture schematic

In order to maintain the relay chain, the Polkadot network establishes four roles: Nominator, Validator, Collator, and Fisherman. Nominators are a group of token (DOT) holders who have the authority to vote for the Validator. Validator nodes have the highest authority in the network, having the ability to create blocks for the whole network. They are elected by Nominator vote and can validate blocks and pack blocks after a sufficient deposit (TOKEN) is mortgaged in the system. If the validators perform their duties, they are rewarded for generating blocks. If the validators don't perform their duties, they are punished by having some or all of their deposit deducted. The collators are a group of nodes that collect information from the Parachain and package it for submission to the validators. They submit candidate blocks to the validators and assist them in creating valid blocks, which are rewarded with a fee. Collators will go and collect as much information as possible in order to get more fees. Fisherman is a relatively independent node in the system. It is only responsible for monitoring the system's illegal activities and reporting their detection. Then it will receive a substantial one-time reward. Moreover, a deposit is required to become a Fisherman, mainly used to prevent Sybil Attack by witches that waste the verifier's computing time and resources.

3.2 Polkadot Cross-Chain Technology

Cross-chain communication is the most critical part of Polkadot, as shown in Fig. 2. Because of the relay chain's security guarantee for the whole system, transactions conducted on one Parachain can be transferred to another Parachain through the relay chain. As a result, cross-chain transactions on Polkadot are simpler and more efficient than other cross-chain methods. Specifically, each Parachain maintains an egress and an ingress transaction queue. The queue uses Merkle trees to ensure data authenticity. When a Parachain (A) initiates a cross-chain transaction to another Parachain (B), the transaction is pushed to Parachain A's egress queue. Then the

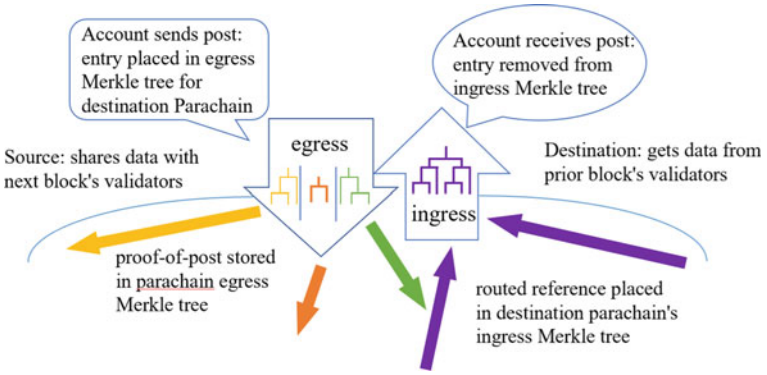

Fig. 2 Schematic diagram of cross-chain transactions on Polkadot This figure from Polkadot White Paper Polkadot (2016)

relay chain transfers the transactions in Parachain A's egress queue to Parachain B's ingress queue, which then processes the transactions in its ingress queue itself (Yuan & Wang, 2019).

4 CBDC Cross-Border Payment Model

This section will introduce a CBDC cross-border payment model based on consortium blockchain technology, referencing Polkadot's Parachain, Relay Chain, and cross-chain technologies. Figure 3 depicts a scalable, high-efficiency, high-security, and privacy-protecting CBDC cross-border payment model.

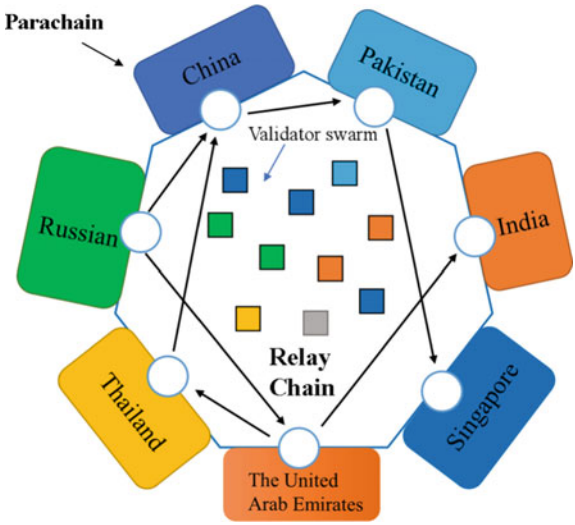

Fig. 3 CBDC cross-border payments model schematic

4.1 Design of Parachain

Every country is a Parachain in this model, and each Parachain is a consortium blockchain. The consortium blockchain is a permissioned blockchain, meaning only the internal designation of several nodes can upload, record, and read data. These nodes act as bookkeeper nodes, and they collectively decide to generate blocks. Using consortium blockchain can significantly improve the blockchain's operational efficiency and reduce network latency, all while ensuring the privacy of each transaction's data. Therefore, Parachain in each country can be adopted in the form of consortium blockchains, which can achieve the purpose of improving efficiency and protecting privacy.

Since Parachain has a high degree of autonomy and flexibility, each country's Parachain can be designed independently according to its own country's conditions. Therefore, the consensus mechanism of each country's Parachain can be chosen according to the country's reality, such as the mainstream consensus mechanisms applied to the consortium blockchain: Raft, PBFT, etc. The Parachain of each country can be divided into several nodes with different authorities according to the actual situation of cross-border payment in the country, including the central bank, trusted financial payment institutions, and regulatory agencies. As a result, three roles are established in the network of this model: Validators, Collators, and Supervisors.

The work of Collators is to collect information on the Parachain, submit candidate blocks to the Validators, and assist the group of validators in creating valid blocks. They also have the authority to vote for the Validators. Consequently, the Collators can be commercial banks and trusted financial payment institutions in every country. Validators are the nodes with the authority to generate blocks and have the highest authority in the system. The Validators nodes are elected by vote of the Collators and are responsible for validating the blocks and packaging them. Each country's central bank or specialized agency can fill this critical role. Supervisors are the nodes that need to be responsible for regulating illegal activities in the system. Thus, it can be held by the regulator of each country.

Taking China as an example, the nodes of Collators can be served by six state-owned commercial banks, including Industrial and Commercial Bank of China (ICBC), Agricultural Bank of China (ABC), Bank of China(BOC), China Construction Bank (CCB), Bank of Communications (BCM), and Postal Savings Bank of China (PSBC), and trusted commercial banks and third-party payment institutions can be added in the future. Validators are the central bank of China, the People's Bank of China, and its affiliated institutions. Supervisors are mainly served by the Ministry of Commerce, the China Banking and Insurance Regulatory Commission, the National Audit Office, and other regulatory authorities.

4.2 Design of Relay chain

In this model, the Relay chain is the same as the consortium blockchain, which contains the protocols of all Parachains, can recognize the transaction format of each country's Parachain, and can be responsible for coordinating consensus and forwarding cross-chain transactions between different Parachains. The Validator nodes, which the Collators vote on in each country, are also added to the relay chain and are responsible for packaging transactions and generating blocks.

Specifically, the Collators in each country's Parachain first elect the Validators in charge of their Parachain, and the Validators are added to the Relay chain. After that, the Collators on each country's Parachain will collect the transactions into the blocks with a Noninteractive Zero-Knowledge Proof, which is used to prove that the father block of this child block is valid, and hand them over to the Validators in charge of their country's Parachain. The Validators from each country involved in this cross-border transaction form a team of Validators to validate the blocks in the order in which the Collators send them and then consensus out the Parachain blocks for that height. When the Validators of each country's Parachain involved in cross-border payments confirm that their country's Parachain has confirmed the transaction, the Validators group then routes the message to the Relay chain and generates the Relay chain blocks. In the next round, the Collators in each country's Parachain vote again to elect new validators and round this cycle.

4.3 Cross-Chain Transaction

The cross-chain transactions of the model are approximately the same as those of Polkadot. Each country's Parachain contains an egress transaction queue and an ingress transaction queue (there can be multiple exports and ingresses if the transaction volume is large). The Relay chain transfers transactions from the egress transaction queue at the source Parachain to the ingress transaction queue at the destination Parachain.

The egress transaction queue contains a list of grids with routing information, each with a concatenated structure of exit submissions. Merkle tree proofs can be provided between verifiers of Parachains so that blocks of one Parachain can be proven to correspond to the egress transaction queue of another Parachain, guaranteeing data authenticity. If the ingress transaction queue of a Parachain exceeds the block processing threshold, it is marked as complete on the relay chain, and no new messages are received until the queue is emptied. The Merkle tree is used to prove that the collector's operations in the Parachain blocks are trustworthy.

For example, the flow of a cross-border transaction between China and Russia is as follows. When a Chinese Parachain launches a cross-chain transaction to a Russian Parachain, this transaction will first be pushed to the Chinese Parachain's egress transaction queue. Then the Relay chain will transfer this transaction from

the Chinese egress transaction queue to the Russian ingress transaction queue. Then the Russian Parachain will process the transaction in the ingress queue. This design can effectively guarantee the security of cross-chain transactions and significantly improve the efficiency of cross-chain transactions.

4.4 Privacy Protection

In Parachain, the blockchain ledger takes advantage of the irreversible nature of the hash algorithm and uses hash digests instead of transaction-sensitive information. At the same time, CoinJoin technology is used to obfuscate transactions and sever the relationship between the input and output addresses of transactions so that the origin and destination of transactions cannot be traced for privacy protection.

In Relay chain, if other countries are not involved, the block can be generated by only the countries involved in cross-border transactions confirming the transactions. The relevant detailed information does not need to be authenticated by the nodes of other countries, which can prevent other countries from knowing the details of the transactions and can effectively protect the privacy of cross-border transaction information for each country.

5 CBDC Cross-Border Payment Model Architecture

The whole cross-border model is divided into four layers, as shown in Fig. 4. They are application layer, architecture layer, blockchain layer, and digital currency issuance layer.

Application layer: This layer mainly faces users and can provide user identity authentication services, system access services, etc. The authentication technology verifies the user's identity through the authentication center to ensure the validity of the trader's identity. Users can access the system if they pass authentication.

Architecture layer: This layer is composed of Parachain, which is designed by each country, and Relay chain, which is responsible for forwarding cross-chain transactions, as shown in Fig. 3. Parachain and Relay chain are consortium blockchains that are ideal for practical applications. Three trusted roles are established in the network of the model: Validators, Collators, and Supervisors, to help the whole system work more effectively.

Blockchain layer: This layer consists of the core technical aspects of blockchain, such as Peer-to-Peer networks, Smart Contracts, Time stamps, and Consensus mechanisms. Distributed ledger technology resolves the problem of storing, transferring, and querying transaction information in cross-border payments. The consensus mechanism solves the agreement between validators on transactions and ledgers (Zhu, 2021). Resolving the problem of double payment through Digital signature and Time Stamp, Smart Contract technology can realize automatic accounting rec-

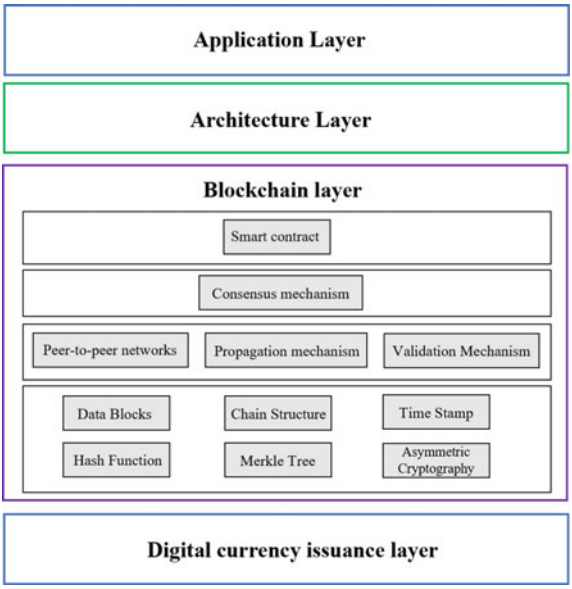

Fig. 4 CBDC cross-border payment model architecture diagram

onciliation and error handling in cross-border payments, ensuring that transactions are trusted and reliable. The smart contract automatically identifies and executes the actual conditions and matches the situations that arise to the relevant processing rules. This way, all information is recorded between the parties synchronously and cannot be tampered with by either party. It can also effectively prevent the loss of information due to technical failures (Huang & Luo, 2021).

Digital currency issuance layer: It mainly corresponds to the issuance and redemption of the digital currency used in this model, as well as the management and maintenance of this digital currency. Since the essence of building a cross-border payment system based on CBDC is to establish a regional economic association, regional economic cooperation is a prerequisite for CBDC to be recognized in cross-border payments. Therefore, a multi-country CBDC anchored to a basket of legal currencies is considered to be issued as the circulating currency in the model; or a stable currency anchored to a basket of legal currencies is issued as the circulating currency. This digital currency should only be used for cross-border payment clearing between Parachains of individual countries. It cannot be freely exchanged or used outside this cross-border payment model between financial institutions on the parallel circulation chain. The intrinsic value and purchasing power of this digital currency can be determined by each party's central basket of traded goods based on historical transaction volumes (or other forms). In this way, it can bypass the US dollar settlement and circumvent the constraints of countries' foreign exchange reserves anchored by the US dollar without challenging the monetary sovereignty of national central banks (Huang & Luo, 2021).

6 Summary and Prospect

This chapter utilizes Polkadot's Parachain, Relay chain, and cross-chain technologies as references and is based on consortium blockchain technology; a scalable, high-efficiency, high-security, and privacy-protecting CBDC cross-border payment model is designed. The purpose of privacy protection is investigated by using hash digest technology and CoinJoin technology to obfuscate the input address and output address of transactions. The construction of a free-floating legal digital currency system bypassing U.S. dollar settlement is discussed so that multi-country CBDCs anchored to a basket of legal currencies can be issued, or stablecoins anchored to a basket of legal currencies can be issued as the circulating currencies in the model in order to contribute to the study of CBDC cross-border payments.

With the increasing perfection of CBDC cross-border payment technology and the more mature development of cross-chain technology, the future is expected to form a regional-centric polycentric pattern. A new pattern of economic development in which "different currencies in the same system" are used within a region and "different currencies in different systems" are used between areas in the future. Nowadays, CBDC cross-border payment has become an international research hotspot. Although some countries are already experimenting with CBDC cross-border payment, the security and scalability of its cross-chain still need time to be proven. Follow-up research can focus on the following two levels: Technically, the focus and difficulty of CBDC cross-border payments lie in cross-chain technology, so the research of more secure and efficient cross-chain technology is a hotspot for future research. At the same time, the research of new consensus algorithms that can be applied to blockchain cross-chain will also greatly improve the efficiency of CBDC cross-border payments. In terms of regulation, it is necessary to strengthen the supervision of CBDC cross-border payments; it is not only essential to identify the regulatory authority of CBDC cross-border payment but also to improve the legal study of CBDC cross-border payment. In addition, we can also learn from the regulatory sandbox model and introduce a new model of the "Chinese regulatory sandbox" to balance risk and innovation.

References

Auer, R., Haene, P., & Holden, H. (2021). Multi-CBDC arrangements and the future of cross-border payments.

Bank of Canada & Monetary Authority of Singapore. (2019). Jasper-Ubin design chapter-enabling cross-border high value transfer using distributed ledger technologies.

Bank of Thailand, Hong Kong Monetary Authority. (2020). Inthanon-LionRock Cleveraging distributed ledger technology to increase efficiency in cross border payments.

BIS. (2022). CPMI, innovation hub, project Dunbar, international settlements using multi-CBDCs.

European Central Bank & the Bank of Japan. (2019). Project Stella-Sync-Hronised cross-border payments.

Huang, G. P. (2022). Cross-border payments of digital currency. *The Chinese Banker, 2022*(4) (in Chinese).
Huang, H. T., & Luo, C. (2021). Constructing the cross-border trade trust mechanism supported by blockchain: A prospective analysis on the trade between China and the Five Central Asian States. *Nankai Journal (Philosophy, Literature and Social Science Edition),2021*(02), 98–110 (in Chinese).
Inthanon-LionRock to mBridge-Building a multi CBDC platform for international payments. (2021).
Polkadot, W. G. (2016). Vision for a heterogeneous multi-chain framework: Draft 1.
Qian, Y. (2021). Some insights and reflections on global central bank digital currency experiments. *Tsinghua Financial Review, 2021*(03), 16–19 (in Chinese).
Saudi Central Bank & Central Bank of The U.A.E. (2020). Project Aber: Saudi Central Bank and Central Bank of The U.A.E. Joint Digital Currency and Distributed Ledger Project. Saudi Central Bank and Central Bank of The U.A.E.
Wan, J. R., & Wu, Y. (2022). Research on the cross-border payment of central bank digital currency. *New Finance, 2022*(01), 58–64 (in Chinese).
Yang, D. (2020). Develop cross-border payment vigorously to enhance international competitiveness. *Economy, 2020*(06), 112–113 (in Chinese).
Yuan, Y., & Wang, F. Y. (2019). Blockchain theory and method (Vol. 2019, p. 97). Tsinghua University Press.
Zhu, L. X. (2021). Challenges and prospects of blockchain for cross-border payment and clearing industry. *Finance Theory and Teaching, 2021*(06), 4–10+17 (in Chinese).

LLE Based K-Nearest Neighbor Smoothing for scRNA-Seq Data Imputation

Yifan Feng, Yutong Ai, and Hao Jiang

Abstract The single-cell RNA sequencing (scRNA-seq) technique allows single cell level of gene expression measurements, but the scRNA-seq data often contain missing values, with a large proportion caused by technical defects failing to detect gene expressions, which is called dropout event. The dropout issue poses a great challenge for scRNA-seq data analysis. In this chapter, we introduce a method based on KNN-smoothing: LLE-KNN-smoothing to impute the dropout values in scRNA-seq data and show that the LLE-KNN-smoothing greatly improves the recovery of gene expression in cells and shows better performance than state-of-the-art imputation methods on a number of scRNA-seq data sets.

Keywords LLE · scRNA-seq · Dropout issue

1 Introduction

Single-cell RNA sequencing (scRNA-seq) was first reported in Tang et al. (2009), and it is a high-throughput sequencing technology of the transcriptome at the single cell level, reflecting the heterogeneity between cells. The technology plays a significant part in many fields, such as developmental biology, microbiology and so on, and has gained a lot of attention in life science research (Kelsey et al., 2017; Stubbington et al., 2019).

The advent of scRNA-seq technology provides great help for revealing hidden biological functions. However, scRNA-seq data is noisy and incomplete, containing

Y. Feng · Y. Ai · H. Jiang (✉)
School of Mathematics, Renmin University of China, No. 59 Zhong guancun Avenue, Haidian District, Beijing, China
e-mail: jiangh@ruc.edu.cn

Y. Feng
e-mail: fengyifannn@163.com

Y. Ai
e-mail: aiyutong1996@qq.com

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_11

a large number of zero values. The zero values caused by failure of signal detection are called dropouts (Liu & Trapnell, 2016). The dropout event results from a failure of amplification of the original RNA transcript, and the generated noise may disrupt potential biological signals and hinder the downstream analysis. Hence it is a great challenge on how to distinguish the true biological zero and the false zero in scRNA-seq data.

A great number of imputation methods have been proposed to solve the dropout issues arisen in bulk RNA-seq data (Moorthy et al., 2019). For example, Kim et al. proposed a local least squares imputation method called LLsimpute (Kim et al., 2004). This method uses least squares optimization to represent the missing genes as a linear combination of its similar genes. Aittokallio (2010) proposed a method based on fuzzy clustering and gene ontology to estimate the missing values in microarray data. However, these imputation methods may not be directly applicable to scRNA-seq data as scRNA-seq data is much more sparse than bulk RNA-seq data.

In the design of imputation methods for scRNA-seq data, some researchers try to interpret the observed data through probability distribution model. Typical models assume that the scRNA-seq data follow Poisson or negative binomial distribution. The analysis of Ziegenhain (2017) and various studies show that in the absence of real expression differences, the mean variance relationship of genes or proteins closely follow Poisson distribution (Grün et al., 2014). The randomness of single-cell sequencing technology leads to excessive zero values in single-cell data, and many studies include zero inflation to explain excessive zero values in scRNA-seq data (Fan et al., 2016; Parekh et al., 2017; Pierson and Yau, 2015; Risso et al., 2017).

MAGIC (Dijk et al., 2017) is a graph imputation method based on Markov affinity matrix. For a given cell, MAGIC first finds its most similar cell and aggregates the gene expression of these highly similar cells, so as to estimate the gene expression of those with dropout events and other noise sources. However, due to the sparsity of scRNA-seq data, the nearest neighbor in the original data may not represent the most biologically similar cells, which may add new bias to the data and eliminate meaningful biological properties. KNN-smoothing (Wagner et al., 2017) is developed by identifying the k-nearest neighbor of cells with average expression update to perform imputation. DrImpute (Gong et al., 2018) is also a smoothing method, which is designed based on the consistency clustering method of scRNA-seq data (Kiselev et al., 2017). In this method, Spearman and Pearson correlation coefficients are used to calculate the distance matrix between cells, while K-means is used to cluster the distance matrix within the expected cluster number range. These representatives form a class of smoothing based imputation methods.

Model based imputation methods constitute a large proportion of imputation methods for scRNA-seq data. scImpute (Wei et al., 2018) uses a mixed model to distinguish dropout zeros from true zeros. However, scImpute assumes that each gene has a dropout rate, but it has been confirmed that the dropout rate of genes depends on many factors, such as cell type and RNA-seq protocols (Kharchenko et al., 2014), so the selection of dropout rate may need further discussion and research. SAVER (Mo et al., 2017) assumes that the original data follow Poisson distribution and form a prediction model for each gene through the observed gene count (UMI) and then

uses the weighted average of the observed count and the predicted value to restore the true expression of each gene in each cell. netNMF-sc (Elyanow et al., 2020) combines the network regularized nonnegative matrix decomposition with zero inflation processes in transcription count matrix. VIPER (Chen and Zhou, 2018) is based on a nonnegative sparse regression model, which predicts the cells to be imputed by actively selecting a set of sparse local neighborhood cells. In addition, VIPER models dropout probability in the way of specific cell types and specific genes and infers all modeling parameters from the data using an efficient quadratic programming algorithm.

Deep learning based imputation methods in the recent years have gained a lot of attention. AutoImpute (Talwar et al., 2018) is based on deep autoencoder and sparse gene expression matrix. DCA (deep count autoencoder) (Eraslan et al., 2019) is based on the negative binomial noise model, which can minimize the reconstruction error without supervision to learn the distribution parameters of specific genes, which can be applied to data sets of millions of cells. DeepImpute (deep neural network imputation) (Arisdakessian et al., 2018) imputes genes by constructing multiple sub-neural networks. The method uses dropout layers and loss functions to learn distribution in the data and constructs a predictive model, with imputation of missing data alone.

Ensemble methods were proposed mainly for fully integrating the advantages of the available methods. Enimpute (Zhang, 2019) combines the basic results of eight different imputation methods (ALRA, DCA, DrImpute, MAGIC, SAVER, scImpute, scRMD, Seurat) and takes trimmed mean to get the robust results. SHARP (Wan et al., 2020) is an algorithm based on ensemble random projection (RP) that is capable to deal with a scale of 10 million cells.

Among the above methods, the smoothing based method mainly imputes the missing values according to the expression of similar cells, which highly relies on distance measures to define similarity. The model-based method can better distinguish the real zeros from the dropouts, but the results largely depend on the assumptions of the models, which may lack generalization ability. Deep learning has high scalability and can process larger data sets, but at the same time, it requires too much time in training and learning steps, and the memory consumption is larger than other methods. In this chapter, we propose LLE (Locally linear embedding) (Zhou, 2016) based on KNN-smoothing for single cell data imputation. While dealing with real data, the global non-linearity of LLE as well as the property of maintaining the manifold structure can better restore the data. Compared with other methods, we believe that LLE-smoothing achieves better results.

Algorithm 1 K-nearest neighbor smoothing for UMI-filtered scRNA-Seq data

Input:
p, the number of genes.
n, the number of cells.
X, a $p \times n$ matrix.
k, the number of neighbors to use for smoothing.
d, the number of principal components to use for determining neighbors.
Output:
S,a $p \times n$ smoothed matrix.

Input: procedure KNN-SMOOTH(p, n, X, k)
$S = COPY(X)$
$steps = \lceil log_2(k+1)) \rceil$
1: **for** $t = 1$ to $steps$ **do**
2: $M =$ MEDIAN-NORMALIZE(S)
3: $F =$ FREEMAN-TUKEY-TRANSFORM(M)
4: $Y =$ LEADING-PC-SCORES(F, d)
5: $D =$ PAIRWISE-DISTANCE(Y)
6: $A =$ ARGSORT-ROWS(D)
7: $k\_step = \text{MIN}\{2^t - 1, k\}$
8: **for** $j = 1$ to n **do**
9: **for** $i = 1$ to p **do**
10: $S_{ij} = 0$
11: **end for**
12: **end for**
13: **for** $j = 1$ to n **do**
14: **for** $v = 1$ to $k\_step + 1$ **do**
15: $u = A_{jv}$
16: **for** $i = 1$ to p **do**
17: $S_{ij} = S_{ij} + X_{iu}$
18: **end for**
19: **end for**
20: **end for**
21: **end for**
22:
23: **return** S

2 Materials and Methods

2.1 The K-Nearest Neighbor Smoothing Algorithm

The k-nearest neighbor smoothing (KNN-smoothing) algorithm realizes imputation by aggregating information from similar cells based on the k-nearest neighbor (KNN) idea. The algorithm is formalized in Algorithm 1. Here, X_{ij} refers to the expression of i'th gene and j'th cell of X. COPY (X) returns an independent memory copy of X. MEDIAN-NORMALIZE (X) returns a new matrix of the same dimension as X, in which the values in each column have been scaled by a constant so that the column sum equals the median column sum of X. FREEMAN-TUKEY-TRANSFORM (X)

returns a new matrix of the same shape as X, in which all values have been Freeman. Tukey transformed (FTT) Freeman and Tukey (1950) $\left(f\left(x\right)=\sqrt{x}+\sqrt{x+1}\right)$. LEADING-PC-SCORES (X, d) returns the principal component scores of the observations in X (contained in the columns) for the first d principal components. PAIRWISE-DISTANCE (X) computes the pair-wise distance matrix D from X, here D_{ij} is the Euclidean distance between the i'th column and the j'th column of X. For a matrix D with n columns, ARGSORT-ROWS (D) returns a matrix of indices A that sorts D in a row-wise manner, i.e., $D_{jA_{j_1}} \leq D_{jA_{j_2}} \leq ... \leq D_{jA_{j_n}}$ for all j.

2.2 Locally Linear Embedding

The k-nearest neighbor smoothing algorithm highly depends on the distance evaluation and hence, LEADING-PC-SCORES (X, d) is a critical step in the realization of the algorithm. Taking into consideration that PCA is a linear embedding method that may neglect the non-linear intrinsic property of scRNA-seq data, we propose LLE-based method for low-dimensional projection of scRNA-seq data.

LLE is a dimensionality reduction method based on the concept of topological manifold. It assumes that each sample point and its neighbor sample point in high-dimensional space are approximately located on a hyperplane, so the sample point can be reconstructed by a linear combination of its neighbor sample points. Since LLE algorithm only considers the k-nearest neighbor information of each point, which is computationally efficient. Assume $X = (x_1, x_2, ..., x_N) \in R^{D\times N}$, for each data point $x_i \in R^{D\times 1}$, it can be represented by the linear combination of its k nearest neighbor:

$$x_i = \sum_{j=1}^{k} w_{ji} x_{ji} \tag{1}$$

where $w_i \in R_{k\times 1}$, w_{ji} is jth of w_i , x_{ji} is the jth nearest neighbor of x_i. i.e.

$$w_i = \begin{bmatrix} w_{1i} \\ w_{2i} \\ \vdots \\ w_{ki} \end{bmatrix} \quad x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Di} \end{bmatrix} \tag{2}$$

Minimize the following loss function:

$$\arg\min_{w} \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} w_{ji} x_{ji} \right\|^2 \tag{3}$$

Solving the above formula, the weight coefficient can be obtained by

$$w = [w_1, w_2, \ldots, w_N] \tag{4}$$

where $w_i \in R_{k\times N}$ corresponds to N data points, $(i = 1, 2, ...N)$.

After reducing the original data from D dimension to d dimension, $x_i \rightarrow y_i$, the reduced representation can still be expressed as the linear combination of its k-nearest neighbors, and the combination coefficient remains unchanged, so the loss function can be written as:

$$\arg\min_{Y} \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{k} w_{ji} y_{ji} \right\|^2 \tag{5}$$

where Y is the data located in the low dimensional space after dimensional reduction is obtained:

$$Y = [y_1, y_2, \ldots, y_N] \tag{6}$$

We can rewrite the optimization objective as follows

$$\begin{aligned} \Phi(w) &= \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} w_{ji} x_{ji} \right\|^2 \\ &= \sum_{i=1}^{N} \left\| \sum_{j=1}^{k} \left(x_i - x_{ji}\right) w_{ji} \right\|^2 \\ &= \sum_{i=1}^{N} \left\| \left(X_i - N_i\right) w_i \right\|^2 \\ &= \sum_{i=1}^{N} w_i^T \left(X_i - N_i\right)^T \left(X_i - N_i\right) w_i \end{aligned} \tag{7}$$

Regarding S_i as the local covariance matrix, we have

$$S_i = \left(X_i - N_i\right)^T \left(X_i - N_i\right)$$

$$\Phi(w) = \sum_{i=1}^{N} w_i^T S_i w_i \tag{8}$$

We can introduce Lagrange multiplier method

$$L\left(w_i\right) = \sum_{i=1}^{N} w_i^T S_i w_i + \lambda \left(w_i^T 1_k - 1\right) \tag{9}$$

to get the optimal solution by derivation

$$\frac{\partial L\left(w_i\right)}{\partial w_i} = 2S_i w_i + \lambda 1_k = 0 \tag{10}$$

$$w_i = \frac{S_i^{-1} 1_k}{1_k^T S_i^{-1} 1_k} \tag{11}$$

where 1_k is the column vector of all 1 elements of $k \times 1$, the local covariance matrix S_i is a matrix of $k \times k$, and its denominator is actually matrix S_i, namely the sum of all elements of the inverse matrix, and its molecule is the column vector obtained by summing rows with the inverse matrix of S_i.

Finally, the optimization problem for the low dimensional embedding becomes

$$\arg\min_Y \Psi(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{k} w_{ji} y_{ji} \right\|^2 \tag{12}$$

$$s.t. \sum_{i=1}^{N} y_i = 0, \sum_{i=1}^{N} y_i y_i^T = N I_{d \times d} \tag{13}$$

where

$$Y = [y_1, y_2, \ldots, y_N] \in R^{d \times N} \tag{14}$$

Let M denote

$$M = (I - W)^T (I - W) \tag{15}$$

The optimization problem can be rewritten as:

$$\arg\min_Y tr(YMY^T), s.t. YY^T = I \tag{16}$$

It can be seen that Y^T is actually a matrix composed of the eigenvector of M, so we only need to take the eigenvector corresponding to the smallest d non-zero eigenvalues of M.

3 Results

3.1 Availability of Data

The scRNA-seq data sets are available from Gene Expression Omnibus (GEO) database. Here, we use three data sets: Brain (Darmanis et al., 2015), Zeisel and Klein for method evaluation. Zeisel and Klein can be downloaded from GEO database with accession numbers GSE60361 and GSE65525 (Table 1).

Algorithm 2 Locally linear embedding neighbor smoothing for UMI-filtered scRNA-Seq data

Input:
p, the number of genes.
n, the number of cells.
X, a $p \times n$ matrix.
k, the number of neighbors to use for smoothing.
d, the dimensions of manifold learning .
Output:
S,a $p \times n$ smoothed matrix.

Input: procedure LLE-smoothing(p, n, X, k)
$S = COPY(X)$
$steps = \lceil log_2(k+1)) \rceil$
for $t = 1$ to $steps$ **do**
  $M =$ MEDIAN-NORMALIZE(S) // a new $p \times n$ matrix
  $F =$ FREEMAN-TUKEY-TRANSFORM(M) // a new $p \times n$ matrix
  $Y =$ LLE(F, d) // a new $d \times n$ matrix
  $D =$ PAIRWISE-DISTANCE(Y) // a new $n \times n$ matrix
  $A =$ ARGSORT-ROWS(D) // a new $n \times n$ matrix
  $k\_step =$ MIN$\{2^t - 1, k\}$
  for $j = 1$ to n **do**
    for $i = 1$ to p **do**
      $S_{ij} = 0$
    end for
  end for
  for $j = 1$ to n **do**
    for $v = 1$ to $k\_step + 1$ **do**
      $u = A_{jv}$
      for $i = 1$ to p **do**
        $S_{ij} = S_{ij} + X_{iu}$
      end for
    end for
  end for
end for

return S

3.2 Data Processing and Visualization

The input of our method is a count matrix X with rows representing genes and columns representing cells. After logarithmic transformation and FTT transformation according to the process of Algorithm 4, X is mapped to a d-dimensional space by LLE. The Euclidean distance between each sample and its k nearest neighbors is calculated to form the distance matrix $X_{n \times n}$ and then smoothed step by step from 1 to k. We use t-distributed neighborhood embedding to visualize the data.

Table 1 Summary of data sets used for imputation

	Data size	Cell clusters
Klein	24175 * 2716	4
Brain	16384 * 420	8
Zeisel	4412 * 3005	9

3.3 Performance Evaluation

For evaluation, we use SC3 to cluster the imputed data to test the imputation effect. The Adjusted Rand index (ARI) is used to evaluate the clustering accuracy between the original cluster label of the data set and the cluster label of SC3. The results show that compared with other imputation methods, LLE-KNN-smoothing provides the best ARI in all three data sets of the experiment (as is shown in Table 2). For the nearest neighbor of parameter k, we can see that the ARI value of LLE-KNN-smoothing method is relatively high under different data sets, and when the value of parameter k changes, the clustering accuracy of LLE-KNN-smoothing changes slowly and remains stable (Figs. 1, 2 and 3).

We use t-SNE visualization to analyze the advantages and disadvantages of various methods under different data sets. We find LLE-KNN-smoothing is better than other methods (Table 2) and that our method performs better between inter-class and intra-class (Fig. 4 shows the result of data set Brain).

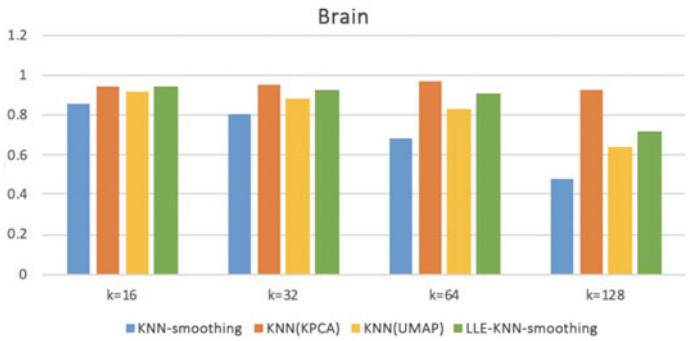

Fig. 1 Different k of the four imputation methods on data set Brain

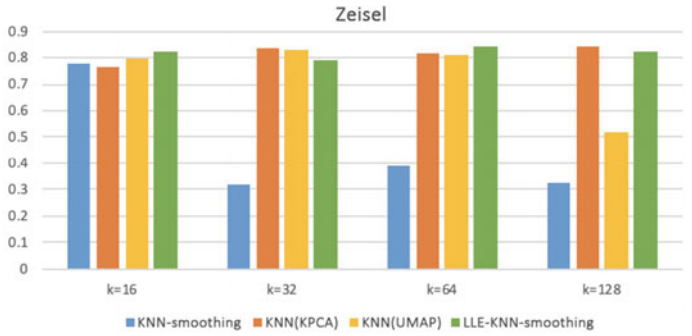

Fig. 2 Different k of the four imputation methods on data set Zeisel

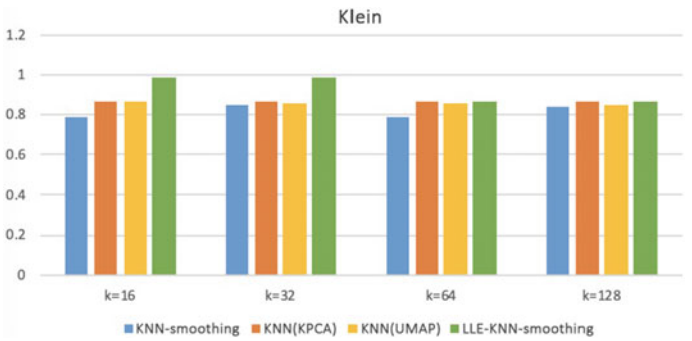

Fig. 3 Different K of the four imputation methods on data set Klein

Table 2 ARI of different imputation methods using SC3 clustering results ($k = 32$)

ARI	Brain	Zeisel	Klein
KNN-smoothing	0.8007	0.3187	0.8508
KNN-KPCA	0.9505	0.8352	0.8649
KNN-UMAP	0.8781	0.8272	0.8594
LLE-KNN-smoothing	0.9425	0.7917	0.9868
MAGIC	0.9239	0.2723	0.3604

4 Conclusions

In this chapter, we have used different data sets to demonstrate that LLE-KNN-smoothing perform better than other methods. In future work, we will continue to study the selection method of parameter k, d, and other manifold learning methods. Other work would be devoted to explore the effect of smoothing for differential expression analysis, gene set enrichment analysis, trajectory inference, etc. We anticipate that LLE-KNN-smoothing algorithm will perform well.

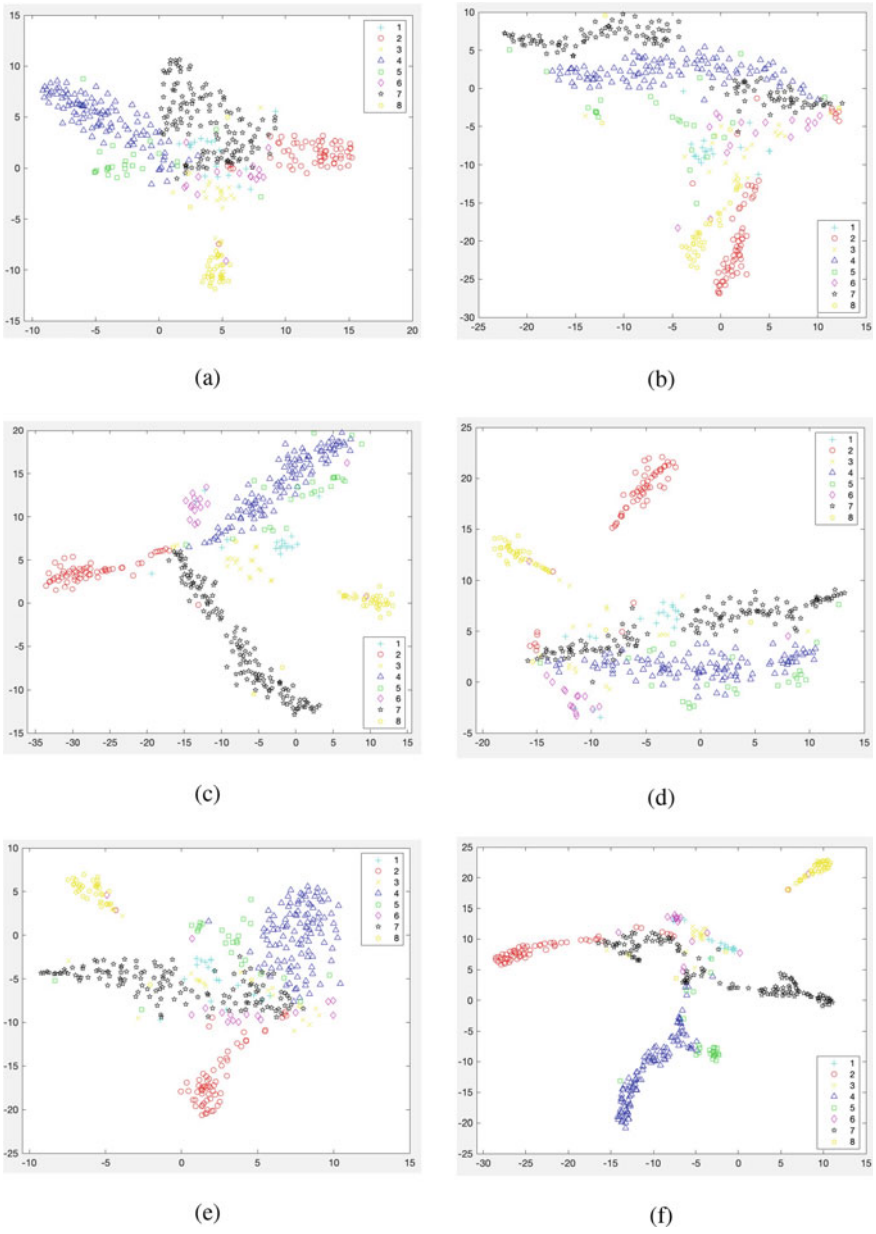

Fig. 4 t-SNE visualization of the reduced dimensions of the five imputation methods on dataset brain. **a** Raw data. **b–f** data after KNN-smoothing, KNN-smoothing (KPCA), KNN-smoothing (UMAP), LLE-KNN-smoothing, MAGIC

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant nos:11901575).

Conflict of interest

The authors declare there is no conflict of interest.

References

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics, 11*(2), 253–264.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., & Garmire, L. X. (2018). Deepimpute: An accurate, fast and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biology*.

Chen, M. J., & Zhou, X. (2018). Viper: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biology*.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences, 112*(23), 7285–7290.

Dijk, D. V., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'Er, D.: Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *BioRxiv*.

Elyanow, R., Dumitrascu, B., Engelhardt, B. E., & Raphael, B. J. (2020). netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Research, 30*(2), gr.251603.119.

Eraslan, G., Simon, L. M., MirCeA, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*.

Fan, J., Salathia, N., Liu, R., Kaeser, G. E, Yung, Y. C., Herman, J. L., Kaper, F., Fan, J. B., Zhang, K., & Chun, J. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square 995 root. *The Annals of Mathematical Statistics*.

Gong, W., Il-Youp, K., Pruthvi, P., Naoko, K. N., & Garry, D. J. (2018). DRIMPUTE: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics, 19*(1), 220.

Grün, D., Kester, L., & Oudenaarden, A. V. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods, 11*(6), 637–40.

Kelsey, G., Stegle, O., & Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science, 358*(6359), 69–75.

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods, 11*(7), 740.

Kim, H., Golub, G. H., & Park, H. (2004). Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics*.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). Sc3: Consensus clustering of single-cell RNA-seq data. *Nature Methods, 15,* 483–486.

Liu, S., & Trapnell, C. (2016). Single-cell transcriptome sequencing: Recent advances and remaining challenges. *F1000 Research, 5*(5), 182.

Mo, H., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., & Zhang, N. R. (2017). Saver: Gene expression recovery for single-cell RNA sequencing. *Nature Methods*.

Moorthy, K., Jaber, A. N., Ismail, M. A., Ernawan, F., & Deris, S. (2019). A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics*.
Parekh, S., Ziegenhain, C., Guillaumet-Adkins, A., Smets, M., & Reinius, B. (2017). Bayesian approach to single-cell differential expression analysis. *Annals of Hematology*.
Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology, 16*(1), 241.
Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., & Vert, J. P. (2017). ZINB-wave: A general and flexible method for signal extraction from single-cell RNA-seq data. *BioRxiv*.
Siddiqui, A. (2009). MRNA-seq whole-transcriptome analysis of a sing cell. *Nature Methods, 6*(5), 377–382.
Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A., Teichmann, S. A. (2019). Single-cell transcriptomics to explore the immune system in health and disease. *Science, 358*(6359), 58–63.
Talwar, D., Mongia, A., Sengupta, D., & Majumdar, A. (2018). Autoimpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific Reports, 8*(1).
Wagner, F., Yan, Y., & Yanai, I. (2017) K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. *BioRxiv*.
Wan, S. B., Kim, J., & Won, K. J. (2016). Sharp: Hyper-fast and accurate processing of single-cell RNA-seq data via ensemble random projection. *Genome Research, 30*(2), gr.254557.119.
Wei, V. L., & Li, J. J. (2018). An accurate and robust imputation method SCIMPUTE for single-cell RNA-seq data. *Nature Communications, 9*(1), 997.
Zhang, X. F. (2019). Enimpute: Imputing dropout events in single-cell RNA-sequencing data via ensemble learning. *Bioinformatics, 35*(22).
Zhou, Z. H. (2016). *Machine learning*. Tsinghua University Press.
Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell, 65*(4), 631-643.e4.

The Application of Time Series Analysis in the Fiscal Budget Variance of China

Guanhua Chen and Xinqi Gong

Abstract During the process of budget planning and execution, irregular behaviors will be reflected in the level of the difference between budgeted and actual figures (named budget variance). Considering that these two processes are both led by Government Of China (hereinafter called GOC), the budget variance is widely used to evaluate the fiscal system. This chapter collects State General Public Budget data from 2000 to 2018 and analyzes their influence on budget variance. Then the forecast for budget variance is completed by modeling the budget execution and budget variance rate separately. The descriptive analysis and AIC (Akaike Information Criterion) contributes to decide the candidate model, the RMSE (Root Mean Square Error) on test data is used to select the final optimal model. The forecast shows that the extent of budget variance will be further controlled in 2011 and 2012, this chapter explains the result with fiscal theories to enhance the credibility of it and thereby provides a couple of policy advice on Chinese budget reform.

Keywords Budget variance · Time series analysis · Fiscal science

1 Introduction

Budgeting is an important administrative process, which reveals both the range and direction of government action, as well as the effectiveness of the monitoring on government activities from National People's Congress (hereinafter called NPC) and private sectors (Chen, 2000). Since the tax sharing reform in 1994, China has

G. Chen · X. Gong (✉)
Institute for Mathematical Sciences, Renmin University of China, No. 59 Zhong guancun Avenue, Haidian District, Beijing, China
e-mail: xinqigong@ruc.edu.cn

G. Chen
e-mail: 13917773472@163.com

X. Gong
Beijing Advanced Innovation Center for Structural Biology, Tsing Hua University, Beijing, China

Z. Zheng (ed.), *Proceedings of the Second International Forum on Financial Mathematics and Financial Technology*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-99-2366-3_12

accomplished extensive and profound reforms in budgeting process. However, there is still some room for improvement, and according to "Decision of the State Council on Deepening the Reform of the Budget Management System" issued in 2014, it is summarized as "the budgeting is not scientific enough, the budget system needs more supervision, the scale of financial carry-over and balance funds is large, and the budget data needs more transparency" etc. Irregular behaviors in budget planning and execution will be reflected in the budget variance. One of the characteristics of a scientific, transparent, and standardized budget system is that the final account income and expenditure are consistent with which planned by the budget.

Since the 21st century, the level of budget variance of the Chinese government has experienced a rise and then a fall. Budget variance generally expanded year by year (Sun & Wu, 2012; Wang, 2009) from 2000 to 2011, peaking at a 15.8% of over-collection and 9% of over-spending. From 2012 to 2018, it has been significantly controlled. The average budget variance in revenue during this period was 6.39%, and that of spending was 4.29%. Figures of United Kingdom and the United States is instructive for further reform: the average public sector recurrent revenue budget deviation in the United Kingdom from 2001 to 2003 was—2.8% (Wang, 2009). In the United States, the average expenditure deviation from the fiscal budget was 2.1% in 2007 (Cui n.d.). China still has a long way to go in terms of reducing budget variance in comparison to countries mentioned above.

The budget is naturally uncertain so a certain degree of variance is permitted. However, large variance can lead to economic and institutional problems. For instance, excessive revenue adds to the burden on tax payers and impacts the market vitality, while excessive under-spending funds leads to inadequate provision of public goods (e.g., infrastructure) by the government, thus affecting social stability and People's livelihood. Institutionally, excessive budget variance implies the weak monitoring of the government. Therefore, it is important for Chinese government to speed up the budget reform and to establish a modern fiscal system to promote the modernization of national governance capacity.

Budget variance contains great research value, so historical data can be analyzed to reveal the reform achievement in China's fiscal system. At the same time, forecasting this indicator can show whether there is still room for improvement and provide a reference for the direction of further reform. The main work of this chapter has three parts. Firstly, we analyze Chinese budget variance, clarifying the linkages between fiscal deviations and the general environment at home and abroad through statistical descriptions. Then the chapter builds a series of time series models and select the best one to predict the variance in the next two years. Finally, the chapter introduces fiscal science theory to explain the prediction and make some reasonable policy advice.

2 A General View on Budget Data

2.1 Introduction to Concept and Data Source

State General Public Budget is one of the four major components of Chinese fiscal system and can be divided into revenue and expenditure. The revenue part mainly includes tax income and non-tax income like confiscation income and transferred funds from the central government to local government. The expenditure part aimed at improving the people's livelihood and maintaining national security, including spending on domestic defense, education, and infrastructure etc. The budgeting process can be divided into two stages, budget planning and budget execution. Former is the annual fiscal revenue and expenditure plan of the state, which is examined and approved by legal procedures, stipulating the sources of national revenue and the purpose of spending, reflecting the scope and direction of government activities. The latter is the annual implementation of budget plan, reflecting the actual economic activities on the state level.

The calculation of budget variance rate is:

$$\text{Budget Variance Rate} = \frac{\text{Budget Execution} - \text{Budget Planning}}{\text{Budget Planning}} \tag{1}$$

The chapter uses the data of State General Public Budget Revenue and Expenditure ranging from 2000 to 2018 (see Table 1), which is obtained from the China Financial Yearbook.

2.2 Descriptive Analysis of Budget Variance

As seen in Fig. 1, the budget deviation is a common phenomenon from 2000 to 2018 on a yearly basis. Most of the situation is over-collection and over-expenditure, under-collection only occurred in 2015 and under-expenditure only occurred in 2014. Over-collection rates peaks in 2007 (16.5%), 2011 (15.8%) and 2010 (12.4%), while over-expenditure rates peaks in 2011 (9%), 2001 (8.9%) and 2007 (7%). The average revenue budget variance rate from 2000 to 2018 is 6.4%, which is higher than that of expenditure which is 4.3%. There is a consistent trend between revenue and expenditure variance rate, and the correlation coefficient of them is 0.77, showing a strong positive correlation.

Trend of budget variance is influenced by domestic fiscal policies as well as the global economic environment. From 1998 to 2004, due to China's two successive active fiscal policies, the revenue budget variance rate stayed around 7%, and finally peaked in 2007. In 2008 and 2009, due to the negative impact of the U.S financial crisis, the budget variance has all fallen back. The new round of fiscal expansion in 2010 and 2011 helped China and the world get out of crisis, however it also led

Table 1 State general public budget execution from 2000 to 2018 (billion yuan; %)

	Revenue			Expenditure		
Year	Budget	Execution	Variance rate (%)	Budget	Execution	Variance rate (%)
2000	12337.77	13395.23	8.6	15136.23	15886.5	4.96
2001	14760.2	16386.04	11.0	17358.3	18902.58	8.9
2002	18014.83	18903.64	4.9	21112.98	22053.15	4.45
2003	20501.32	21715.25	5.9	23699.62	24649.95	4.0
2004	23570.34	26396.47	12.0	26768.64	28486.89	6.4
2005	29255.03	31649.29	8.2	32255.03	33930.28	5.2
2006	35423.38	38760.2	9.4	38373.38	40422.73	5.3
2007	44064.85	51321.78	16.5	46514.85	49781.35	7.0
2008	58486	61330.35	4.9	61386	62592.66	2.0
2009	66230	68518.3	3.5	76235	76299.93	0.1
2010	73930	83101.51	12.4	84530	89874.16	6.3
2011	89720	103874.43	15.8	100220	109247.79	9.0
2012	113600	117253.52	3.2	124300	125952.97	1.3
2013	126630	129209.64	2.0	138246	140212.1	1.4
2014	139530	140370.03	0.6	153037	151785.56	–0.8
2015	154300	152269.23	–1.3	171500	175877.77	2.6
2016	157200	159604.97	1.5	180715	187755.21	3.9
2017	168630	172592.77	2.3	194863	203085.49	4.2
2018	183177	183359.84	0.1	209830	220904.13	5.3
Average	80492.67	83684.87	6.39	90320.05	93563.22	4.29

Over-collection and over-expenditure is common phenomenon from 2000 to 2018. Over-collection rates peaks in 2007(16.5%), 2011 (15.8%) and 2010 (12.4%), while over-expenditure rates peaks in 2011 (9%), 2001 (8.9%) and 2007 (7%)

to a high budget variance (Chen & Lv, 2019). In the following years, the economy of China entered a new normal, the budget deviation began to decline due to the slowdown of GDP growth, under-spending appears in 2014 and under-revenue does so in 2015. The Ministry of Finance has been promoting structural tax cuts and fee reductions since 2014 and appropriately expanding the fiscal deficit. Corresponding to this policy, the spending budget variance rate exceeded the revenue budget variance rate for the first time in 2015, and remained higher than the latter in the next four years, growing slowly continuously.

Budget variance was controlled in general after 2015 mainly because the new budget law came into effect in that year, deepening the fiscal reform. The reduction of the budget variance is the latest achievement in establishing a modern fiscal system in China, indicating that the budgeting process is progressing toward a more scientific direction. The fact that the average expenditure budget variance is smaller than that of revenue reflects that the budget review system is more stringent in expenditure management than revenue. The lack of systematic auditing and monitoring of over-

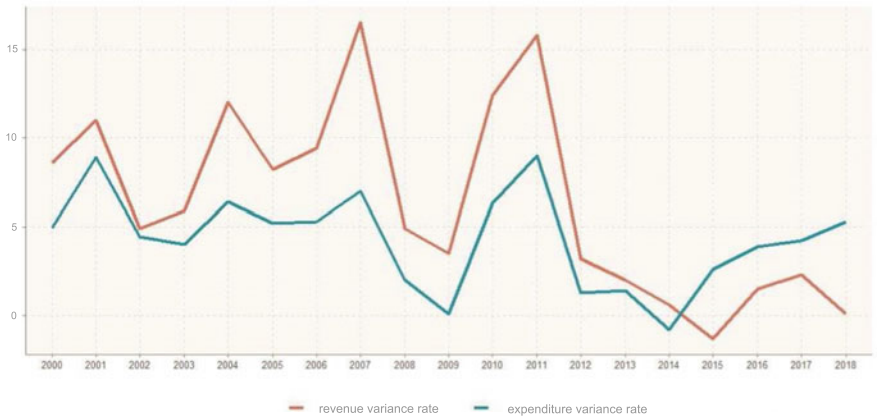

Fig. 1 State General Public Budget variance rate from 2000 to 2018 (%). The revenue and expenditure variance rate peaked in 2007 and 2011 and were significantly controlled after 2014

collection funds leads to the less frugal use of it, therefore over-collection can partially explanation for over-expenditure (Focus on budget deviations, 2008).

2.3 Descriptive Analysis of Budget Execution

As seen in Fig. 2, the budget execution is cyclical on a quarterly basis. In terms of revenue, it peaks in the second quarter and then falls back in the rest of the year while in terms of expenditure the largest percentage of annual spending occurs in the last quarter. The average revenue for the four quarters from 2000 to 2018 is 20789.60, 23961.65, 19442.53, 19476.90 billion yuan, while that of expenditure is 17888.10, 23212.16, 21828.40, 30602.28 billion yuan. The cause of spending surging at the end of the year is the lack of scientific budgeting and monitoring of the over-collected funds. The difference between revenue and expenditure is usually positive in the first part of the year and negative in the second part of the year.

3 Overview of Time Series Analysis Techniques

Time series are defined as ordered random variable sequences. The most common time series are discrete stochastic processes obtained at successively equally spaced time points. Time series describes the intrinsic structure of the series and the target of modeling is to make predictions. In brief, time series forecasting is the art of predicting the future by understanding the past.

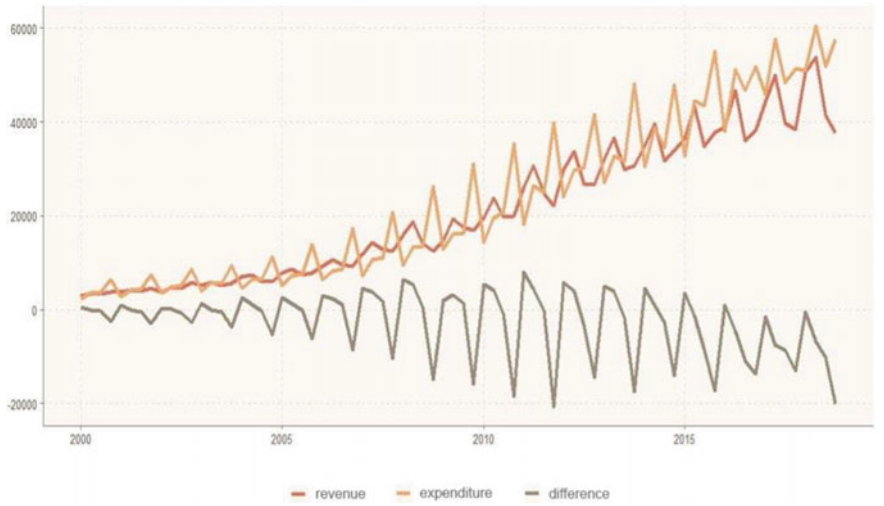

Fig. 2 The State General Public Budget execution from 2000 to 2018 (season; billion yuan). Revenues and expenditures have a tendency to grow over time, while making seasonal changes in cycles of 4. The difference between revenue and expenditure shows that there is often a slight surplus in national income in the first half of the year, which is offset by a sudden increase in expenditures in the last quarter

3.1 Decomposition of Time Series

The time series can be decomposed into long-term trend variation T (a trend in a long period of time), seasonal variation S (regular variation due to seasonal changes), cyclical variation C (longer, more irregular cyclical variation) and irregular variation L (change caused by many contingent factors). The time series Y can be expressed as a function of the above four factors i.e. $Y = F(T, S, C, L)$ like additive model ($Y = T + S + C + L$) and the multiplicative model ($Y = T * S * C * L$). This chapter focuses on modeling T, S, L of the time series.

3.2 ARIMA(p, d, q)

In $AR(p)$ the value at time t is a linear combination of the intercept, past p period observations, and a random error obeying normal distribution.

$$X_t = \psi_0 + \sum_{i}^{p} \psi_i X_{t-i} + \varepsilon_t. \tag{2}$$

Here X_t represents the value at moment t, ε_t represents random error obeying normal distribution at moment t with a variance of σ^2 which is mutually independent from X_t and ψ_i represents the weight of lag order i. Generally a sufficient and necessary condition for $AR(p)$ to be stationary is that all roots of the characteristic equation falling outside the unit circle. Specially, the $AR(1)$ model is a Markov process, and X_t only relates to X_{t-1}. For example, the sufficient condition for $AR(1)$ to be stationary is $|\psi_i| \leq 1$. Its mathematical properties when smooth are shown below, which shows that the auto-correlation of $AR(1)$ model is long-tailed.

$$E(X_t) = \frac{\phi_0}{1 - \phi_1}. \tag{3}$$

$$\mathrm{Var}(X_t) = \frac{\sigma^2}{1 - \phi_1^2}. \tag{4}$$

$$\gamma_k = \begin{cases} \dfrac{\sigma^2}{1 - {\phi_1}^2} & k = 0 \\ \phi_1 \gamma_{k-1} & k > 0 \end{cases}. \tag{5}$$

$$\rho_k = \begin{cases} 1 & k = 0 \\ \phi_1 \rho_{k-1} & k > 0 \end{cases}. \tag{6}$$

In $MA(q)$ model the value at t is a linear combination of the past q period random error, an intercept and the random error obeying normal distribution.

$$X_t = \theta_0 + \sum_{i}^{p} \theta_i \varepsilon_{t-i} + \varepsilon_t. \tag{7}$$

Here θ_i represents the weight of lag order i. The MA process of finite order does not require any precondition to be stationary. The mathematical properties of $MA(1)$ are shown below, from which can be seen that the auto-correlation of the finite order model is truncated-tailed.

$$E(X_t) = \theta_0. \tag{8}$$

$$\mathrm{Var}(X_t) = \sigma^2(1 + {\theta_1}^2). \tag{9}$$

$$\gamma_k = \begin{cases} \sigma^2(1 + {\theta_1}^2) & k = 0 \\ \sigma^2 \theta_1 & k = 1 \\ 0 & k > 1 \end{cases}. \tag{10}$$

$$\rho_k = \begin{cases} 1 & k = 0 \\ \dfrac{\theta_1}{1+{\theta_1}^2} & k = 1 \\ 0 & k > 1 \end{cases}. \tag{11}$$

$ARMA$ model combines $AR(p)$ with $MA(q)$ which can be written as:

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \varepsilon_{t-i}. \tag{12}$$

$ARMA(1, 1)$ also requires $|\phi_1| \leq 1$ to be stationary, and the mathematical properties of the stationary $ARMA(1, 1)$ is shown below. Its auto-correlation coefficient is similar to the $AR(1)$ model, and the partial auto-correlation coefficient is similar to the $MA(1)$ model which are both long-trailed, decaying since two order lag.

$$E(X_t) = \frac{\phi_0}{1-\phi_1}. \tag{13}$$

$$\mathrm{Var}(X_t) = \sigma^2 \frac{1+{\theta_1}^2+2\phi_1\theta_1}{1-{\phi_1}^2}. \tag{14}$$

$$\gamma_k = \begin{cases} \sigma^2 \dfrac{1+{\theta_1}^2+2\phi_1\theta_1}{1-{\phi_1}^2} & k = 0 \\ \sigma^2 \dfrac{(\phi_1+\theta_1)(1+\phi_1\theta_1)}{1-{\phi_1}^2} & k = 1 \\ \phi_1\gamma_{k-1} & k > 1 \end{cases}. \tag{15}$$

$$\rho_k = \begin{cases} 1 & k = 0 \\ \dfrac{(\phi_1+\theta_1)(1+\phi_1\theta_1)}{1+{\theta_1}^2+2\phi_1\theta_1} & k = 1 \\ \phi_1\rho_{k-1} & k > 1 \end{cases}. \tag{16}$$

$ARIMA(p, d, q)$ model introduces the differential operation to $ARMA(p, q)$. In $ARIMA(p, d, q)$, a time series is firstly transformed into a stationary one through d difference and then the $ARMA$ model is established on it. Using B to denote the lagging operator, $ARIMA(p, d, q)$ can be defined as:

$$X_t = (1-B)^d X_t. \tag{17}$$

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \varepsilon_{t-i}. \tag{18}$$

3.3 $SARIMA(p, d, q)(P, D, Q)_s$

SARIMA model introduces seasonal trend to *ARIMA*, using B to denote the lagging operator, B^s to denote s-order lag operator $SARIMA(p, d, q)$ can be defined as:

$$Y_t = (1 - B^s)^D (1 - B)^d X_t. \tag{19}$$

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{P} \phi_i^s Y_{t-si} + \varepsilon_t + \sum_{i=1}^{q} \varepsilon_{t-i} + \sum_{i=1}^{Q} \varepsilon_{t-si}. \tag{20}$$

4 Modeling of Budget Variance

4.1 Prediction of Budget Execution

4.1.1 Exploratory Data Analysis

The fiscal revenue and expenditure was increasing between 2000 and 2018, and the volatility was small. In Figs. 3 and 4 the auto-correlation decreases slowly, so a unit root non-stationary model can be applied.

The unit root is detected by Adf test, after the first-order difference the serial auto-correlation decreases slowly, and the auto-correlation plot shows that there is likely to be a seasonal trend with a period of four for both fiscal revenue and expenditure. Revenue and expenditure items are often similar to those of the same period last year. For example, some expenditure items in education, defense, and public facilities field is fixed every year. Therefore, the seasonal effect is reasonable. Hence further make fourth-order differences on the data.

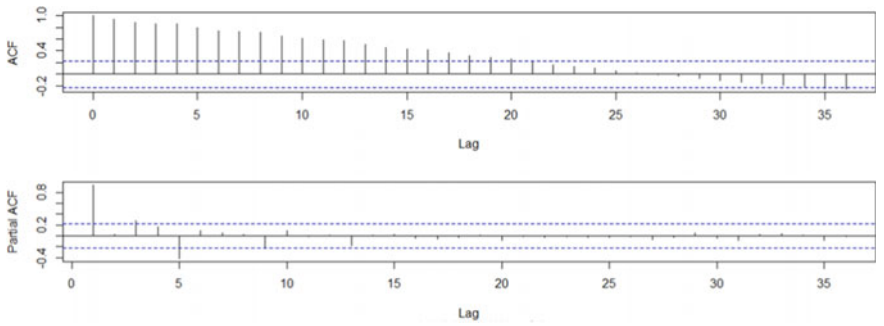

Fig. 3 The auto-correlation plot of revenue budget execution. The auto-correlation trails off, but the partial auto-correlation truncates after the third order

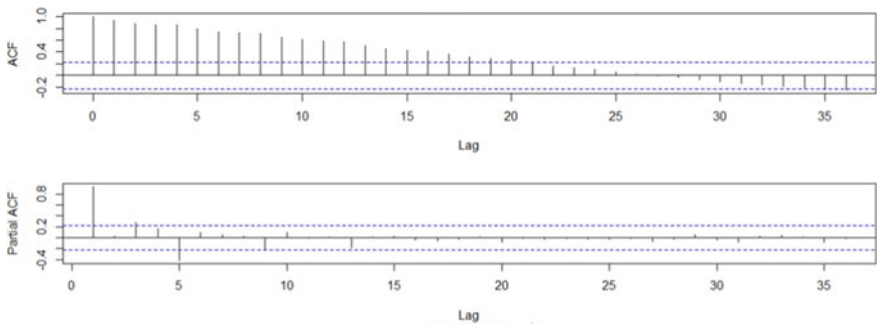

Fig. 4 The auto-correlation plot of expenditure budget execution. The auto-correlation trails off, but the partial auto-correlation truncates after the fifth order

The data after fourth-order difference to extract seasonal effects are smooth non-white noise series, which meet the prerequisites of modeling. Then we take the data from 2000 to 2016 as training set and the data from 2017, 2018 as test set.

4.1.2 Unit Root Non-stationary Model

The seasonal parameter is decided as 4, establishing $AR(3)$ based on the AIC. The $SARIMA(3, 1, 0)(0, 1, 0)_4$ model shows that all the parameters are statistically significant with AIC of 1082.82. Using MA, the auto-correlation plot indicates a lag order of three, so $SARIMA(0, 1, 3)(0, 1, 0)_4$ is established. The model shows that only coefficient ma_1 was not statistically significant and the AIC is 1079.42. Setting ma_1 to 0 and AIC becomes 1080.63, so we choose the first MA model.

Trying $SARIMA(3, 1, 3)(0, 1, 0)_4$, AIC is 1083.23 and coefficients ar_1, ar_2, ar_3, ma_1, ma_3 are not statistically significant. Fix ar_1 to 0, the new model shows an AIC of 1081.28 with ar_2, ma_1, ma_3 being not statistically significant. Then fix ma_1 to 0, the AIC becomes 1079.38, with all remaining coefficients being statistically significant.

Comparing all models based on AIC, it is concluded that $SARIMA(3, 1, 3)(0, 1, 0)_4, ar_1 = 0, ma_1 = 0$ is the optimal unit root non-stationary model for revenue execution.

For expenditure budget execution the seasonal parameter is also decided as 4. Using AR model first, deciding the order as 6 according to AIC, the results of $SARIMA(6, 1, 0)(0, 1, 0)_4$ shows that all parameters are statistically significant and the AIC is 1132.89. Then consider MA model, $SARIMA(0, 1, 5)(0, 1, 0)_4$ model shows that coefficients ma_2, ma_3, ma_4, ma_5 are not statistically significant and the AIC is 1136.75. Fix ma_2, ma_3, ma_4, ma_5 to 0 and it shows that all the remaining parameters are statistically significant and AIC becomes 1131.5.

Considering $ARIMA$, $SARIMA(6, 1, 5)(0, 1, 0)_4$ model shows that the AIC is 1133.54 and coefficients ar_1, ar_3, ar_5, ma_2, and ma_3 are not statistically significant. After fixing ar_5, ma_1, ma_3, ma_4 to zero the AIC becomes 1129.38 and all remaining

coefficients are statistically significant. $SARIMA(6, 1, 5)(0, 1, 0)_4, ar_5 = 0, ma_1 = 0, ma_3 = 0, ma_4 = 0$ is the optimal unit root non-stationary model by comparing AIC of all the models mentioned above.

4.1.3 Fixed Trend Model

Since the data has a increasing trend, fixed trend model is another candidate model. Considering the nonlinear trend, the regression of revenue to time shows that both one-order and two-order term coefficients of time are statistically significant, with an R-squared of 0.95. In Fig. 5, overlaying the estimated trend on the original series, it is found in Fig. 6 that the fitted values are consistent with the actual ones.

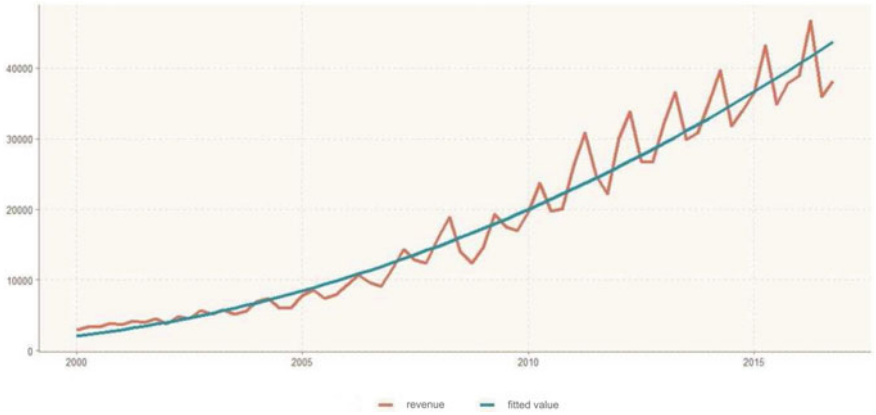

Fig. 5 Fitting effect of revenue budget execution (billion yuan). The fitted curve is basically consistent with data

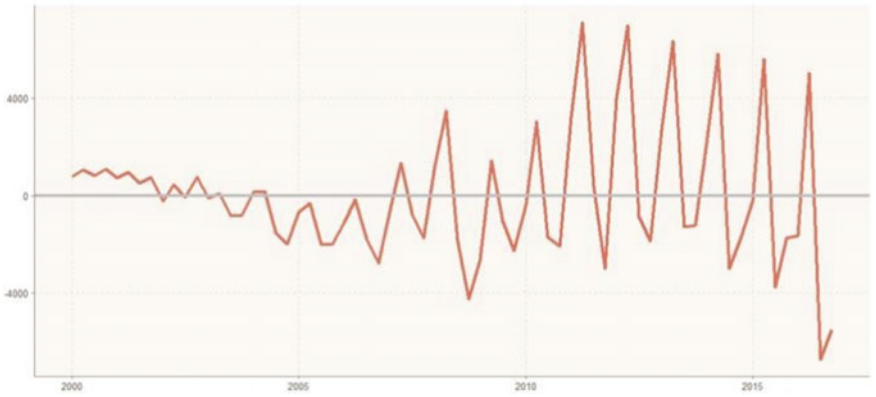

Fig. 6 Fitting residuals of revenue (billion yuan). Obviously it is non-stationary time series

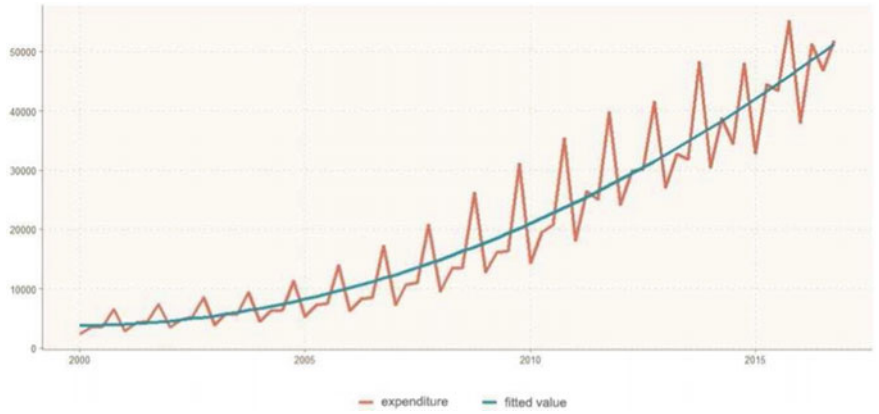

Fig. 7 Fitting effect of expenditure budget execution (billion yuan). The fitted curve is basically consistent with data

After removing the fixed trends, the remaining part are clearly non-stationary series and the auto-correlation plot of it shows that *AR* model can be applied. According to AIC $AR(5)$ is chosen whose AIC is 1173.7. Considering that the data have a strong seasonal correlation so the fourth-order difference is applied and the model $SARIMA(1, 0, 0)(0, 1, 0)_4$ is built according to the auto-correlation plot, whose AIC is 1092.85.

The Regression of expenditure to time shows that the first-order and the second-order coefficients are statistically significant, and the R-squared is about 0.9. Figures 7 and 8 shows that the fitted values are consistent with data, and the residual meets the modeling requirement.

$AR(4)$ model is applied and the AIC is 1223.14. Considering the seasonal trend, fourth-order difference operation is done and the optimal model is obtained which is $SARIMA(0, 0, 0)(0, 1, 0)_4$ whose AIC is 1142.3.

4.1.4 Model Comparison and Testing

The optimal unit root non-stationary model as well as the fixed trend model are selected based on AIC and their performance on test data is compared to make the final choice (see Table 2). It is found that the *SARIMA* model outperforms the *AR*, *MA* model with seasonal parameters on both revenue and expenditure. In terms of revenue budget execution, the fixed trend model outperforms the unit root non-stationary model on test data, although its AIC is slightly higher than it. In terms of expenditure budget execution, the fixed trend model outperforms the unit root non-stationary model on both training and test data. In conclusion, $SARIMA(1, 0, 0)(0, 1, 0)_4$ with a fixed trend is the optimal model for predicting expenditures execution,

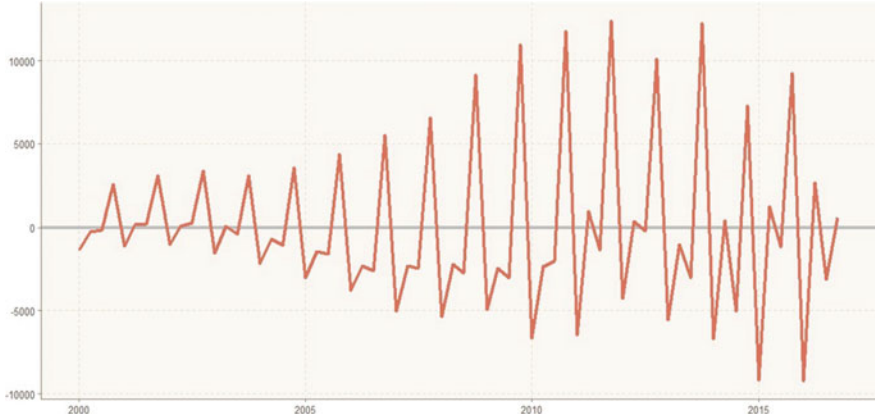

Fig. 8 Fitting residuals of expenditure (billion yuan). Obviously it is non-stationary time series

Table 2 Comparison of revenue and expenditure forecasting models

	Model type	Model	AIC	RMSE
Revenue	Unit root non-stationary model	$SARIMA(3,1,0)(0,1,0)_4$	1082.82	/
		$SARIMA(0,1,3)(0,1,0)_4$	1079.42	/
		$SARIMA(3,1,3)(0,1,0)_4$, $ar_1 = 0, ma_1 = 0$	1079.38	4411.12
	Fixed trend model	$SARIMA(1,0,0)(0,1,0)_4$ with quadratic trend term	1092	4212.779
Expenditure	Unit root non-stationary model	$SARIMA(6,1,0)(0,1,0)_4$	1132.89	/
		$SARIMA(0,1,5)(0,1,0)_4$, $ma_2 = 0$, $ma_3 = 0$, $ma_5 = 0$	1131.5	/
		$SARIMA(6,1,5)(0,1,0)_4$, $ar_5 = 0, ma_1 = 0, ma_3 = 0, ma_4 = 0$	1129.38	3242.246
	Fixed trend model	$SARIMA(0,0,0)(0,1,0)_4$ with quadratic trend term	1142.3	4376.246

and $SARIMA(6,1,5)(0,1,0)_4$, $ar_5 = 0, ma_1 = 0, ma_3 = 0, ma_4 = 0$ is the optimal model for predicting revenue execution. Validating these two models, the residual of their prediction meet the white noise requirement.

4.1.5 Forecast and Policy Advice

The forecast shows that the revenue and expenditure execution will continue growing: the state general public budget revenue will reach 246,765.0 and 267,782.89 billion yuan in 2021 and 2022 while expenditure will reach 258,529.00 and 273,031.33 billion yuan (see Table 3). Figure 9 shows that expenditure will still exceed the

revenue in the next two years, but the difference between revenue and expenditure will be significantly controlled. According to the model, from 2021 to 2022, it will be reduced to –1176.396 billion yuan and then to –5248.44 billion yuan.

The excess of expenditure over revenue is the result of China's proactive fiscal policy in recent years. Premier Li Keqiang said in the 2020 government work report that "the current international situation is more unstable and uncertain, the world economic situation is complex and severe; the domestic economy is not yet a solid foundation for recovery, consumer spending is still constrained, investment growth is not strong enough, small and medium-sized enterprises and individual entrepreneurs have more difficulties, and the pressure on stable employment is greater. In this situation, the government still needs to give a hand", so the policy will continue as a heartening agent for market vitality.

On the other hand, the decrease in the fiscal deficit indicates that China's fiscal sustainability will be improved. Although China's fiscal deficit ratio is always at a low

Table 3 Forecast of the state general public budget revenue and expenditure

Season	Revenue	Expenditure	Difference
2021Q1	59138.40	56870.24	2268.16
2021Q2	68150.72	75785.07	–7634.35
2021Q3	58292.60	62349.92	–4057.32
2021Q4	61183.32	63523.78	–2340.46
2022Q1	64315.78	60857.28	3458.49
2022Q2	73379.50	79660.89	–6281.39
2022Q3	63572.76	65720.32	–2147.56
2022Q4	66514.86	66792.84	–277.98

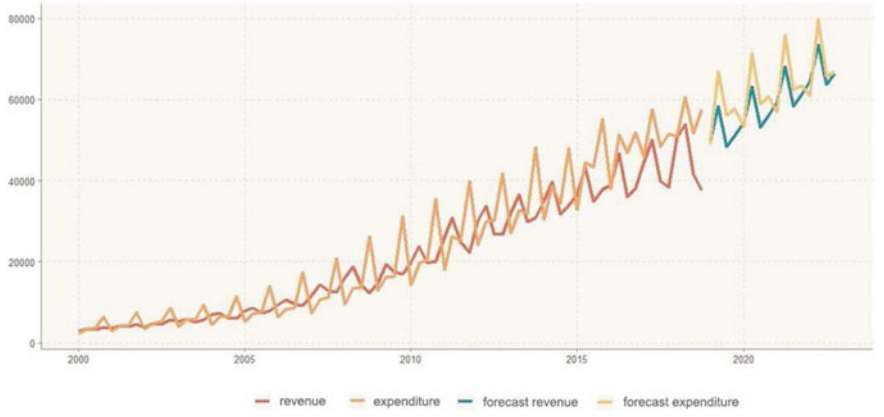

Fig. 9 Forecast of revenue and expenditure budget execution from 2021 to 2022 (season; billion yuan). In the next two years, the expenditure execution will still exceed the revenue, while the difference between income and expenditure will continue to narrow

level compared to Japan, the U.S. and other developed countries, China still needs to be wary of the rapid expansion of government debt, which will rapidly climb once the government becomes debt-dependent, creating a "snowball" effect, i.e., "issuing new debt to pay off old one". In addition, Chinese government also has a hidden debt problem, which is not yet accurately counted, but is generally considered to be of a large scale (Liu & Huang, 2008).

All in all, the future trend of fiscal revenue and expenditure reflects the trade-off between controlling fiscal deficit and restoring the vitality of the market economy, which is a reflection of the "no sharp turn" fiscal policy.

By 2020, China has already implemented a massive tax and fee reduction policy. Finance Minister Liu Kun stated at a meeting of the NPC on March 5, 2021: "During the 13th Five-Year Plan period, tax cut and fee reduction is unprecedented, reaching a total of 7.6 trillion yuan, thus effectively promoting the development of market players and the real economy." However, tax cuts and fee reductions can also make it more difficult to balance fiscal revenue and expenditure. Therefore it is suggested that the finance department should make appropriate adjustments on policy, optimize the tax structure, achieve increases and decreases in different tax sections rather than large and general tax cuts. At the same time, the existing tax policy should be implemented, especially the "precise policy", which means the cuts must be applied to the most difficult areas and enterprises of small and medium-size in the industry. In terms of expenditure, we should keep the strategy of reducing expenditure, especially those for going abroad, vehicle purchase and operation and official reception.

4.2 Prediction of Budget Variance

Based on Eq. 1 the chapter will predict budget variance using the following formula on the grounds that the factors affecting it can be divided into those which affects the level of budget execution and those who affects budget variance rate, separately modeling them can help to better capture the serial correlation in data. This section will complete the modeling of the rate of budget variance.

$$\text{Budget Variance} = \text{Budget Execution} * \left(\frac{\text{Budget Variance Rate}}{\text{Budget Variance Rate} + 1} \right). \quad (21)$$

4.2.1 Exploratory Data Analysis

Over-collecting and over-spending are common between 2000 and 2018, and it can also be seen in Fig. 10 that the budget variance of revenue and expenditure are both non-stationary time series. After first-order difference the data passes the stationary test, thus a unit root non-stationary model can be applied.

4.2.2 Unit Root Non-stationary Model

MA model is applied and the lag order is set as 2 according to the auto-correlation plot. $ARIMA(0, 1, 2)$ shows an AIC of 110.63 and the parameter ma_1 is not statistically significant. After setting it to zero the AIC decreases to 109.6, which is better than the previous model.

In *AR* model, according to the auto-correlation plot, the lag order should be set to 2. $ARIMA(2, 1, 0)$ shows an AIC value of 108.19 and the parameter ar_1 is not statistically significant. The AIC value rises to 108.63 after setting ar_1 to zero, so the previous model is chosen.

$ARIMA(2, 0, 2)$ shows an AIC value of 116.64. Setting the insignificant parameters ma_1 and ma_2 to zero degrades the model to the AR model, so *ARMA* is not included as a candidate model.

MA model is applied for the budget variance rate of expenditure after first-order difference. The lag order should be set as 2 according to the auto-correlation plot. The $ARIMA(0, 1, 2)$ model shows an AIC of 90.89, with parameter ma_1 and ma_2 being statistically significant. These two parameters were retained because the AIC increases after removing them.

According to the auto-correlation plot $ARIMA(2, 1, 0)$ is set with AIC as 92.92 and ar_1 being not statistically significant. After setting it to zero AIC decreases to 92.37.

Considering *ARMA*, $ARIMA(2, 0, 2)$ model shows an AIC of 93.83. Setting ar_1 and ma_2 to zero, AIC decreases to 90.41.

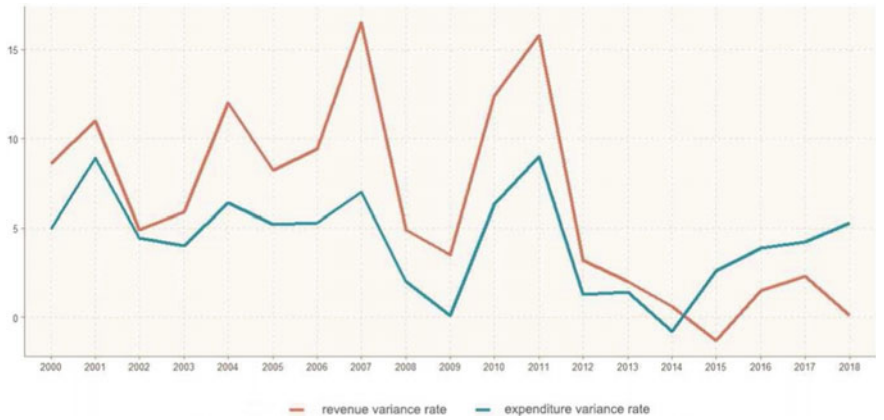

Fig. 10 The state public budget data variance rate (%)

Table 4 Comparison of the model on budget variance rate

	Model	AIC
Revenue	$ARIMA(0, 1, 2), ma_1 = 0$	109.6
	$ARIMA(2, 1, 0)$	108.19
	$ARIMA(0, 1, 2)$	90.89
Expenditure	$ARIMA(2, 1, 0), ar_1 = 0$	92.37
	$ARIMA(2, 0, 2), ar_1, ma_2 = 0$	90.41

4.2.3 Model Comparison and Testing

The optimal model is selected from those mentioned above (see Table 4): $ARIMA(2, 1, 0)$ is the optimal model for revenue budget variance rate, $ARIMA(2, 0, 2), ar_1 = 0, ma_2 = 0$ is the optimal model for expenditure. The residual of both models pass the white noise test.

4.2.4 Forecast and Policy Advice

The revenue budget variance is forecast to be 1.0377 and 0.4684% in 2021 and 2022, and 3.685 and 3.8845% for those of expenditure. Figure 11 tells that the budget variance rate in the next two years will be further controlled compared with the previous years, while the phenomenon of over-collection, over-spending and the general trend that expenditure variance rate exceeding revenue variance rate will remain unchanged. Taking the forecast results into Eq. 21, the revenue variance in 2021 and 2022 will be 2534.38 and 1248.44 billion yuan while expenditure variance will be 9188.20 and 10209.32 billion yuan. Compared with the data of 2017 and 2018, the absolute value of budget variance remains unchanged, which indicates that it is under control, considering the continuous growth of national economy. The forecast results can be explained from four perspectives:

The first factor is economy. Economic factors affect both budget variance rate and the level of budget execution: the faster the economy develops, the faster the budget execution grows, while the uncertainty of budgeting process also increases, leading to an increasing variance. In 2010s, China has experienced a shift from "speed" to "quality" in development priorities and started to consciously control its economic growth rate, while at the same time launching a supply-side reform. These macro factors have made it easier to estimate China's growth prospects and thus make it much less difficult to control budget deviations in the future. From a revenue perspective, it is common practice for the Chinese government to set revenue budgets by adding a few percentage points to the current year's GDP growth rate (Sun & Wu, 2012). Considering that governments at all levels have always been more inclined to leave room while budgeting (Wang, 2009), the over-collection phenomenon will

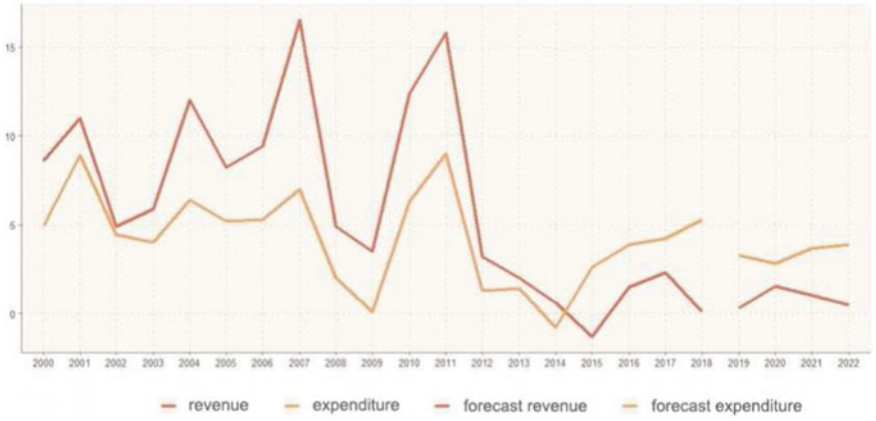

Fig. 11 The forecast budget variance rate (year; %). The revenue and expenditure variance rate will be further controlled in 2021 and 2022, and the rate of expenditure will be higher than that of revenue

remain constant in the long run. From the perspective of expenditures, more fiscal spending items will occur during proactive fiscal policy, leading to more difficulties in controlling variance, which constitutes a partial explanation for the fact that expenditure variance will exceed revenue variance in the future.

The second factor is the soft constraints of budget execution (Chen & Lv, 2019). This factor mainly affects the level of budget variance rate, and the stricter it is, the lower the rate will be. Before 2007, the over-collected funds were neither included in the supervision of the NPC nor excluded from the next year's budget. The inadequate regulation becomes an incentive for over-collection (Ma, 2009). Similarly, the lack of supervision on under-spent funds will also encourage expenditure variance to increase. In order to solve the soft constraint problem, the GOC has made unremitting efforts since 2007: in 2007, the central government established the Central Budget Stabilization and Adjustment Fund (CBSAF) to save and subsidize the over-collected funds to the short-collection year. The Budget Law of the People's Republic of China (2014 Revision) (hereinafter referred to as the new budget law), which was implemented in 2014, has increased the control over budgeting processes by increasing the transparency of the budget, establishing a system to control the inter-year budget and balance the budget across years; In the second revision of the Regulations on the Implementation of the Budget Law in 2018, the reform guideline of improving the budget performance management system and constructing a new pattern of all-round budget performance management was proposed. These moves target strengthening soft constraints, which will help restrain the budget variance in the long run.

The third factor is related to fiscal management system which mainly affects the level of budget variance rate. As local governments rely on the central government's transfer payments, they are motivated to fight for more financial aids than they need, which encourages the phenomenon of "fighting more than spending" and oppor-

tunistic behaviors (Chen & Lv, 2019). Furthermore, local governments tend to raise their spending, especially in the last quarter, to fulfill budget tasks, which leads to irregular use of funds and other risks. These two factors contributed to the variance in fiscal expenditure in the past. However, the Regulations on the Implementation of the Budget Law, which were revised for the second time in 2018, point to improve the transfer payment process, clarify the fiscal relationship between governments by clarifying the types and scope of transfer payments, improving the evaluation of special transfer payments and regulating the process of fund transfer more strictly. These actions have significantly controlled the fiscal expenditure variance, which is reflected in the forecast.

The fourth factor is external supervision, which mainly affects the level of budget variance rate. In China, the NPC is charged with overseeing the budget (Chen & Lv, 2019), while the private sector also plays a supervisory role. Before the revision of the new budget law, the NPC's supervision was not sound enough, lacking professionalism, while it was also difficult for the private sector to form a strong supervision on the budget due to the insufficient information disclosed by the government. In 2014, the New Budget Law made efforts to strengthen the NPC's supervision and audit power over the use of budget funds, and the budget report was also required to be more detailed. In 2018, the Implementation of the Budget Law further requires more transparency, disclosing more data about government debt, agency operating expenses, government procurement and financial earmarked funds. Furthermore, it stipulated explicitly that special transfer payments should be disclosed by region and project, and expenditures by item. These measures strengthen the supervision function of the NPC and society, which is a reflection of the people's ownership and will continuously curb the budget variance rate.

All in all, the decrease of budget variance rate in the next two years is the result of the continuous reform of China's fiscal budget system over years. It is one of the most important achievement of the GOC's modernization of national governance capacity.

Based on the results of modeling and analysis, in order to further control the budget variance, we can focus on the following aspects. The first is to develop more scientific budgeting methods like adopting more mathematical models (e.g., time series techniques and uncertainty theory) and big data techniques (e.g., deep learning) to establish a more predictive budget planning system; The second is to strengthen the management of budget performance evaluation, such as implementing medium-term financial budgeting management and gradually integrating it into the existing budget performance evaluation. The third is to harden the soft constraints on budget planning and execution. For example, by eliminating the misuse of over-collected and under-spent funds through legislation. The fourth is to further improve the budget law, clarify the fiscal relationship between central government and local government, focus on controlling scale of transfer payments, especially special transfer payments, enhance the efficiency of capital flow. Finally, GOC should further implement the supervision over the budget planning and execution to create a more transparent budget system. For this purpose, the NPC's power must be reinforced and more details of fiscal data needs to be published.

5 Conclusion

This chapter completes a descriptive analysis of budget data from 2000 to 2018 in China, pointing out that the level of history budget variance is affected by the global economy and the macroeconomic policies of China. A series of budgetary management actions and fiscal reforms implemented by the Chinese government since 2014 were effective which is reflected in the fact that the budget variance has been well controlled in recent years. The chapter chooses unit root non-stationary model and fixed trend model to model for budget execution and budget variance rate data. The future budget variance in the next two years is calculated from the forecast of the two models. According to the prediction, the budget variance in 2021 and 2022 will be further controlled, and this positive trend is the result of the combination of economic, soft constraints, institutional and regulatory factors.

There is still much room for improvement. Theoretical forecast errors for 2021 and 2022 may exist (Zhao & Wu, 2013) mainly because the impact of the epidemic on China's economy is not taken into account. There is also room for improvement in the combination of fiscal theory and modeling results. Other researchers can make further analysis, like forecasting the level of budget variance of a province or municipality, thus drawing conclusions with local characteristics (Lin & Ma, 2013). China's fiscal data are increasingly abundant, so the research difficulty of this topic will decrease over time. Finally, there are more time series methods that can be used to accomplish the tasks accomplished in this chapter, such as time series multiplicative models (Jiang & Cheng, 2018), more advanced mathematical tools such as uncertainty theory (Liu & Peng, 2005), data mining techniques (Qin, 2018) and Markov Chains (Hou et al., 2010; Li & Yi, 1997) can be used during modeling, which may lead to more accurate conclusions.

References

Chen, G. (2000). *Fiscal science*. Renmin University of China Press.

Chen, Z. G., & Lv, B. Y. (2019). Chinese government budget variance: A typical fiscal phenomenon. *Fiscal Research, 01,* 24–42.

Cui, Z. D. Study on the deviation of China's government budget and accounts. Capital University of Economics and Business.

Focus on budget deviations. (2008). *Foreign-Related Taxation,* (1), 5–6.

Hou T. T., Yang C., Zhan B. H. (2020) Fiscal revenue forecasting in China based on ARIMA and Markov chain model. *Pingdingshan College Journal,35*(02), 6–11+42.

Jiang, Z. D., & Cheng, M. L. (2018). Application of ARIMA multiplicative seasonal model in fiscal revenue forecasting. *Journal of Suzhou University of Science and Technology (Natural Science Edition), 35*(01), 28–32.

Lin, M. H., & Ma, J. (2013). Study on budgetary supervision of local people's congresses in China. *Chinese Social Science, 6,* 78–103.

Liu, B. D., & Peng, J. (2005). *Uncertainty theory tutorials*. Tsinghua University Press.

Liu, S. B., & Huang, W. Q. (2008). A study on the status of invisible debt of local Governments in China. *Fiscal Research, 000*(009), 64–68.
Li, H. X., & Yi, Y. W. (1997). Research and application of financial forecasting models. *Information and Control, 03,* 56–61.
Ma, C. C. (2009). The formation mechanism and management of over-collected funds in China's government budget. *Finance and Trade Economy, 04,* 18–22.
Qin P. (2018). Forecasting and analysis of revenue in Wuhan city based on data mining technology. Huazhong University of Science and Technology.
Sun, Y. D., & Wu, Z. F. (2012). The formation mechanism and governance of over-collection and over-spending in China's budget execution. *Journal of Nanjing Audit University, 9*(004), 1–12.
Wang, X. Z. (2009). An investigation on the budget variance of China. *Exploration of Economic Issues, 000*(009), 164–167.
Zhao, H. L., & Wu, M. M. (2013). Analysis of the accuracy of China's fiscal revenue forecast. *Economic Research Reference, 45,* 41–47.