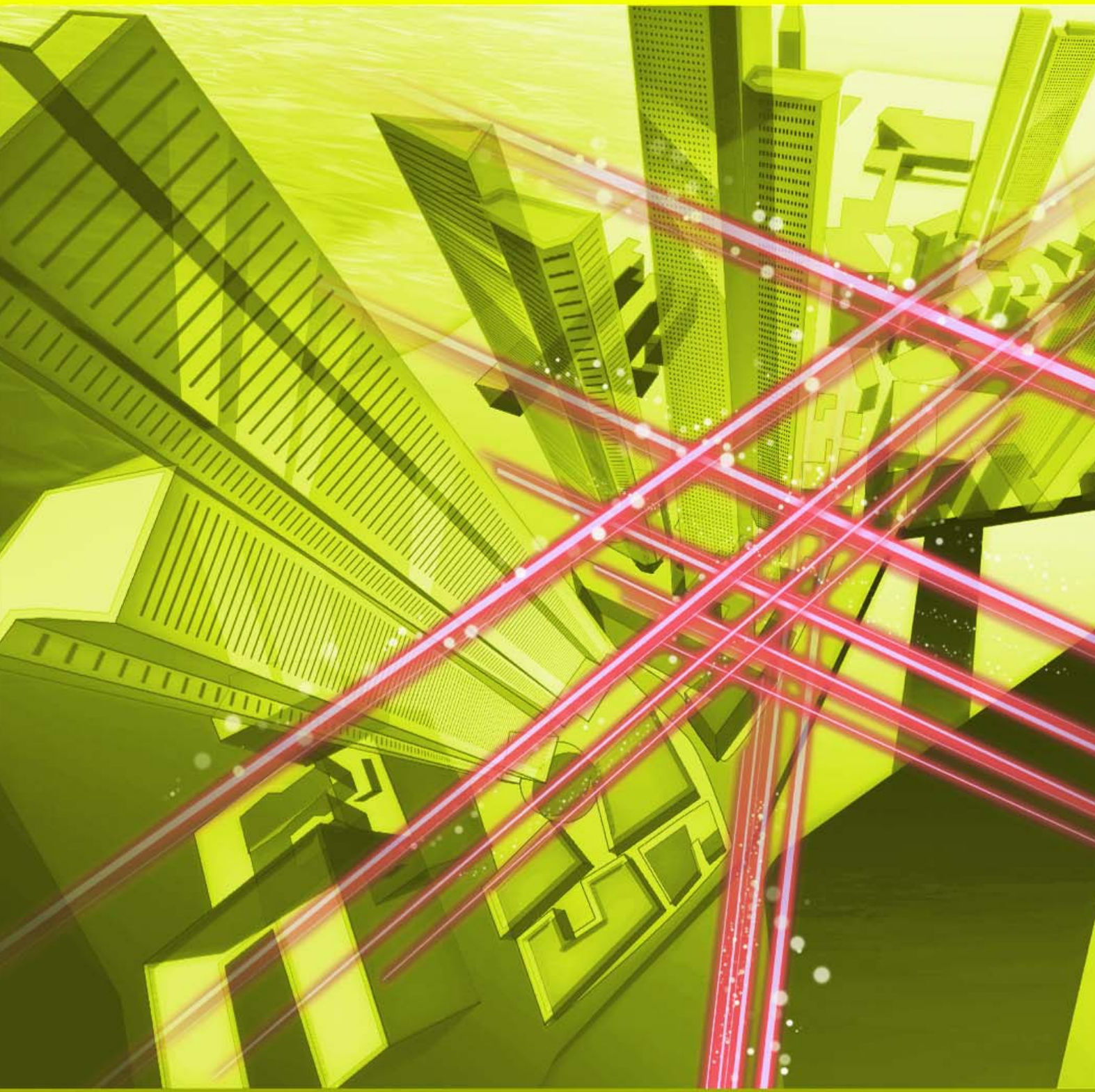
NTT Technical Review

9
2024



September 2024 Vol. 22 No. 9

NTT Technical Review

September 2024 Vol. 22 No. 9

Becoming an Organization of Intuition by Being Exposed to a Shower of Peripheral Information from the Frontlines and Market

Kei Ikeda
Senior Vice President,
Head of Technology Planning,
NTT Corporation

Abstract

The NTT Group is committed to solving social issues as a leading corporate group in the telecommunications business with a public responsibility. Its goal is to create a well-being society by building the world's most-advanced and sustainable social systems and infrastructures. We interviewed Kei Ikeda, NTT senior vice president, head of technology planning, who also serves as the co-chief artificial intelligence officer and chief information officer, about the NTT Group's technology strategies and his beliefs as a top executive.

Keywords: intuition, peripheral information, IOWN, green transformation

Promote implementation of new technologies in-house and confidently propose our accumulated results and expertise

—What is the main role that the Technology Planning Department plays in the NTT Group's technology strategies?

Resolving social issues is the mission of the NTT Group, and the mission of the Technology Planning Department is to formulate technology strategies and incorporate them into the business operations of our operating companies. When talking about NTT technology, some people may think of our research laboratories, but the laboratories and the Research and Development Planning Department that oversees them are engaged in research themes ranging from basic research to practical application of future-oriented technologies. The Technology Planning Department formulates strategies for developing and introducing new technologies in the market as well as our laboratories' technologies for NTT's overall network and information technology systems; implements such technologies at each NTT Group company; and refines and introduces the technologies into society by using the peripheral information possessed by our operating companies' employees. Our frontline employees are exposed to a shower of peripheral information and accumulate experience and expertise, which enables us to hone our technologies to the

point at which we can confidently propose to customers and thereby help solve their problems.

Making IOWN truly marketable

—You mentioned "a shower of peripheral information." What does that phrase mean?

My frequent use of the term "shower of peripheral information" was inspired by the book, "The Structure of Intuition" [1], written by Masakazu Nakayama, a former researcher of the Nippon Telegraph and Telephone Public Corporation. The book explains human's senses and intuition from the viewpoint of cerebral physiology held at the time. To summarize the book's point in simple terms, there are tens of thousands of times more information on the periphery of the information we want to know, and we absorb it without even knowing it. We also have a mechanism that suppresses the information unconsciously absorbed and stored so that it does not overflow uncontrollably, but the information that overrides this mechanism leads to so-called intuition. In other words, when a person is curious and working on a wide range of things, it may not make sense at the time, but if they are consciously thinking about a problem in the first person, those things will combine and be expressed as a meaningful intuition. I believe this is a great strength that humans possess. If the various peripheral information possessed by individual employees in an organization could be bundled together, and if those individuals share the same awareness of a problem and work together, the organization evolves into one that has instinct. The shower of peripheral information is the very source of this intuition. The Technology Planning Department plays a central role in formulating the NTT Group's technology strategies, and we conduct our daily work while aiming to be this ideal organization with intuition.

With this vision of an organization and sense of mission in mind, we are pursuing the widespread use of NTT's Innovative Optical and Wireless Network (IOWN) as well as strengthening of our networks to counter communication failures and natural disasters by using a shower of peripheral information.

—The APN (All-Photonics Network) IOWN1.0 service has been launched. Can you tell us about the importance of a shower of peripheral information in spreading IOWN? Also, how are your efforts to create a circular economy society going?

Simply put, the aim of IOWN is to apply the optical technology that has been cultivated in the world of telecommunications to the world of computers

(signal processing) and bring ultra-low-power-consumption servers to the market. Devices used in current computers, such as central processing units (CPUs), graphics processing units (GPUs), and memory, operate using electrical signals, and these devices communicate with each other using electrical signals. By replacing the electrical signals with optical signals, we aim to significantly reduce power consumption. The key to achieving this is photonics-electronics convergence devices, which requires software called a network operation system (NOS). Against this background, in December 2023, NTT agreed to invest in ACCESS Co., Ltd., which is a parent company of IP Infusion that develops OcNos, the highly evaluated NOS in the global market.

One of the advantages of investing in ACCESS was that we are now able to draw a development roadmap in conjunction with marketers who are active at the forefront of the global market. There is a great benefit of having NTT researchers directly access information on the frontlines of the market, that is, being exposed to a shower of peripheral information. Although we received harsh comments such as "the market doesn't want that kind of functionality," we have been able to form solutions that those global marketers say will sell. We are currently preparing for the first release of the network solutions.

As IOWN advances from the concept stage to implementation, we will strive to make IOWN truly marketable by receiving a shower of peripheral information.

The NTT Group announced its environment and energy vision "NTT Green Innovation toward 2040" in 2021, declaring to achieve carbon neutrality by FY2040 by reducing approximately half of its carbon dioxide (CO_2) emissions with IOWN and reducing the remaining half by increasing the use of renewable energy sources.

In line with this vision, NTT Anode Energy acquired the wind-power generation company Green Power Investment Corporation (GPI) in 2023. GPI has the capacity to generate two-million kW of renewable energy (equivalent to power generated by two nuclear power units), including those under development. With this acquisition, we have almost secured the renewable energy we need within the NTT Group. Going forward, we plan to supply renewable energy to customers outside the NTT Group.

Renewable energy is, however, highly unstable and requires the capability of balancing supply and demand of electricity. In response to this requirement, NTT Anode Energy is working on stabilizing an electricity grid by combining an energy-management system that uses the latest information and communication technology of the NTT Group and storage batteries. I also believe that we can contribute to the local production for local consumption of renewable energy through IOWN initiatives that I mentioned earlier. If technology can be established to link computer devices by using optical technology, for example, computing resources such as CPUs and GPUs in a datacenter in Hokkaido or Kyushu and storage in a datacenter in the Tokyo metropolitan area could be linked and operated as a single large computer. If the Kyushu area is sunny, computing resources in Kyushu can be used; conversely, if the Hokkaido area is sunny, computing resources in Hokkaido can be used. Through such utilization of resources, it becomes possible to match energy demand to fluctuating supply of renewable energy.

To strengthen our efforts in the field of green transformation (GX), we have launched a group-wide brand "NTT G×Inno (pronounced 'geeno')." G×Inno stands for "GX times Innovation" and expresses our desire to create innovation in the GX field and

contribute to achieving Japan's goal of carbon neutrality by 2050. We will first decarbonize the NTT Group and our value chains. We will then propose GX solutions that leverage the expertise and achievements gained from these efforts to corporate customers, local governments, and other parties to contribute to achieving carbon neutrality throughout society. We aim to grow our business in the GX field and achieve over one trillion yen in sales by FY2030.

IOWN and GX will play a critical role in maintaining the irreplaceable global environment. Although we face many challenges ahead, we will continue to promote IOWN and GX initiatives while being motivated by the desire to preserve the global environment.

The source of the NTT Group's strength is the collective power of our employees

—Connectivity is a lifeline of our lives. What kind of measures are you taking against communication failures and natural disasters and to enhance network resiliency?

I had only been in the Technology Planning Department for little more than a week when another company experienced a major communication failure. This was not someone else's problem, and just as we were starting to consider countermeasures within the NTT Group, NTT WEST caused a large-scale communication failure. This incident prompted us to establish the System Failure Recurrence Prevention Committee consisting of chief technology officers, chief digital officers, and other executives from NTT operating companies, and the committee began investigating ways to create a more resilient network that would prevent such incidents from occurring or enable quick recovery if they did occur. However, even after the establishment of the committee, large-scale communication failures continued to occur one after another at our operating companies.

In response to this situation, in addition to implementing our existing measures of conducting a comprehensive group-wide review of the apparent risks, we have embarked on measures to further improve reliability on the premise that unforeseen events are bound to occur no matter how many measures we implement to prevent recurrence.

For example, we assumed extreme, abnormal circumstances that would not normally occur, such as two-thirds of the equipment suddenly going down or three times the amount of traffic flowing in, and we

asked each company to estimate what kind of impact such circumstances would have on their services. Through this activity, I was very encouraged by the fact that the answer to this difficult problem was found within the NTT Group. For example, the initiatives of one company were applied to address certain issues, while those of another company were applied to address other issues; in this way, a robust network was created throughout the NTT Group. Each operating company is exposed to different showers of peripheral information and seeks solutions on the basis of that information. The holding company is responsible for creating an organization that can generate new intuition by combining these intuitions in each company and rolling it out across the NTT Group.

—What is important to you as a top executive? What is your message to researchers, engineers, and customers?

To take advantage of the shower of peripheral information, I practice a bottom-up management style rather than a top-down approach. I believe that the source of the NTT Group's strength lies in how it brings together the information possessed by each individual. The NTT Group can become a stronger organization if we can collect knowledge and share it effectively even under the governance with some centrifugal force.

I joined NTT in 1992 as an engineer with the goal of becoming a specialist. However, I ended up somewhat of a generalist, and I even worked in the Human Resources Department at NTT EAST. Since I had more opportunities to meet with young employees through training work, I began to understand the organization as a whole. I was attracted to work in the field, and after consulting with my boss at the time, I was transferred to the Plant Department of a branch office. At that workplace, I witnessed highly motivated senior colleagues. I was impressed by those who had developed their skills while being exposed to a shower of peripheral information in the field, who worked with pride at being promoted to the Tokyo division, and who were acutely aware that they were the ones supporting the facilities in the field with their extensive expertise.

I realized that the company was supported by people who were willing to work hard, and I was convinced that the company would be stronger if we can mobilize people with this kind of spirit. When I served as the head of the Plant Department at a branch office and branch manager, I visited the frontline at least once a week to receive showers of peripheral information. This experience made me think about how to increase the psychological safety of our employees—without putting up walls—so that they can provide us with peripheral information without hesitation.

Last but not least, to engineers and researchers, let us keep in mind how our frontline employees, customers, and the market view our technologies and achievements, and hone our intuitions by interacting with such people and receiving a shower of peripheral information.

I would also like to ask our customers and business partners to give us your honest opinions about the activities of our researchers and engineers. Receiving a shower of peripheral information from all of you sharpens our intuition and sensibilities. We will do our utmost to contribute to all your businesses through the technology we have refined in this way.

Reference

[1] M. Nakayama, "The Structure of Intuition: What Inspires Ideas," Chuokoron-Shinsha, 1968 (in Japanese).

Interviewee profile

■ Career highlights

Kei Ikeda joined Nippon Telegraph and Telephone Corporation in 1992. In his career at NTT EAST, he served as the director of the Plant Department of Kanagawa Branch in 2012, the director of the Chiba Division in 2017, and the deputy director of the Network Business Headquarters in 2020. He has been in his current position since June 2022.

Creating Bio-digital Twins by Using Crossmodal Representation Learning

Kunio Kashino
NTT Fellow, NTT Communication Science Laboratories

Abstract

Announced in November 2020, NTT's Medical and Health Vision states that the company will strive to put biological simulations using bio-digital twins (BDTs) into practical use to make effective use of medical resources, alleviate physical constraints of medical resources, provide continuous care from prevention to post-treatment, and provide precise care personalized for each individual. A BDT is a digital representation of a living organism, and the key to creating a BDT is how to express (quantify or symbolize) the functions of the living organism—and the physical and chemical mechanisms behind them—as digital information. We interviewed NTT Fellow Kunio Kashino of NTT Communication Science Laboratories, who has taken the challenge of creating a BDT, about the use of artificial intelligence in the biomedical field, stimulation and awareness in collaborative research across different disciplines, and his thoughts on the importance of striking a balance between what should be done and how it should be done.

Keywords: representation learning, crossmodal approach, bio-digital twin

Generating new knowledge through crossmodal approach and applying it to the biomedical field

—Can you tell us about the research you are currently conducting?

I am currently focusing on representation learning and basic research on biomedical informatics. I have been researching sound and image/video recognition and retrieval for a long time. The key point of this research is how to represent media information accurately and efficiently as digital information, which involves representation learning, my first focus. I have recently expanded my research beyond accurate and efficient information representation to include information representation that can uncover hidden structures in data and support the discovery of new knowledge. My second focus, basic research on biomedical informatics, involves the application of representation learning in the medical and health fields.

Generative artificial intelligence (AI) based on large language models has been attracting attention. Generative AI also operates on the basis of the information representation (a digital representation of information) of something and generates output appropriate to a given condition (e.g., a question in sentence form) while referring to the information representation. How to represent information is learned from large amounts of data. The power of large amounts of data is so overwhelming that it is beyond even the imagination of experts, and AI systems are being created that produce useful outputs,

such as text, images, and video, that would have been unthinkable not long ago. However, some aspects of these systems are not fully understood; for example, it is not known why a system behaves in the way that it does or how specific information is represented within a system. The above-mentioned research on representation learning acknowledges this problem, and it can be said that it aims to achieve both the advantages of learning based on large amounts of data and the transparency of information representation. Therefore, if this research is successful, it should provide answers to questions such as how to optimize the size of AI systems and ensure the reliability of their output.

My research topics on representation learning include elucidation, analysis, construction, optimization, and application of information representation. My goal is to establish a method of configuring an optimized information representation with a clear mechanism of action. To achieve this goal, I'm now focusing on extracting and using the relationships between different pieces of information. Let's consider the problem of image recognition as an example. If image data are the only reference, the usual approach is to prepare a large amount of training data consisting of pairs of images labelled with the names of the objects depicted in each image and train the AI with that data. Although this approach is very powerful when it works, it cannot always easily create suitable training data and is not suitable when the way objects are named changes over time. However, when an image and its associated audio information are available, such as online videos, television programs, and everyday scenes, it is unnecessary to manually prepare training data; instead, it is possible to use the relationship between the image and audio as a clue to identify the image and even determine the semantic relationship (closeness of meaning) between the objects shown in the image.

Our experiments have shown that if an AI system is given hundreds of hours of recorded sumo-wrestling broadcasts without any prior knowledge, it can identify frequently occurring (e.g., top 10) winning moves from only image and audio information with a reasonable degree of accuracy. As in this example, it is often the case that something is difficult to understand by referencing just one type of data, but it becomes clear by referencing multiple types of data. The types of information are called "modalities," and I believe that one of the keys to representation learning in the future will be to study these modalities and analyze their relationships with each other. We call this the "crossmodal approach."

It is now possible to collect a large amount of various types of information. In the field of medicine and biology, recent research and technological advances are gradually making it possible to analyze and cross-check large amounts of different types of information, such as genetic information, the cellular basis of behavior, and clinical test results. The increased deployment of the Innovative Optical and Wireless Network (IOWN) will further strengthen these advances. By clarifying the hidden relationships between information collected in this crossmodal manner and reflecting them in inferences and simulations, I believe that—in the not-too-distant future—it will be possible to predict the state of a person's health several years into the future and estimate the efficacy and side effects of drugs and treatment methods to a certain extent. I also believe that through this type of crossmodal research, we can create AI that generates new knowledge that humans have not been aware of. AI may also be able to demonstrate its creativity as a useful partner to humans in scientific and technological research.

—Applying crossmodal approach to biological information leads to bio-digital twins, right?

Yes. Modeling of living organisms has been around for a long time, and you could even say that the history of medicine and biology is the history of modeling. Most such modeling was based on individual experiments and insights of experts. Living organisms are, however, complex subjects, so creating detailed models by hand is naturally limited. To harness the potential power of large volumes of or diverse data, technology that can automatically learn information representations is therefore important.

In November 2020, NTT announced its Medical and Health Vision to contribute to the effective use of medical resources, alleviation of physical constraints of medical resources, provision of continuous care from prevention to post-treatment, and provision of precise care personalized for each individual. The core concept is simulating living organisms using bio-digital twins (BDTs). Our AI telestethoscope introduced in the previous interview (August 2021 issue) is one of the sensing tools used for actualizing this concept. Later, the Biomedical Informatics Research Group was newly established in the Media Information Laboratory of NTT Communication Science Laboratories to support BDTs from the perspective of basic research on informatics, and I was

appointed as its leader. Although the basic activities of this group are concerned with creating new basic technologies for machine learning, signal processing, and pattern processing for a wide variety of information, the group is also actively applying these technologies to biological information and developing new socially useful fields of application. We are collaborating with not only the Bio-Medical Informatics Research Center of NTT Basic Research Laboratories and the Alliance Department of the Research and Development Market Strategy Division at NTT but also universities and hospitals that have unique strengths.

Let me introduce specific studies undertaken by my research team. We are collaborating with Sakakibara Heart Institute on high-precision simulation of the cardiopulmonary function. Heart disease ranks first or second among causes of death in many countries, and early detection and treatment as well as post-treatment rehabilitation (exercise therapy) are known to be particularly effective. Exercise therapy, in particular, can dramatically improve five-year survival rates by having patients exercise at an intensity appropriate to the individual. However, the question is how to set the appropriate intensity of the exercise. Too little exercise will be ineffective; too much exercise could be counterproductive or even dangerous. When exercise is prescribed, a test called cardiopulmonary exercise testing (CPX) is therefore used to set the exercise intensity. However, CPX is not widely used because it requires the subject to exercise close to their limit. Given that situation, using data from Sakakibara Heart Institute, which has conducted the largest number of CPX in Japan, we used machine learning to create a model that can estimate CPX results from physical findings without actually conducting a CPX. We hope that this model will enable more people across the country and around the world to receive appropriate exercise therapy. We are currently preparing to practically apply the model in the near future through various verifications and trials.

We are also working with the Premium Research Institute for Human Metaverse Medicine (PRIMe) at Osaka University on performance measurement and modeling of cardiac muscle cells by using induced pluripotent stem (iPS) cells. As many readers know, iPS cells have the ability to differentiate into cells of any type of tissues and organs by manipulating cells taken from human skin, blood, and other sources in a specific manner and culturing them. Osaka University is using this capability of iPS cells to study disease models using artificial cardiac muscle tissue. In other words, this study aims to explain the mechanism of heart disease and develop treatment methods by culturing cells taken from heart disease patients with genetic factors, growing the cells into cardiac muscle tissue, and measuring their properties as cardiac muscle, such as contractility and diastolic force. This is a physical disease model, but we are working on digitizing the model in collaboration with Osaka University while measuring contraction and expansion forces. Among the many advantages of digital modeling, I believe the most important is the possibility of creating a model of the heart (i.e., an organ) by synthesizing cardiac muscle tissue in digital space. The model connects the numerous pieces of available information about the heart with microscopic biological information about cells and genes (**Fig. 1**). It is currently difficult to physically construct organs with complex structures because the cell mass that can be produced in culture is small compared with an organ, i.e., a few millimeters to a few centimeters in diameter. In digital space, however, it may be possible to synthesize cell masses together and estimate their performance as an organ under certain conditions and assumptions. I believe this will be a significant step forward in improving treatment options.

Regarding research on multi-channel, multi-modal biomeasurement using the AI telestethoscope, which uses AI to analyze data collected from electrocardiogram electrodes, microphones, pressure sensors, acceleration sensors, etc. as well as sounds (such as heart and internal sounds) to infer physical conditions remotely (**Fig. 2**), we are aiming to create new use cases of the AI telestethoscope and improve its accuracy by inputting the measured data into a crossmodal encoder/decoder and displaying explanatory text on the basis of the acquired information representation. In a joint research project with Kitasato University Hospital, we are currently verifying the practicality of a stethoscope-type sensor device that is placed on the chest. The findings that we obtained have recently been published in a medical journal.

I am also pursuing research on the fundamentals of machine learning. During the cell-differentiation process, for example, cells obtain different functions in accordance with the biological tissue, and it is important to analyze, estimate, and model the original state of the cells (before cell-differentiation process) from observations at certain points in time. From a machine-learning perspective, the challenge is how to construct a model with high probability in a situation with many uncertainties. We are developing and verifying such a method.

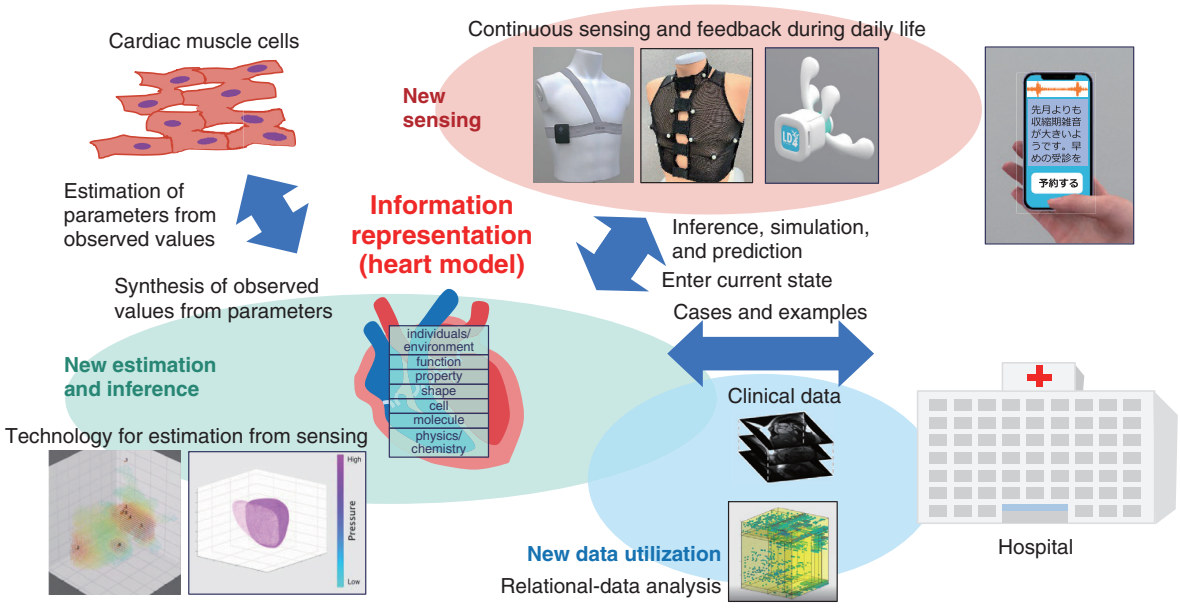


Fig. 1. Overview of a cardiovascular BDT.

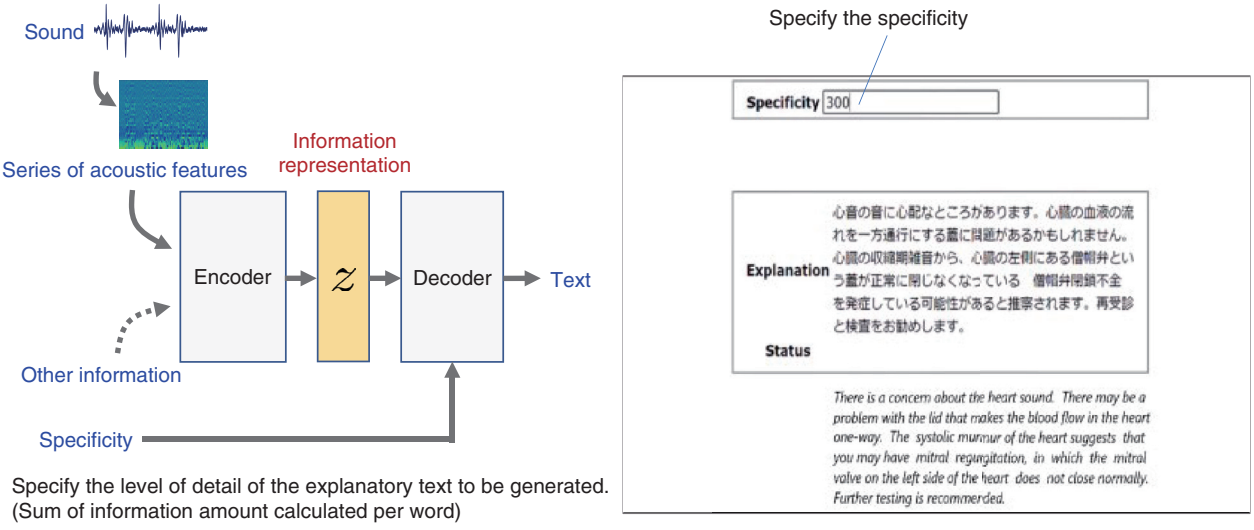


Fig. 2. Use case of the AI telestethoscope (generating explanatory text from heart sounds).

Overcoming barriers to collaborative research across different disciplines

—While the work you have described is basic research on technologies. What are the challenges toward practical use of these technologies?

First, from the perspective of research, I think the fact that many members of my research team—including me—were not experts in medicine or biology was a challenge. Collaboration across disciplines requires a certain amount of basic knowledge of disciplines different from us, and since the language and behavioral styles are often completely different, imagination and a willingness to compromise are required for successful collaboration. Although my

team members still face many challenges, they also enjoy surprises and new discoveries.

From the perspective of practical application, when we are dealing with living body and medical care, we must ensure reliability in accordance with established rules, which is an area outside our expertise. Fortunately, we have the cooperation of many people in this area, and I am grateful that we are gradually making progress.

By understanding each other's differences in common practice and culture, important things that people have in common will become apparent

—What do you keep in mind as a researcher?

It may sound surprising, but I try to be very conscious of social perspectives in my work, which is centered on basic research. NTT Group is becoming an ever more global corporate entity, but I have found that most of my contact with the world has been through participation in international conferences and meetings and other contacts within the research community. Recently, however, as I have come into contact with the field of biological information, I have come to think more about the lives of ordinary people living in various environments around the world.

Another point I try to keep in mind is the balance between what should be done and how it should be done. Which of these two questions is more important depends on the situation, but as far as the creation of new technology is concerned, I believe that both are important. I believe that the dynamism of a situation in which new methodologies increase what can be done and the pursuit of what should be done accelerates the birth of new methodologies is the driving force for change in the world.

To add to my recent impressions, I have had more opportunities to come into contact with medical professionals and been impressed by their sense of mission and sincere attitude toward matters. I feel that by sharing what is important with people from different disciplines and mutually understanding each other's common practice and cultural differences, important things that people have in common will become apparent. This approach is similar to a mechanism in which essential information is brought to light through crossmodal information processing.

—What is your message to future researchers?

I would be happy to share with you the challenges and joys of creating something new. Let's think about and focus on what is important and challenge ourselves without being bound by common practice.

■ **Interviewee profile**

Kunio Kashino received a Ph.D. in electrical engineering from the University of Tokyo in 1995 and joined NTT Basic Research Laboratories the same year. He was a visiting scholar in Cambridge University in 2002. His research interests include audio and crossmodal information processing, media search, and biomedical informatics. He is a fellow of the Institute of Electronics, Information and Communication Engineers (IEICE), and a member of the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery, Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, and the Acoustic Society of Japan. He received IEEE Transactions on Multimedia Paper Award in 2004, Maejima Hisoka Award in 2010, Kenjiro Takayanagi Achievement Award in 2016, IEICE Achievement Award in 2017, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2019.

A New World Emerging from Interactions between Humans and AI: Building Human-AI Cooperative Infrastructure

Makoto Nakatsuji
Distinguished Researcher, NTT Human Informatics Laboratories

Abstract

The artificial intelligence (AI) field offers a variety of services such as chatbots and image generation, including NTT's proprietary large language model "tsuzumi." In the midst of the dizzying pace of change, one of NTT's goals is to create a world in which AI grows autonomously and cooperates with humans. NTT Distinguished Researcher Makoto Nakatsuji believes that building such a world "enables productive endeavors that are more creative," which are very different from conventional AI-based endeavors. In this interview, we asked him about his current project, "building human-AI cooperative infrastructure with generative AI agents."

Keywords: AI, human-AI cooperation, attention model

AI agents that grow together in cooperation with humans enable support for more advanced creative endeavors

—First, please tell us about your ongoing research.

I am currently working on "building human-artificial intelligence (AI) cooperative infrastructure with generative AI agents." Simply put, this research aims to build a human-AI cooperative social infrastructure in which AI cooperates with humans in productive endeavors, just like humans. In the past, AI's main task has been to provide people with an interactive experience and to perform people's tasks on their behalf. In this research, however, AI provides a very different role from conventional AI, in which AI cooperates as a partner with people (doing things by joining forces) (**Fig. 1**). As a concrete example, an alter ego AI agent (Another Me [1]) that thinks like you and interacts with people and the environment is generated in cyberspace to grow and cooperate. This will enable productive endeavors that are more creative, such as the development of new services and research and development by or with AI agents in place of or in conjunction with people.

I believe there are three key technical challenges (**Fig. 2**) that must be addressed in order to achieve such a cooperative relationship. First, Technical Challenge 1 (AI understands the situation it is in now) is to enable AI to understand complex situations in

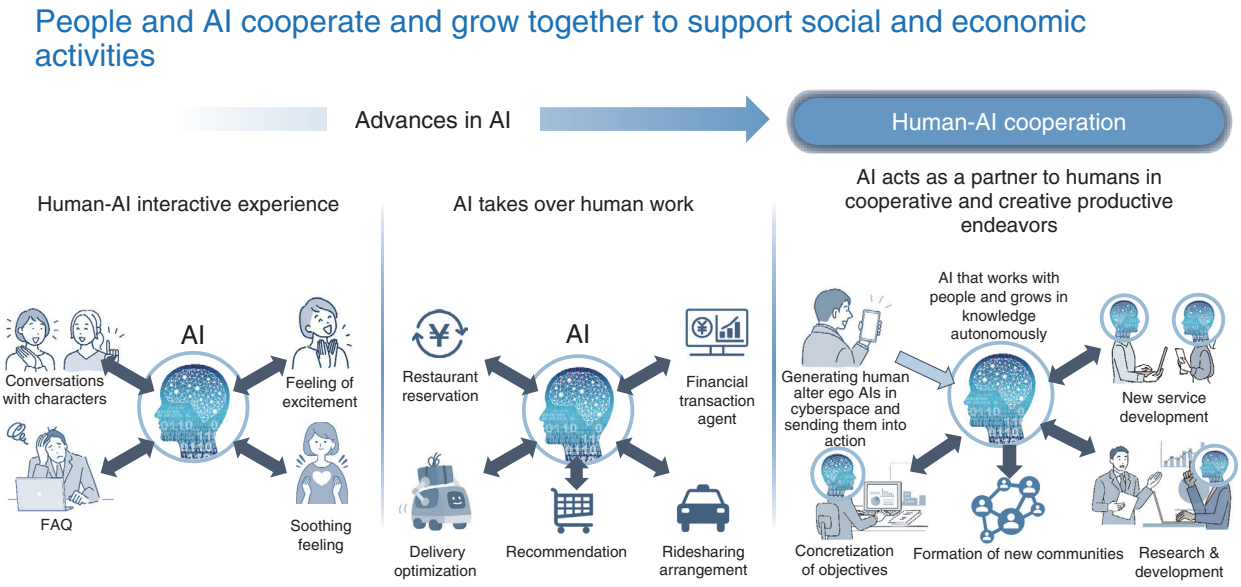


Fig. 1. Transition from conventional AI to human-AI cooperation.

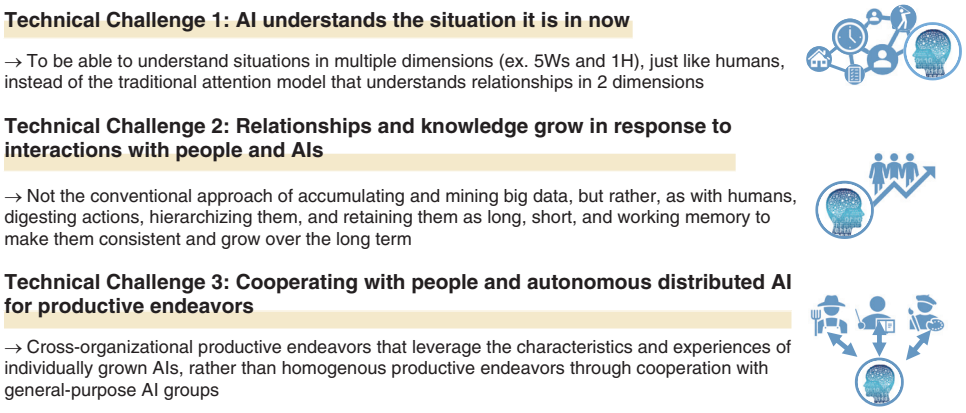


Fig. 2. Three Technical Challenges to achieve human-AI cooperation.

multiple dimensions, e.g., 5Ws and 1H, just like humans. Conventional AI is based on understanding the two-dimensional relationship between "response" and "speech" (two-dimensional attention model [2]), such as learning from a vast history of interactions with humans. In the future, however, I believe that AI can learn the relationship between speech and response in the context of knowledge topics, time, and location, thereby creating a multi-dimensional attention model that better fits human knowledge. In addition, AI's knowledge distribution can be structured and visualized along a contextual axis, making knowledge sharing with people much easier. Specific results to date include improved precision, with response selection precision on multiple data sets 10 to 30% higher than conventional methods.

Next, Technical Challenge 2 (relationships and knowledge grow in response to interactions with people and AIs) aims for AI agents to acquire and grow relationships and knowledge through autonomous communication using memories and knowledge. The key is to digest and hierarchize actions like a human, rather than using the conventional AI approach of accumulating and mining big data. For

example, just like the human brain, memories and knowledge obtained from conversation records, etc., can be organized and managed in order of abstraction. This allows for reusability for future actions, learning and predicting the next action, and dynamically reflecting it in prompts.

Technical Challenge 3 (cooperating with people and autonomous distributed AI for productive endeavors) aims to support productive endeavors that are more creative through the autonomous and distributed growth and cooperation of generative AI agents, which are Another Me or a substitute for humans. As an example of my past efforts, I have developed a group chat model for AI characters and am still working on the same type of research. In this model, the AI character learns the user's habits and preferences, absorbs trends and general knowledge gathered from news and search engines, and attempts to exchange knowledge gathered from other AI characters possessed by other humans. The AI character would then engage with humans and other AI characters to grow autonomously and create cooperative relationships. In other words, the vision of this technology is to leverage the experience and knowledge of individually grown AIs and apply it to current productive endeavors. In doing so, we aim to offer a wide variety of diverse and creative productive endeavors, which are very different from conventional AI-based homogenous productive endeavors.

—What services have you been involved with?

At the time I started this research, I belonged to NTT Resonant. There, we were developing AI technology and providing services at the same time, so I have a lot of experience in creating systems to achieve a business plan. Specific services include "Love consultation AI Oshieru" [3] on "Teach! goo" and TV character AI "AI Nana-chan" [4] and "AI Kahoko" [5] that I helped produce for Nippon Television Network Corporation in one of their projects. These have thankfully been very well received and used by many users, and have been featured in numerous media outlets. In response to this, there is a trend to provide services with interactive AI. The "AI x Design" [6] service was created, and the "AI suite" [7] service was developed from it. This was an attempt to stay ahead of the times and incorporate not only language, but also audio and video, and personalized AI technology to take advantage of business opportunities. About two years after its launch, I was reassigned to NTT laboratories, but AI suite is still provided by NTT DOCOMO.

Thus, during my time at NTT Resonant, I was engaged in research and development while providing a variety of services. In this context, I was trying to devise my own way to explore how to capture the market with AI technology. I was also trying to devise a system that would involve an autonomous cycle of quickly incorporating new technologies into our services and generating revenue from them while conducting further research. Consequently, for about seven years, I have been able to do both the AI business and research and development, and several of my services have been used by a very large number of users. Behind the scenes, the algorithm, which was also accepted by top conferences such as International Semantic Web Conference (ISWC), Association for the Advancement of Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI), was in action, achieving a high level of results in terms of both service and technology. In addition, some of the techniques still have the world's highest response precision for response selection models in evaluations on Chinese data sets.

I am currently developing a new model of human-AI cooperation, called the "miniature garden model" (**Table 1**). This model allows AI agents to deepen their knowledge through team discussions and conversations on behalf of humans, ultimately outputting business proposals and other documents. Each agent that has been generated has individual expertise and propagates knowledge by discussing with each other. In this way, there is a collaborative process that aligns the individual objective of each agent and the overall goal, and ultimately produces output across services.

Leveraging NTT laboratories' extensive research results in the field of AI, where "speed" is important

—What are some of the approaches that you value in your research?

In conducting research in the field of AI, I believe it is very important to keep up with the latest trends and seek out the next trend in research. For example, one of the reasons why the ChatGPT service has spread so much is because they were the first to introduce their service to the market. In addition, we are able to find an answer to the question of whether the direction of the research is really correct by introducing a service to the market. If the service is embraced

Table 1. Characteristics of the miniature garden model.

How to generate prompts for multiple people	Generate multiple people at individual prompts
Open domain	Possible
Strengthening of expertise	Possible. Each agent possesses knowledge individually
Moving the generative agent	Possible
Information transfer between generative agents	Possible. Each agent possesses knowledge individually and propagates knowledge
Existence of roles/objectives of generative agent	Generative agents have individual roles and objectives
Output production	Coordinated role-based work between generative agents, aligning individual objectives and overall objectives, to produce output across services
Refining the generative agent	Refine agent knowledge and service/team knowledge based on collaborative work of agents within and outside the service team

by the world, we can use it as evidence of the value of further research. When I was a member of NTT Resonant, I would immediately launch AI services in line with the trends of the times, receive feedback, and think about the next research topic on a daily basis. In an environment where I could hear directly from the public and customers, I wondered daily what direction I should take my research in. In addition, AI suite was released very smoothly with the help of several supervisors and team members involved. Meanwhile, as a researcher, I was also facing a very difficult time with the emergence of ChatGPT and how to compete with a huge research organization.

This situation has made us think again about the importance of creating a mechanism to systematically bring products to market in a seamless and market-oriented manner, utilizing NTT laboratories' broad range of researchers and diverse research findings. In research where time is limited, new ideas come out one after another, but unless they are built quickly, experimented on, and patented as soon as possible, other competitors will get a head start. Especially in the AI field with its dizzying pace of change, such risks are always at hand, and I believe that there is a common understanding among researchers that they must proceed with their research with speed. In order not to fall behind in these circumstances, I am always mindful of keeping up with trends through papers and discussions with other researchers.

—Finally, do you have a message for researchers, students, and business partners?

When I joined NTT in 2003, research on AI, search systems, and the Semantic Web was gaining momentum, and I felt that this field would be truly transformative. I was especially interested in NTT, which runs its business with research at the core. NTT laboratories have an extensive pool of researchers in machine learning and AI, and in fact, NTT researchers make many presentations every year at top conferences such as AAAI and IJCAI. AI technology is at the core of future business opportunities, and I feel that we have a great advantage in pioneering the AI business in Japan. NTT also has a wide range of operating companies that can create services that make use of this advantage, and above all, NTT's network services and its ability to collaborate with end users are major advantages for the company. I believe it has great potential to be the next game changer, depending on how it is done. In addition, by being able to hear directly from the users of the service, I am able to summarize the results of my research in a paper, which motivates me as a researcher, so I feel that I am

in a very fortunate environment in this respect.

I rejoined NTT laboratories in July 2023 and am currently conducting research through discussions with a diverse group of people. In doing so, I have noticed things from a new perspective and also felt the importance of taking seriously the results of research from the research field's community. I think it applies to people around my age that if we can keep in mind to find the value of what other parties want to do in the community, I believe that the discussion will be productive and produce positive results. I also believe that it is necessary to take a broad perspective and move forward with research in a direction that produces results as a whole. I intend to do so and approach my research with utmost sincerity.

I believe it is important to continue to demonstrate real-world implementation ahead of the times, while also making an impact from the academic side by aligning milestones with top conferences and other events. I also believe that it is important to conduct creative research activities and exciting work that emerge from collaboration among researchers, whether in research or real-world implementation, and to build such relationships. If anyone reading this is interested, I hope we can collaborate at some point on research that will enhance the AI field.

References

[1] A. Ohtsuka, C. Takayama, F. Nihei, R. Ishii, and T. Nishimura, "Technologies for Achieving Another Me," NTT Technical Review, Vol. 20, No. 3, pp. 21–25, Mar. 2022. https://doi.org/10.53829/ntr202203fa3

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," 2017. https://doi.org/10.48550/arXiv.1706.03762

[3] Press release issued by NTT Resonant on Jan. 29, 2019 (in Japanese). https://cdn.kyodonewsprwire.jp/prwfile/release/M101547/201901282607/_prw_PR1fl_QRWJFcAz.pdf

[4] Press release issued by NTT Resonant on Aug. 5, 2019 (in Japanese). https://cdn.kyodonewsprwire.jp/prwfile/release/M101547/201908029369/_prw_PR1fl_71K67PCD.pdf

[5] K. Kawakami, H. Yoshida, M. Nakatsuji, S. Okui, S. Sagara, and Y. Nishimuro, "A Development of the Character Chatbot 'AI Kahoko' After the Lead Character of the TV Drama," The Journal of The Institute of Image Information and Television Engineers, Vol. 73, No. 1, pp. 173–175, 2019. https://www.jstage.jst.go.jp/article/itej/73/1/73_173/_pdf/-char/ja

[6] AI x Design (in Japanese), https://aixdesign.goo.ne.jp/

[7] AI suite (in Japanese), https://aisuite.jp/

■ **Interviewee profile**

Makoto Nakatsuji received his B.S. in systems science from the Graduate School of Informatics, Kyoto University in 2003. In the same year, he joined NTT. In 2010, he completed a doctoral program in social informatics at the Graduate School of Informatics, Kyoto University. He was a visiting scholar at Rensselaer Polytechnic Institute in 2013. He joined NTT Resonant in 2015 and has been with NTT Human Informatics Laboratories since 2023. He received the 2015 Annual Paper Award from the Japanese Society for Artificial Intelligence and the Best Paper Award from the Institute of Electronics, Information and Communication Engineers. After advancing research and development of dialogue systems based on deep learning as well as commercialization projects, he is currently engaged in the research and development of autonomous growth of AI and cooperative endeavors between humans and AI.

A Mathematical World Woven by Number Theory, Algebraic Geometry, and Representation Theory

Kaoru Sano, Hiroyasu Miyazaki, and Masato Wakayama

Abstract

Through basic research in mathematics, the NTT Institute for Fundamental Mathematics aims to enrich the "fountain of knowledge" that nourishes science and technology. In this article, we first provide an overview of the research being carried out at the Institute then introduce the Institute's core research areas: number theory, especially arithmetic dynamics; algebraic and arithmetic geometry; and representation theory and automorphic forms.

Keywords: elliptic curve, Galois theory, zeta function

1. A mathematical world woven by number theory, algebraic geometry, and representation theory

It is quite impressive that research into prime numbers was being carried out in Greece approximately 2500 years ago. Two early important achievements were the proofs of the infinitude of prime numbers (in a constructive way), and the unique factorization of natural numbers into products of prime numbers. The original motivations for this research remain unclear. Until the invention of the RSA public-key cryptosystem by Rivest, Shamir, and Adelman in 1977, there was no expectation for applications of number theory to engineering or society. In fact, it took more than 300 years to establish RSA since the discovery of Fermat's little theorem [1] (proved by Leibniz): "if p is a prime number and a is an integer, then $a^p \equiv a \pmod{p}$ holds," which is one of the key technical results needed for public-key cryptosystems.

Number theory is the branch of mathematics that studies the properties of numbers, especially integers and number systems and structure derived from them [2]. These systems include algebraic number fields such as the rational number field, made up of all fractions with the four basic operations (addition, subtraction, multiplication, and division); finite fields; local fields, such as the real number field; the set of numbers obtained as limits of sequences of rational numbers; and the fields of p-adic numbers. Number theory is said to have originated in the study of Diophantine equations during the Roman Empire. Diophantine equations are defined using polynomials with rational coefficients. Although it was desirable to completely solve them, this is generally difficult. Therefore, the interest was directed toward solutions that are rational numbers. This is probably because irrational numbers were thought to be incomplete numbers, being defined as limits of rational numbers.

Although there is an infinite number of rational numbers, most real numbers are in fact irrational. Lumping them together would be like ignoring the dark matter in the universe. It would be unsatisfactory as science. For example, the equation $x^2 - 2 = 0$ has an irrational solution, $\sqrt{2}$. Irrational numbers, such as $\sqrt{2}$ and $\sqrt[3]{3}$, are called algebraic numbers and distinguished from transcendental numbers such as π, Napier's number e, and $2^{\sqrt{2}}$. For $\sqrt{2}$ and $\sqrt[3]{3}$, it is possible to compute $\sqrt{2} \times \sqrt[3]{3} = \sqrt[6]{72}$ without going back to their definition as limits of rational numbers. In other words, the set of algebraic numbers defines a closed system of numbers within itself. Other types of numbers, such as $\sqrt{-1}$, have also been introduced and has enabled the expansion of the concept of

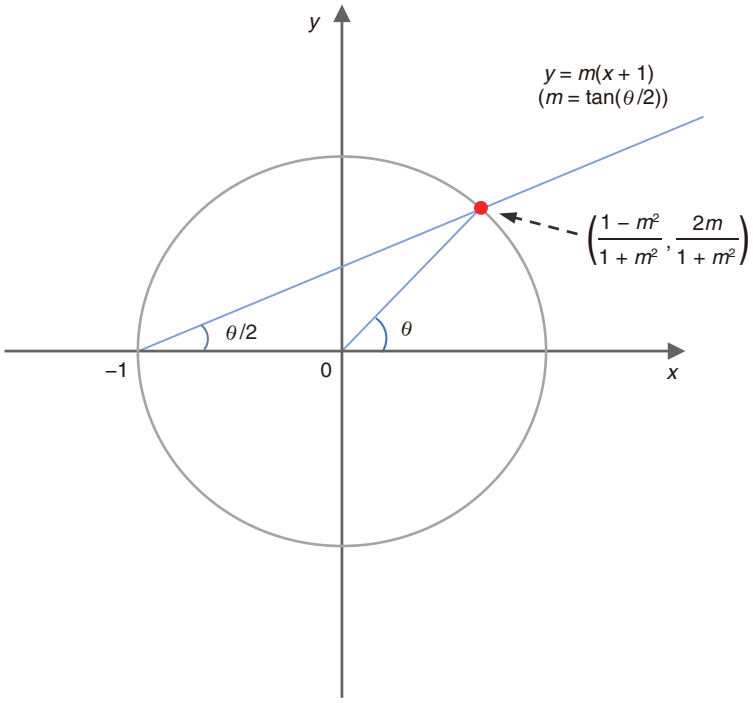


Fig. 1. Unit circle (the double angle formula for trigonometric functions).

numbers. Even today, mathematicians are conscious of the problem of finding new classes of numbers that are broader than algebraic numbers. There is hope that we can find a broad, algebraically controlled class of transcendental numbers that have integral expressions called "periods," such as pi, $\pi = \int_{x^2+y^2<1} dxdy$. This is known as the Kontsevich–Zagier conjecture. Although it seems difficult to solve the conjecture affirmatively, it is a sort of ideal guiding research. These attitudes toward numbers are also closely related philosophically to the three major classic Greek drawing problems of doubling the cube, angle trisection, and squaring a circle, which were proven unsolvable.

To determine whether a given equation has a rational solution is a very delicate issue. For instance, the equation $x^2 + y^2 = 1$ (unit circle) has an infinite number of rational points, that is, with coordinates given by rational numbers. It is easy to confirm this by considering the intersection of a circle with a straight line that passes through the point (–1,0) and has a slope of tan $\theta/2$ on a plane (using the double angle formula for trigonometric functions) (**Fig. 1**). However, there are no rational points on the circle $x^2 + y^2 = 0.999999$, which is a slightly shrunk version of the unit circle. Speaking of subtleties, it took about 350 years to solve Fermat's conjecture (Last Theorem), which states that there are no rational points other than the obvious ones on $x^n + y^n = 1$ (n=3,4,5...). This could not have been achieved without rich theories that make full use of the best of modern mathematics to go beyond a pure algebraic perspective and transcend the circle (which has the structure of an abelian group with addition as a product operation according to the addition theorem) to recognize the agreement between the zeta function determined locally from a geometric structure of an abelian group called an elliptic curve (**Fig. 2**) and a zeta function defined globally and related to representation theory.

The statement of problems in number theory, including the Fermat conjecture, are usually easier to understand than problems in other fields of mathematics. The same can be said for combinatorics and graph theory. Some of the problems in combinatorics and graph theory, such as the construction of Ramanujan graphs, are deeply related to number theory and representation theory, and play a prominent role in applications, including the construction of efficient networks. To solve these problems, a wide variety of methods are used, including using highly advanced modern mathematical tools (rather than just formulas and equations, they use abstract concepts

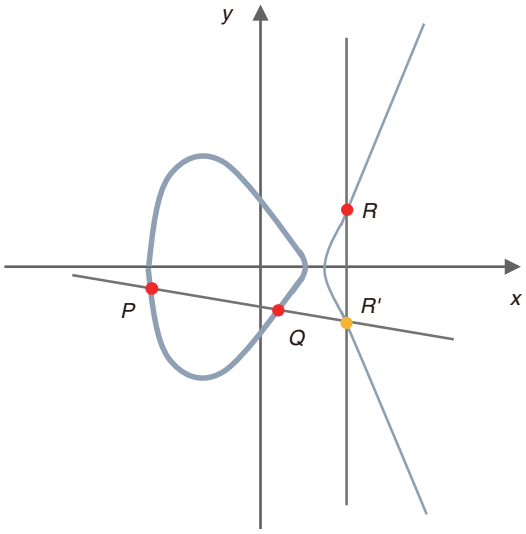


Fig. 2. Elliptic curve: $R = P + Q = -R'$.

and techniques, e.g., considering the relations of objects with "arrows →"). For this reason, many conjectures (statements with supporting evidence but without mathematical proof) that appear sometimes simple remain unsolved. Two of the most well-known and challenging mathematical problems (both in number theory) are the Riemann Hypothesis, which has remained inaccessible for 165 years since its conception, and the Birch and Swinerton-Dyer (BSD) conjecture, which describes the set of rational solutions defining an elliptic curve. The Ramanujan and Weil conjectures, which were solved by Deligne thanks to Grothendieck's innovations in algebraic geometry, led number theory research in the 20th century. The Taniyama–Shimura conjecture, which was the key to the Fermat conjecture, and the Langlands program, which aims to construct a non-commutative class field theory, give us grand dreams for the future. What is particularly noteworthy is that many of these can be expressed as the correspondence of zeta functions and L-functions with different origins, including class field theory.

There is also active research into deriving number-theoretic properties from abstract geometric objects, e.g., investigating rational points and integer points on orbits under dynamical systems determined by repeated composition of polynomials and rational functions and regarding finite-order points on figures with an abelian group structure as periodic points. For example, the far-unexplored Vojta conjecture in Diophantine geometry, which includes the Mordell conjecture (Faltings' theorem), may be understood from this perspective.

It is thus not an exaggeration to say that number theory is voracious. It will use anything, including geometry, analysis, and probability theory, if necessary to solve a problem. It actively draws on geometric inspiration to derive number-theoretic properties from the geometric properties of abstract figures, it makes extensive use of functions that precisely incorporate infinity while retaining invariance, and even incorporates measure theory into its scope of discussion by closely observing distributions and density patterns that are familiar in probability theory. Therefore, it does involve many fields of mathematics to advance its research but contributes to the development of each of them. Bernhard Riemann's research into analytic number theory, which began with the distribution of prime numbers, promoted the development of the theory of complex functions with one variable. It was Carl Friedrich Gauss, the greatest mathematician of the 19th century, who said, "Mathematics is the queen of science, and number theory is the queen of mathematics." This is usually taken to be a succinct expression of the beauty of number theory, but the true meaning may be about the brilliant use of a variety of mathematics to actualize this beauty.

The mathematical research covered in the Feature Articles in this issue [3–9] has as its background theories of arithmetic geometry, which explores problems in number theory using methods from algebraic geometry, arithmetic dynamics, dynamical systems of complex and p-adic fields, and automorphic forms, as well as representation theory. Representation theory is directly or indirectly related to number theory, mathematical physics, combinatorics and graph theory, special function theory, and differential equations and topology. For this reason, the NTT Institute for Fundamental Mathematics, to which the authors belong, is an organization that promotes research that is both cohesive and expansive in the field of mathematics, not only because of its appetite for number theory but also because of the central role of representation theory, a field that deals with symmetry and is at the intersection of algebra, geometry, analysis, and probability theory and has a wide range of applications.

We have put together these articles in the hope that readers will be provided a glimpse into some of the major trends in modern mathematics.

Fig. 3. (From left) Sphere, torus, and double torus.

2. Number theory and arithmetic dynamics

2.1 Fermat's Last Theorem

Pierre de Fermat wrote "Cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet." (I have discovered a truly marvelous proof of this theorem, which however the margin is too small to contain.) in the margin of his copy of Diophantus' "Arithmetica." After the annotated edition of "Arithmetica," in which the problem he noted was published in 1670, many mathematicians challenged it for over 300 years until Andrew Wiles finally proved it in 1995. This theorem laid the foundation for the development of arithmetic geometry.

Fermat's Last Theorem: For any integer n greater than 2, there are no rational pairs (x, y) satisfying the equation (1) $x^n + y^n = 1$ with $x, y \neq 0$.

When $n = 2$, the equation defines a circle with radius 1 centered at the origin. This circle has infinitely many rational points (= points with rational coordinates). All these rational points can be represented as intersections of the circle with lines having rational slopes passing through $(-1, 0)$. This means the rational points on the curve can be parameterized by a single parameter t, the slope of the line. The space of all such slopes is called the projective line. If we extend the coordinates to complex numbers, the figure defined by (1) becomes a real surface. The number of holes in this surface is called the genus. Geometrically, the projective line is a smooth curve of genus 0, while a smooth curve of genus 1 is known as an elliptic curve. Over complex numbers, an elliptic curve looks like a torus (**Fig. 3**).

For $n = 3$ and $n = 4$, the curves defined by (1) are elliptic curves. Fermat proved the theorem for $n = 4$ using a method called infinite descent, which was later extended to the proof of Mordell–Weil theorem. For $n = 3$, the proof was achieved by extending the world of numbers from rational numbers to numbers including the cubic root of unity and using the uniqueness of prime factorization in this larger number system. Today, this involves considering extensions of a number field. Although unique factorization does not hold in general number fields, Kummer's theory of ideal numbers, developed to overcome this, became today's ideal theory. The concept of field extensions laid the foundation for Galois theory and is crucial in modern number theory and arithmetic geometry. For $m \geq 5$, the curve defined by (1) is of genus greater than or equal to 2, and Mordell's conjecture (Faltings' theorem) implies that such a curve has only finitely many rational solutions.

Mordell's conjecture (Faltings' theorem): A smooth curve defined by polynomials with rational coefficients has only finitely many rational points if its genus is 2 or more.

This theorem is a remarkable connection between geometric information (the genus of the curve) and an arithmetic phenomenon (the finiteness of rational points). This result earned Faltings the Fields Medal in 1986. Although there is much more history to be discussed regarding Fermat's Last Theorem, we conclude this section.

2.2 Elliptic curves

We mentioned that smooth curves of genus 0 can be parameterized by rational points, but it is not as simple for genus 1 elliptic curves. However, elliptic curves allow an "addition" where rational points can

Table 1. Dictionary between arithmetic geometry and dynamical systems.

Arithmetic geometry	Dynamical systems
Space	Orbit
Rational/Integral points on a space	Rational/Integral points on an orbit
Torsion points on elliptic curves	(Pre)Periodic points of rational maps
Mazur's theorem	Morton–Silverman's uniform boundedness conjecture

be added together to produce new rational points. This means the set of rational points on an elliptic curve forms a group. This operation enables us to create new rational points from known rational points. One of the key results is the Mordell–Weil theorem.

Mordell–Weil theorem: There exist finitely many rational points P_1, P_2, …, P_r, Q_1, Q_2, …, Q_s on an elliptic curve E such that any rational point P on E can be uniquely expressed as $P = n_1P_1 + n_2P_2 + \cdots + n_rP_r + Q_t$ (n_1, n_2, …, n_r are integers, $1 \leq t \leq s$), where Q_1, Q_2, …, Q_s are points that become O (the identity element, i.e., $P + O = P$ for any rational point P) under multiplication by some positive integer.

The number r is called the rank of the elliptic curve E, and Q_1, Q_2, …, Q_s are called the torsion points of E. While Mazur has completely analyzed the group structure when restricted to torsion points, there remain many unresolved issues concerning the rank. One of the Millennium Prize Problems is the BSD conjecture, which examines the coincidence between the rank and order of the zero of the L-function. The existence or non-existence of elliptic curves with arbitrarily large ranks also remains an open problem. The current world record for the highest known rank, achieved by Elkies, is at least 28. It may be surprising at how small this record is, given the question of whether it is infinite.

2.3 Arithmetic dynamics

Problems in the field of arithmetic dynamics can be traced back to Northcott's theorem in 1950, which states that a morphism defined over a number field on projective space has only finitely many rational periodic points. However, the term "arithmetic dynamics" started being used from 2000. It was clearly recognized as one field of study when the 2010 Mathematics Subject Classification (MSC2010) included 11S82 Non-Archimedean dynamical systems and 37Pxx Arithmetic and non-Archimedean dynamical systems. Broadly speaking, arithmetic dynamics studies the behavior of rational points under the iteration of polynomials or rational maps defined over fields of arithmetic interest (such as p-adic fields or the field of rational numbers). Depending on whether the emphasis is more on number theory or dynamical systems, the nature of the research varies. From a number-theoretic perspective, a large goal might be to create a dictionary of analogies (or generalizations) between results about abelian varieties in number theory and their dynamical system analogs or to obtain new insights into arithmetic geometry through these analogies. Although detailed terminology cannot be explained due to space limitations, the following analogies are being considered (**Table 1**).

Problems regarding arithmetic dynamics over number fields are detailed in the article "Arithmetic Problems in Dynamical Systems" in this issue [3]. When the emphasis is placed on dynamical systems, the field appears somewhat more descriptive. There is an effort to trace similarities between non-Archimedean dynamical systems (such as those on p-adic fields) and complex dynamical systems, with applications to both complex dynamics and arithmetic dynamics. These topics are introduced in the article "How Number Theory Elucidates the Mysteries of Complex Dynamics—Viewed through Non-Archimedean Dynamics" in this issue [4] (**Fig. 4**).

Many books and surveys on arithmetic dynamics have been published, and extensive bibliographies [10] have been compiled. Simply glancing at the titles of the papers listed in these bibliographies reveals the rapid growth of this new field.

3. Algebraic geometry and arithmetic geometry

3.1 A bridge between algebra and geometry

Solving equations is a fundamental but difficult task in mathematics. One of the ultimate goals in the field of algebra is to understand the behaviors of all equations of the form "polynomial(s) = 0," called algebraic equations. Taking their "graphs" is a very important technique when studying algebraic equations.

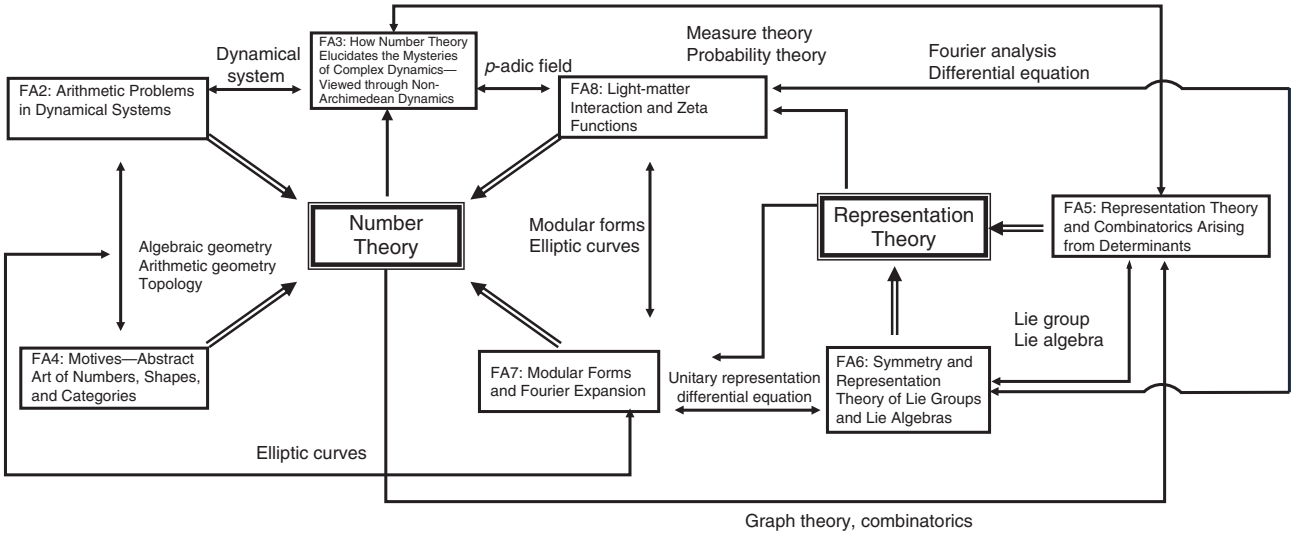


Fig. 4. Overview of research mentioned in this issue.

The graphs are "shapes," for example, the graph of the algebraic equation $x^2 + y^2 = 1$ is nothing but the circle with radius 1 centered at the origin. The graph of $y = m(x + 1)$ is the straight line with slope m passing through the point $(-1, 0)$. The common solutions of the two equations are also expressed as the intersection of the two graphs (Fig. 1). Therefore, graphs transform algebraic problems into geometric ones, enabling us to benefit from the very rich intuition, tools, and ideas from geometry.

The method of graphs was established by René Descartes in his book "Discours de la méthode" published in 1637, and it is now taught in primary education worldwide. However, there are limits to the intuitive method. If we increase the number of variables in the equations, their graphs are (usually) not able to be embedded into the three-dimensional spaces where we live; hence, we cannot "see" the graphs directly. Even if the graphs are (luckily) embedded in the three-dimensional space, their shapes could be too complicated to study just by seeing them with our eyes. After Descartes, many mathematicians made tremendous efforts over the centuries to overcome these difficulties. Finally, we reached the huge theoretical system called "algebraic geometry." The development of the theory of algebraic geometry throughout the 19th to the 20th centuries was so rapid and innovative, and it had many irreversible effects on mathematics afterwards.

3.2 A mathematical paradigm shift—Which came first, functions or spaces?

A breakthrough in mathematics is often accompanied with an important paradigm shift—in the case of algebraic geometry, it came from the relationship between spaces and functions.

In modern mathematics, geometric objects are called "spaces." The graphs of algebraic equations are also spaces. A function is a rule assigning a value to each point on a space. For example, we have a function $f(x) = x + 1$ on the real number line. The values of a function are just numbers, so we can define the addition, subtraction, and multiplication of functions on a space (we cannot define the division of functions in general since the value of a function could be zero, and the division by zero is not defined). An algebraic structure consisting of addition, subtraction, and multiplication is generally called a "ring." In the above discussion, we have seen that the set of functions on a space has a natural structure of a ring.

Let us consider the formula $f(x) = 1/x$, which associates to each x its reciprocal. In fact, this does not define a function on the entire real number line. Indeed, $x = 0$ does not have a reciprocal. However, if we consider the space obtained by removing the origin from the real number line, then $f(x) = 1/x$ defines a function on it. This in turn shows that, if a function $f(x) = 1/x$ lives on a space, the space cannot contain zero.

In this way, a space determines the ring of functions,

and on the contrary, the structure of the ring of functions reveals the property of the space. This phenomenon can be expressed, metaphorically, as follows: if we regard a space as a kind of a nation, then the functions on the space could be thought of as the people living in the nation. If we have a nation, there are people living there and they are interacting with each other via "+, –, ×." Conversely, if we want to know about the nation, it would be very effective to see the people there and how they interact.

On the basis of this observation, Alexander Grothendieck, one of the greatest mathematicians in the history of modern mathematics, boldly claimed that, "Any ring is the ring of the functions on a space." In other words, he claimed that starting from any ring (which could be purely algebraic and could have nothing to do with geometry a priori), we can always find a certain space and regard each member of the ring as a function on the space. In fact, this is a vast generalization of Descartes' idea of "taking graphs." Given an algebraic equation, we can form a ring called the "residue ring" by a purely algebraic procedure, and the space associated with this residue ring recovers the graph of the algebraic equation (more precisely, the space is an algebraic variety, which is a geometric object that has richer information than the classical graph).

Grothendieck established the above philosophy as a rigorous mathematical theory called the "scheme theory," which rewrote the entire framework of classical algebraic geometry. His theory was developed on the basis of many methods and concepts from abstract mathematics developed in the 20th century, including categories and functors.

3.3 Arithmetic geometry

The main purpose of Grothendieck's scheme theory was to apply the method of algebraic geometry to number theory. One of the ultimate goals of number theory is to understand the properties of the ring consisting of all integers. (It is a ring since the set of integers is closed under the addition, subtraction, and multiplication.) Thanks to the scheme theory, we can regard the ring of integers as the ring of functions on a space, hence can translate number-theoretic problems into geometric ones. The research field in which we study number theory using scheme theory is generally called "arithmetic geometry" (for more details about arithmetic geometry, see the article "Motives—Abstract Art of Numbers, Shapes, and Categories" in this issue [5]).

Using the scheme theory, we can construct the theory of geometry on the basis of a system of non-intuitive numbers. For example, we often encounter a system of numbers in which $1 + 1 = 0$ holds. Of course, this property does not hold in the world of real numbers. However, such a system of numbers is inevitable in the study of number theory, and even in many applications in science technologies. The scheme theory states that even in such a "strange" world of numbers, we can naturally consider nice geometry, making it possible to apply the method of algebraic geometry to information theory and cryptography. Algebraic geometry and arithmetic geometry stemmed from purely mathematical motivation and have been developed using many methods in abstract mathematics. However, surprisingly, they became a fountain of concrete applications in society.

Arithmetic geometry developed closely with algebra, geometry, and analysis and became a mainstream of number theory. Interestingly, arithmetic geometry has provided many important concepts that have unexpected applications in different fields of science. For example, the theory of "weights," which was a key to the proof of the Weil conjecture (an analogue of the Riemann hypothesis), became an essential basis of certain fields of theoretical physics, including string theory and mirror symmetry. The theory of elliptic curves, which played a fundamental role in the proof of the Fermat conjecture, has been widely used in constructing post-quantum cryptography. Arithmetic geometry is relatively young in the history of mathematics, and many innovations continue to occur. It will undoubtedly give us unexpected value in and outside mathematics in the next few centuries.

4. Representation theory and automorphic forms

4.1 Group action

When mathematicians hear about representation theory, the first thing they think of is the action of a group on another object. When we think of groups, Évariste Galois, who died in a duel at the age of 20, comes to mind. He greatly simplified and generalized the proof of the Abel–Ruffini theorem, which states that "there is no formula for a general algebraic solution (a solution that can be expressed using the four arithmetic operations and roots) for equations of degree 5 or higher," and used the forerunner of the group concept to characterize when a given equation has an algebraic solution. This theory is known today as Galois theory. Based on Galois theory is the monumental achievement in class field theory of Teiji

Takagi, who studied under Frobenius in Berlin then under Hilbert and Klein in Göttingen at the turn of the 19th century, for actualizing "Kronecker's dream of youth (Kronecker's Jugendtraum)" based on Gauss's law of quadratic reciprocity. It can also be said that Galois theory is the pillar of the magnificent theoretical system now known as the Langlands program (conjecture/philosophy), which is aimed at developing a non-commutative class field theory. However, the definition of a group is quite simple, it is a set such that 1) there is a binary operation called "multiplication" satisfying the associative law, 2) an identity element exists, and 3) each element has an inverse.

Examples of finite groups are the familiar groups of symmetries of regular polyhedrons, the crystallographic point groups, symmetric groups that appear in the definitions of determinants, as well as general linear group GL(V), which is formed by regular (i.e., having the inverse) linear transformations on a finite-dimensional vector space V. The series of operations required to align the puzzle pieces on a Rubik's Cube can also be thought of as physical group of operations performed by the "hand."

4.2 Representation theory of groups

When we speak of representation theory, we are sometimes asked, "Do you mean literature or art? Or is there something like that in the field of mathematics?" It is influenced by expressionism, the art movement that originated in Germany in the early 20th century. It generally refers to the tendency to express emotions by reflecting them in works, as opposed to classical forms of representation*. Representation theory [11], in its most basic form, is the branch of mathematics that studies Vs on which elements of a group act as linear transformations of V (i.e., matrices, once a basis is fixed). Historically, the opportunity for representation theory to become an independent research subject was the letter from Dedekind to Frobenius in 1886 regarding the problem of factorization of the group determinants. This is the beginning of the character theory of finite groups. A character is the trace of a representation (of a matrix-valued function). In fact, a representation is essentially determined by its character. Sophus Lie also conducted research aimed at developing a Galois theory for differential equations and founded the current notion of Lie algebras and Lie groups. However, when it comes to these representations, it is indispensable to consider infinite dimensional Vs, in which case we need to consider the topology of the V. However, what decisively advanced the development of modern representation theory are the revolutionary theories in physics known as "relativity" and "quantum mechanics," as well as the dramatic progress made in number theory on the road to the Langlands program.

From a technical standpoint only, and although they vary somewhat depending on the algebraic system, representation theory can be summarized and broadly divided into the following three goals:

- Construction and classification of irreducible representations (creating a complete list with no omissions or duplicates). Irreducible representations play a fundamental role, analog to the prime numbers in number theory or elementary particles in particle physics.
- Decomposing a given representation into a "sum" of irreducible representations (division or reduction of complexity).
- Study of various characteristics and/or geometric realizations of equivalence classes of irreducible representations. The various elements of the equivalence classes can be constructed, for example, using interesting geometric objects that are useful in applications.

Some people may wonder why groups such as matrix groups, which already seem simple, must be expressed as difficult linear transformations on infinite dimensional spaces. However, the opposite is true. Even if something seems very complicated, if one unravels it correctly (decomposition), one will find that each part is simple (the action of an "easy" or "simple" group), thus one will be able to reach a true understanding of the object. These studies make full use of differential equations, functional analysis, the theory of special functions, combinatorics [12], and category theory, which has been well suited since Galois.

4.3 Automorphic forms

The Langlands program is often discussed purely in algebraic terms today, but the idea originated in Selberg's theory of analytic continuation of non-holomorphic Eisenstein series and Harish-Chandra's research on the representation theory of reductive Lie groups. In fact, representation theory is a strong bridge for solving questions in number theory that are formulated purely algebraically by replacing them with analytical notions such as Fourier transforms

* In Japanese, the word for both representation and expression is "hyogen (表現)," which is where the confusion comes from.

and *q*-series (via automorphic forms [13]). The natural actions of continuous and discrete groups are behind it, and the description of invariance with respect to these actions often clarifies the problems. For example, after Fermat's Last Theorem, the Sato–Tate conjecture, which was considered to be many times more difficult, was (partially, i.e., a certain important class) solved by Richard Taylor and others in 2011. The Ramanujan conjecture, which led number theory in the 20th century, was that the absolute value distribution of the zeros of the *L*-function of an elliptic curve satisfies an analog of the Riemann hypothesis, but it was further formulated by Mikio Sato in 1963 that the argument (angle) distribution of the zeros follows a $\sin^2$-distribution. The difficulty lies in the fact that a problem that was solved with a single *L*-function in the Fermat conjecture must now be solved for an *L*-function associated with a representation determined by a symmetric product of *n* numbers (n = 1,2,3...). This solution is groundbreaking, but many more important issues remain unsolved. The challenge to the Sato–Tate conjecture is at the heart of the Langlands conjecture, and the solution awaits progress in non-holomorphic automorphic forms, which are deeply related to representation theory from the spectral viewpoint of invariant differential operators. The two articles in this issue on automorphic forms/representations [8] and representation theory [7] focus on research that goes to the core of this problem. The relationship between representation theory and quantum mechanics is deep and extends to number theory (e.g., [9]). It is also deeply connected to problems in invariant theory, combinatorics, special function's theory, probability theory, and statistics (e.g., [6]). Some of these issues are introduced in the articles in this issue.

References

[1] J. Dieudonné, "History of Algebraic Geometry," Wadsworth, ISBN 978-0-534-03723-9, MR 0780183, 1985.

[2] P. Cartier, "A Mad Day's Work: from Grothendieck to Connes and Kontsevich. The Evolution of Concepts of Space and Symmetry," Bull. Amer. Math. Soc., Vol. 38, No. 4, pp. 389–408, 2001. https://doi.org/10.1090/S0273-0979-01-00913-2

[3] K. Sano, "Arithmetic Problems in Dynamical Systems," NTT Technical Review, Vol. 22, No. 9, pp. 26–29, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa2.html

[4] R. Irokawa, "How Number Theory Elucidates the Mysteries of Complex Dynamics—Viewed through Non-Archimedean Dynamics," NTT Technical Review, Vol. 22, No. 9, pp. 30–38, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa3.html

[5] H. Miyazaki, "Motives—Abstract Art of Numbers, Shapes, and Categories," NTT Technical Review, Vol. 22, No. 9, pp. 39–44, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa4.html

[6] C. Reyes-Bustos and M. Wakayama, "Representation Theory and Combinatorics Arising from Determinants," NTT Technical Review, Vol. 22, No. 9, pp. 45–52, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa5.html

[7] R. Nakahama, "Symmetry and Representation Theory of Lie Groups and Lie Algebras," NTT Technical Review, Vol. 22, No. 9, pp. 53–58, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa6.html

[8] S. Horinaga, "Modular Forms and Fourier Expansion," NTT Technical Review, Vol. 22, No. 9, pp. 59–64, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa7.html

[9] C. Reyes-Bustos and M. Wakayama, "Light-matter Interaction and Zeta Functions," NTT Technical Review, Vol. 22, No. 9, pp. 65–72, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa8.html

[10] J. H. Silverman, "Bibliography for Arithmetic Dynamics," 2022. https://www.math.brown.edu/johsilve/ADSBIB.pdf

[11] W. Fulton and J. Harris, "Representation Theory: A First Course," Graduate Texts in Mathematics, Vol. 129, Springer, 2013. https://doi.org/10.1007/978-1-4612-0979-9

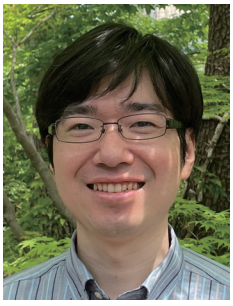
[12] R. Berndt, "Representations of Linear Groups: An Introduction Based on Examples from Physics and Number Theory," Vieweg, 2007. https://doi.org/10.1007/978-3-8348-9401-4

[13] N. Koblitz, "Introduction to Elliptic Curves and Modular Forms," Graduate Texts in Mathematics, Vol. 97, Springer, 1993. https://doi.org/10.1007/978-1-4612-0909-6

Kaoru Sano

Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a B.E., M.E., and a Ph.D. in science from Kyoto University in 2014, 2016, and 2019. He worked as an assistant professor in the Faculty of Science and Engineering at Doshisha University before joining NTT in March 2023. His current interest is the arithmetic aspects of dynamical systems.



Hiroyasu Miyazaki

Senior Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in mathematical sciences from the University of Tokyo in 2011, 2013, and 2016. Before joining NTT in March 2022, he worked at Institut de Mathématiques de Jussieu-Paris Rive Gauche as a postdoctoral fellow of Fondation Sciences Mathématiques de Paris then at RIKEN as a senior research scientist. He received FY2021 RIKEN BAIHO Award. He specializes in arithmetic geometry and algebraic geometry.

Masato Wakayama

Research Principal and Head of NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a Ph.D. in mathematics from Hiroshima University in 1985. Before joining NTT in 2021, he had held various academic positions: an associate professor at Tottori University, visiting fellow at Princeton University, visiting professor at the University of Bologna, distinguished lecturer at Indiana University, professor of mathematics, distinguished professor, dean of the Graduate School of Mathematics, the founding director of the Institute of Mathematics for Industry and executive vice president of Kyushu University, and vice president and professor at Tokyo University of Science. He is now also a professor emeritus of Kyushu University. He specializes in representation theory, number theory, and mathematical physics.

Arithmetic Problems in Dynamical Systems

Kaoru Sano

Abstract

In discrete dynamical systems, the ultimate goal is to understand the asymptotic behavior of all points under iterated compositions of a certain transformation (self-map) of a certain space. In arithmetic dynamics, the asymptotic behavior of points with coordinates of arithmetic interest (algebraic numbers or p-adic numbers) is examined. In connection with arithmetic dynamics, some problems are reduced to the determination of rational points on curves. This article introduces the issues in arithmetic dynamics related to these problems.

Keywords: number theory, dynamical systems, arithmetic dynamics

1. Introduction

A system in which points move according to a certain rule over time is called a dynamical system. Given a polynomial or a rational map f, we consider the orbit of each point under iterated composition, that is,

$$z \mapsto f(z) \mapsto f(f(z)) = f^2(z) \mapsto f(f(f(z))) = f^3(z) \mapsto \cdots.$$

Regarding this sequence as a discrete time series, we obtain a dynamical system. The questions of when this sequence diverges to infinity or converges to a certain value are fundamental yet challenging. Arithmetic dynamics studies arithmetic phenomena in such dynamical systems and was established around 2000 by Silverman. Depending on whether the focus is more on number theory or dynamical systems, the nature of the research varies. This article introduces arithmetic dynamics from a number theoretic perspective, particularly problems related to the determination of rational points on curves. Problems from the dynamical-systems perspective are introduced in another article [1] in this issue.

One major goal in arithmetic dynamics is to complete the dictionary of analogies between the theory of elliptic curves or their higher-dimensional analogs, Abelian varieties in number theory, and their dynamical system counterparts. Through the dictionary, one often obtains new insights into arithmetic geometry.

2. Morton–Silverman conjecture

A torsion point on an elliptic curve is a point that becomes the identity element O under repeated addition. This is equivalent to a point where the orbit under the iterated composition of the doubling map is a finite set. When an elliptic curve is defined over rational numbers, Mazur proved that there are at most 16 such rational points (more precisely, he completely determined the possible group structures) [2].

What about, for example, the iteration of the map z^2 on the complex plane? The points, the orbits under z^2 of which are finite in the complex domain, are the roots of unity and 0. Among these, the rational points (rational preperiodic points) are only 0, 1, and –1. What about the map $z^2 - \frac{3}{4}$? The rational preperiodic points in this case are only $\frac{1}{2}$, $-\frac{1}{2}$, $\frac{3}{2}$, and $-\frac{3}{2}$. Is the finiteness of rational preperiodic points special to these maps? In fact, it can be proven that the number of rational preperiodic points is finite for any polynomial with rational coefficients of degree $d \geq 2$. However, is the number of such points as small as 3 or 4? When restricted to rational periodic points, i.e., rational points with periodic orbits, what periods are possible? The following conjecture addresses this.

Morton–Silverman uniform boundedness conjecture (special case): For any integer $d \geq 2$, there exists a constant N_d such that the number of rational preperiodic points of any rational function f of degree d is at most N_d.

This conjecture remains largely open even for quadratic polynomials $z^2 + c$ (with c a given rational number). It is relatively easy to prove that there are infinitely many cs for which there are rational periodic points z of periods 1, 2, and 3. However, it has been shown that there are no cs for which $z^2 + c$ allows rational periodic points of periods 4, 5, or 6 (with the 6-period case requiring the assumption of the Birch–Swinnerton-Dyer (BSD) conjecture) [3–5]. Assuming a generalized *abc* conjecture, it has been proven that $z^2 + c$ does not have rational periodic points of period 4 or higher. One might think of solving the equation $f_c^n(z) = z$ to find rational periodic points of period n. For $n = 4$, for example, one would consider the solutions of the polynomial obtained by dividing $f_c^4(z) - z$ by $f_c^2(z) - z$. The situation is similar for general period n. The resulting polynomials are called the n-th dynatomic polynomials. The figure X_n^{dyn} defined by these polynomials is a curve, and the problem reduces to determining the rational points on this curve. These problems are familiar to those aware of Fermat's Last Theorem. In fact, for $n = 4$, 5, 6, the results are proven using theories developed for determining rational points on specific curves, fully using techniques built up until the solution of Fermat's Last Theorem. However, it should be noted that the key theory that led to the final proof of Fermat's Last Theorem is not about determining rational points on specific curves. By using a non-trivial rational solution of Fermat's Last Theorem, a too-nice elliptic curve called the Frey curve is defined. According to the Taniyama–Shimura conjecture, which is now a theorem, any elliptic curve corresponds to a modular form [6, 7]. However, due to the properties of the original elliptic curve, the corresponding modular form has such too-nice properties that it can be shown not to exist, leading to a contradiction. Therefore, Fermat's Last Theorem's non-trivial solution does not exist. While it is natural to explore if this surprising method can be applied to the Morton–Silverman conjecture, no method, such as for defining a Frey curve, has been developed.

As mentioned above, Mazur proved that the number of rational torsion points on an elliptic curve is at most 16, but what about the higher-dimensional case, such as Abelian varieties? This problem remains open even for the two-dimensional case. Fakhruddin has shown that this conjecture follows from the Morton–Silverman conjecture, indicating a significance beyond merely following analogies [8].

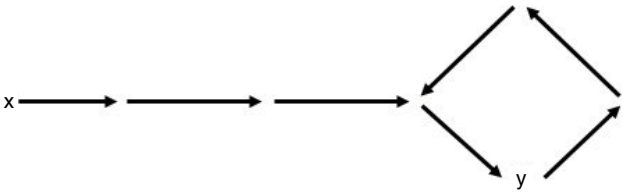


Fig. 1. Preperiodic orbit.

3. Dynamical cancellation

Consider the following scenario related to the Morton–Silverman conjecture. Let f be a polynomial of degree d. Suppose f has a rational preperiodic point x. Such a point will eventually enter a periodic orbit after certain iterations of f. Suppose it enters a periodic orbit of period 4 at time 3, as illustrated in the orbit diagram (**Fig. 1**). In this case, let $y = f^4(x)$. Then, x and y collide at time 3. That is, $f^2(x) \neq f^2(y)$ and $f^3(x) = f^3(y)$. For a fixed rational map f, how many rational pairs (x, y) satisfy $f^{n-1}(x) \neq f^{n-1}(y)$ and $f^n(x) = f^n(y)$? This question is called dynamical cancellation. To answer this, one could examine the existence of solutions (x, y) to the equation $\frac{f^n(x) - f^n(y)}{f^{n-1}(x) - f^{n-1}(y)} = 0$ for each integer n (≥ 1). This is again reduced to the problem of determining rational points on curves, which is difficult. However, in 2023, Bell, Matsuzawa, and Satriano proved that for any rational function f of degree 2 or higher, there are no rational pairs (x, y) satisfying $f^{n-1}(x) \neq f^{n-1}(y)$ and $f^n(x) = f^n(y)$ for sufficiently large n [9]. In joint work with Matsuzawa, I have generalized this result to two dimensions [10], and Zhong obtained results for higher dimensions [11]. Although these results are about determining rational points on curves, their proofs use algebraic geometry and p-adic analysis.

In a different direction from this generalization to higher dimensions, another interesting question is whether the bound on n in dynamical cancellation is independent of f when the degree d is fixed. If this uniform version of dynamical cancellation holds, by considering examples such as those mentioned at the beginning of this section, we can determine the maximum length of the tail of preperiodic orbits,

contributing to the Morton–Silverman conjecture.

4. Preimages of 0

Returning to the topic of elliptic curves, let us consider the problem of finding torsion points that become O under repeated multiplication by a prime p. How many such torsion points exist? If it can be shown that there are no such points other than O for all but finitely many p, it would yield a result comparable to Mazur's theorem. Similar considerations are applied to Abelian varieties. In the context of dynamical systems, consider the analogous problem for the map $f_c(z) = z^2 + c$. How many pairs of rational numbers (c, z) and positive integers n satisfy $f_c^n(z) = 0$? This is the problem of determining rational points on the curve X_n^{pre} defined by $f_c^n(z) = 0$. Faber, Hutz, and Stoll have shown, assuming the BSD conjecture, that for $n \geq 4$, there are no rational points (c, z) with $c \neq -1, 0$. The map sending a point (z, c) on X_n^{pre} to $(f_c(z), c)$ on X_{n-1}^{pre} deeply connects these curves, resembling the modular curves describing torsion points on elliptic curves.

5. Arboreal Galois representations

Shifting direction from the problem of determining rational points on curves, let us consider problems related to extensions of number fields. As in Kummer's approach to Fermat's Last Theorem, the uniqueness of prime factorization becomes a crucial issue when extending the world of numbers (i.e., considering number fields). The extent to which unique factorization fails is described by the quantity called the class number. Computing the class number of a given number field remains a central, challenge in modern number theory. For example, Iwasawa's theory studying the field extension called $\mathbb{Z}_p$-extension is one of the great theories in this direction.

In arithmetic dynamics, a similar problem to Iwasawa's theory arises. Fix an f and rational number x, and consider the tree of points formed by the preimages under iterated composition of f (**Fig. 2**). The problem of determining the number of rational points in this tree was discussed in section 4, where it was noted that rational points typically disappear early. The field obtained by adding these points to the field of rational numbers is called an iterated Galois extension. How does this extension change as the number of iterations increases? Consider the preimages of 1 under $f(z) = z^p$. This corresponds to considering all p-th roots of unity. Adding these to the field of rational numbers yields a cyclotomic $\mathbb{Z}_p$-extension. When this extension is stopped at the n-th stage, a number field is obtained. In Iwasawa's theory, the Iwasawa class number formula describes the asymptotic behavior of the class number, which is a remarkable theorem. What about the iterated Galois extensions arising from the preimages of 0 under $z^2 + 1$? Is there an asymptotic formula for the class number like Iwasawa's class number formula? In Iwasawa's theory, class field theory is used as a fundamental tool, and the commutativity of the Galois group (describing the symmetries of number fields) is an essential assumption. In most cases, however, the Galois group of iterated Galois extensions is non-Abelian and is expected to realize a large part of the symmetry of the tree (the automorphism group). In $\mathbb{Z}_p$-extensions, they realize very little of the symmetry of the tree, which is a rare situation. When the Galois group realizes very little of the tree's symmetry (i.e., when it has an infinite index in the automorphism group of the tree), it is considered that f has special dynamical properties. For example, if the dynamical system has an automorphism, all critical points are pre-periodic points, or the orbits of multiple critical points intersect, the Galois group has an infinite index. However, it is an open question whether these situations exhaust all possibilities for an infinite index. Solving these problems would contribute to the non-Abelian generalization of Iwasawa's theory.

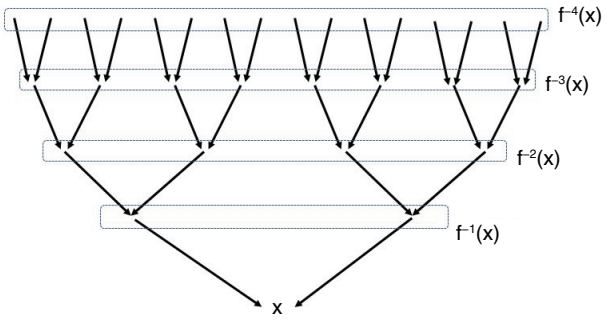


Fig. 2. Preimage tree.

6. Conclusion

I have introduced several number-theoretic problems arising from the iteration of polynomials and rational functions. These problems not only follow the analogy with the theory of elliptic curves and Iwasawa's theory but also extend techniques from

complex dynamics and reveal new arithmetic phenomena. Arithmetic dynamics is still a young field but developing rapidly, involving researchers from various fields such as algebraic geometry, complex dynamics, and arithmetic geometry. I look forward to future research developments.

References

[1] R. Irokawa, "How Number Theory Elucidates the Mysteries of Complex Dynamics—Viewed through Non-Archimedean Dynamics," NTT Technical Review, Vol. 22, No. 9, pp. 30–38, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa3.html

[2] B. Mazur, "Modular Curves and the Eisenstein Ideal," Publications Mathématiques de l'IHÉS., Vol. 47, pp. 33–186, 1977. https://doi.org/10.1007/BF02684339

[3] P. Morton, "Arithmetic Properties of Periodic Points of Quadratic Maps, II," Acta Arithmetica, Vol. 87, No. 2, pp. 89–102, 1998.

[4] E. V. Flynn, B. Poonen, and E. F. Schaefer, "Cycles of Quadratic Polynomials and Rational Points on a Genus-2 Curve," Duke Math. J., Vol. 90, No. 3, pp. 435–463, 1997. https://doi.org/10.1215/S0012-7094-97-09011-6

[5] S. Michael, "Rational 6-Cycles Under Iteration of Quadratic Polynomials," LMS Journal of Computation and Mathematics, Vol. 11, pp. 367–380, 2008. https://doi.org/10.1112/S1461157000000644

[6] A. Wiles, "Modular Elliptic Curves and Fermat's Last Theorem," Annals of Mathematics, Vol. 141, No. 3, pp. 443–551, 1995.

[7] A. Wiles, "Modular Forms, Elliptic Curves, and Fermat's Last Theorem," Proc. of the International Congress of Mathematicians, Zürich, Switzerland, 1994, pp. 243–245, Birkhäuser Verlag, Basel, 1995. https://doi.org/10.1007/978-3-0348-9078-6_18

[8] N. Fakhruddin, "Questions on Self Maps of Algebraic Varieties," J. Ramanujan Math. Soc., Vol. 18, No. 2, pp. 109–122, 2003.

[9] J. P. Bell, Y. Matsuzawa, and M. Satriano, "On Dynamical Cancellation," International Mathematics Research Notices, Vol. 2023, No. 8, pp. 7099–7139, 2023. https://doi.org/10.1093/imrn/rnac058

[10] Y. Matsuzawa and K. Sano, "On Preimages Question," preprint, arXiv:2311.02906.

[11] X. Zhong, "Preimages Question for Surjective Endomorphisms on $(\mathbb{P}^1)^n$," preprint, arXiv: 2311.04349.

Kaoru Sano

Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a B.E., M.E., and a Ph.D. in science from Kyoto University in 2014, 2016, and 2019. He worked as an assistant professor in the Faculty of Science and Engineering at Doshisha University before joining NTT in March 2023. His current interest is the arithmetic aspects of dynamical systems.

How Number Theory Elucidates the Mysteries of Complex Dynamics—Viewed through Non-Archimedean Dynamics

Reimi Irokawa

Abstract

The theory of complex dynamics is an area of pure mathematics. Even though the theory of dynamical systems belongs to analysis, it is studied in relation to various fields of mathematics, including algebra and geometry. I study it using the theory of non-archimedean numbers from number theory, which seems distant from dynamics. In this article, we discuss how attractive these theories are.

Keywords: dynamical systems, mathematics, weather forecast

1. Introduction

The theory of complex dynamics is a field in mathematics that studies asymptotic behaviors of dynamical systems defined from recurrence formulae with complex coefficients. For instance, the formula $x_{n+1} = x_n^2 + c$, parametrized by a complex number c, represents the dynamics over the complex plane. That is, if we set an initial state x_0, which is a complex number, the states x_1, x_2, x_3, ... are determined sequentially with the formula. The question is what happens to x_n when we consider $n \to \infty$. Despite the simple form of the formula, the answer is deeply related to rich and intriguing structures such as the Mandelbrot set (**Fig. 1**).

Research in this field is fascinating because of its crucial connections with other areas, such as real dynamics, arithmetic dynamics, non-archimedean dynamics, algebraic geometry, and Arakelov geometry. These areas have their methods and purposes but affect each other, i.e., if one field has significant progress, it stimulates and is applied to different fields. In this article, we explore this connection from the example of a relation between complex dynamics and non-archimedean dynamics, which is the author's field of study, after we look at what precisely complex dynamics involves.

2. Complex dynamical systems

Let us consider the example $x_{n+1} = x_n^2 + c$ that appeared in the previous section. The simplest case is when $c = 0$, that is, $x_{n+1} = x_n^2$. Since the recurrence formula has a solution $x_n = (x_0)^{2^n}$, we can see the asymptotic behavior depending on the absolute value of the initial state x_0 as follows:

(a) Case $|x_0| < 1$: $\{x_n\}$ converges to 0;
(b) Case $|x_0| > 1$: $\{x_n\}$ diverges to ∞;
(c) Case $|x_0| = 1$: $|x_n|$ stays 1 for any n.

In cases (a) and (c), the asymptotic behaviors are stable under small fluctuation of x_0 while it is not in case (b). That is, in case (a) (resp. (c)), if we give a slight change to x_0 so that $|x_0| < 1$ (resp. $|x_0| > 1$) still holds, the asymptotic behavior, converging to 0 (resp. diverging to ∞), does not change. In case (b), however, small fluctuation of an initial state affects its asymptotic behavior; if the absolute value $|x_0|$ becomes smaller (resp. greater) than 1, case (b) becomes (a) (resp. (c)). The set of unstable points under slight fluctuation is called the Julia set. In this

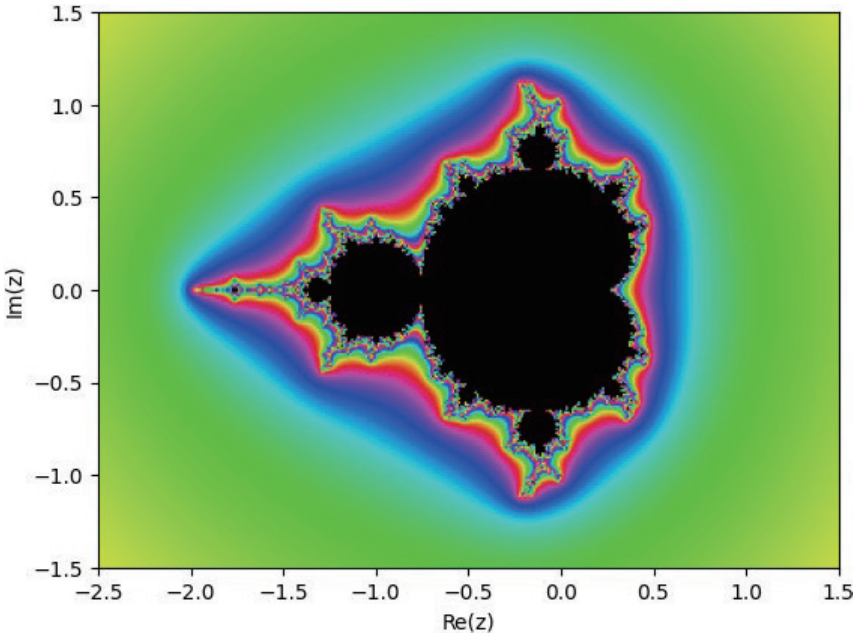


Fig. 1. The Mandelbrot set.

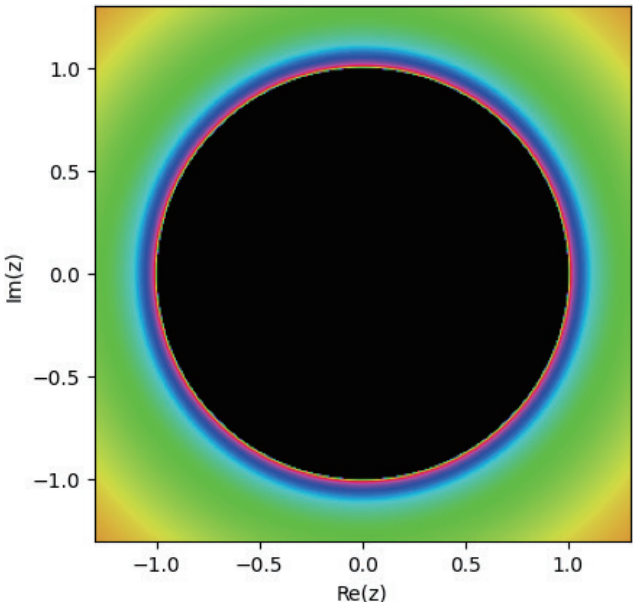


Fig. 2. The Julia set at $c = 0$.

specific case of the dynamical system defined by $x_{n+1} = x_n^2$, the Julia set is the set of point x_0 with $|x_0| = 1$, namely the unit circle in the complex plane $\mathbb{C}$. It has a simple shape, a circle as in **Fig. 2**, but this is special. For instance, **Fig. 3** shows the Julia set of the dynamics $x_{n+1} = x_n^2 + c$ with $c = -1$. This set is called "Basilica" because of its shape. **Fig. 4** is the Julia set "Rabbit" of the dynamics $x_{n+1} = x_n^2 + c$ with $c = 0.123 + 0.745i$, and **Fig. 5** is "Airplane" that appears when $c = -1.75488$. We see how myriad shapes of the Julia

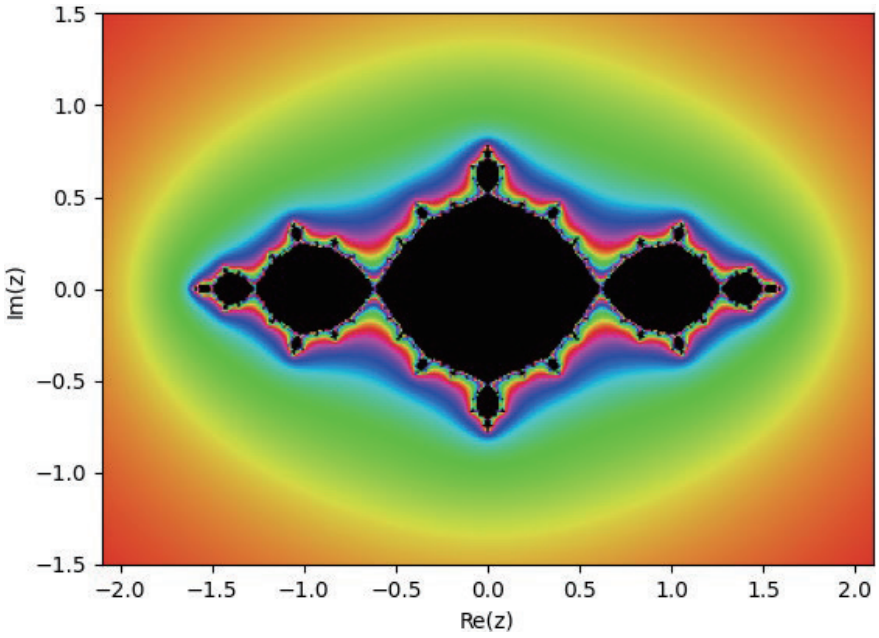


Fig. 3. The Julia set at $c = -1$, "Basilica."

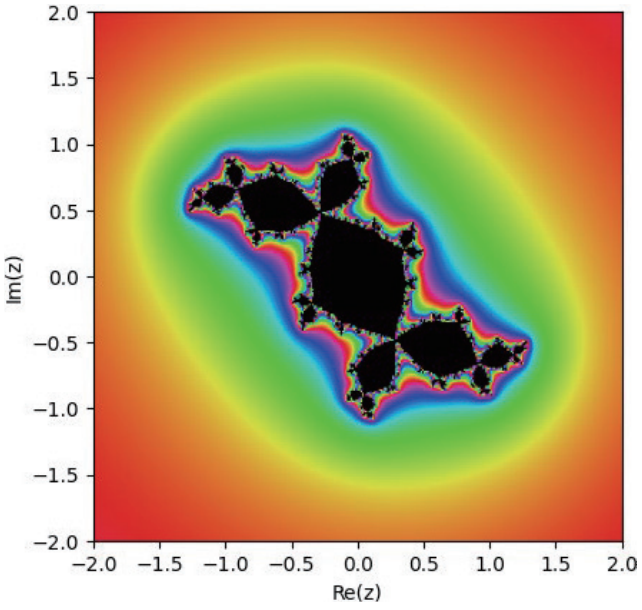


Fig. 4. The Julia set at $c = -0.123 + 0.745i$, "Rabbit."

set appear by simply taking various c's in the formula $x_{n+1} = x_n^2 + c$.

The Mandelbrot set controls the shape of Julia sets attached to the dynamics $x_{n+1} = x_n^2 + c$. Let us take a look at it more closely. It is a subset of the parameter space of $x_{n+1} = x_n^2 + c$, i.e., the space of c, which has a complicated shape with many areas divided by it. The meaning of this set is as follows. Each time you take point c from the complex plane, where the Mandelbrot set lives, it corresponds to a recurrence formula

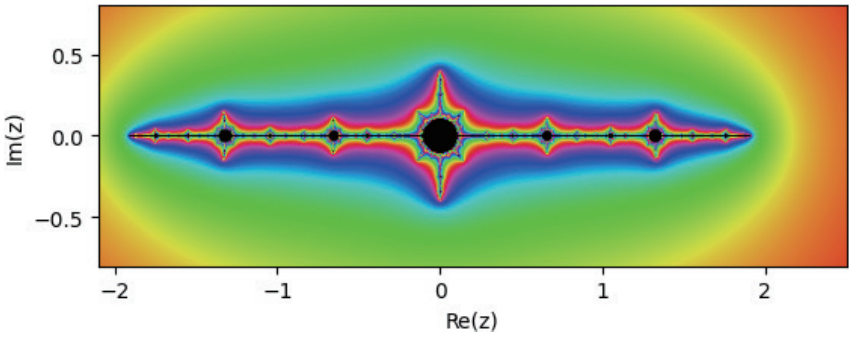


Fig. 5. The Julia set at $c = -1.75488$, "Airplane."

$x_{n+1} = x_n^2 + c$, the dynamics defined from it, and the Julia set of it. If one moves c continuously, it is reasonable to expect that the Julia set moves continuously along the motion of c. Intriguingly, it was found that this was different and the true description was as follows. The motion is continuous only if one takes the parameters inside an area divided by the Mandelbrot set. In this case, the Julia sets' shapes look similar. However, once one crosses the border, it is no longer the case. For example, $c = 0$ lies inside the area that looks like a cardioid (the widest area), $c = -1$ inside the area to the left of the cardioid that looks like a disk, and $c = 0.123 + 0.745i$ inside the area above the cardioid that looks like a disk, which means they lie in areas different to each other. The Julia sets keep their shape inside each area. However, the shape changes completely when the area moves from one area to another. The motion of the Julia set is unstable under small fluctuations on the boundary of the Mandelbrot set. Such a phenomenon, the drastic change in shapes of the Julia sets with slight fluctuation of parameters, is called bifurcation, and a parameter at which the bifurcation occurs is called a bifurcation point. The Mandelbrot set is the set of bifurcation points of c in the recurrence formula $x_{n+1} = x_n^2 + c$. The Julia set controls the stability of asymptotic behaviors for fixed c while the set of bifurcation points (or the Mandelbrot set in this specific example) controls the stability of motion of the Julia sets when we fluctuate the dynamical systems along parameters.

The Julia sets and Mandelbrot set have quite rich and interesting properties. As shown in Figs. 2–5, the various Julia sets look complicated but have rules inside the structure, called self-similarity. The Mandelbrot set has a similar property. If one zooms in on part of it, another Mandelbrot-look-alike set appears inside the Mandelbrot set, as in **Fig. 6**. It is fascinating, but at the same time, quite difficult. A case in point is the property of the so-called local connectivity of the Mandelbrot set, which is still open.

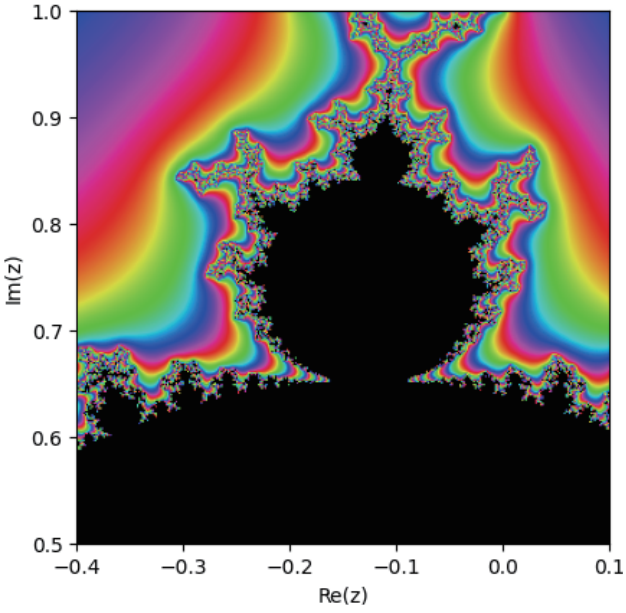


Fig. 6. The Mandelbrot set zoomed at [–0.4, 0.1] × [0.5, 1.0].

We can define similar stabilities, i.e., the notion of Julia sets and the sets of bifurcation points for different recurrence formulae. For instance, when we consider a polynomial f of more than two degrees in the formula $x_{n+1} = f(x_n)$, its parameter space can be higher dimensional. We can also consider dynamical systems with higher dimensional phase spaces. An Hénon map defines a crucial example of such dynamics that has been studied actively. It is represented by the recurrence formula $(x_{n+1}, y_{n+1}) = (x_n^2 - ax_n + by_n,$

x_n), where a and b are complex numbers with $b \neq 0$. Hénon devised it as real dynamics, i.e., x_n, y_n, a, b are all real numbers, to initially study the chaos phenomena generated from a differential equation called the Lorentz equation that simulates climate. Mathematicians later observed interesting phenomena unique to complex versions of the dynamics of the Hénon maps, and even several applications to real Hénon maps. This is a part of an interaction between real and complex dynamics.

3. Non-archimedean numbers and non-archimedean dynamics

Just as complex dynamics is a theory of dynamical systems over complex numbers, non-archimedean dynamics treats dynamical systems over non-archimedean numbers. Non-archimedean numbers is a name for several types of numbers with a common property called non-archimedes. The most typical example of a non-archimedean number is the p-adic number, which we take as an example to see what non-archimedes is. The "p" inside the p-adic number is a prime. There are 2-, 3-, 5-adic numbers, …, and they are lateral to complex and real numbers. Another typical example is called the Laurent series, which we discuss later. Let us look at the properties of non-archimedean numbers from an instance of 2-adic numbers.

Consider a metric on the set of integers, which is different from the usual one and called 2-adic distance, given by the following rule. The more times the difference of two numbers is divisible by 2, the closer they are to each other. We observe this rule by an example. In a usual metric, 4 is closer to 2 and further to 8. However, the difference between 2 and 4 is 2, which is divisible by 2 once, while that between 4 and 8 is 4, which is divisible by 2 twice. We conclude that 8 is closer to 4 than 2 with respect to the 2-adic metric.

The 2-adic distance is deeply related to the 2-adic expansion, i.e., the binary representation of numbers. The 2-adic expansions of the above three numbers are

$2 = (10)_2$;
$4 = (100)_2$;
$8 = (1000)_2$.

When we compare their distance, we need to read the 2-adic expansions from right to left. We observe that the first two digits coincide in the 2-adic expansions of 4 and 8, while only one digit is in those of 2 and 4. Comparing the above rule of the 2-adic distance with the definition of the 2-adic expansion, we see an alternative, but the equivalent rule of the 2-adic distance is the more digits of the 2-adic expansions of two numbers coincide, the closer they are to each other when we read the expansion from right to left. More precisely, it is conventional that the distance of two numbers, the 2-adic expansion of which shares the same first n digits, is defined as 2^{-n}. We write the 2-adic distance of two positive integers a and b as $d_2(a, b)$.

Let us now extend the metric. We consider a 2-adic expansion $(\cdots 111)_2$ that continues infinitely to the left. By definition, it is the limit of the sequence $(1)_2$, $(11)_2$, $(111)_2$, …, i.e., 1, 3, 7, …. They look to diverge to infinity but converge to −1 in the 2-adic distance. By adding 1 to this sequence, we obtain a sequence $(10)_2$, $(100)_2$, $(1000)_2$, …, which converges to 0. More specifically, the 2-adic distance between $(10)_2 = 2$ and 0 is $1/2$, that between $(100)_2 = 4$ and 0 is $1/4$, and that between $(1000)_2 = 8$ and 0 is $1/8$, …. We see that the distance converges to 0, i.e., $(\cdots 111)_2 + 1 = 0$, which means $(\cdots 111)_2 = -1$. The 2-adic distance can naturally extend to numbers written as a 2-adic expansion that continues infinitely to the left, such as $(\cdots 111)_2 = -1$ above.

We can consider decimals as we do in real numbers. In the 2-adic expansion, there are halves place, quarters place, $1/8$ place, etc. It is natural, by the exponential law, to regard these as being divisible "1 times," "−2 times," and "−3 times," respectively, by 2. Namely, we have $d_2(1/2, 0) = 2$, $d_2(1/4, 0) = 4$, $d_2(1/8, 0) = 8$, …. and clearly this sequence diverges to ∞. Recall that in a standard distance, a sequence, where the number of digits to the left of the decimal point increases, diverges, while a sequence, where the number of digits after the decimal point increases, does not. The opposite is true for the 2-adic distance.

Now let us define the 2-adic numbers. The 2-adic numbers are the 2-adic expansions with an infinite digit to the left of the decimal point and finite digit to the right. Note that a number with infinite digits to the left of the decimal point can be obtained as the limit of the finite-digit numbers obtained by truncating the n-th digit. The sequence of these truncated numbers converges by the above argument. We extend the 2-adic distance considered above to the set of 2-adic numbers, and it is possible to assess convergence and divergence. With a few arguments, we can show that the set of 2-adic numbers contains all the rational numbers, including the positive integers we considered first and the negative integers such as $-1 = (\cdots 111)_2$. In this sense, we can say that the set of

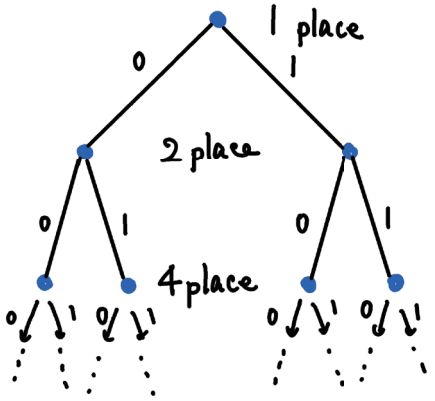


Fig. 7. Binary tree representation of 2-adic expansion.

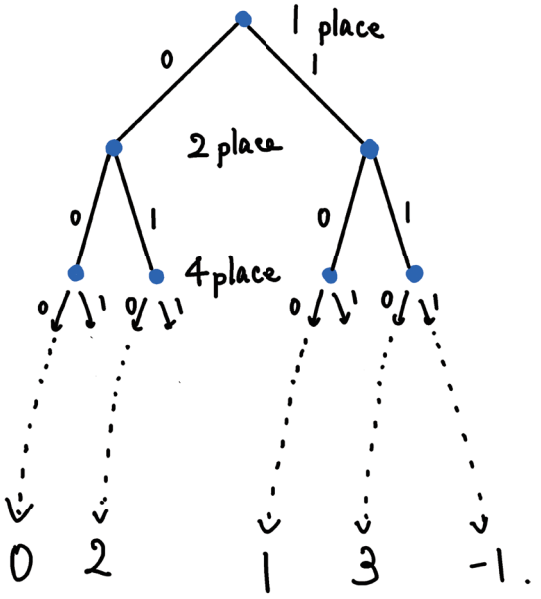


Fig. 8. Binary tree representation of 2-adic expansion with endpoints.

2-adic numbers is similar to that of real numbers: they both have distances and limits (so-called "completeness") and include all the rational numbers. However, we can also say they differ from other perspectives, most of which come from the property of the 2-adic distance, i.e., non-archimedes. The 2-adic numbers are non-archimedean, which means the strong triangle inequality holds. For any 2-adic numbers a, b, and c, we have $d_2(a + b, c) \leq \max(d_2(a, c), d_2(b, c))$. Because 2 is divisible by 2 once and 4 twice, 2 + 4 is divisible once, which is the minimum number of times that they are divisible by 2. Since this "1" in "once" appears in the 2-adic distance of 2 and 4 $d_2(2, 4) = 2^{-1}$, with −1 multiplied, the "minimum" appears as maximum in the inequality of distance. By putting any positive real numbers in a, b, and c, this inequality does not hold for the usual distance of real and complex numbers (remark: the triangle inequality $d(a + b, c) \leq d(a, c) + d(b, c)$, the weaker form of the strong triangle inequality holds instead). That is, the strong triangle inequality is unique to 2-adic numbers. Let us next discuss the difference the strong triangle inequality makes between real and 2-adic numbers.

Let us look at the difference in "shapes" of the sets of real and 2-adic numbers. Real numbers can be seen as points on a number line, which means the shape of the real numbers is a line. The question is, what is the shape of 2-adic numbers? We derive the answer from the 2-adic expansion. Consider the 2-adic numbers without decimals and describe their 2-adic expansion as a binary tree. This tree starts from one root, with two edges extending, each corresponding to the first digit (counted from the right) 0 and 1. In the same manner, from each node at the end of the edges, two more branches extend, corresponding to the second digit of the 2-adic expansion. Since 2-adic numbers have 2-adic expansion with an infinite digit to the left, we repeat this procedure infinite times. Note that if the number has only a finite digit, we take 0 infinite times. We obtain a rooted tree (**Figs. 7** and **8**) with infinite depth from this operation. The 2-adic numbers without decimals can then be seen as the set of "endpoints" that is not the first root of this tree. We also acquire the tree for all 2-adic numbers, possibly with decimals, by adding another tree above this tree. We create another rooted tree containing the original one by adding another root corresponding to the halves place, from which two edges extend, and each node at the end of the edge corresponds to one place from each of which the original tree appears. We continue this infinite times to obtain the tree of whole 2-adic numbers, which no longer has a root. The shape of 2-adic numbers is then the set of endpoints of this tree. Not only does each endpoint correspond to a 2-adic number but the tree structure also reflects the distance. As mentioned above, with respect to the 2-adic distance, the more digits of two 2-adic numbers coincide, the closer they are. By tree representation, it is equivalent to saying that the deeper they share edges in the tree, the closer they are.

We also see a difference between complex (or real) and 2-adic numbers if we consider the dynamics over

them. We have already observed the dynamical system $x_{n+1} = x_n^2 + c$ over the complex numbers above. Now, let us look at the same dynamical system, with both phase and parameter spaces 2-adic; x_0 and c are both 2-adic numbers. Because of the limited space, we only consider the formula with $c = 0$, i.e., $x_{n+1} = x_n^2$. We see that even such a simple dynamical system is good enough to observe the difference. It has a solution $x_n = x_0^{2^n}$, and we can see the asymptotic behavior depending on the absolute value of the initial state $|x_0|$ (that is, defined by $d_2(x_0, 0)$), (a)–(c) in Section 1, which is the same as the complex case. However, the strong triangle inequality makes the behavior completely different; case (b) is no longer unstable under slight fluctuation of x_0. Let us discuss this in more detail. We take any number ε such that $|\varepsilon| < 1$ and consider $x_0 + \varepsilon$. In the complex case, we saw that the asymptotic behavior was unstable under small fluctuation; for instance, if we take $x_0 = 1$, then any $\varepsilon > 0$ makes $|x_0 + \varepsilon| > 1$, which is case (a). In the 2-adic case, however, the conditions $|x_0| = 1$ and $|\varepsilon| < 1$ with the strong triangle inequality indicate that $d_2(x_0 + \varepsilon, 0) \leq \max(d_2(x_0, 0), d_2(\varepsilon, 0))$, i.e., $|x_0 + \varepsilon| \leq \max(|x_0|, |\varepsilon|) = 1$. Case (a) can never occur with small fluctuation. The condition $|\varepsilon| < |x_0|$ also indicates $|x_0 + \varepsilon| = |x_0|$, which is deduced simply by the strong triangle inequality, though we do not present its proof. This means that case (c) cannot occur, either. As a consequence, small fluctuation in the sense of the 2-adic distance does not break the condition in case (b). The dynamics $x_{n+1} = x_n^2$ over the 2-adic numbers does not have a Julia set, which is known to be impossible in complex dynamics. We can see how much the non-archimedean property affects the behavior of dynamical systems.

We go back to the first paragraph of this section from the specific number, the 2-adic one. By replacing 2 in the 2-adic distance with other prime numbers, such as 3, 5, and 7, we obtain 3-, 5-, and 7-adic numbers, with the 3-, 5-, and 7-adic distances, respectively. As mentioned above, the p-adic number is a collective term for numbers defined for each prime number. Each set of p-adic numbers has the p-adic distance for which the strong triangle inequality holds as 2-adic distance. Note that they are different from each other; that is, the sets of 2-adic and 3-adic numbers have several common properties, but not the same sets. The p-adic numbers exist as much as the number of the prime numbers, which are infinitely many. As mentioned above, the strong triangle inequality affects the nature of the numbers such as the theory of dynamical systems. Non-archimedean numbers are numbers with a distance that satisfies the strong triangle inequality, which clearly includes p-adic numbers. Even though the examples of the difference between real numbers and non-archimedean numbers (2-adic numbers there) mentioned above may seem peculiar and unintuitive, the p-adic numbers are one of the most fundamental tools in modern number theory, the area of mathematics studying the properties of integers and rational numbers. For instance, p-adic numbers are necessary to prove Fermat's Last Theorem. In the next section, we examine another application of non-archimedean numbers that are not p-adic.

4. Non-archimedean dynamics and hybrid dynamics

The last topic in this article is devoted to the theory of hybrid dynamics—the application of non-archimedean dynamics to complex dynamics. We take an example as a recurrence formula $x_{n+1} = tx_n^2$ parametrized by a complex number t. This recurrence formula is unique compared with the above example because we have $x_n = 0$ for any n when $t = 0$. We observe that a drastic change, called degeneration, occurs at $t = 0$, which is not the above-mentioned bifurcation. Bifurcation is a change in asymptotic behavior, not a form of formulae, as in degeneration. Degeneration is related to significant problems associated with the so-called compactification of the moduli spaces. Hybrid dynamics is a strong tool for studying degeneration by means of non-archimedean dynamics. As complex (resp. non-archimedean) dynamics involves the study of dynamical systems over complex (resp. non-archimedean) spaces, hybrid dynamics involves the study of dynamical systems over "hybrid" spaces. The hybrid space was introduced by Boucksom et al. [1] to investigate degeneration phenomena in algebraic geometry. Favre later imported it to study degeneration phenomena in complex dynamics [2].

The non-archimedean number considered in this theory is not a p-adic one but one that is called a complex Lauren series. A (one-dimensional) complex Laurent series is a series $\sum_{n=m}^{\infty} c_n t^n$, an infinite sum of $c_n t^n$ with variable t and complex coefficients c_n, permitting a finite number of negative powers. By complex analysis, every meromorphic function defined around the origin can be written as an infinite sum of such series, i.e., it is an example of the complex Laurent series. Because we do not require any condition

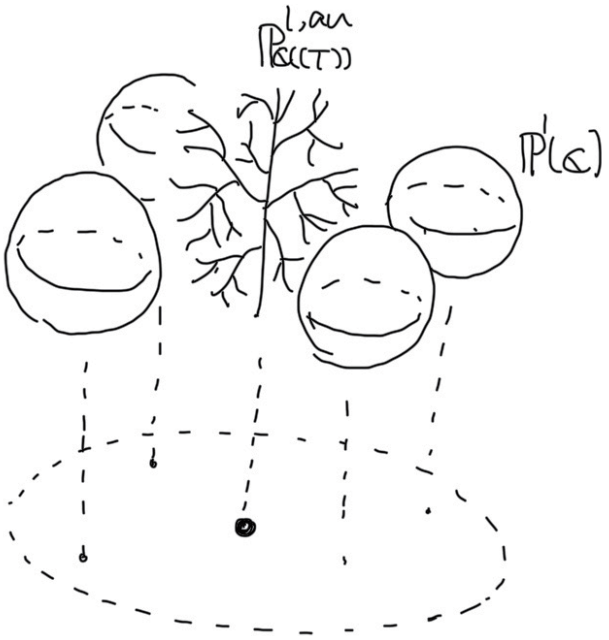


Fig. 9. A hybrid space.

on convergence, the Laurent series includes more than just meromorphic functions around the origin. Recalling that a 2-adic number has a binary representation, an infinite sum of 2^n permitting a finite number of negative powers, we can see a similarity between 2-adic numbers and complex Laurent series. It is possible to introduce a distance to the set of complex Laurent series similar to the 2-adic one. If we define that the distance between two complex Laurent series $\sum_{n=m}^{\infty} c_n t^n$ and $\sum_{n=m}^{\infty} c'_n t^n$ is e^{-n_0} if $c_n = c'_n$ holds for any $n \leq n_0$, then the set of the complex Lauren series is a kind of non-archimedean numbers by this distance. Looking at the above example $x_{n+1} = tx_n^2$ again, the coefficient $t = 0 \cdot 1 + 1 \cdot t + 0 \cdot t^2 + \cdots$ is a complex Lauren series. With t regarded as a non-archimedean number, we may consider the recurrence formula $x_{n+1} = tx_n^2$ as a dynamical system over the set of Lauren series, i.e., non-archimedean dynamics. In other words, we may consider the dynamics of formulas themselves. In general terms, hybrid dynamics describes the relation between the original complex dynamics $x_{n+1} = tx_n^2$ and induced non-archimedean dynamics.

I drew a rough diagram of the hybrid space in **Fig. 9**. It is a family of spaces parametrized by a complex number t. Any non-zero t gives just a complex plane, while $t = 0$ gives the space of the complex Laurent series. The space of the complex Laurent series is the so-called Berkovich space, the details of which are too technical to explain in this short article. Roughly speaking, by considering the Berkovich space, we take into account the whole tree in Fig. 8, not just the set of endpoints where the actual complex Laurent series lives. It is sufficient to have the image in which as the endpoints of the tree structure move according to a recurrence relation, the internal points of the tree also move accordingly. This rough diagram of the hybrid space enables us to regard spaces of non-archimedean numbers as the degenerating "limit" of a family of complex planes, or if it converts into the language of dynamics, hybrid dynamics enables us to regard non-archimedean dynamics as the degenerating "limit" of complex dynamics.

There are several reasons we need such a complicated space. One reason comes from the natural way mathematicians look at mathematical phenomena. As mentioned above, the formula $x_{n+1} = tx_n^2$ degenerates to a constant when $t = 0$. Mathematicians consider this phenomenon a wrong consequence by looking at it in an inadequate space. The next question we need to consider is what the "correct" space is to glean a proper degeneration feature. The hybrid space is one possible answer, where the degeneration limit still defines a non-trivial (or non-constant in this case) dynamics.

Favre showed that, for a family of complex dynamical systems whose recurrence formulae are one-variable with one-dimensional parameter t and possibly degenerating at $t = 0$, the family of Julia sets determined for each non-zero t "converges" to the Julia sets of the induced non-archimedean dynamics, which is a dynamical system over the space of complex Laurent series when we consider everything over the hybrid space [2]. The author presented a similar result for the dynamics of Hénon maps [3], and it is natural to expect to observe similar results in various degenerating complex dynamical systems. Even though non-archimedean numbers may seem odd at first glance, they are essential not only in number theory, as mentioned above, but also in complex dynamics, where non-archimedean dynamics uniformly describe the degeneration of complex dynamical systems.

5. Conclusion

The term "complex dynamics" does not indicate a study method but an area to be studied. Numerous mathematicians study the mysteries in complex

dynamical systems with various tools from miscellaneous viewpoints. I gave an example of non-archimedean dynamics. However, mathematicians are constantly presenting more results from multiple approaches every day. I would be delighted if I could share even a glimpse of its importance and allure.

References

[1] S. Boucksom and M. Jonsson, "Tropical and Non-Archimedean Limits of Degenerating Families of Volume Forms," Journal de l'École Polytechnique—Mathématiques, Vol. 4, pp. 87–139, 2017. https://doi.org/10.5802/jep.39

[2] C. Favre, "Degeneration of Endomorphisms of the Complex Projective Space in the Hybrid Space," Journal of the Institute of Mathematics of Jussieu, Vol. 19, No. 4, pp. 1141–1183, 2020. https://doi.org/10.1017/S147474801800035X

[3] R. Irokawa, "Hybrid Dynamics of Hénon Mappings," arXiv preprint: 2212.10851.

Reimi Irokawa

Postdoctoral Fellow, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

She received a B.S. from Waseda University, Tokyo, in 2017 and M.S. and Ph.D. in science from Tokyo Institute of Technology in 2019 and 2022. She joined the NTT Institute for Fundamental Mathematics in 2022 as a postdoctoral fellow and currently studies complex and non-archimedean dynamics.

Motives—Abstract Art of Numbers, Shapes, and Categories

Hiroyasu Miyazaki

Abstract

In arithmetic geometry, we study problems of numbers by transforming them into problems of shapes called algebraic varieties (geometric objects). Cohomology theories extract the information of algebraic varieties as linear data. Various cohomology theories have been developed to study different aspects of algebraic varieties. However, it is widely believed that there is a universal theory called the motive theory, which unifies these cohomology theories. This article gives an overview of the motive theory and presents attempts by the author and his collaborators to generalize it.

Keywords: arithmetic geometry, cohomology, motive

1. What is motive?

In mathematics, it is often the case that two different phenomena/objects show a surprising relationship. Behind such a surprising connection, mathematicians sometimes find a new mathematical concept. Motive is a very good example of this—it is a universal mathematical object that should exist behind many cohomology theories appearing in arithmetic geometry.

2. Arithmetic geometry

Using the framework of arithmetic geometry, a large part of the study of number theory is replaced with research on geometric objects called 'algebraic varieties.' A simple example of algebraic varieties is the graph of an (system of) algebraic equation(s). For example, the graph of the equation $y = x^2$, a parabola, is an algebraic variety. When an equation contains only a small number of variables, its graph is 'visible' to our eyes. However, if an equation contains many variables, its graph is often of higher dimension, making it 'invisible' to us. Even if the graph has a lower dimension, its shape could be too complicated to study just by directly seeing it.

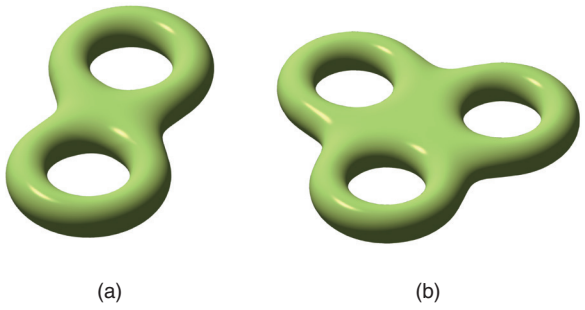
3. Invariants—How to quantify shapes

When it is difficult to investigate a shape by seeing it, the notion of 'invariant' helps us. An invariant transforms a property of shapes into a certain quantity. A typical and useful example is the genus of surfaces (**Fig. 1**), i.e., the number of holes. Of course, the genus captures just one aspect of surfaces, but it has the following important property:

> Theorem: The genus does not change after any continuous deformation.

A continuous deformation means regarding a surface as a 'soft rubber' and transforming it without tearing. As an application of this theorem, let us do the following exercise: can we continuously deform the surface in Fig. 1(a) into the one in Fig. 1(b)? Apparently, the genus of Fig. 1(a) is 2, and that of Fig. 1(b) is 3. Thanks to the theorem, any surface obtained by a continuous deformation of Fig. 1(a) remains having genus 2, not 3. This shows that Fig. 1(a) can never be continuously deformed to Fig. 1(b).

This conclusion could be intuitively obvious since these surfaces have simple structures. However, what if a surface has a trillion holes? At least to me, it is completely non-obvious that the number of holes will not change after any continuous deformation of the



(a) (b)

Fig. 1. The genus of surfaces: (a) surface of genus 2 and (b) surface of genus 3.

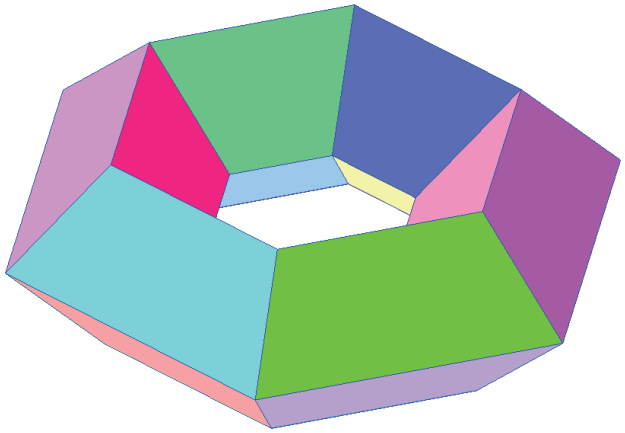


Fig. 2. Polyhedron of genus 1.

surface. The point of the above theorem is that the result is mathematically proven true for any extreme examples outside our imagination.

4. How to 'see' the invisibles

In the previous example, we could count the number of holes by directly seeing the figures. However, we will not be able to do this for 'invisible' shapes, which often appear in the study of mathematics. Therefore, let us think of another method of calculating the genus of surfaces. As the above theorem says, the genus is unchanged by any continuous deformation. Therefore, we can replace the surface of a donut with a polyhedron, as in **Fig. 2**. We then have the following surprising theorem:

> Theorem: If the genus of a polyhedron is g, then we have #(vertices) – #(edges) + #(faces) = 2–2g.

Here, #(vertices) means the number of vertices on the polyhedron, and similarly for edges and faces. Also, the alternating sum #(vertices) – #(edges) + #(faces) is called the Euler characteristic. By direct counting, we can check if the theorem is true for the surface in Fig. 2: it has 24 vertices, 48 edges, and 24 faces. And the genus is $g = 1$. Substituting them into the equation in the theorem, both sides have the same value 0. This works. If one has a pencil and piece of paper, it would be a fun exercise to try other examples, e.g., a hexahedron. In this case, the genus is $g = 0$.

In the above example, we could easily and directly count the number of holes since we could see the entire surface structure. If we live on the surface (like we live on the earth, a sphere), however, counting the number of holes would be much more difficult. Even in this situation, the above theorem ensures that we can 'compute' the genus by dividing the surface into a polyhedron and by counting the numbers of vertices, edges, and faces (which should be possible by moving around on the surface without seeing it from the universe).

This approach can be applied not only to surfaces but also to geometric objects (shapes) of higher dimensions. A (two-dimensional) polyhedron consists of three types of 'parts'—vertices, edges, and faces. These are also called cells. A shape that can be continuously deformed to an n-dimensional disk is generally called an n-cell. A vertex is a zero-dimensional disk, so it is a zero-dimensional cell. Similarly, an edge is a one-dimensional cell, and a face is a two-dimensional cell (the faces of a polyhedron are angular, but they are continuously deformed to a disk). A shape constructed by combining cells is called a cell complex[*1] (a polyhedron is a two-dimensional cell complex). Just as a surface could be continuously deformed to a polyhedron, a large part of higher dimensional shapes can be deformed to cell complexes. Cell complexes contain concrete information, such as the number of cells in each dimension and how two cells are connected (or non-connected) by another cell (e.g., we can ask whether two vertices are connected by an edge). Such information reveals important properties of 'invisible' shapes living in higher dimensions.

*1 Cell complex: If a shape (topological space) can be continuously deformed to a cell complex, it is called a CW complex, where the C stands for closure finite and the W for weak topology. Many topological spaces appearing in applications are CW complexes.

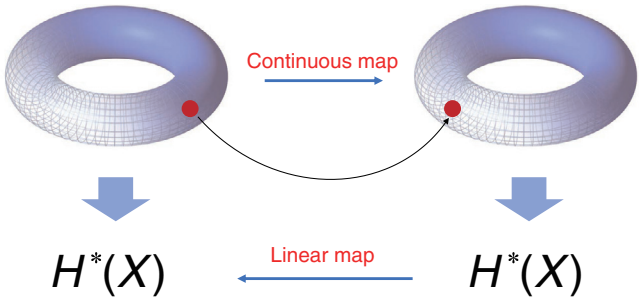


Fig. 3. Functoriality of cohomology.

5. Cohomology

The Euler characteristic depends only on the number of cells in each dimension appearing in the polyhedron and does not use the information of the relationship between the cells. By using this extra information, we can construct the cellular cohomology[*2], which drastically upgrades the Euler characteristic of a surface. The Euler characteristic assigns values to shapes, while the cellular cohomology assigns vector spaces to shapes.

Let us use the letter X to denote the shape we want to study, and let d be the dimension of X. Suppose also that X is a cell complex (by applying continuous deformation). Then there are $(d + 1)$-types of cells appearing on X—cells of 0, 1, 2, ⋯, d dimensions. The cellular cohomology is given as $d + 1$ vector spaces[*3] corresponding to the dimensions of cells, which are usually written as

$$H^0(X), H^1(X), H^2(X), \cdots, H^d(X).$$

Usually, we abbreviate the collection of these $d + 1$ vector spaces as $H^*(X)$ to simplify the notation. When X has dimension 2 (i.e., if X is a surface), then the cellular cohomology of X consists of three vector spaces $H^0(X)$, $H^1(X)$, $H^2(X)$. Any vector space has dimension, and if X is a surface, then the Euler characteristic of X coincides with the alternating sum of the dimensions of these three vector spaces. Thus, we can regard the cellular cohomology as a generalization of the Euler characteristic.

6. Functoriality of cohomology

The cellular cohomology has much richer information than the Euler characteristic. To see this, we should consider not only the shapes but also the continuous maps between them. By the cellular cohomology, a linear map $H^*(Y) \to H^*(X)$ is assigned to a continuous map[*4] $X \to Y$. This property is called the functoriality of the cellular cohomology (**Fig. 3**). If we are given the data of 'objects' and 'maps (morphisms) between objects' satisfying suitable conditions, they are generally called a category. If we have two categories and a rule to assign objects and morphisms of one of them to those of the other, it is called a functor. Using these terminologies, we can say that the cellular cohomology is a functor from the category of CW complexes to the category of (graded) vector spaces.

The functoriality of the cellular cohomology extracts much information from continuous maps. Indeed, it transforms any continuous map to a linear map, and any linear map can be represented by a matrix after fixing the basis of the vector spaces. A matrix is simply a table of numbers, which is nothing but numerical data. By applying the theory of linear algebra such as determinant, trace, and eigen values, we can obtain essential information of the matrix, hence of the continuous map we started from.

The functoriality is also useful in the study of the symmetry of shapes, which is mathematically an action of a group on a shape (e.g., rotation of a circle). If the action on a shape X is continuous, then each member of a group gives a continuous map $X \to X$, and the functoriality of the cellular cohomology

*2 Cellular cohomology: Cellular cohomology is defined only for CW complexes, but we can generalize this to another theory called 'singular cohomology', which can be applied to all topological spaces.

*3 Vector space: A set equipped with addition, subtraction, and scalar multiplication satisfying suitable conditions. Any vector space can be identified with a set of numerical vectors, i.e., tuples of numbers by fixing a basis.

*4 Continuous map: Intuitively, a map f between shapes (topological spaces) is continuous if the change in the value $f(x)$ is small whenever the change in x is small.

induces a linear map $H^*(X) \to H^*(X)$. This is nothing but a representation of the group.

7. Cohomology in arithmetic geometry

Now, let us go back to arithmetic geometry. The aim of arithmetic geometry is to study the properties of algebraic varieties, i.e., the graphs of algebraic equations. If we consider the solutions in real (or complex) numbers, then the graphs have continuous nature since the set of real or complex numbers has a continuous geometric structure. Hence, we can apply the cellular cohomology to study graphs.

However, the main target of number theory is the solutions in integers, rational numbers, etc. These numbers are non-continuous, hence, so are the graphs. It is not a good idea to apply cellular cohomology to study such non-continuous graphs since it was developed to capture the continuous nature of shapes.

To overcome this difficulty, Alexander Grothendieck, a founder of arithmetic geometry, introduced the étale cohomology as an analog of cellular cohomology in the context of arithmetic geometry. He and his collaborators proved and published the fundamental results on étale cohomology [1]. Similarly to the cellular cohomology, the étale cohomology assigns vector spaces to algebraic varieties and has a certain functoriality (we must replace 'continuous maps' with 'morphisms of algebraic varieties'). The main idea of cellular cohomology is to regard a shape as a structure built with small pieces (cells) and extract global information from those pieces. The idea of étale cohomology is similar—we split algebraic varieties into small pieces (in a suitable sense) and glue them to recover the global structure—but its actual construction uses many abstract concepts such as categories, functors, and sheaves developed in the 20th century. This abstract approach is not so-called abstract nonsense. Grothendieck used these abstract concepts to upgrade the concepts from the usual geometry.

The theory of étale cohomology is abstract and complicated but very powerful and has continuously provided many applications in arithmetic geometry, including the proof of the Weil conjecture (an analog of the Riemann hypothesis) by Deligne [2] and the proof of Fermat's last theorem by Wiles [3, 4]. It might be fair to say that arithmetic geometry cannot even exist without the theory of étale cohomology.

Modern mathematics has created many new concepts, including categories, functors, and sheaves. The extremely abstract nature of those concepts often gives the impression that mathematicians are deliberately trying to make things difficult. However, these abstract concepts were created to achieve simple goals, such as 'to create a meaningful geometry even in a discontinuous world'. Throughout history, new mathematical concepts were often considered abstract and without substance but were widely accepted by societies afterwards. Negative numbers and complex numbers are good examples, and cohomology is becoming one of them. Cohomology has been a powerful tool for capturing structures and patterns in data, opening a new field of topological data analysis providing new applications.

8. Motive

In addition to étale cohomology, various other cohomologies have been developed for different applications. Examples include de Rham cohomology, which extracts the differential geometrical structure of algebraic varieties, and crystalline cohomology, which extracts the analytic structure in the world of numbers with positive characteristics. These are created by focusing on different aspects of algebraic varieties and are seemingly unrelated to each other at first glance. Nevertheless, these different cohomologies share common properties. Various comparison theorems also hold. In other words, in certain settings, different cohomologies can be isomorphic.

Why is there such a deep relationship between cohomologies coming from very different contexts? Is it simply a coincidence? Grothendieck's answer was 'no'. He conjectured that 'behind the cohomologies of algebraic varieties, there must be a universal object unifying them' and named this hypothetical object 'motive' [5].

The term motive (motif in French) originally meant the 'driving force' of the creation of art works, such as music or paintings. Grothendieck seems to have used this term to mean a driving force creating various cohomologies. In fact, Grothendieck developed his theory by constructing the motive theory for cohomologies of algebraic varieties under the assumption that algebraic varieties are projective[*5]

*5 An algebraic variety is called projective if it is identified with the set of solutions of homogeneous equations inside a projective space. A projective space is obtained by adding infinity point(s) to the usual coordinate spaces (affine spaces). Also, if an algebraic variety has a self-intersection point or a sharp point, they are called singular points. An algebraic variety is smooth if it does not have singular points.

and smooth. His theory is now called the theory of pure motive.

9. Mixed motive

However, Grothendieck's theory can be applied only to cohomologies for projective smooth varieties. In fact, most of the cohomologies for projective smooth varieties are generalized for smooth varieties that are not necessarily projective, and they are very important in arithmetic geometry. Thus, after Grothendieck, there were many attempts to generalize the theory of pure motive by removing the projectivity condition. The result was the theory of mixed motives, which was constructed independently by Masaki Hanamura, Marc Levine, and Vladimir Voevodsky in different formulations. Let us discuss Voevodsky's method [6].

Roughly speaking, Voevodsky's idea is to construct an analogue of cellular cohomology (or, more generally, singular cohomology) in the framework of algebraic geometry. As mentioned above, it is difficult to capture number-theoretic information (e.g., solutions in rational numbers) of algebraic equations by using the usual cellular cohomology. This is because the continuous deformation kills such information—whether a point on the graph is a solution in rational numbers is completely lost if the point is moved even slightly.

Voevodsky constructed the concept of continuous deformation that makes sense even in the discontinuous world of integers and rational numbers. Mathematically, the usual continuous deformation inside a space X is formalized as a continuous map from the product of X and the real number line to X. In other words, the real number line plays the role of the space of deformation parameters (i.e., time axis). However, as explained above, the real number line (which is a continuous space) cannot be used to capture number-theoretic information. Therefore, instead of the real number line, Voevodsky used the affine line, which is a convenient algebraic variety that represents a 'one-dimensional coordinate axis' regardless of the range in numbers under consideration. It corresponds to the real number line in the world of real numbers and to the complex plane in the world of complex numbers (the complex plane is a one-dimensional space represented by one complex variable, though it is two-dimensional from the standpoint of real numbers).

Voevodsky's idea is very simple, but there were many technical difficulties to overcome. He successfully established his theory in a very satisfactory way. As naturally expected from the design, his theory produces an algebro-geometric analogue of cell complex (and its generalization, singular complex), which is called the mixed motive. The mixed motive has the information of various cohomologies of algebraic varieties. For example, singular cohomology, étale cohomology[*6], and de Rham cohomology can all be derived from the mixed motive. In other words, the mixed motive is the 'seed' of the various cohomologies. Voevodsky used his theory to prove a new comparison theorem for cohomologies, called the Milnor conjecture (and its generalization, the Bloch–Kato conjecture), for which he received the Fields Medal.

10. Towards a further generalization of motive

One of the most important and fundamental properties of the mixed motive is homotopy invariance. In the usual theory of continuous deformation, we use the real number line as the space of the deformation parameter. This automatically implies that a real number line can be continuously deformed to a single point. If we consider a continuous deformation that transfers a point x on the real number line to the point $(1 - t)x$ at time t, the point at the initial position $(1 - 0)x = x$ will be moved to the origin $(1 - 1)x = 0$ at time $t = 1$.

In the theory of mixed motives, the affine line is used as a replacement of the real number line. Therefore, the affine line is 'continuously deformed' to a single point for the same reason as above. This, in turn, means that in the theory of mixed motives, there is no distinction between the affine line and a single point. This property is called the homotopy invariance of mixed motive.

Homotopy invariance is powerful, implying various useful facts about mixed motives. However, it also imposes a fundamental restriction: the cohomology captured by the theory of mixed motive is limited to those satisfying homotopy invariance, while many cohomologies in arithmetic geometry do not satisfy homotopy invariance.

My collaborators and I have therefore constructed the theory of 'motives with modulus' that generalizes the theory of mixed motive by replacing homotopy invariance with a 'weaker' property and recasting the whole theory from scratch [7–9]. Many useful cohomologies appearing in arithmetic geometry are

*6 Étale cohomology: Precisely, we refer to the l-adic étale cohomology.

expected to be controlled by this new framework. Cohomologies that do not satisfy homotopy invariance, including cohomology of the structure sheaf, Hodge cohomology, cyclic cohomology, and Hodge–Witt cohomology have been proven to be controlled by the theory of motives with modulus.

11. Future perspectives

The theory of mixed motives is expected to control a wide class of cohomologies not captured by the classical motive theory. Our future aim is to control the theory of *p*-adic cohomologies, which has made remarkable progress. The étale cohomology referred to in this article is precisely what is often referred to as *l*-adic étale cohomology (*l*-adic cohomology for simplicity). The slogan is that *l*-adic cohomology captures the topological aspects of algebraic varieties, whereas *p*-adic cohomologies focus on the analytic aspects. Despite this difference, it is observed that there are interesting similarities and correspondences between the two. Therefore, it is naturally expected that there could be a hidden 'motive' behind them. An obvious problem is that *p*-adic cohomologies (at least part of them) are not homotopy invariant and cannot be captured using the classical motive theory. However, if *p*-adic cohomologies can be controlled by the theory of motives with modulus, comparing these theories on a common ground will become possible. The future success of our attempt will elucidate the unknown mechanism by which mysterious similarities between cohomologies are produced and will significantly impact the entire study of number theory.

References

[1] M. Artin, A. Grothendieck, and J.-L. Verdier (eds.), "Théorie des Topos et Cohomologie Etale des Schémas. Séminaire de Géométrie Algébrique du Bois-Marie 1963-64 (SGA4)," Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[2] P. Deligne, "La Conjecture de Weil. I," Publications Mathématiques de L'Institut des Hautes Études Scientifiques, Vol. 43, pp. 273–307, 1974. https://doi.org/10.1007/BF02684373

[3] A. Wiles, "Modular Elliptic Curves and Fermat's Last Theorem," Annals of Mathematics, Vol. 141, No. 3, pp. 443–551, 1995.

[4] R. Taylor and A. Wiles, "Ring-Theoretic Properties of Certain Hecke Algebras," Annals of Mathematics, Vol. 141, No. 3, pp. 553–572, 1995.

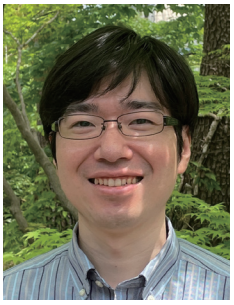
[5] J. Milne, "Motives — Grothendieck's Dream," 2012. https://www.jmilne.org/math/xnotes/MOT.pdf

[6] V. Voevodsky, "Triangulated Categories of Motives over a Field," Cycles, Transfers, and Motivic Homology Theories, Vol. 143, pp. 188–238, Princeton Univ. Press, 2000. https://doi.org/10.1515/9781400837120.188

[7] B. Kahn, H. Miyazaki, S. Saito, and T. Yamazaki, "Motives with Modulus, I: Modulus Sheaves with Transfers for Non-proper Modulus Pairs," Épijournal de Géométrie Algébrique, epiga:5979, 2021. https://doi.org/10.46298/epiga.2021.volume5.5979

[8] B. Kahn, H. Miyazaki, S. Saito, and T. Yamazaki, "Motives with Modulus, II: Modulus Sheaves with Transfers for Proper Modulus Pairs," Épijournal de Géométrie Algébrique, epiga: 5980, 2021. https://doi.org/10.46298/epiga.2021.volume5.5980

[9] B. Kahn, H. Miyazaki, S. Saito, and T. Yamazaki, "Motives with Modulus, III: The Categories of Motives," Ann. K-Theory, Vol. 7, No.1, pp. 119–178, 2022. https://doi.org/10.2140/akt.2022.7.119



Hiroyasu Miyazaki

Senior Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in mathematical sciences from the University of Tokyo in 2011, 2013, and 2016. Before joining NTT in March 2022, he worked at Institut de Mathématiques de Jussieu-Paris Rive Gauche as a postdoctoral fellow of Fondation Sciences Mathématiques de Paris then at RIKEN as a senior research scientist. He received FY2021 RIKEN BAIHO Award. He specializes in arithmetic geometry and algebraic geometry.

Representation Theory and Combinatorics Arising from Determinants

Cid Reyes-Bustos and Masato Wakayama

Abstract

Originating from the unified treatment of probability distributions, the α-determinant is a one-parameter interpolation of the determinant and permanent of a matrix. While it generally does not have the invariance properties of the determinant, the irreducible representations of the associated general linear (Lie) groups define interesting invariants that unveil a new and rich mathematical theory with abundant unsolved problems, related to symmetric functions, representation theory, combinatorics, number theory, probability and other areas.

Keywords: α-determinant, wreath determinant, irreducible decomposition of representations

1. Introduction

In this article, we focus on matrix Lie groups, that is, groups of matrices with the operation of the usual matrix product (i.e., composition of linear forms), the most basic example of which is the general linear group over the complex numbers. Since Lie groups are also geometrical objects (differentiable manifolds), both the group structure and differentiable structure coexist in a single object and must be accounted for. Starting from certain special representations of these groups, we introduce research on several areas of mathematics, including new research directions and open problems.

1.1 Groups

A group G is a set with an operation (generically called "product") $G \times G \ni (g, h) \to gh \in G$ that is associative, i.e., $(gh)k = g(hk)$ for $g, h, k \in G$. In addition, G must have an identity element $e \in G$ such that $ge = eg = g$ holds for all $g \in G$, and for all $g \in G$ there must be an inverse element g^{-1} such that $gg^{-1} = g^{-1}g = e$.

The main examples of groups in this article are the symmetric group of m elements $\mathfrak{S}_m$, i.e., the groups of permutations of m letters with the operation of composition, and the general linear group $GL_n(\mathbb{R})$ (resp. $GL_n(\mathbb{C})$), the group of n dimensional square nonsingular (invertible) matrices with real (resp. complex) entries with the usual matrix multiplication.

1.2 Lie groups, Lie Algebras and their representations

A representation (π, V) of G is a group homomorphism π from G to the group of linear transformations $GL(V)$ of a vector space V. We normally consider only vector spaces over the complex numbers. In this setting, we say that G acts on V, or that V is a G-module, and if it is clear from the context, we omit the notation π. To consider an infinite dimensional V, it is indispensable to introduce the notion of topology; thus, we only consider finite dimension. By introducing a Hermitian inner product to the n-dimensional V, we may consider the group $U(n)$ of unitary matrices, the group of matrices that leave the inner product invariant. The development of representation theory took place in parallel with the revolutionary physical theories of relativity and quantum physics, resulting in a notation and terminology that mixes mathematical and physical subtleties and may be confusing at first but that is not without merit.

As mentioned above, Lie groups are manifolds with

geometric structure and notions such as "curvature" and "connectedness," so they cannot be described faithfully only with linear algebra. Because of this, and to the extent possible, instead of G, we consider the associated "differentiated" Lie algebra $\mathfrak{g}$. Since $\mathfrak{g}$ is equivalent as a vector space to the set of tangent vectors at the identity $e \in G$, its action is regarded as the infinitesimal version of the Lie group action. In this article, we only consider finite-dimensional representations, so Lie algebras suffice, but we note that this is not the case for general Lie groups and representations.

The Lie algebra $\mathfrak{gl}_n(\mathbb{C})$ of the Lie group $GL_n(\mathbb{C})$ is the ring of all n-dimensional matrices. For a matrix $X \in \mathfrak{gl}_n(\mathbb{C})$, the exponential map satisfies $\det(\exp X) = e^{\mathrm{tr}X} \neq 0$, thus $\exp X \in GL_n(\mathbb{C})$ is the corresponding Lie group element. Therefore, if the action of $GL_n(\mathbb{C})$ on a polynomial f over Mat_n is $(g.f)(x) = f(g^{-1}x)$, the infinitesimal action of $\mathfrak{gl}_n(\mathbb{C})$ is given by

$$(X.f)(x) = \frac{d}{dt} f(\exp(-tX)x)|_{t=0}.$$

1.3 Motivation for the representation theoretical study of α-determinant

The α-determinant is a generalization of both the determinant and permanent*. In the definition of the α-determinant, given below, the sign changes appearing in the definition of the determinant are replaced with certain weights depending on the parameter α. By setting $\alpha = -1$ we recover the definition of the usual determinant and with $\alpha = 1$, that of the permanent. It was originally introduced as α-permanent by Vere-Jones [1] in the context of probability theory. The name α-determinant follows the convention of Tomoyuki Shirai and Yoichiro Takahashi [2]. They used the α-determinant to construct point processes that generalize the ones used in finance and time-series analysis, namely, the boson point process, fermion point process, and Poisson point process.

Surprisingly, when $\alpha \neq -1$ the α-determinant loses the multiplicative property of the usual determinant, that is, $\det(AB) = \det(A)\det(B)$. Thus, we immediately ask the following problem question.

Problem 1: Where did the multiplicative property go?

The determinant is a one-dimensional representation of $GL_n(\mathbb{C})$. While the permanent does not define a one-dimensional representation, it defines an irreducible representation of the vector space spanned by the symmetric tensor product. With irreducible representation, we refer to a subspace of V, different from $\{0\}$ and V itself, that is invariant under the action of G. Informally, we may think of irreducible representations as prime numbers or elementary particles in particle physics. Therefore, we may also ask the following problem question.

Problem 2: For general α, what is the structure of the spaces spanned as the group acts on the α-determinant?

These two questions sparked the interest of the second author of this article in the representation theory of the α-determinant [3].

2. The α-determinant

For a fixed α, the α-determinant of a square matrix $A = (a_{ij}) \in \mathrm{Mat}_n$ is a modification of the usual determinant defined as

$$\det{}^{(\alpha)}A := \sum_{\sigma \in \mathfrak{S}_n} \alpha^{n-\nu_n(\sigma)} a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

Here, $\nu_n(\sigma)$ is the cycle number of a permutation $\sigma \in \mathfrak{S}_n$, i.e., the number of cycles in the cycle decomposition of σ. Hereafter, we write $\nu(\sigma) := n - \nu_n(\sigma)$ and note that $\nu(\sigma)$ is the minimum number of transpositions required to express σ.

A theorem of Vere-Jones [1] is that $\det(I_n - \alpha TA)^{-\frac{1}{\alpha}}$ can be expanded in terms of the α-determinant as

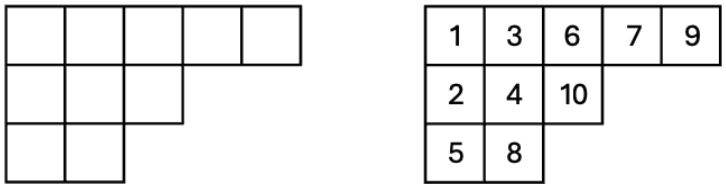
$$\det(I_n - \alpha TA)^{-\frac{1}{\alpha}} = \sum_{k \geq 0} \frac{1}{k!} \sum_{i_1, \ldots, i_k \geq 0} t_{i1} \cdots t_{ik} \det{}^{(\alpha)} \left(a_{i_p, i_q}\right)_{1 \leq p,q \leq k},$$
$$T = (t_{ij}) \in \mathrm{Mat}_n.$$

We refer the reader to a previous study [4] for details on the discovery of the formula and its applications.

2.1 Preliminaries from representation theory

It is well known that the finite dimensional irreducible representations (irreducible $\mathfrak{gl}_n$-modules) of $GL_n(\mathbb{C})$ and its Lie algebra $\mathfrak{gl}_n(\mathbb{C})$ are parametrized by the so-called highest weights. With respect to the standard action of $GL_n(\mathbb{C})$ on $\mathbb{C}^n$, the natural action of $GL_n(\mathbb{C})$ on the tensor product $(\mathbb{C}^n)^{\otimes m}$ commutes with the action of $\mathfrak{S}_m$ permuting the m spaces in the tensor product, because there is a duality between the two actions. This is the well-known Schur–Weyl duality [5]. From this perspective, the highest weights are

* Permanent: A matrix function like the determinant but with no sign changes in the definition.



(a) Young diagram of the partition (5,3,2) ⊢ 10 (b) Example of standard tableau

Fig. 1. Young diagrams: visual representations of partitions.

identified with partitions of integers. We say that the vector of non-negative integers $\lambda = (\lambda_1, \lambda_2, ..., ..)$ is a partition of n, written as $\lambda \vdash n$, when it satisfies $n = \lambda_1 + \lambda_2, + \cdots (\lambda_1 \geq \lambda_2 \geq \cdots)$. The length $l(\lambda)$ is the number non-zero parts of the partition λ.

A λ is often identified with the associated Young diagram, defined by

$$\{(i, j) \in \mathbb{Z}^2 \mid 1 \leq j \leq \lambda_i, i \geq 1\}.$$

It is customary to illustrate the Young diagrams as left-aligned arrangements of boxes with λ_i boxes in the i-th row, as shown in **Fig. 1(a)**. If we arrange the n numbers 1, 2, ..., n in a Young diagram λ such that numbers in a row increase from left to right and in each column the numbers increase from top to bottom we obtain a standard tableau of shape λ (**Fig. 1(b)**). The set $\mathrm{Stab}(\lambda)$ is the set of all standard tableaus of shape λ and is a fundamental notion in the representation theory of the symmetric group.

2.2 Cyclic modules arising from α-determinant

The action of $\mathfrak{gl}_n(\mathbb{C})$ on the ring of polynomial functions $\mathcal{A}(\mathrm{Mat}_n)$ on the variables $\{x_{ij}\}_{1 \leq i,j \leq n}$ is given by the differential operators

$$E_{ij}.f = \sum_{k=1}^{n} x_{ik} \frac{\partial f}{\partial x_{jk}}.$$

Here, $\{E_{ij}\}$ $(1 \leq i, j \leq n)$ is the standard (matrix elements) basis of $\mathfrak{gl}_n(\mathbb{C})$.

For $X = (x_{ij})$, we denote as $V_n(\alpha)$ the cyclic $\mathfrak{gl}_n(\mathbb{C})$-module generated by $\det^{(\alpha)}X$. In other words, $V_n(\alpha)$ is the vector space (module) generated by the (repeated) action of elements of $\mathfrak{gl}_n(\mathbb{C})$ on $\det^{(\alpha)}X \in \mathcal{A}(\mathrm{Mat}_n)$. For $\alpha = -1$, we have $\det^{(\alpha)}X = \det X$; thus, $V_n(-1) = \mathbb{C}\cdot\det X$. Therefore, the study of the structure of $V_n(\alpha)$ is an important step towards the answer to both **Problems 1** and **2**.

Theorem 2.1 [3]. Let $\mathbf{E}_n^{\lambda}$ be an irreducible $\mathfrak{gl}_n(\mathbb{C})$-module of highest weight λ. Then, $V_n(\alpha)$ has the irreducible decomposition

$$V_n(\alpha) \cong \bigoplus_{\substack{\lambda \vdash n \\ f_\lambda(\alpha) \neq 0}} (\mathbf{E}_n^{\lambda})^{\oplus f^{\lambda}}.$$

Here, $f^{\lambda} := |\mathrm{Stab}(\lambda)|$ and $f_\lambda(x) := \prod_{(i,j)\in\lambda} (1 + (j - i)x)$ is the content polynomial of λ.

This theorem shows that if $\alpha \notin \left\{\pm 1, \pm\frac{1}{2}, ..., \pm\frac{1}{n-1}\right\}$, $V_n(\alpha)$ is equivalent to the standard representation of the tensor product $(\mathbb{C}^n)^{\otimes n}$ of $\mathbb{C}^n$, and that the structure degenerates when α is the reciprocal of an integer with an absolute value smaller than n.

It is also natural to generalize this situation and consider the cyclic $\mathfrak{gl}_n$-module generated by the power $(\det^{(\alpha)}X)^l$. In this case, the highest weight modules appearing in the decomposition have nontrivial multiplicity given by a combinatorial quantity, and computing this multiplicity is generally a difficult problem. When $n = 2$, the pair $(\mathfrak{S}_{2l}, \mathfrak{S}_l \times \mathfrak{S}_l)$ of a symmetric group and Young subgroup is a Gelfand pair, which enables one to write the multiplicity in terms of zonal spherical functions given explicitly by hypergeometric polynomials [6]. The general case is still open and currently there is not even a conjectural form for the multiplicity. Moreover, one of the classic and fundamental notions in the theory of symmetric functions is plethysms [7]. In a previous study [8], for the power of $\det^{(1)}(X)$ (permanent) and the generated cyclic module structure, a conjecture that is related with plethysms for the symmetric tensor product was presented. The case of general n appears to be difficult, but it has been verified for $n = 2$ and arbitrary l, as well as for $n = 3$ and $l = 2$. Note that analogous research has been conducted for quantum groups [9]. For $n = 2$, the multiplicity is given both by hypergeometric polynomials and q-hypergeometric polynomials. This is a fascinating and promising line of research [10, 11].

3. Wreath determinant

As discussed in the previous section, when α takes the values $\pm\frac{1}{k}$, the structure of the cyclic module structure arising from the α-determinant changes drastically. A careful analysis shows that when $\alpha = -\frac{1}{k}$, there is a weak form of the alternating property of the usual determinant.

Let K denote the subgroup of $\mathfrak{S}_n$ consisting of permutations that fix any elements except from $k+1$ taken (arbitrary) from $1, 2, \ldots, n$. Then, for $A \in \mathrm{Mat}_n$ we observe that the sum $\sum_{\sigma\in K} \det^{(\alpha)}(AP(\sigma))$ contains the factor $(1+\alpha)(1+2\alpha)\ldots(1+k\alpha)$ ($A \in \mathrm{Mat}_n$), where $P(\sigma)$ is the permutation matrix of $\sigma \in \mathfrak{S}_p$. In other words, the sum $\sum_{\sigma\in K} \det^{(-\frac{1}{k})}(AP(\sigma))$ vanishes and if $k+1$ columns of A are equal we have $\det^{(-\frac{1}{k})} A = 0$, generalizing the alternating property of the usual determinant.

Since $\det^{(\alpha)}(AP(\sigma)) = \det^{(\alpha)}(P(\sigma)A)$, the foregoing discussion holds similarly for rows. A consequence of the weak alternating property is that we have analogs of Vandermonde and Cauchy determinant formulas [12].

3.1 Invariant theory of wreath determinants

We denote by $1_{p,q}$ the matrix of size $p \times q$ with all entries equal to 1. For $A \in \mathrm{Mat}_{n,kn}$, the k-wreath determinant is defined by

$$\mathrm{wrdet}_k A := \det^{(-\frac{1}{k})}(A \otimes 1_{k,1}).$$

For example, we have

$$\mathrm{wrdet}_2 \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \end{pmatrix}$$
$$= \det^{(-\frac{1}{2})} \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \end{pmatrix}$$
$$= \frac{a_1a_2b_3b_4}{4} - \frac{a_1a_3b_2b_4}{8} - \frac{a_2a_3b_1b_4}{8} - \frac{a_1a_4b_2b_3}{8} - \frac{a_2a_4b_1b_3}{8} + \frac{a_3a_4b_1b_2}{4}.$$

Like the usual determinant, the wreath determinant is characterized by its left-GL_k, right-$\mathfrak{S}_k \wr \mathfrak{S}_n$ relative invariance. Here, $\mathfrak{S}_k \wr \mathfrak{S}_n$ is the wreath product of the groups $\mathfrak{S}_k$ and $\mathfrak{S}_n$ and is defined using the concept of the semidirect product of groups (a technique used to obtain a new group from two known groups).

Theorem 3.1 [12]. A map f: $\mathrm{Mat}_{n,kn} \to \mathbb{C}$ with properties

1. f is a multilinear map with respect to the columns,
2. $f(QA) = (\det Q)^k f(A)$ holds for any $Q \in \mathrm{Mat}_n$, that is, f is left-GL_k relative invariant, and
3. $f(AP(\sigma)) = \pm f(A)$ holds for any $\sigma \in \mathfrak{S}_k \wr \mathfrak{S}_n$ (if $\sigma \in \mathfrak{S}_k^n$ we have $f(AP(\sigma)) = f(A)$)

is equal to the map $A \mapsto \mathrm{wrdet}_k A$ up to a scalar multiple.

While this theorem may be proved in an elementary manner, it is also an easy consequence of the invariant theory of the underlying (GL_n, GL_{kn})-duality.

3.2 An analog of group determinant for group-subgroup pairs using wreath determinants

Ferdinand Georg Frobenius developed the theory of group characters in his study of the group determinant for non-abelian groups. For a finite G, consider an indeterminate x_g for each group element $g \in G$, then the group determinant is given by

$$\Theta(G) := \det(x_{uv^{-1}})_{u,v\in G}.$$

The group determinant is a complete invariant with respect to the isomorphism classes of groups. Concretely, it holds that

$$\Theta(G) = \Theta(G') \Leftrightarrow G \cong G'.$$

The main result of Frobenius is the factorization of the group determinant into irreducible polynomials with coefficients given by the group characters, in modern terms this is equivalent to the decomposition of the regular representation of a group (the action of G on its group ring) into irreducible representations.

The origin of the work of Frobenius in the group determinant is a letter by Richard Dedekind in 1896. Dedekind had previously discovered how to factorize the group determinant for abelian groups and posed the problem for general non-abelian finite groups to Frobenius. It is not an exaggeration to say that this letter marked the beginning of representation theory of finite groups.

Recall that the k-wreath determinant is defined for $n \times kn$ matrices; therefore, by considering a pair (G, H) of a finite G and subgroup H of index k, by analogy we may define

$$\Theta(G, H) := \mathrm{wrdet}_k(x_{hg^{-1}})_{h\in H, g\in G}.$$

Note that since the wreath determinant does not have an invariance with respect to column transpositions, this definition depends on the ordering of the matrix columns and rows. Therefore, it is necessary to consider a fixed numbering (or naming) of the group elements. For a fixed bijection ϕ: $\{0, 1, \ldots, kn-1\} \to G$, we say that $g_i = \phi(i)$ is a numbering of G. For a

$\mathbb{C}$-algebra A (e.g. a polynomial ring in $kn-1$ variables) a map f: $g_i \mapsto q^i$ is called specialization. When G is an abelian group, for a reasonable numbering of the elements of G, we may consider the specialization f: $g_i \mapsto q^i$ to see that the determinant $\Theta(G, H)$ factors into products of polynomials of the form $q^i - 1$ [13].

Let us give a concrete example. From the partition $(k, \ldots, k) \vdash kn$ and corresponding irreducible characters $\chi^{(k^n)}$ of $\mathfrak{S}_k^n$, we define a $\mathfrak{S}_k^n$-bi-invariant function $\omega^{(k^n)}$ on $\mathfrak{S}_{kn}$ by averaging over $\mathfrak{S}_k^n$. Concretely, we define

$$\omega^{(k^n)}(x) = \frac{1}{(k!)^n} \sum_{g \in \mathfrak{S}_k^n} \chi^{(k^n)}(xg) \quad (x \in \mathfrak{S}_{kn}).$$

Example 3.2. For $G = \mathbb{Z}_n \times \mathbb{Z}_n$ and $H = \mathbb{Z}_n \times \{0\}$, we have

$$\Theta(G, H) = \omega^{(n^n)}(\sigma\tau^{-1}) \left(\frac{n!}{n^n}\right)^n q^{\frac{n^3(n-1)}{2}} (q^n - 1)^{n(n-1)}.$$

Here, $\sigma, \tau \in \mathfrak{S}_{n^2}$ and we define $P(\tau) = P((1\ 2\ \ldots\ n)) \otimes P((1\ 2\ \ldots\ n))$, $1_{1,n} \otimes I_n = (I_n \otimes 1_{1,n})P(\sigma)$ to obtain

$$\omega^{(n^n)}(\sigma\tau^{-1}) = \frac{1}{n!^n} \times \left\{ \text{coefficient of square-free elements } \prod_{i,j=1}^{n} x_{ij} \text{ in } \left(\det(x_{ij})_{1 \le i,j \le n}\right)^n \right\}.$$

Note that if $n = 3$, we have $\Theta(G, H) = 0$.

A comprehensive theory of group-subgroup wreath determinant has not yet been developed, but two related developments have surfaced, one is related to the Alon–Tarsi conjecture by Kazufumi Kimoto and the other is the construction of new graphs based on group and subgroups.

3.3 Alon–Tarsi conjecture

A Latin square is a square arrangement (i.e. a $n \times n$ matrix) filled with the numbers $1 \sim n$ in such a way that in each row and column, each number appears exactly once. For instance, the 12 Latin squares of size 3 are

$$\begin{pmatrix}1&2&3\\2&3&1\\3&1&2\end{pmatrix}, \begin{pmatrix}1&2&3\\3&1&2\\2&3&1\end{pmatrix}, \begin{pmatrix}2&3&1\\1&2&3\\3&1&2\end{pmatrix}, \begin{pmatrix}2&3&1\\3&1&2\\1&2&3\end{pmatrix},$$
$$\begin{pmatrix}3&1&2\\1&2&3\\2&3&1\end{pmatrix}, \begin{pmatrix}3&1&2\\2&3&1\\1&2&3\end{pmatrix}, \begin{pmatrix}1&3&2\\2&1&3\\3&2&1\end{pmatrix}, \begin{pmatrix}1&3&2\\3&2&1\\2&1&3\end{pmatrix},$$
$$\begin{pmatrix}2&1&3\\1&3&2\\3&2&1\end{pmatrix}, \begin{pmatrix}2&1&3\\3&2&1\\1&3&2\end{pmatrix}, \begin{pmatrix}3&2&1\\1&3&2\\2&1&3\end{pmatrix}, \begin{pmatrix}3&2&1\\2&1&3\\1&3&2\end{pmatrix}.$$

The sign sgn L of a Latin square L is the product of the signs of all column and row permutations. If sgn $L = 1$, we say L is even and odd if sgn $L = -1$.

If n is an odd number, the formula $\text{sgn}(P((1,2))L) = (-1)^n \text{ sgn } L$ holds for the transposition $(1,2)$, giving a bijection between odd and even Latin squares and showing that the number of odd and even Latin squares of odd size n is equal. For even numbers the problem is still open.

Conjecture 3.3 (Alon–Tarsi conjecture 1992). For an even number n, the number of even and odd Latin squares of size n is different.

While the conjecture originated in graph coloring problems, it has interesting and nontrivial equivalent formulations such as the Rota basis conjecture on a certain special choice of basis for an n dimensional vector spaces, and the non-vanishing of certain integrals over the special unitary group $SU(n)$ with respect to the square-free coordinate product Haar measure ($SU(n)$ bi-invariant measure). Kimoto also discovered the equivalence of the Alon–Tarsi conjecture in terms of the wreath determinant [12].

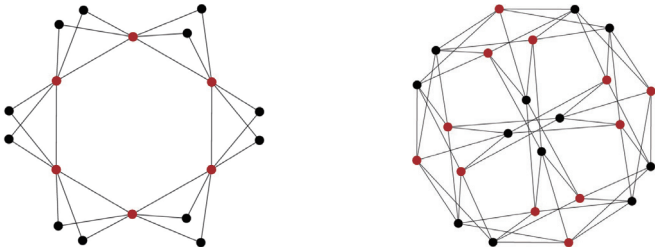
Theorem 3.4. Let n be an even number. The Alon–Tarsi conjecture is equivalent to any of the following three statements below.

1. $\text{wrdet}_n\left(\overbrace{I_n\ I_n\ \ldots\ I_n}^{n}\right) \neq 0$,
2. $\omega^{(n^n)}(g_n) \neq 0$,
3. $\sum_{y \in \mathfrak{S}_n^n} \left(-\frac{1}{n}\right)^{\nu(g_n y)} \neq 0$.

Here, $g_n \in \mathfrak{S}_{n^2}$ is a permutation satisfying $g_n((i-1)n + j) = (j-1)n + i (1 \le i,j \le n)$ and $\mathfrak{S}_n^n$ is the set of permutations of $\mathfrak{S}_{n^2}$ that fix the set $\{(i-1)n + j | 1 \le j \le n\}$ $(i = 1, \ldots, n)$.

The Alon–Tarsi conjecture is easy to verify for specific cases, but as usual in number theory and combinatorics, the proof appears to be a formidable challenge. The best results to date are those of Drisko (1997) for $n = p + 1$, where p is a prime number, and of Glynn (2010) for $n = p - 1$ (see [14] for a review up to 2019). It is worth mentioning that an alternative proof of the latter result using the wreath determinant has been given by Kimoto [15, 16].

For the number of Latin squares $ls(n)$, although it is a weak result, we remark that the asymptotic formula $ls(n)^{1/n^2} \sim e^{-2}n$ is known [17]. One of the reasons that explains that the Alon–Tarsi conjecture may be difficult to solve is that the number of even and odd Latin squares appear to be asymptotically equal.



(a) Irregular group-subgroup pair graph (b) Ramanujan (regular) pair-graph

Fig. 2. (a) Pair-graph of cyclic group of 18 elements, cyclic group of 6 elements. (b) Pair-graph of symmetric group $\mathfrak{S}_4$, alternating group A_4.

Let us propose a new research problem.

Problem 3.5. Find a suitable definition of a Latin square zeta function $\zeta_{LS}(s)$ such that we have the following equivalence:

best estimate in the prime number theorem ($\leftrightarrow$ Riemann hypothesis): $\zeta(s)$ = Alon–Tarsi conjecture $\zeta_{LS}(s)$.

In other words, such that the Alon–Tarsi conjecture is equivalent to a Riemann hypothesis' analog for $\zeta_{LS}(s)$, it might be helpful to compare the Alon–Tarsi conjecture with the phenomenon of Chebyshev's bias for prime numbers.

3.4 Group-subgroup pair graphs

In graph theory, there is an important class of graphs known as Ramanujan graphs. Ramanujan graphs are theoretically the best expanders, i.e., graphs with rapid mixing and good diffusion properties. The properties of Ramanujan graphs make them particularly important for applications, including the construction of cryptographic hash functions. A graph being Ramanujan is equivalent to the analog of Riemann Hypothesis for the Ihara zeta function of the graph, making it also interesting from the point of view of number theory.

The main problem in this area is the construction of explicit families of Ramanujan graphs of fixed degree. Some known examples are the Lubotzky–Phillips–Sarnak graphs (1994), constructed with Cayley graphs of projective linear groups over finite fields using the theory of Hamiltonian quaternion algebras and their maximal orders, and Pizer graphs (1990), based on the theory of isomorphism classes (isogenies) of hyperelliptic curves over finite fields. In the former, for a prime number p congruent to 1 modulo 4, the resulting graphs are a family of $(p + 1)$-regular Ramanujan graphs. A Cayley graph is a regular graph constructed from the elements of a finitely generated G and where the edges are defined using the (group) multiplication using a generating set S.

A new research direction is defining hash functions for group-subgroup pair graphs, which are a generalization of Cayley graphs (see **Fig. 2**), and studying their cryptographic properties [18]. For a G, subgroup H and a subset $S \subset G$ such that $S \cap H$ is a symmetric set, the group-subgroup pair graph, or simply pair-graph, $\mathcal{G}(G, H, S)$ is a graph defined in a manner analogous to Cayley graphs, but such that the resulting graph is not generally regular. Concretely, $\mathcal{G}(G, H, S)$ is a graph, the vertices of which are the elements of G and such that $h \in H$ and $g \in G$ are connected by an edge when they satisfy the relation $h \sim g \Leftrightarrow g = hs (s \in S)$ [19]. While examples of Ramanujan group-subgroup pair graphs are known for the regular case (see Fig. 2(b)), infinite families have not yet been constructed. The construction of infinite families of Ramanujan graphs is one of the major goals in the theory of pair-graphs from the point of both mathematics and applications

4. Perspectives on positive singular values: Wishart distribution, Wallace sets, and positivity of α-determinant

A natural problem is to consider the positive case $\alpha = \frac{1}{k}$ as a generalization of the permanent in a similar manner to the wreath determinant for negative singular values. We conclude this article with hints that may serve as a starting point for this research.

The study of positive definite functions on convex cones on Euclidean spaces has applications to several

areas, including optimization and probability. In the analysis of Hilbert spaces of holomorphic functions on symmetric cones (symmetric spaces of Lie groups) and its unitary representation, there is an important notion called Wallach set. For the integral kernel of symmetric cones, the set is basically given by reciprocals $\alpha = \frac{1}{k}$ of the positive singular values [20].

In statistics, there is multivariate generalization of the χ-square distribution for positive definite matrices, called Wishart distribution. Let n mutually independent p-variate random vectors $\{x_1, x_2, ..., x_n\}$, with $n \geq p$, follow a multivariate normal distribution $N(0, \Sigma)$ with mean 0 and covariance matrix Σ. Then, the random variable $X = \sum_{k=1}^{n} x_i \, {}^t x_i$ is said to follow the Wishart distribution. The Wishart distribution is used to model the distribution of a sample covariance matrix for data following a multivariate normal distribution, after scaling by sample size. The Wishart distribution, due to the intimate relation with symmetric cones of positive definite matrices, also has applications to the representation theory of Jordan algebras [20]. On the other hand, it is known that the α-determinant has an interpretation in terms of the Wishart distribution [21]. Using this relation, Shirai proved that if α is an element of $\{2/k\}_{1 \leq k < n}$, or $\{-1/k\}_{1 \leq k < n}$, then the α-determinant is non-negative for non-negative definite Hermitian real symmetric matrices. The proof uses Jack polynomials [7], an important example of symmetric functions. As is evident, these values also correspond to the positive singular values described in this article and the values of the Wallach set. It is difficult to expect that all of this is a coincidence, so an important research direction is to clarify this situation and explain the relation between these distinct areas.

References

[1] D. Vere-Jones, "A Generalization of Permanents and Determinants," Linear Algebra Appl., Vol. 111, pp. 119–124, 1988. https://doi.org/10.1016/0024-3795(88)90053-5

[2] T. Shirai and Y. Takahashi, "Random Point Fields Associated with Certain Fredholm Determinants I: Fermion, Poisson and Boson Point Processes," J. Funct. Anal., Vol. 205, No. 2, pp. 414–463, 2003. https://doi.org/10.1016/S0022-1236(03)00171-X

[3] S. Matsumoto and M. Wakayama, "Alpha-Determinant Cyclic Modules of $\mathrm{gl}_n(\mathbb{C})$," J. Lie Theory, Vol. 16, No. 2, pp. 393–405, 2006.

[4] D. Vere-Jones, "Alpha-Permanents and Their Applications to Multivariate Gamma, Negative Binomial and Ordinary Binomial Distributions," New Zealand J. Math., Vol. 26, No. 1, pp. 125–149, 1997.

[5] W. Fulton and J. Harris, "Representation Theory: A First Course," Graduate Texts in Mathematics, Vol. 129, Springer, 1991.

[6] K. Kimoto, S. Matsumoto, and M. Wakayama, "Alpha-Determinant Cyclic Modules and Jacobi Polynomials," Trans. Amer. Math. Soc., Vol. 361, No. 12, pp. 6447–6473, 2009.

[7] I. G. Macdonald, "Symmetric Functions and Hall Polynomials," Oxford Univ. Press, UK, 1979.

[8] K. Kimoto and M. Wakayama, "Invariant Theory for Singular α-Determinants," J. Combin. Theory Ser. A, Vol. 115, No. 1, pp. 1–31, 2008. https://doi.org/10.1016/j.jcta.2007.03.008

[9] K. Kimoto and M. Wakayama, "Quantum α-Determinant Cyclic Modules of $\mathcal{U}_q(\mathrm{gl}_n)$," J. Algebra, Vol. 313, No. 2, pp. 922–956, 2007. https://doi.org/10.1016/j.jalgebra.2006.12.015

[10] K. Kimoto, "Quantum Alpha-Determinants and q-Deformed Hypergeometric Polynomials," Int. Math. Res. Not., Vol. 2009, No. 22, pp. 4168–4182, 2009. https://doi.org/10.1093/imrn/rnp083

[11] K. Kimoto, "Quantum α-Determinants and q-Deformations of Hypergeometric Polynomials," RIMS Kokyuroku Bessatsu, Vol. B36, pp. 97–111, 2012.

[12] K. Kimoto, "Zonal Spherical Functions on Symmetric Groups and the Wreath Determinant," RIMS Kokyuroku, Vol. 2031, pp. 218–234, 2017 (in Japanese).

[13] K. Hamamoto, K. Kimoto, H. Tachibana, and M. Wakayama, "Wreath Determinants for Group–Subgroup Pairs," J. Combin. Theory, Ser. A, Vol. 133, pp. 76–96, 2015. https://doi.org/10.1016/j.jcta.2015.02.002

[14] B. Friedman and S. McGuinness, "The Alon–Tarsi Conjecture: A Perspective on the Main Results," Discrete Math., Vol. 342, No. 8, pp. 2234–2253, 2019. https://doi.org/10.1016/j.disc.2019.04.018

[15] K. Kimoto, "The Alon-Tarsi Conjecture on Latin Squares and Zonal Spherical Functions on Symmetric Groups," RIMS Kokyuroku, Vol. 2039, pp. 193–210, 2017 (in Japanese).

[16] K. Kimoto, "Wreath Determinants, Zonal Spherical Functions on Symmetric Groups and the Alon-Tarsi Conjecture," Ryukyu Math. J., Vol. 34, pp. 5–19, 2021.

[17] J. H. van Lint and R. M. Wilson, "A Course in Combinatorics (2nd ed.)," Cambridge University Press, Cambridge, UK, 2001.

[18] C. Reyes-Bustos, "Towards Hash Functions Based on Group–Subgroup Pair Graphs," Mathematical Foundations for Post-Quantum Cryptography, Springer, 2024 (to appear).

[19] C. Reyes-Bustos, "Cayley-type Graphs for Group–Subgroup Pairs," Linear Algebra Appl., Vol. 488, pp. 320–349, 2016. https://doi.org/10.1016/j.laa.2015.09.049

[20] J. Faraut and A. Koranyi, "Analysis on Symmetric Cones," Oxford Mathematical Monographs, Clarendon Press, Oxford, UK, 1995.

[21] T. Shirai, "Remarks on the Positivity of α-Determinants," Kyushu J. Math., Vol. 61, No. 1, pp. 169–189, 2007. https://doi.org/10.2206/kyushujm.61.169

Cid Reyes-Bustos
Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He obtained a Ph.D. in functional mathematics from Kyushu University in 2018. He held the position of specially appointed assistant professor at the Tokyo Institute of Technology from 2019 to 2022. He joined the NTT Institute for Fundamental Mathematics (NTT Communication Science Laboratories) in 2022 as a research associate and started his current position in 2024. His main research interests are number theory, representation theory, mathematical physics, graph theory, and the interaction between these areas.

Masato Wakayama
Research Principal and Head of NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.

He received a Ph.D. in mathematics from Hiroshima University in 1985. Before joining NTT in 2021, he had held various academic positions: an associate professor at Tottori University, visiting fellow at Princeton University, visiting professor at the University of Bologna, distinguished lecturer at Indiana University, professor of mathematics, distinguished professor, dean of the Graduate School of Mathematics, the founding director of the Institute of Mathematics for Industry, executive vice president of Kyushu University, and vice president and professor at Tokyo University of Science. He is now also a professor emeritus of Kyushu University. He specializes in representation theory, number theory, and mathematical physics.

Symmetry and Representation Theory of Lie Groups and Lie Algebras

Ryosuke Nakahama

Abstract

Representations of Lie groups are abstractions of continuous symmetries of linear spaces. By considering their differentials (linear approximations), we obtain representations of Lie algebras. These are regarded as generalizations of the Fourier analysis and important tools not only in mathematics but also in physics. In this article, I present fundamental and important examples on representations of Lie groups and Lie algebras and present some of my latest results.

Keywords: fundamental mathematics, Lie groups, Lie algebras

1. Introduction

Lie groups are abstractions of continuous symmetries of spaces, and linear symmetries are abstracted by representations of Lie groups [1]. These are useful for analyzing functions on spaces with symmetries. However, Lie groups are non-linear objects, and it is not easy to treat their representations directly. To overcome this non-linearity, it is useful to consider the representations of Lie algebras by taking the differentials (linear approximations) of representations of Lie groups. These differentials preserve much information on the original representations and are easier to treat.

2. Representations of Lie groups

First, a Lie group is defined as a subset of the set of all $n \times n$ invertible matrices with complex entries (which is denoted as $GL(n, \mathbb{C})$ and called the general linear group), closed by the products, the inverses, and taking the limits[*1]. For example, $GL(n, \mathbb{C})$,

$$GL(n, \mathbb{R}) = \{n \times n \text{ invertible matrices with real entries}\},$$
$$SL(n, \mathbb{C}) = \{g \in GL(n, \mathbb{C}) \mid \det(g) = 1\},$$
$$O(n) = \{g \in GL(n, \mathbb{R}) \mid g^t g = I_n\}, \text{ and}$$
$$U(n) = \{g \in GL(n, \mathbb{C}) \mid g^t \bar{g} = I_n\}$$

are typical examples of Lie groups (where I_n is the identity matrix). Next, let G be a Lie group, and X be a space such that "convergence can be defined" (i.e., a topological space). If a transformation $\tau(g): X \to X$ is given for each element $g \in G$ and if it satisfies the associative law and the continuity in a suitable sense, we say that G acts on X. For example, rotations of the unit disk (the disk of radius 1) around the origin are regarded as the action of the Lie group $U(1)$. Similarly, conformal transformations of the unit disk, i.e., transformations that preserve angles of two intersecting curves, are almost regarded as the action of the Lie group $SU(1, 1)$ (**Fig. 1**). These examples show that an action of G on X controls the symmetries of X.

When the space $X = V$ on which G acts is a linear space and $\tau(g)$ is a linear map on V (i.e., it preserves additions and scalar multiplications), (τ, V) is called a representation of G. For example, if G acts on X, then G acts automatically on the space of functions on X (e.g., $V = L^2(X) = \{f: X \to \mathbb{C} \mid \int_X |f(x)|^2 dx < \infty\}$: the

*1 More generally, Lie groups are defined as sets equipped with group and manifold structures such that the group operations are differentiable. Groups that are not realizable as closed subgroups of matrices but locally isomorphic to those are also called Lie groups.

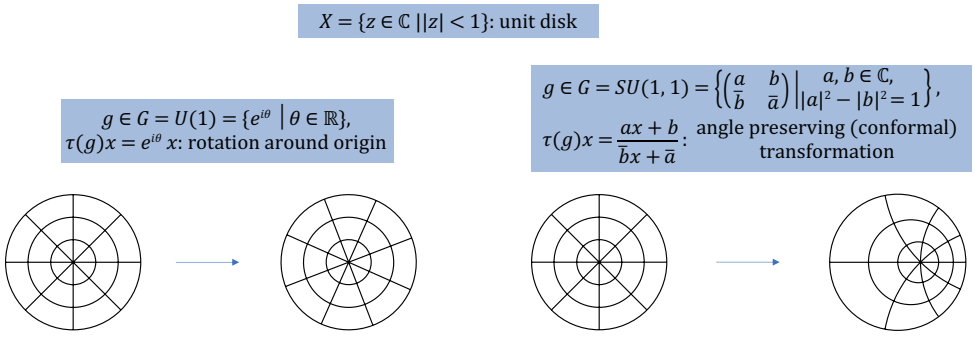


Fig. 1. Actions of $U(1)$, $SU(1,1)$ on unit disk.

space of square integrable functions). This action on $L^2(X)$ is linear and becomes a representation of G. Such representation is in general infinite-dimensional and looks difficult, but in many cases, it consists of a sum of simpler representations. Therefore, to understand function spaces, it is important to understand simpler representations in detail.

The most elementary example of representations of a Lie group $G \subset GL(n, \mathbb{C})$ is $V := \mathbb{C}^n$ (the space of column vectors), with $\tau(g)$ defined by the product of matrices

$$\tau(g)\colon \mathbb{C}^n \to \mathbb{C}^n, \quad \tau(g)v := gv$$

for each $g \in G$. As more non-trivial examples, for a non-negative integer k, the action of the Lie group $G = U(n)$ on the linear space

$$V = \mathcal{P}_k(\mathbb{C}^n) := \{f(x) = f(x_1, \ldots, x_n)\text{: polynomial of } n \text{ variables} \mid f(tx) = t^k f(x)\ (t \in \mathbb{C})\}$$

(the space of homogeneous polynomials of n variables, degree k), with $\tau(g)$ defined by the product of matrices on the variables becomes a representation. Similarly, the action of the Lie group $G = O(n)$ on the linear space

$$V = \mathcal{H}_k(\mathbb{C}^n) := \left\{f(x) \in \mathcal{P}_k(\mathbb{C}^n) \,\middle|\, \sum_{j=1}^n \frac{\partial^2 f}{\partial x_j^2} = 0\right\}$$

(the space of homogeneous harmonic polynomials of n variables, degree k), with $\tau(g)$ defined similarly also becomes a representation. The representation of $U(n)$ on $\mathcal{P}_k(\mathbb{C}^n)$ and that of $O(n)$ on $\mathcal{H}_k(\mathbb{C}^n)$ are examples of irreducible representations. "Irreducible" means that the representation can no longer be decomposed, or there are no linear subspaces $W \subset V$ satisfying $\tau(G)W \subset W$ other than $\{0\}$ and V.

3. Irreducible decompositions of representations—Generalization of Fourier analysis

One of the most fundamental problems in representation theory is to decompose a given representation into a sum of irreducible representations. The irreducible decomposition of the function space (e.g. $L^2(X)$) on X with an action of G is useful for understanding X. For example, $G = O(n)$ acts on the $n-1$-dimensional sphere $S^{n-1} := \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 = 1\}$ by the usual rotations and acts on the function space $L^2(S^{n-1})$ linearly. The space of the restriction of homogeneous harmonic polynomials of degree k on the sphere S^{n-1} (spherical harmonic functions, denoted as $\mathcal{H}_k(\mathbb{C}^n)|_{S^{n-1}} =: \mathcal{H}_k(S^{n-1})$) is then preserved by this action, and becomes a sub-representation. Namely,

> If $f(x) \in \mathcal{H}_k(S^{n-1})$, then $\tau(g)f(x) \in \mathcal{H}_k(S^{n-1})$ for all $g \in O(n)$.

Moreover, each $f(x) \in L^2(S^{n-1})$ is expressed uniquely by the form

$$f(x) = \sum_{k=0}^{\infty} f_k(x), \quad f_k(x) \in \mathcal{H}_k(S^{n-1}).$$

Hence, $L^2(S^{n-1})$ is decomposed into the direct sum

$$L^2(S^{n-1}) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^{n-1}).$$

Since each $\mathcal{H}_k(S^{n-1})$ is irreducible, this gives the irreducible decomposition. When $n = 2$, by the coordinate $(\cos\theta, \sin\theta)$ of S^1, we have

$$\mathcal{H}_k(S^1) = \mathbb{C}e^{ik\theta} + \mathbb{C}e^{-ik\theta} = \mathbb{C}\cos k\theta + \mathbb{C}\sin k\theta = \{ae^{ik\theta} + a'e^{-ik\theta} = (a + a')\cos k\theta + i(a - a')\sin k\theta \mid a, a' \in \mathbb{C}\},$$

and the above decomposition coincides with the

Fourier series expansion

$$f(\cos\theta, \sin\theta) = \sum_{k=-\infty}^{\infty} a_k e^{ik\theta} = a_0 + \sum_{k=1}^{\infty}(b_k \cos k\theta + c_k \sin k\theta)$$

($b_k = a_k + a_{-k}$, $c_k = i(a_k - a_{-k})$). Similarly, the (inverse) Fourier transform

$$f(x) = \int_{\mathbb{R}} \hat{f}(\xi) e^{ix\xi}\, d\xi$$

is regarded as the decomposition of the function space $L^2(\mathbb{R})$ on the real line $\mathbb{R}$ into the sum of one-dimensional sub-representations $\mathbb{C}e^{ix\xi}$ as the representation of the additive group $\mathbb{R}$,

$$L^2(\mathbb{R}) = \int_{\mathbb{R}}^{\oplus} \mathbb{C}e^{ix\xi}\, d\xi.$$

Note that this is a sum of uncountably many spaces and called the direct integral instead of the direct sum. Needless to say, the Fourier analysis is important in many fields such as signal processing, and the theory of spherical harmonics is important in the treatment of rotation-invariant systems in quantum physics. Irreducible decompositions of general representations are regarded as generalizations of these important theories.

Among irreducible decompositions, the decomposition of a representation of G under its subgroup $G' \subset G$ is called the branching law. For example, we consider the restriction of the representation $\mathcal{P}_k(\mathbb{C}^n)$ of $G = U(n)$ to the subgroup

$$G' = \left\{ \begin{pmatrix} g & 0 \\ 0 & 1 \end{pmatrix} \middle| \, g \in U(n-1) \right\} \simeq U(n-1).$$

Then as a representation of $U(n)$, $\mathcal{P}_k(\mathbb{C}^n)$ is irreducible but as a representation of $U(n-1)$, the subspaces

$$\mathcal{P}_m(\mathbb{C}^{n-1})x_n^{k-m} := \{f(x_1, \ldots, x_{n-1})x_n^{k-m} \mid f(x_1, \ldots, x_{n-1}) \in \mathcal{P}_m(\mathbb{C}^{n-1})\} \subset \mathcal{P}_k(\mathbb{C}^n)$$

($m = 0, \ldots, k$) are clearly sub-representations. Therefore, we find that the irreducible decomposition (branching law) of $\mathcal{P}_k(\mathbb{C}^n)$ under $U(n-1)$ is given by

$$\mathcal{P}_k(\mathbb{C}^n)|_{U(n-1)} = \bigoplus_{m=0}^{k} \mathcal{P}_m(\mathbb{C}^{n-1})x_n^{k-m}.$$

Similarly, it is known that the irreducible decomposition of $\mathcal{P}_k(\mathbb{C}^n)$ under $G' = O(n)$ is given by

$$\mathcal{P}_k(\mathbb{C}^n)|_{O(n)} = \bigoplus_{m=0}^{[k/2]} \mathcal{H}_{k-2m}(\mathbb{C}^n)\|x\|^{2m}$$

(where $\|x\|^2 := \sum_{i=1}^n x_i^2$). As a more non-trivial example, we consider the branching law of the representation $\mathcal{H}_k(\mathbb{C}^n)$ of $G = O(n)$ under the subgroup $G' \simeq O(n-1)$. For $m = 0, 1, \ldots, k$, let $\tilde{P}_k^m(y)$ be a polynomial of degree at most $k - m$ satisfying $\tilde{P}_k^m(-y) = (-1)^{k-m}\tilde{P}_k^m(y)$, and we consider the linear space

$$W_m := \left\{ \|x\|^{k-m}\tilde{P}_k^m\left(\frac{x_n}{\|x\|}\right) f(x_1, \ldots, x_{n-1}) \middle| \, f(x_1, \ldots, x_{n-1}) \in \mathcal{H}_m(\mathbb{C}^{n-1}) \right\} \subset \mathcal{P}_k(\mathbb{C}^n).$$

This is then an irreducible representation of $O(n-1)$, and if we suitably choose $\tilde{P}_k^m(y)$, then we can make $W_m \subset \mathcal{H}_k(\mathbb{C}^n)$. In this situation, $\mathcal{H}_k(\mathbb{C}^n)$ is irreducibly decomposed under $O(n-1)$ as

$$\mathcal{H}_k(\mathbb{C}^n)|_{O(n-1)} = \bigoplus_{m=0}^{k} W_m \simeq \bigoplus_{m=0}^{k} \mathcal{H}_m(\mathbb{C}^{n-1}).$$

The polynomial $\tilde{P}_k^m(y)$ is obtained by solving a differential equation concerning the Laplacian and given explicitly by $\tilde{P}_k^m(y) = C_{k-m}^{(m-1+n/2)}(y)$ using the Gegenbauer polynomial $C_k^{(\alpha)}(y)$. When $n = 3$, this is given by a constant multiple of the associated Legendre polynomials. With these polynomials, we can explicitly construct a basis of $\mathcal{H}_k(\mathbb{C}^n)$ inductively on n (**Fig. 2**: $n = 3$).

4. Representations of Lie algebras—Linear approximation of those of Lie groups

Lie groups are generally not linear spaces, and it is not easy to treat their representations directly. To overcome this non-linearity, we consider the Lie algebras associated with the Lie groups instead. For $X \in M(n, \mathbb{C})$ (an $n \times n$ matrix with complex entries) and $t \in \mathbb{R}$, we consider the exponential function $\exp(tX) := \sum_{j=0}^{\infty}(tX)^j/j!$. This satisfies the usual law of exponents $\exp((s+t)X) = \exp(sX)\exp(tX)$ and $\frac{d}{dt}\exp(tX)|_{t=0} = X$. By using this, we define the Lie algebra $Lie(G) \subset M(n, \mathbb{C})$ associated with the Lie group $G \subset GL(n, \mathbb{C})$ by

$$Lie(G) := \{X \in M(n, \mathbb{C}) \mid \exp(tX) \in G \text{ holds for all } t \in \mathbb{R}\}.$$

This then becomes a linear space, and $[X, Y] := XY - YX \in Lie(G)$ holds for all $X, Y \in Lie(G)$. Next, for a finite-dimensional representation (τ, V) of G, we define the representation $(d\tau, V)$ of the Lie algebra $Lie(G)$ by

$$d\tau(X)v := \frac{d}{dt}\tau(\exp(tX))v|_{t=0}$$

$(X \in Lie(G), v \in V)$.

Then $d\tau(aX + bY) = ad\tau(X) + bd\tau(Y)$ and $d\tau([X, Y]) = d\tau(X)d\tau(Y) - d\tau(Y)d\tau(X)$ hold for all $X, Y \in Lie(G)$, $a, b \in \mathbb{R}$. This representation $(d\tau, V)$ preserves most information on the original representation (τ, V). For

$$\mathcal{H}_k(S^2) = \bigoplus_{m=0}^{k} \tilde{P}_k^m(\cos\theta)\,\mathcal{H}_m(S^1) = \bigoplus_{m=0}^{k} (\mathbb{C}\cos m\phi + \mathbb{C}\sin m\phi)\,P_k^m(\cos\theta),$$

$$P_k^m(\cos\theta) = \sin^m\theta\,\tilde{P}_k^m(\cos\theta) = \frac{\sin^m\theta}{2^k}\sum_{j=0}^{[(k-m)/2]} \frac{(-1)^j\,(2k-2j)!}{j!\,(k-j)!\,(k-2j-m)!}\,(\cos\theta)^{k-2j-m}$$

$\mathcal{H}_0(S^2) = \mathbb{C}$

$\mathcal{H}_1(S^2) = \mathbb{C}P_1^0(\cos\theta) + (\mathbb{C}\cos\phi + \mathbb{C}\sin\phi)P_1^1(\cos\theta)$

$\mathcal{H}_2(S^2) = \mathbb{C}P_2^0(\cos\theta) + (\mathbb{C}\cos\phi + \mathbb{C}\sin\phi)P_2^1(\cos\theta) + (\mathbb{C}\cos 2\phi + \mathbb{C}\sin 2\phi)\,P_2^2(\cos\theta)$

$\mathcal{H}_3(S^2) = \mathbb{C}P_3^0(\cos\theta) + (\mathbb{C}\cos\phi + \mathbb{C}\sin\phi)P_3^1(\cos\theta) + (\mathbb{C}\cos 2\phi + \mathbb{C}\sin 2\phi)P_3^2(\cos\theta) + (\mathbb{C}\cos 3\phi + \mathbb{C}\sin 3\phi)P_3^3(\cos\theta)$

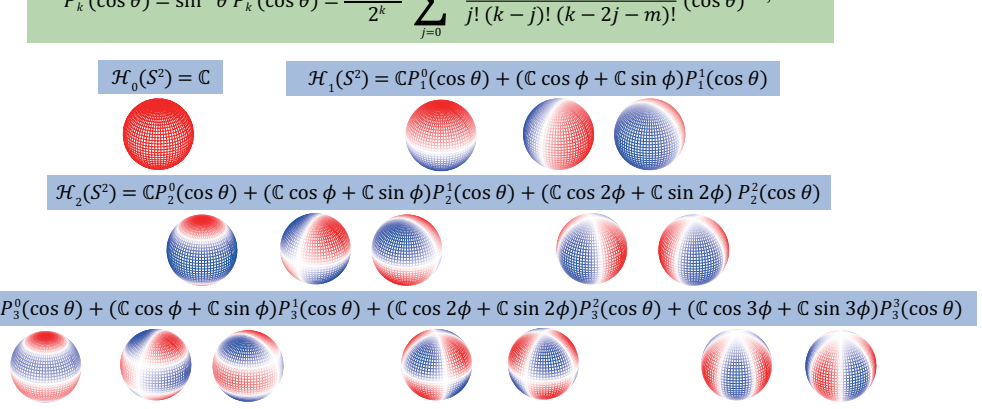


Fig. 2. Express $\mathcal{H}_k(S^2) = \mathcal{H}_k(\mathbb{C}^3)|_{S^2}$ by sum of $\mathcal{H}_m(S^1) = \mathcal{H}_m(\mathbb{C}^2)|_{S^1}$.

$$d\tau(H)f(x,y) = \frac{d}{dt}f\big((x,y)\exp(tH)\big)\Big|_{t=0} = \frac{d}{dt}f\left((x,y)\begin{pmatrix} e^t & 0 \\ 0 & e^{-t}\end{pmatrix}\right)\Big|_{t=0} = \left(x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y}\right)f(x,y),$$

$$d\tau(E)f(x,y) = \frac{d}{dt}f\big((x,y)\exp(tE)\big)\Big|_{t=0} = \frac{d}{dt}f\left((x,y)\begin{pmatrix} 1 & t \\ 0 & 1\end{pmatrix}\right)\Big|_{t=0} = x\frac{\partial}{\partial y}f(x,y),$$

$$d\tau(F)f(x,y) = \frac{d}{dt}f\big((x,y)\exp(tF)\big)\Big|_{t=0} = \frac{d}{dt}f\left((x,y)\begin{pmatrix} 1 & 0 \\ t & 1\end{pmatrix}\right)\Big|_{t=0} = y\frac{\partial}{\partial x}f(x,y).$$

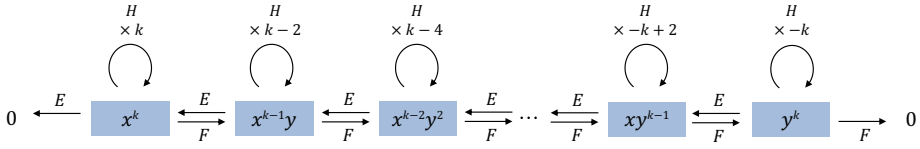


Fig. 3. Representation of Lie algebra of $SU(2)$ on $\mathcal{P}_k(\mathbb{C}^2)$.

example, if G is connected, then the irreducibility of a finite dimensional representation (τ, V) under G is equivalent to the irreducibility of the differential representation $(d\tau, V)$ under $Lie(G)$.

Let us consider the representation of $G = SU(2) := U(2) \cap SL(2, \mathbb{C})$ on $V = \mathcal{P}_k(\mathbb{C}^2)$ as an example. First, the Lie algebra $Lie(G)$ associated with $G = SU(2)$ and its complexification $Lie(G) \otimes \mathbb{C}$ are given by

$$Lie(G) = \mathfrak{su}(2) := \left\{\begin{pmatrix} a & b \\ c & -a\end{pmatrix} \,\middle|\, \begin{matrix} a, b, c \in \mathbb{C}, \\ a = -\bar{a},\ b = -\bar{c}\end{matrix}\right\},$$

$$Lie(G) \otimes \mathbb{C} = \mathfrak{sl}(2, \mathbb{C}) := \left\{\begin{pmatrix} a & b \\ c & -a\end{pmatrix} \,\middle|\, a, b, c \in \mathbb{C}\right\}.$$

We take a basis $H = \begin{pmatrix} 1 & 0 \\ 0 & -1\end{pmatrix}$, $E = \begin{pmatrix} 0 & 1 \\ 0 & 0\end{pmatrix}$, $F = \begin{pmatrix} 0 & 0 \\ 1 & 0\end{pmatrix}$ of $\mathfrak{sl}(2, \mathbb{C})$. Then their actions on $\mathcal{P}_k(\mathbb{C}^2)$ are given by

$$d\tau(H)f(x,y) = \left(x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y}\right)f(x,y),$$

$$d\tau(E)f(x,y) = x\frac{\partial}{\partial y}f(x,y),$$

$$d\tau(F)f(x,y) = y\frac{\partial}{\partial x}f(x,y)$$

(see **Fig. 3** for the intermediate process). In particular, the actions on the basis $\{x^k, x^{k-1}y, x^{k-2}y^2, \ldots, y^k\}$ of $\mathcal{P}_k(\mathbb{C}^2)$ is written as

$$d\tau(H)x^{k-j}y^j = (k - 2j)x^{k-j}y^j,$$
$$d\tau(E)x^{k-j}y^j = jx^{k-j+1}y^{j-1},$$
$$d\tau(F)x^{k-j}y^j = (k - j)x^{k-j-1}y^{j+1}.$$

Hence, $d\tau(H)$ has the eigenvalues $\{k, k-2, k-4, \ldots, -k\}$, $d\tau(E)$ raises the eigenvalue of an eigenvector of $d\tau(H)$ by 2, and $d\tau(F)$ lowers the eigenvalue by 2. This structure is equivalent to those of the spin representations appearing in quantum physics. In fact, a general irreducible representation of $SU(2)$ always has such a structure, especially equivalent to $\mathcal{P}_k(\mathbb{C}^2)$ for a non-negative integer k. As a higher example, we consider the representations of the Lie group $U(n)$. Again, by looking in detail at the action of (the complexification of) the associated Lie algebra for diagonal, upper-triangular, and lower-triangular matrices, we can then show that the set of all irreducible representations of $U(n)$ has a one-to-one correspondence with the set of n-tuples of decreasing integers

$$\{(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{Z}^n \mid \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\}.$$

As we have seen above, representations of Lie algebras are helpful for the classification of representations of Lie groups. It is also known that the dimension of each irreducible representation of $U(n)$ is characterized by the number of combinatoric objects called the semistandard Young tableaux.

5. Example of infinite-dimensional representations

Next, we consider an example of infinite-dimensional representations. Let us consider the Lie group $G = SL(2, \mathbb{R})$. Its Lie algebra is then given by

$$Lie(G) = \mathfrak{sl}(2, \mathbb{R}) := \left\{ \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \middle| \, a, b, c \in \mathbb{R} \right\}.$$

We take the basis $H, E, F \in \mathfrak{sl}(2, \mathbb{R})$ as before and define the action of $\mathfrak{sl}(2, \mathbb{R})$ on (some dense subspace of) $V = L^2(\mathbb{R})$ by

$$d\tau(H)f := x\frac{df}{dx} + \frac{1}{2}f, \quad d\tau(E)f = -\frac{i}{2}x^2 f,$$
$$d\tau(F)f = -\frac{i}{2}\frac{d^2 f}{dx^2}.$$

This does not lift to a representation of the Lie group $SL(2, \mathbb{R})$ but lifts to that of the double covering group $\widetilde{SL}(2, \mathbb{R})$ (that is, there exists a representation $(\tau, L^2(\mathbb{R}))$ of $\widetilde{SL}(2, \mathbb{R})$, the differential of which coincides with the above $d\tau$ (on some dense subspace)). The action of $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \exp\frac{\pi}{2}(E - F) \in \widetilde{SL}(2, \mathbb{R})$ also coincides with a constant multiple of the Fourier transform.

$$\left(\tau\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} f\right)(x) = \left(\tau\left(\exp\frac{\pi}{2}(E - F)\right)\right)$$
$$= \frac{e^{-\pi i/4}}{\sqrt{2\pi}} \int_{\mathbb{R}} f(\xi) e^{-ix\xi} \, d\xi.$$

This is proved by observing the eigenfunctions of $d\tau(E - F) = -\frac{i}{2}\left(x^2 - \frac{d^2}{dx^2}\right)$ (the quantum harmonic oscillator). While the representation $(\tau, L^2(\mathbb{R}))$ is infinite-dimensional, this is nearly irreducible[*2] and called the Weil representation or the metaplectic representation. This construction is generalized to the representation $L^2(\mathbb{R}^n)$ of the larger Lie group $Sp(n, \mathbb{R})$ ($Sp(1, \mathbb{R}) = SL(2, \mathbb{R})$ for $n = 1$ case). By taking the irreducible decomposition of the representation of $Sp(nm, \mathbb{R})$ on $L^2(\mathbb{R}^{nm}) \simeq L^2(M(n, m; \mathbb{R}))$ (the function space on $n \times m$ matrices) under the subgroup $Sp(n, \mathbb{R}) \times O(m) \subset Sp(nm, \mathbb{R})$, we can also obtain various representations of $Sp(n, \mathbb{R})$; thus, this Weil representation is considered quite important in representation theory.

6. Branching law of infinite-dimensional representations

In recent research, I was interested in explicitly determining the branching laws of infinite-dimensional representations (see [2] and references therein). When we restrict a "good" representation (τ, V) of G to a subgroup $G' \subset G$, V is decomposed into a direct sum (or a direct integral) of (in general infinite number of) irreducible representations of G'. Even if we know which representation V' of G' appears abstractly in the decomposition of V, it is generally a difficult problem to determine how V' is included explicitly in V (corresponding to the determination of $\tilde{P}_k^m(y)$ in the example of $\mathcal{H}_k(\mathbb{C}^n)$). I determined the explicit inclusion map (intertwining operator) from V' into V for "good" tuples (G, G', V, V') [3, 4] (**Fig. 4**). There then appear special functions such as hypergeometric functions (and their multivariate generalizations). In the future, I aim to obtain analogous results for more general tuples (G, G', V, V').

*2 The space of all even or odd functions is irreducible, and $L^2(\mathbb{R})$ is a sum of these two irreducible representations.

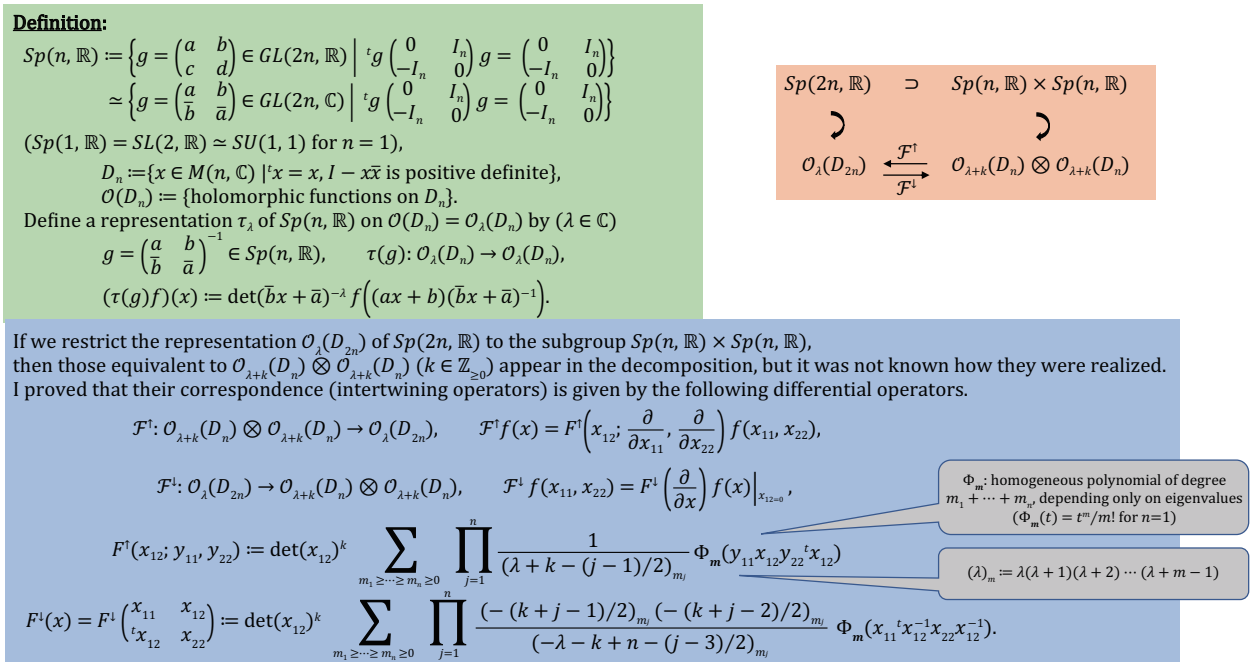


Fig. 4. Example of my results.

References

[1] A. W. Knapp, "Lie Groups Beyond an Introduction," Progr. Math., Vol. 140, Birkhauser Boston, Inc., Boston, MA, USA, 1996.

[2] T. Kobayashi, "A Program for Branching Problems in the Representation Theory of Real Reductive Groups," Representations of Reductive Groups, Progr. Math. Vol. 312, pp. 277–322, Birkhäuser/Springer, Cham, 2015. https://doi.org/10.1007/978-3-319-23443-4_10

[3] R. Nakahama, "Construction of Intertwining Operators between Holomorphic Discrete Series Representations," Symmetry, Integrability and Geometry: Methods and Applications (SIGMA), Vol. 15, 036, 2019. https://doi.org/10.3842/SIGMA.2019.036

[4] R. Nakahama, "Computation of Weighted Bergman Inner Products on Bounded Symmetric Domains and Restriction to Subgroups," SIGMA, Vol. 18, 033, 2022. https://doi.org/10.3842/SIGMA.2022.033

Ryosuke Nakahama
Research Associate, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.
He received a B.S., M.S., and Ph.D. in mathematical sciences from the University of Tokyo in 2011, 2013, and 2016 and joined the NTT Institute for Fundamental Mathematics in 2022. His research interests include the representation theory of Lie groups and mathematical physics. He is a member of the Mathematical Society of Japan.

Modular Forms and Fourier Expansion

Shuji Horinaga

Abstract

Fourier analysis is an indispensable technology, but so is mathematics. In this article, we review the history of modular forms and give an overview of the relationship among representation theory, Fourier analysis, and modular forms. The explanation of difficult terms is confined to footnotes, and we focus on the relationship between the concepts. Finally, we discuss the remaining difficulties in the modern theory of modular forms, challenges, and the author's research.

Keywords: modular forms, Fourier analysis, representation theory

1. History of modular forms and remaining problems

One of the origins of modular forms[*1] is the study of elliptic functions that began in the 1800s when the accuracy of stargazing methods became more precise and astronomy improved rapidly. Since orbits of astronomical objects are generally ellipses, we need to measure the circumference of ellipses. Of course, it is easy to compute the circle's circumference in terms of π, but with ellipses, it is challenging and written as the (second) elliptic integral. Through the studies of Legendre, Gauss, and Abel, elliptic integrals became not only the circumference of ellipses but also interesting objects connecting to modern mathematics. A notable example is the theta function. Theta functions are used in various areas of mathematics (**Fig. 1**). They are typical examples of modular forms, a central theme of this article, and play crucial roles in elliptic curves and number theory. As another application, Kronecker constructed class fields of imaginary quadratic fields. This study is called Kronecker's Jugendtraum and is a significant result relating to a part of Hilbert's 23rd problem[*2]. The study of modular forms started in this way with typical examples.

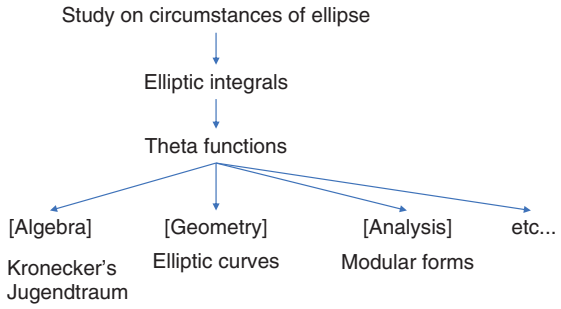


Fig. 1. The study on the circumference of ellipses reveals important interrelated concepts and objects in modern mathematics.

1.1 Developments of the theory of modular forms and obstructions

To further developments in the theory of modular forms, we needed to wait for the study of modular forms of Hecke, a student of Hilbert. Of course, there are many known results for modular forms, but Hecke arranged such results and initiated the theory

*1 Modular forms: Functions with quite strong automorphy. Due to the automorphy, it is highly nontrivial that modular forms exist. The modular forms we define below have several deep and interesting arithmetic properties.

*2 Hilbert's 23rd problem: German mathematician David Hilbert proposed 23 problems in 1900. These problems played a huge role in constructing the basics of modern mathematics. Although it has been more than 100 years since Hilbert's proposal, up to half the problems have been proved.

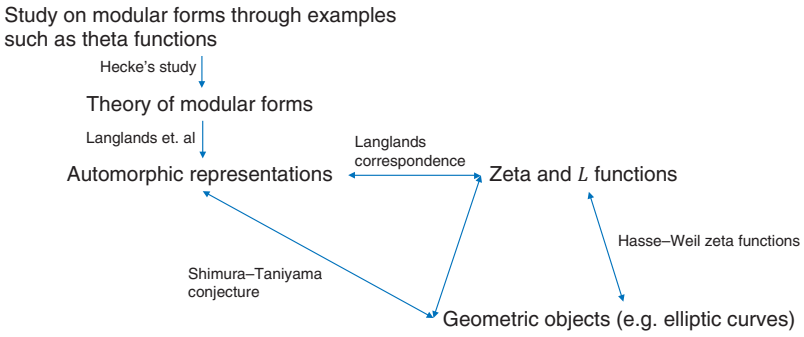


Fig. 2. Modular forms, zeta and L functions, geometric objects.

of modular forms. Hecke defined the zeta functions and L functions for modular forms on the basis of the Riemann zeta function. These results opened the way for the theory of modern modular forms. This theory became the theory of automorphic representations through the works of Langlands and other mathematicians. To review the theory of automorphic representations, the Shimura–Taniyama conjecture is one of the most significant results. It is a profound conjecture that connects automorphic representations and elliptic curves. In 1995, Wiles solved the "semistable" case of the conjecture, proving Fermat's conjecture completely. The Shimura–Taniyama conjecture has now been completely proven. The paramodular conjecture, a generalization of the Shimura–Taniyama conjecture, has been partially proved. To formulate these conjectures, it is necessary to use automorphic representations, which also play a vital role in these conjectures.

We stated that the Shimura–Taniyama conjecture connects geometric objects, such as elliptic curves, and analytic objects such as holomorphic modular forms. Through the pioneering studies of many researchers, for example, Shimura and Langlands, the Shimura–Taniyama conjecture exceeds the original formulation and became a theory to unify algebra, analysis, and geometry (**Fig. 2**). However, there is a remaining problem in the theory of modular forms, i.e., a study of non-holomorphic modular forms. Shimura and Taniyama formulated the conjecture on the basis of many numerical computations of Fourier coefficients of modular forms. However, no known examples of Fourier coefficients of non-holomorphic modular forms exist. The lack of such examples is an obstruction of further development. Recall that we prove the Shimura–Taniyama conjecture and hypersphere packing problem[*3] using holomorphic modular forms. It is easy to imagine that non-holomorphic modular forms have several applications similar to holomorphic modular forms, but there is room for improvement in modular forms. In this article, we first discuss the relationship between Fourier expansion and representation theory then discuss the Langlands conjecture and Arthur conjecture, which may be viewed as a generalization of the relationship, and introduce a joint study with myself and Narita, a professor of Waseda University, about Fourier expansion of non-holomorphic modular forms.

2. Fourier expansion and representation theory

2.1 Fourier expansion

Many people may have heard of Fourier expansion and Fourier transform. These are indispensable techniques for modern society; for example, they are frequently used to process sound signals. No matter how complicated sounds are, we may construct complex sounds using simple ones such as the time signals on television and radio. This point of view is the method of Fourier transform and Fourier expansion. In mathematics, one may regard such simple sounds as $\sin x$ and $\cos x$ functions. In pure mathematics, Fourier expansion means an expansion of periodic functions as $\sin x$ and $\cos x$, and Fourier coefficients are the coefficients in such an expansion. Fourier expansion plays a significant technical and theoretical role in modern mathematics. We first discuss the relationship between Fourier analysis and representation theory.

The philosophy of Fourier expansion and Fourier

*3 Hypersphere packing problem: An analog of the ball-packing problem in 3 dimensions, known as the Kepler conjecture. Viazovska solved the problem for 8 and 24 dimensions and won the Fields medal in 2022.

transform is to understand the space of periodic functions via the translations for periodic functions. To understand this, we discuss the mathematical details. Let f be a complex-valued function on the space of real numbers $\mathbb{R}$. We say that f has the period 1 if $f(x + 1) = f(x)$ for any $x \in \mathbb{R}$. Thus, we may regard f as a function on $\mathbb{R}/\mathbb{Z}$. The theory of Fourier expansion tells us to rewrite f as a sum of sin x and cos x. More precisely, by $e^{2\pi\sqrt{-1}\,ny} = \cos(2\pi nx) + \sqrt{-1}\sin(2\pi nx)$, we have an infinite sum

$$f(x) = \sum_{n=-\infty}^{+\infty} a_n e^{2\pi\sqrt{-1}\,ny}.$$

Such an expression is called Fourier expansion, and coefficients a_n are the Fourier coefficients. It is known that a_n equal $Ff(n)$, where Ff is the Fourier transform.

In representation theory, we divide mathematical objects, such as periodic functions, into smaller objects with more precise conditions such as periodic functions. A typical example of representation theory is the action of matrices on vector spaces. We can easily understand the action of matrices by dividing them into eigenvalues and eigenvectors. Such a framework is one fundamental aspect of representation theory.

Next, we discuss the relationship between Fourier analysis and representation theory. In the above representation theoretical method, we consider the vector spaces and the matrices acting on them as the space of periodic functions and the "translation," respectively. For a periodic function f and real number $y \in \mathbb{R}$, we define the translation r_y by $r_y f(x) = f(x + y)$. Then, the n-th Fourier coefficient of $r_y f$ equals $a_n e^{2\pi\sqrt{-1}\,ny}$. Therefore, the action defined by the translation r_y by y has the function $x \mapsto e^{2\pi\sqrt{-1}\,ny}$ as an eigenfunction and $e^{2\pi\sqrt{-1}\,ny}$ as the eigenvalue of it. Summarizing thus far, by combining the period function, Fourier transform, and the translation, we conclude that the following two objects relate:

- Representations on $\mathbb{R}/\mathbb{Z}$ defined by $y \mapsto e^{2\pi\sqrt{-1}\,ny}$
- n-th term of Fourier expansion of period functions

We thus grasp one face of representation theory by connecting a function and representations, a mysterious object. We may find this a surprising correspondence in a broad framework rather than an easy object $\mathbb{R}/\mathbb{Z}$. The central theme of the next section is a generalization of such surprising correspondence.

2.2 From Fourier analysis to Langlands conjecture

We deeply observe the relationship between functions and representations for $\mathbb{R}/\mathbb{Z}$ and period functions. A fundamental property is the "compactness" of $\mathbb{R}/\mathbb{Z}$. For a non-compact object such as the real numbers $\mathbb{R}$, such correspondence becomes more difficult due to technical difficulties, for example, a convergence of integrals. We may find differences between Fourier analysis of period functions and a function on $\mathbb{R}$ in certain literature on Fourier analysis. The nature of these differences comes from the topological property of $\mathbb{R}$ and $\mathbb{R}/\mathbb{Z}$, i.e., the non-compactness of $\mathbb{R}$.

Harish-Chandra produced a breakthrough in the representation theory of reductive groups, one of the most essential classes of non-compact groups. He mainly considered the Lie groups, containing $\mathbb{R}$ and $\mathbb{R}/\mathbb{Z}$. His pioneering work is the classification of discrete series representations. Recall that we consider the space of functions for $\mathbb{R}$ and $\mathbb{R}/\mathbb{Z}$. For reductive groups G, he considered the space $L^2(G)$ of square-integrable functions[*4] and translations on it. We may naturally find discrete series representations in $L^2(G)$. A realization of discrete series representations on $L^2(G)$ is done by matrix coefficients[*5]. Like this, we highly develop the representation theory through a space of certain functions and analysis. On the basis of Harish-Chandra's study, Knapp and Zuckerman classified the tempered representations, and Langlands classified all the irreducible representations of Lie groups (**Table 1**). This classification is due to Langlands and is called the Langlands classification. Since Lie groups are a theory for $\mathbb{R}$ or complex numbers $\mathbb{C}$, number theorists need a similar theory for p-adic groups.

On the basis of various trials and errors, the local Langlands conjecture, the classification theory of irreducible representations on p-adic groups, was becoming clear. The local Langlands conjecture states the correspondence of the following two objects for a connected reductive group G[*6]:

$$\{\text{Irreducible representations of } G\} \to \{L \text{ parameters of } G\}.$$

*4 Square-integrable function: A function f on G such that $\int_G |f(g)|^2\, dg < +\infty$.

*5 Matrix coefficients: A representation ρ is a homomorphism ρ of a group to the group of matrices, possibly infinitely columns and rows. An entry of a matrix in the image of ρ is called the matrix coefficient.

*6 Connected reductive group: For an algebraic group, we mean a group and algebraic variety. Connected is the connectedness as an algebraic variety, and reductive is a class of groups. For example, general linear, orthogonal, and unitary groups are connected reductive groups, but upper unipotent groups are not reductive.

Table 1. Classification and construction of representations.

	Classification	Properties of matrix coefficients	Construction
Discrete series	Harish-Chandra	Square-integrable	Realization on L^2 space via matrix coefficients
Tempered representations	Knapp–Zuckerman	Tempered	Parabolically induced representation
Unitary representations	Unknown	Definable	Unknown in general
Irreducible representations	Langlands	Non-definable	Langlands quotient of parabolically induced representations

L parameters on the right side are arithmetic objects and define an L function.

As in another article [1] in this issue, one aspect of the local Langlands conjecture is a non-commutative class field theory. Such an aspect appears in the L parameters. The local Langlands conjecture has made significant progress and become more precise. It is now called the endoscopic classification.

The representation theory of p-adic groups and modular forms or automorphic representations are inseparable. They have a history of developing together while compensating for each other's weaknesses. Finally, we discuss the author's results for the Fourier expansion of non-holomorphic modular forms.

3. Modular forms and representation theory

3.1 Fourier expansion of holomorphic modular forms

We first consider the Fourier expansion and coefficients of holomorphic modular forms. Let $\mathfrak{H}$ be the upper half plane[*7] and $\mathrm{SL}_2(\mathbb{R})$ be the special linear group of degree two[*8]. The group $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathfrak{H}$ by the linear fractional transformation[*9]. Let $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ be the subgroup of $\mathrm{SL}_2(\mathbb{R})$ with integer entries. Set j to be the factor of automorphy[*10]. Take an integer k and holomorphic function f on $\mathfrak{H}$. We say that f is a modular form[*11] of weight k with respect to Γ if $f(\gamma(z)) = j(\gamma, z)^k f(z)$ for any $z \in \mathfrak{H}$ and $\gamma \in \Gamma$. Thus, modular forms are not entirely invariant under Γ other than $k = 0$ but is invariant under Γ with certain modified factors due to k and j. In particular, one would obtain $f(z + 1) = f(z)$ for a modular form f. Since f is holomorphic, one obtains the following Fourier expansion by Cauchy's integral formula:

$$f(x + \sqrt{-1}y) = \sum_{n=-\infty}^{+\infty} a_n e^{2\pi\sqrt{-1}\, n(x+\sqrt{-1}\, y)}.$$

Surprisingly, a_n are independent of the imaginary part y under this expression. This expression is usually called the Fourier expansion of f, and a_n are called the Fourier coefficients of f.

3.2 Modular forms and representation theory

We observed a strong relationship between translations on function spaces and representation theory. Similar phenomena may occur for modular forms. More precisely, we may lift modular forms to functions on a Lie group. Let φ_f be the lift of f. We then may regard φ_f as a function on $C^\infty(\Gamma\backslash \mathrm{SL}_2(\mathbb{R}))$. As we have seen in the section discussing the Fourier expansion, one can define the right translation by $\mathrm{SL}_2(\mathbb{R})$ on the space $\Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. Thus, modular forms and representation theory of $\mathrm{SL}_2(\mathbb{R})$ relate. With a similar method, modular forms would become a function φ_f on an adele group $\mathrm{SL}_2(\mathbb{A})$ using the adele ring $\mathbb{A}$ of $\mathbb{Q}$. If f is square-integrable or more strongly f is a cusp form, φ_f is a function on $L^2(\mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A}))$. We summarize that from a square-integrable modular form f, the function φ_f becomes a function on $L^2(\mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A}))$ and relates representations of $\mathrm{SL}_2(\mathbb{R})$ and $\mathrm{SL}_2(\mathbb{Q}_p)$. This phenomenon resembles Fourier analysis and representation theory (**Fig. 3**).

In the modern modular form theory, we generalize SL_2 in $L^2(\mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A}))$ to a connected reductive group. Like Harish-Chandra's study on discrete series, we may consider the discrete spectrum of $L^2(\mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A}))$. A recent study gives us a description of such a discrete spectrum. This study is based on the research of many researchers. Arthur, a student of

*7 Upper half plane: The set of complex numbers with positive imaginary part.

*8 Special linear group of degree two: The group of invertible real 2 × 2 matrices with determinant one.

*9 Linear fractional transformation: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z) = \dfrac{az+b}{cz+d}$, $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$

*10 Factor of automorphy: For $z \in \mathfrak{H}$ and $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$, we put $j(\gamma, z) = cz + d$.

*11 Strictly speaking, this definition states that the function is a weak modular form. For a weak modular form f, we say that f is a modular form if the Fourier coefficient a_n defined below is zero for $n < 0$.

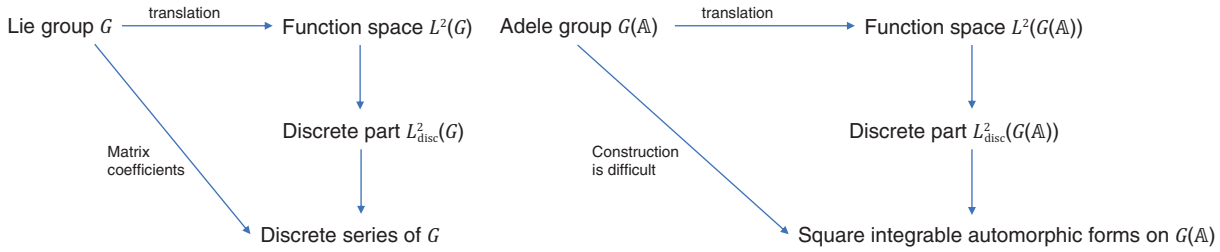


Fig. 3. Comparison of local and global.

Langlands, established the Arthur conjecture, which describes the discrete spectrum and proves his conjecture for orthogonal and symplectic groups under appropriate modification. This study is one of the highest peaks in modern theory of modular forms. Many researchers now consider generalizations and applications of Arthur's study.

3.3 Toward the generalization of Fourier expansion of modular forms

We saw that the Fourier coefficients of holomorphic modular forms f are constant. This fact is based on the holomorphy of f. Thus, if we remove the holomorphy assumption, the Fourier expansion of f is expressed as

$$f(x+\sqrt{-1}y)=\sum_{n=-\infty}^{+\infty}a_n(y)e^{2\pi\sqrt{-1}\,n(x+\sqrt{-1}\,y)}.$$

The coefficients $a_n(y)$ depend on the imaginary part y. Therefore, it is not easy to consider $a_n(y)$. In the joint study with Narita [2], we treat such Fourier expansion of non-holomorphic modular forms. A typical example of a non-holomorphic modular form is a Maass form, but we do not treat Maass forms. Our main target is a modular form naturally arising from representation theory*12. Discrete series representations are a key idea in our joint study. We recall Harish-Chandra's classification of discrete series representations to understand the idea. The modular form on the upper half plane corresponds to a function on $\mathrm{SL}_2(\mathbb{R})$. In a certain sense, there is essentially only one discrete series representation of $\mathrm{SL}_2(\mathbb{R})$. One of the difficulties of Maass form is that the corresponding representation is not a discrete series representation. An example of a group with non-holomorphic discrete series representations is the symplectic group $\mathrm{Sp}_4(\mathbb{R})$ of degree two. For $\mathrm{Sp}_4(\mathbb{R})$, there are essentially two discrete series representations in a certain sense. One is holomorphic and the other one is non-holomorphic. Also, $\mathrm{Sp}_4(\mathbb{R})$ is a minimal with such a property. We may define a modular form on $\mathrm{Sp}_4(\mathbb{R})$. Such a modular form is a function on the Siegel upper half plane $\mathfrak{H}_2$ of degree two*13 satisfying the Fourier expansion:

$$f(x+\sqrt{-1}y)=\sum_{h\in\mathrm{Mat}_2(\mathbb{Q}),\ {}^th=h}a_h(y)e^{2\pi\sqrt{-1}\,\mathrm{tr}(hx)},$$

$$x+\sqrt{-1}y\in\mathfrak{H}_2.$$

The $a_h(y)$ are called the generalized Whittaker function. Unlike holomorphic modular forms, $a_h(y)$ are never a constant. We can introduce a differential equation to evaluate $a_h(y)$ when considering the discrete series representations. In our joint study, we explicitly compute the solution of the differential equation and prove several properties of $a_h(y)$. As an application, we explicitly describe the space of all the non-cuspidal automorphic forms, generating discrete series representations of $\mathrm{Sp}_4(\mathbb{R})$. As in **Fig. 4**, this joint study is the first to describe all the non-cuspidal modular forms, including their construction. In our next joint study, we will consider an explicit computation of Fourier coefficients. We will extend our research to provide an arithmetic property of L function through explicit computation of modular forms corresponding to discrete series representations.

*12 Modular forms are, of course, related to a function on groups. If a modular form relates to a representation σ, we say that a modular form generates σ.

*13 Siegel upper half plane $\mathfrak{H}_2$ of degree two:
$\mathfrak{H}_2=\{z\in\begin{pmatrix}a & b\\ b & c\end{pmatrix}\in\mathrm{Mat}_2(\mathbb{C})\mid \mathrm{Im}(z)\text{ is positive definite.}\}$

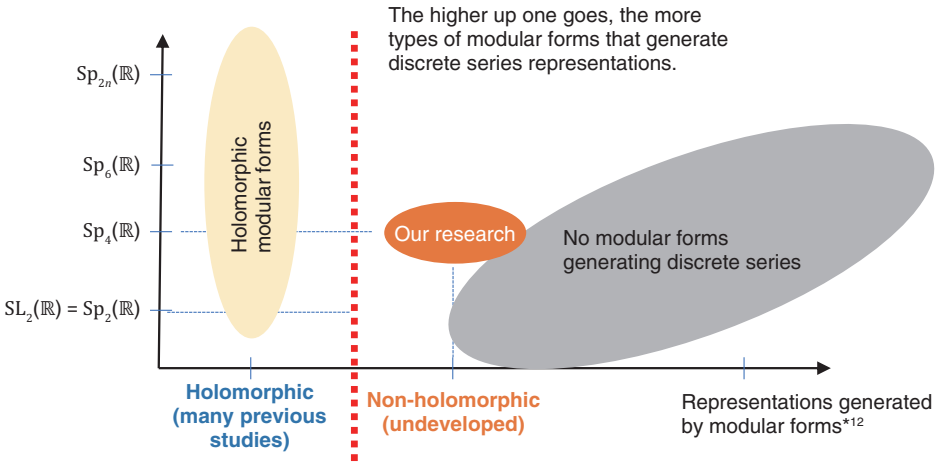


Fig. 4. Current status of mine and Narita's.

References

[1] K. Sano, H. Miyazaki, and M. Wakayama, "A Mathematical World Woven by Number Theory, Algebraic Geometry, and Representation Theory," NTT Technical Review, Vol. 22, No. 9, pp. 16–25, Sept. 2024. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202409fa1.html

[2] S. Horinaga and H. Narita, "Cuspidal Components of Siegel Modular Forms for Large Discrete Series Representations of $Sp_4(\mathbb{R})$," manuscripta math., Vol. 174, pp. 159–202, 2024. https://doi.org/10.1007/s00229-023-01513-3

Shuji Horinaga
Research Associate, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.
He received a B.S. in 2015, an M.S. in 2017, and a Ph.D. in science in 2020, all from Kyoto University. In 2022, he joined the NTT Institute for Fundamental Mathematics and studied modular forms and automorphic representations. He is a member of the Mathematical Society of Japan.

Light-matter Interaction and Zeta Functions

Cid Reyes-Bustos and Masato Wakayama

Abstract

Knowledge of the partition and spectral zeta function of a quantum system is fundamental for both physics and mathematics, and the positions these functions occupy in their respective fields share a common philosophy. In this article, we describe the number theoretic structures hidden behind light and matter interaction models, focusing on the partition function and special values of the spectral zeta function, highlighting how modern mathematical research is involved.

Keywords: quantum Rabi model, non-commutative harmonic oscillator, partition function

1. Introduction

The historical origin of mathematics was the process of counting. Two fundamental results of early mathematics, believed to be discovered at the Pythagorean school and documented by Euclid around 2500 years ago, are 1) the number of prime numbers is infinite, and 2) any natural number can be uniquely factorized into prime factors. One of the most fascinating aspects of number theory is the stark contrast between the simplicity of the integers and the complex and seemingly irregular distribution of prime numbers. The distribution of prime numbers is still a subject of research and takes a concrete form in the Riemann hypothesis, the most well-known open problem in mathematics, still unsolved after 165 years.

In the modern world, the development of quantum information technologies requires the understanding and control of light and matter interactions. The most fundamental theoretical model of this type of quantum interaction is the quantum Rabi model (QRM) [1]. In applications, the systems are always subject to the passing of time, so it is fundamental to understand the time evolution of the system, mathematically controlled by the partition function and heat kernel of the system. Informally speaking, the essence of the partition function is to allow the discernment of properties of ensembles of particles (macro level) having independent states (micro level). More concretely, the partition function is defined as the sum of certain weighted values depending on all the possible states of the system.

Similarly, the Riemann zeta function is also defined as the product of geometric series defined for all prime numbers, each prime number existing independently from the others. The Riemann zeta function enables us to perfectly understand the distribution of the prime numbers, impossible only looking at the individual prime numbers, in the Riemann hypothesis. Surprisingly, physics and number theory share a similar philosophy, and the connections do not end here. For instance, the partition function of the quantum harmonic oscillator and Riemann zeta function are explicitly connected. It is generally equivalent to consider the partition function and the spectral zeta function (the Dirichlet series associated with the eigenvalues) of a quantized physical system. The mathematical theory that bridges the two worlds is representation theory, historically developed alongside relativity theory and quantum mechanics.

In this article, we give an overview of the theory of the partition function and spectral zeta function of quantum interaction models with the hope that the reader will discover and appreciate the bonds between quantum physics and number theory. Hereafter, we denote the ring of integers, field of rational numbers, field of real numbers, and field of complex

numbers as $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, respectively.

2. Special values of the Riemann zeta function and automorphic forms

The harmonic series, the sum of the reciprocals of all positive integers, was known to diverge since the Middle Ages, with an ingenious proof by Nicole Oresme in the 14th century. What about the sum of the reciprocal of the squares? This question, later known as the Basel[*1] problem, was posed by the Bolognese mathematician Pietro Mengoli in 1644. The problem remained unsolved for almost 90 years, until Leonhard Euler discovered in 1735 that it converges to the exact value $\frac{\pi^2}{6}$. At the time, it was a surprise that the irrational number π appeared in the answer.

These results on the harmonic series may be written as $\zeta(1) = +\infty$, and $\zeta(2) = \frac{\pi^2}{6}$, where $\zeta(s)$ is the Riemann zeta function (called *Riemann zeta*), defined by

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p=2,3,5,7,\ldots\,(primes)} \frac{1}{1-p^{-s}}.$$

The middle equality connecting the series and the Euler product manifest the fact that any integer can be factored uniquely into prime factors. The domain of absolute convergence of the series and the infinite product is the half-plane $\Re(s) > 1$, and it is known that the Riemann zeta $\zeta(s)$ can be extended analytically into a meromorphic function defined in the whole complex plane with a unique simple pole[*2] at $s = 1$. Euler solved the Basel problem by comparing infinite product expansion of $\sin(\pi x)$ with the Taylor series of second order and discovered that the values of the zeta function $\zeta(2n)$ at the even integers are given by

$$\zeta(2n) = \frac{(-1)^{n+1}(2\pi)^{2n} B_{2n}}{2(2n)!},$$

where B_n are the Bernoulli numbers defined by the generating series $\frac{x}{e^x - 1} =: \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n$.

In contrast, the passing of time has not illuminated the question of irrationality (or rationality) of the special values at odd integers. The first nontrivial odd value, $\zeta(3)$, had to wait until 1979 to be shown to be irrational by Apéry [2]. Apéry defined mysterious sequences of numbers, now called Apéry numbers, and used them in an inventive way to prove that $\zeta(2)$ and $\zeta(3)$ are irrational numbers.

Even today, not much else is known about the properties of the remaining odd special values. At the turn of the 21st century, Rivoal proved that the sequence $\zeta(2n + 1)(n = 2, 3, \ldots)$ contains infinite number of irrational numbers, and that there is at least one irrational among the numbers $\zeta(5)$, $\zeta(7)$, $\zeta(9)$, ..., $\zeta(21)$. Shortly after, in 2001 Zudilin improved the result to show that the numbers $\zeta(5)$, $\zeta(7)$, $\zeta(9)$, $\zeta(11)$ contain at least one irrational. This is the present state of knowledge about this question, at least here on planet Earth.

2.1 The prime number theorem

The prime number theorem, a result describing the distribution of prime numbers, was conjectured by Carl Friedrich Gauss and Adrien-Marie Legendre in the 18th century and proved independently by Charles de la Vallé Poussin and Jacques Hadamard in 1896 using the ideas introduced by Bernhard Riemann in his seminal work in number theory.

If $\pi(x)$ is the function describing the number of primes less than $x(> 0)$, then the prime number theorem is precisely stated as

$$\pi(x) \sim Li(x) := \int_2^x \frac{dy}{\ln(y)} \sim \frac{x}{\ln x}.$$

Here, $f(x) \sim g(x)$ means that the limit $f(x)/g(x) \to 1$ holds as $x \to \infty$. At the heart of the proof is Riemann's idea that $\zeta(s) \neq 0$ for $\Re(s) = 1$.

The revolutionary contribution of Riemann of recognizing that the seemingly random distribution of prime numbers is intimately related to the analytical properties of $\zeta(s)$ may even be a greater achievement that a future proof of the Riemann hypothesis itself.

2.2 Functional equation of the Riemann zeta function

One of the main features of $\zeta(s)$ is the functional equation. Let $\Gamma(s)$ be the gamma function, and set $\tilde{\zeta}(s) := \pi^{-s/2}\Gamma(s/2)\zeta(s)$, then the functional equation is

$$\tilde{\zeta}(1 - s) = \tilde{\zeta}(s).$$

The essential idea of the functional equation was discovered by Euler in its computations aimed to assign values to divergent series, including

$$\text{“}1 + 8 + 27 + 64 + 127 + \cdots\text{”} = \frac{1}{120}, \textit{ etc.}$$

The computational results obtained by Euler are correct even though the concept of analytical continuation (or even complex function theory) did not exist at the time.

Let us give an outline of the proof of the functional equation to introduce some of the ideas used later for

*1 Basel is the birthplace of Leonhard Euler (1707–1783).

*2 We might regard a pole as a situation similar to when the denominator of a fraction becomes 0.

the partition function. To avoid technical complications, we assume that all series and integrals converge and behave in a reasonable manner.

The Mellin transform $\mathcal{M}f$ of a function f is defined by

$$\mathcal{M}f(s) := \int_0^\infty f(t)t^{s-1}\,dt.$$

A fundamental example is given by $f(t) = e^{-nt}$ with $t > 0$ with Mellin transform $\mathcal{M}f(s) = n^{-s}\Gamma(s)$, verified directly from the definition of $\Gamma(s)$. Similarly, for a series $g(z) = \sum_{n=0}^{\infty} a_n z^n$, the Mellin transform of $h(t) = g(e^{-t})$ is easily verified to be $\sum_{n=0}^{\infty} a_n n^{-s} = \Gamma(s)^{-1}\mathcal{M}h(s)$. The main point to note here is that the Mellin transform relates the series of exponential with the Dirichlet series.

Let us define the series $\theta(z)$ in the upper-half complex plane $\mathbb{H} := \{z \in \mathbb{C}\ \Im(s) > 0\}$ as

$$\theta(z) := \sum_{n\in\mathbb{Z}} e^{i\pi n^2 z}.$$

From the foregoing discussion, we verify that by setting $h(t) = \frac{1}{2}\theta(it)$, we obtain $\tilde{\zeta}(s) = \mathcal{M}f(s/2)$. The function $\theta(t)$ is an automorphic form called the Jacobi theta function.

Automorphic functions (resp. forms) are functions that are invariant (resp. almost invariant) under certain actions of non-commutative groups. Trigonometric functions are well-known to be invariant under translations (i.e. are periodic functions), in other words, they are invariant under the action of the abelian group $\mathbb{Z}$. Thus, in this sense automorphic functions may be thought as non-commutative versions of trigonometric functions.

For $z \in \mathbb{H}$, in addition to the translation invariance $\theta(z+2) = \theta(z)$, the theta function satisfies the relation $\theta(-1/z) = \sqrt{-iz}\theta(z)$. If $\hat{f}$ is the Fourier transform of a function f, then by the Poisson summation formula

$$\sum_{m\in\mathbb{Z}} \hat{f}(m) = \sum_{n\in\mathbb{Z}} f(n)$$

and replacing by the rapidly decreasing function $f_t(x) = e^{-\pi t x^2}$, we obtain the desired relation. Finally, for $z = it\,(t > 0)$, we obtain $\frac{1}{t}\,\theta\left(\frac{i}{t}\right) = \theta(it)$ and applying the Mellin transform to both sides we obtain the functional equation for $\tilde{\zeta}(s)$.

Let us give an interpretation of the Poisson summation used above. Similar to the idea of the hands on a clock[*3], we consider two real numbers to be equivalent if they have the same fractional part and write this set as $\mathbb{Z}\backslash\mathbb{R}$. With this in mind, the right side of the Poisson summation formula is the sum over the lengths of a circumference (i.e., number of turns), and the left side is the sum over the irreducible representations that appear in the Fourier transform, that is, representations $x \mapsto e^{2\pi i y x}$ of the abelian group $\mathbb{R}$ that are trivial on $\mathbb{Z}$. In other words, we might think of it as a sum over all $y = m \in \mathbb{Z}$. The left side may also be interpreted as the sum over the eigenvalues of the Laplacian $\Delta = -\frac{d^2}{dx^2}$ of $\mathbb{R}$. The extension of this idea for non-commutative groups is the celebrated Selberg trace formula[*4].

2.3 Modular and automorphic forms

Let $SL_2(\mathbb{Z})$ (respectively $SL_2(\mathbb{R})$), be the group of 2 × 2 matrices with integer (respectively real) entries and determinant 1. The group $SL_2(\mathbb{Z})$ is generated[*5] by matrices

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Now, $g = \begin{pmatrix} a & b \\ c & a \end{pmatrix}$ acts on the upper half-plane by the linear fractional transformation $z \mapsto g.\,z := \frac{az+b}{cz+d}$ for $z \in \mathbb{H}$. This action preserves the Poincaré (hyperbolic) metric on $\mathbb{H}$ induced by y^{-2} dxdy. Since the matrices $\pm\mathbf{1}$ ($\mathbf{1}$ is the identity matrix) define the same action, we consider the (projective) modular group $\Gamma = PSL_2(\mathbb{Z}) := SL_2(\mathbb{Z})/\pm\mathbf{1}$ and the space $\Gamma\backslash\mathbb{H}$ of points of $\mathbb{H}$ that are not equivalent under the action of Γ.

This is the type of stage where automorphic forms (modular forms) reside. Note that in this case we do not make a distinction between $\pm\mathbf{1}$. The theta function $\theta(z)$ above, is almost invariant under the action of the subgroup $\Gamma(2)\,(\cong \Gamma_0(4))$ of $SL_2(\mathbb{Z})$ generated by S and T^2, and is therefore called a $\Gamma(2)$-automorphic form.

3. Quantum interaction models

3.1 QRM

In quantum optics, the QRM is the most fundamental model to describe light-matter interaction. Its Hamiltonian is given by

$$H_{\text{Rabi}} := a^\dagger a + \Delta\sigma_z + g\sigma_x(a^\dagger + a).$$

Here, $a^\dagger = \frac{1}{\sqrt{2}}\left(x - \frac{d}{dx}\right)$, $a = \frac{1}{\sqrt{2}}\left(x + \frac{d}{dx}\right)$ are the creation and annihilation operators for the quantum harmonic

*3 For instance, adding 8 hours at 20:00 (8 PM), we have 4:00 (4 AM) of the next day and not 28:00 hours of the same day, and similarly with minutes.

*4 For commutative groups, the degree of irreducible representations on a complex vector space is always 1, and the trace (or index) of the representation is the representation itself.

*5 In other words, any matrix in $SL_2(\mathbb{Z})$ is written as a finite product of S and T.

Table 1. Number-theoretical properties of the spectral zeta function.

	Spectral zeta function of NCHO			Riemann zeta function			
	$\zeta_Q(2)$	$\zeta_Q(3)$	$\zeta_Q(n)$	$\zeta(2)$	$\zeta(3)$	$\zeta(2n)$	$\zeta(2n+1)$
Special values (positive integers)	Elliptic integrals (Hypergeometric)	Integral of algebraic functions	Sums of integrals of algebraic functions	$\frac{\pi^2}{6}$	Irrational	Benoulli number $\times \pi^{2n}$	Unknown
Geometric period	Picard-Fuchs ODE of family of elliptic curves with $\Gamma(2)$-torsion	?	?	Picard-Fuchs ODE of family of elliptic curves with $\Gamma_1(5)$-torsion	Picard-Fuchs ODE of K3-surfaces	Not considered	Not considered
Apéry(-like) numbers	✓	✓	Defined from a part of anomaly	✓	✓	Undefined	Undefined
i) Binomial expression	✓			✓	✓	Undefined	Undefined
ii) p-ary congruence relation	✓			✓	✓	Undefined	Undefined
iii) Hierarchy of recurrence relations	✓			Unknown			
iv) Modular interpretation of generating functions	$\Gamma(2)$-modular forms	?	Eichler forms for $n = 4$	$\Gamma_1(5)$-modular forms			
v) Metagenerating functions	Modular Mahler Measure expression			Unknown			
Special values (negative integers)	$0 \quad (-2n)$ NC Bernoulli number $(-(2n+1))$			$0 \quad (-2n)$ Bernoulli number $(-(2n+1))$			

oscillator (bosonic mode, photon or "light") with angular frequency $\omega(=1)$, the matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

are the Pauli matrices for a two-level system (particle, qubit or "matter"), $g > 0$ is the coupling strength between the two-level system and the photon, and $2\Delta > 0$ is the energy difference between the levels of the two-level system. We may assume that the Hamiltonian H_{Rabi} acts on the Hilbert space $L^2(\mathbb{R}) \otimes \mathbb{C}^2$ of integrable two-dimensional vector-valued functions.

The addition of a bias term to the Hamiltonian H_{Rabi} results in the model $H_{\text{Rabi}}^{\epsilon} := H_{\text{Rabi}} + \epsilon\sigma_x$ ($\epsilon \in \mathbb{R}$), called the asymmetric quantum Rabi model (AQRM). The AQRM appears naturally in the experimental realization of deep strong coupling accomplished using the theory of cavity quantum electrodynamics [3].

3.2 Non-commutative harmonic oscillator

The non-commutative harmonic oscillator (NCHO) [4, 5] is defined as a system of ordinary differential equations having a Hamiltonian

$$Q := \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix}\left(-\frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{2}x^2\right) + \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}\left(x\frac{d}{dx} + \frac{1}{2}\right) \quad (\alpha, \beta \in \mathbb{R})$$

acting on $L^2(\mathbb{R}) \otimes \mathbb{C}^2$. When the parameters $\alpha, \beta > 0$ satisfy $\alpha\beta > 1$, Q is a self-adjoint operator having only the discrete spectrum $(0 <)\lambda_0 \le \lambda_1 \le \cdots \le \lambda_n \le \cdots \uparrow \infty$ with multiplicity of at most 2.

The spectral zeta function $\zeta_Q(s)$ of the NCHO is given by

$$\zeta_Q(s) := \sum_{n=0}^{\infty} \lambda_n^{-s} \quad (\Re s > 1).$$

Note that when $\alpha = \beta$, Q is unitarily equivalent to a couple of harmonic oscillators; thus, $\zeta_Q(s) = 2\sqrt{\alpha^2 - 1}\,\zeta(s)$. The spectral zeta function $\zeta_Q(s)$ has many interesting and fascinating properties. For instance, $\zeta_Q(s)$ can be analytically continued to the complex plane with a unique simple pole at $s = 1$ and similarly to $\zeta(s)$, $\zeta_Q(s)$ has "trivial zeros" at negative even integers [6].

It is also possible to define analogs of the Apéry numbers from the special values $\zeta_Q(2)$, $\zeta_Q(3)$, $\zeta_Q(4)$ that unveil a rich mathematical structure. For instance, for $\zeta_Q(2)$, there are explicit relations with automorphic forms and elliptic curves (**Table 1**). For $\zeta_Q(4)$, one has to venture beyond the usual modular forms and consider natural extensions of Eichler forms (given by generalized Abel integrals) [7] associated with a new cohomology [8, 9]. In the explicit description of the Apéry-like numbers one also encounters integrals of generalized Eisenstein series [10], deeply related to the research started by Shimura

in 1982 on holomorphic modular forms for one variable [11].

3.3 Covering models

One of the motivations for the introduction of the NCHO was to slightly weaken the symmetries imposed on the quantum harmonic oscillator to rise above the Gauss hypergeometric functions appearing in classic representation theory and to consider the resulting spectral zeta function as an extension of $\zeta(s)$.

In practice, by using representation theory, we can see that the eigenvalue problem of the NCHO does goes beyond the Gaussian hypergeometric differential equations[*6] and corresponds to the existence of holomorphic solutions of a Heun ordinary differential equation (ODE), with four singular points $\{0, 1, \alpha\beta, \infty\}$, in a region containing $\{0, 1\}$ but not $\alpha\beta$ [12, 13]. By joining the singular point $\alpha\beta$ and ∞ via a confluence process, we obtain a confluent Heun ODE, which corresponds directly to the eigenvalue problem of the QRM.

We may thus say that the NCHO is a covering model of the QRM by looking at the corresponding Heun ODE pictures [13]. The same covering relation holds for the η-NCHO, a shifted version of the NCHO and AQRM [14]. It was discovered [15] that the eigenvalue problem of the NCHO gives rise to a long-established physical model called the two-photon quantum Rabi model (tpQRM) [16]. This shows that the interaction between one photon and a two-level atom (QRM) can be obtained from that between two photons and a two-level atom (tpQRM) via the covering relation. It would be interesting to confirm the mathematical concept of covering between these physical models through actual experiments. It is also worth remarking that in a previous study [15], using representation theory, the covering relation takes a simpler and clearer form. Further exploring the physical and number theoretical implications of the covering relations is one of the promising research directions in this area.

3.4 Partition function and spectral zeta functions

In this section, we consider a quantum system with self-adjoint Hamiltonian H. As mentioned in the introduction, we are interested in knowing the action of the unitary operator $\exp(-itH)$ (propagator/heat kernel) and its trace, the partition function $Z_H(\beta)$ of H. The partition function is defined as the sum of the Boltzmann factors $\exp(-\beta E(\mu))$, where $E(\mu)$ is the energy (eigenvalue) of the state μ, that is, it is given by

$$Z_H(\beta) := \mathrm{Tr}[\exp(-tH)] = \sum_{\mu\in\Omega} \exp(-\beta E(\mu)),$$

where Ω is the set of all possible eigenstates of H. The partition function is one of the fundamental tools of statistical mechanics for the study of entropy and other properties of a system in thermodynamical equilibrium.

On the other hand, the spectral zeta function $\zeta_H(s)$ is defined as the Dirichlet series determined by the eigenvalue sequence $E(\mu)$. We assume for simplicity[*7] that $E(\mu) \neq 0$. Concretely, $\zeta_H(s)$ is given by

$$\zeta_H(s) := \sum_{\mu\in\Omega} E(\mu)^{-s} \quad (\Re(s) \gg 1).$$

Therefore, by the definition of $\Gamma(s)$, the two functions are connected via the Mellin transform

$$\zeta_H(s) = \frac{1}{\Gamma(s)} \int_0^\infty t^{s-1} Z_H(t) e^{-t\tau}\, dt.$$

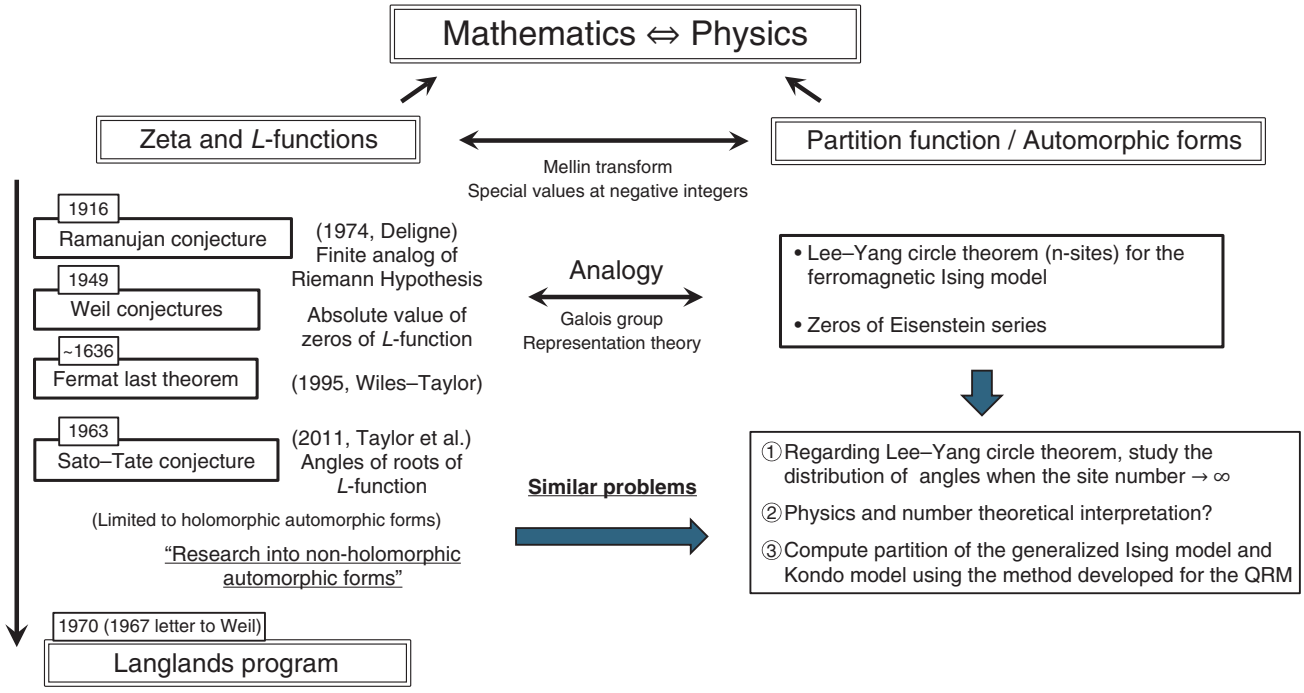
It is important to mention that the long awaited explicit formulas of the heat kernel and partition function of the QRM and AQRM were finally obtained [17–19]. The technique for the computation is based on the Trotter–Kato product formula, regarded as the mathematical formulation of the Feynman path integral, multivariate Gaussian integrals and the Fourier transform in $\mathbb{F}_2^n$ (n = 1, 2, ...) interpreted as a Weyl representation of $SL_2(\mathbb{F}_2)$, where $\mathbb{F}_2$ denotes the field with 2 elements. The series expression for the heat kernel corresponds to the sum of irreducible representations of the decomposition of the action of the infinite symmetric group $\mathfrak{S}_\infty$ over $\mathbb{F}_2^\infty$, and each summand is an orbital integral of $\mathfrak{S}_\infty$. This is a surprising discovery that gives further hints on the computation of the heat kernel for more general models, and in general to the structure of interaction models.

The partition function also provides a short proof of the analytic continuation of the corresponding spectral zeta function using a path integral expression going from infinity to the origin, then circling the origin and back to infinity[*8], and extensions of Bernoulli numbers for the spectral zeta function (e.g. Rabi-Bernoulli polynomials for the QRM). Note that the partition function can be recovered from the Rabi-Bernoulli polynomials since the Laurent expansion at the origin of the generating functions of the Rabi-Bernoulli

*6 The Gaussian hypergeometric ODE (or function) has $\{0, 1, \infty\}$ as regular singular points in standard form.

*7 For general systems, the spectral zeta functions are usually of Hurwitz type.

*8 This type of path is known as Hankel contour.



* Ramanujan conjecture reduces to the Weil conjecture for certain algebraic varieties in 11 dimensions (discovered by Mikio Sato in 1962).

Fig. 1. Partition function and spectral zeta function.

polynomials is equal to the partition function. Although integral expressions for the positive integer points for the spectral zeta function of the NCHO are known, an explicit expression for the partition function has not been obtained. The values at negative integers may also be regarded as a generalization of Bernoulli numbers, called NC-Bernoulli numbers. Therefore, we might expect that the partition function is given by the Laurent series at the origin of the generating function of the NC-Bernoulli numbers. Unfortunately, at present, even with strong supporting evidence, it remains as a conjecture [20].

Nevertheless, the research towards conjecture has illuminated several aspects of the theory. For instance, to work with the formal expressions of the special values of the spectral zeta functions arising from the partition functions, it is useful to consider Borel-summation and non-Archimedian methods to deal with certain divergent series. In particular, certain expressions of special values of zeta functions that are divergent in $\mathbb{R}$ may be interpreted as special values of zeta (Hurwitz type) functions [21] defined in the p-adic fields $\mathbb{Q}_p$ [20].

4. *L*-functions and the structure of zeros of partition function

Table 1 shows a comparison between the zeta functions $\zeta_Q(s)$ and $\zeta(s)$. As mentioned above, if we let $\alpha/\beta \to 1$ in $\zeta_Q(s)$, we essentially obtain $\zeta(s)$. However, what it is important is that $\zeta_Q(s)$ reveals a structure that is not visible in $\zeta(s)$. It is also worth remarking that we cannot expect $\zeta_Q(s)$ to have an Euler product expression or functional equation. In fact, the integral expressions at positive integer points obtained in a previous study [9] suggest that $\zeta_Q(s)$ may be expressed as a sum of number theoretical L-functions (zeta functions associated to certain representations). If this conjecture is correct, then even if the individual L-functions have functional equations, $\zeta_Q(s)$ may not have one. Similarly, the lack of a functional equation is not a contradiction since the axis of symmetry (the line $\Re(s) = \frac{1}{2}$ for $\zeta(s)$) for each L-function in the summands may be different. In any case, further research is needed to clarify these questions.

On the right of **Fig. 1**, one of the points involves the partition function of the ferromagnetic Ising model, given by

$$Z(\beta) := \sum_{X \subset \Lambda} e^{\beta |X|} \prod_{x \in X} \prod_{y \in \Lambda \setminus X} a_{xy}$$
$$(a_{xy} = a_{yx} \in [-1, 1]).$$

In particular, the Lee–Yang circle theorem (1952) states that all of its zeros are imaginary numbers (in other words, that $z = e^{-\beta}$ lie in the unit circle) [22]. Of course, the zeros are important physically because phase transition precisely occurs around these points.

The study on the zeros of zeta functions and L-functions motivated several research problems that inspired and led research in number theory in the 20th century, including the Weil conjectures, a finite field analog of the Riemann hypothesis that led to the innovations in algebraic geometry by Alexander Grothendieck and the Ramanujan "$\Re(s) = \frac{11}{2}$" conjecture on the absolute value of the zeros (of the reciprocal) of the L-function

$$L(s, \Delta) := \sum_{n=1}^{\infty} \tau(n) n^{-s}$$
$$= \prod_{p:\text{primes}} (1 - \tau(p) p^{-s} + \mathrm{p}^{11-2s})^{-1}$$

associated to the automorphic form $\Delta(z) := e^{2\pi i z} \prod_{n=1}^{\infty} (1 - e^{2\pi i n z})^{24} = \sum_{n=1}^{\infty} \tau(n) e^{2\pi i n z}$. On the side of partition functions and automorphic forms, the study of zeros of Eisenstein series has seen progress, but the significance of the results is still not clear. We expect that significant research on this area may also be conducted from the perspective of phase transition.

Another interesting point is that the research on the distribution of angles (complex argument) of the Ising model with infinite site number $|\Lambda|$ appears to be similar to the Sato–Tate conjecture regarding the distribution of angles of the zeros of $L(s, \Delta)$. For the Sato–Tate conjecture, it is also desirable to advance beyond the holomorphic automorphic forms and arithmetic geometry considered up to now, into non-holomorphic and Maass forms, which appear in the Selberg trace formula.

In our institute, Horinaga investigated non-holomorphic automorphic forms [23], and Nakahama is working on the representation theoretical aspects underlying the NCHO, with the aim of defining a multivariate version of the NCHO [24]. This latter research is an application of a particular case of Howe's theory of dual pairs [25] (i.e., the theory of spherical harmonics), which forms the basis of the modern invariant theory and has applications to automorphic forms. Higher dimensional constructions of the NCHO may be obtained using general dual pairs [25], thus we may expect the appearance of Sigel modular forms in the study of spectral zeta functions for these generalized constructions.

References

[1] E. T. Jaynes and F. W. Cummings, "Comparison of Quantum and Semiclassical Radiation Theories with Application to the Beam Maser," Proc. IEEE, Vol. 51, No. 1, pp. 89–109, 1963. https://doi.org/10.1109/PROC.1963.1664

[2] R. Apéry, "Irrationalité de $\zeta(2)$ et $\zeta(3)$,"Astérisque, Vol. 61, pp. 11–13, 1979.

[3] T. Fuse, F. Yoshihara, K. Kakuyanagi, and K. Semba, "Interaction between an Artificial Atom and an Electromagnetic Field~Beyond the Strong Coupling~," Nihon Butsuri Gakkaishi, Vol. 73, No. 1, pp. 21–26, 2018 (in Japanese).

[4] A. Parmeggiani and M. Wakayama, "Oscillator Representations and Systems of Ordinary Differential Equations," Proc. Natl. Acad. Sci. USA, Vol. 98, No. 1, pp. 26–30, 2001. https://doi.org/10.1073/pnas.98.1.26

[5] A. Parmeggiani, "Spectral Theory of Non-commutative Harmonic Oscillators: An Introduction," Lecture Notes in Math., Vol. 1992, Springer, 2010.

[6] T. Ichinose and M. Wakayama, "Zeta Functions for the Spectrum of the Non-commutative Harmonic Oscillators," Comm. Math. Phys., Vol. 258, pp. 697–739, 2005. https://doi.org/10.1007/s00220-005-1308-7

[7] R. C. Gunning, "The Eichler Cohomology Groups and Automorphic Forms," Trans. Amer. Math. Soc., Vol. 100, No. 1, pp. 44–62, 1961. https://doi.org/10.2307/1993353

[8] K. Kimoto and M. Wakayama, "Elliptic Curves Arising from the Spectral Zeta Function for Non-commutative Harmonic Oscillators and $\Gamma_0(4)$-modular Forms," Proc. of the Conference on L-Functions, Fukuoka, Japan, Feb. 2006, pp. 201–218, 2007. https://doi.org/10.1142/9789812772398_0011

[9] K. Kimoto and M. Wakayama, "Apéry-like Numbers for Non-commutative Harmonic Oscillators and Automorphic Integrals," Ann. Inst. H. Poincaré D, Vol. 10, No. 2, pp. 205–275, 2023. https://doi.org/10.4171/AIHPD/129

[10] B. C. Berndt, "Generalized Eisenstein Series and Modified Dedekind Sums," J. Reine Angew. Math., Vol. 1975, No. 272, pp. 182–193, 1975. https://doi.org/10.1515/crll.1975.272.182

[11] S. Horinaga, "On the Representations Generated by Eisenstein Series of Weight $\frac{n+3}{2}$," J. Number Theory., Vol. 201, pp. 206–227, 2019. https://doi.org/10.1016/j.jnt.2019.02.007

[12] H. Ochiai, "Non-commutative Harmonic Oscillators and Fuchsian Ordinary Differential Operators," Comm. Math. Phys., Vol. 217, pp. 357–373, 2001. https://doi.org/10.1007/s002200100362

[13] M. Wakayama, "Equivalence between the Eigenvalue Problem of Non-commutative Harmonic Oscillators and Existence of Holomorphic Solutions of Heun Differential Equations, Eigenstates Degeneration, and the Rabi Model," Int. Math. Res. Notices, Vol. 2016, No. 3, pp. 759–794, 2016. https://doi.org/10.1093/imrn/rnv145

[14] C. Reyes-Bustos and M. Wakayama, "Covering Families of the Asymmetric Quantum Rabi Model: η-Shifted Non-commutative Harmonic Oscillators," Comm. Math. Phys., Vol. 403, pp. 1429–1476, 2023. https://doi.org/10.1007/s00220-023-04825-3

[15] R. Nakahama, "Equivalence between Non-commutative Harmonic Oscillators and Two-Photon Quantum Rabi Models," Preprint 2024. arXiv:2405.19814 [math-ph].

[16] D. Braak, "Spectral Determinant of the Two-Photon Quantum Rabi Model," Ann. Phys., Vol. 535, No. 3, p. 2200519, 2023. https://doi.org/10.1002/andp.202200519

[17] C. Reyes-Bustos and M. Wakayama, "The Heat Kernel for the Quantum Rabi Model," Adv. Theor. Math. Phys., Vol. 26, No. 5, pp. 1347–1447, 2022. https://doi.org/10.4310/ATMP.2022.v26.n5.a8

[18] C. Reyes-Bustos and M. Wakayama, "Heat Kernel for the Quantum Rabi Model: II. Propagators and Spectral Determinants," J. Phys. A:

Math. Theor., Vol. 54, p. 115202, 2021. https://doi.org/10.1088/1751-8121/abdca7
[19] C. Reyes-Bustos, "The Heat Kernel of the Asymmetric Quantum Rabi Model," J. Phys. A: Math. Theor., Vol. 56, p. 425302, 2023. https://doi.org/10.1088/1751-8121/acfbc8
[20] K. Kimoto and M. Wakayama, "Partition Functions for Non-commutative Harmonic Oscillators and Related Divergent Series," Indag. Math., 2024. https://doi.org/10.1016/j.indag.2024.05.011
[21] H. Cohen, "Number Theory Volume II: Analytic and Modern Tools," Springer, 2007.
[22] D. Ruelle, "Statistical Mechanics: Rigorous Results," Addison-Wesley, 1989. https://doi.org/10.1142/4090
[23] S. Horinaga and H. Narita, "Cuspidal Components of Siegel Modular Forms for Large Discrete Series Representations of $Sp_4(\mathbb{R})$," Manuscripta Math., Vol. 174, pp. 159–202, 2024. https://doi.org/10.1007/s00229-023-01513-3
[24] R. Nakahama, "Representation Theory of $\mathfrak{sl}(2, \mathbb{R}) \simeq \mathfrak{su}(1, 1)$ and a Generalization of Non-commutative Harmonic Oscillators," in Mathematical Foundation for Post-Quantum Cryptography, "Mathematic for Industry," Springer, 2024 (in print), arXiv:2310.17118 [math-ph].
[25] R. Howe, "Remarks on Classical Invariant Theory," Trans. Ameri. Math. Soc., Vol. 313, No. 2, pp. 539–570, 1989. https://doi.org/10.2307/2001418

Cid Reyes-Bustos
Research Scientist, NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.
He obtained a Ph.D. in functional mathematics from Kyushu University in 2018. He held the position of specially appointed assistant professor at the Tokyo Institute of Technology from 2019 to 2022. He joined the NTT Institute for Fundamental Mathematics (NTT Communication Science Laboratories) in 2022 as a research associate and started his current position in 2024. His main research interests are number theory, representation theory, mathematical physics, graph theory, and the interaction between these areas.

Masato Wakayama
Research Principal and Head of NTT Institute for Fundamental Mathematics, NTT Communication Science Laboratories.
He received a Ph.D. in mathematics from Hiroshima University in 1985. Before joining NTT in 2021, he had held various academic positions: an associate professor at Tottori University, visiting fellow at Princeton University, visiting professor at the University of Bologna, distinguished lecturer at Indiana University, professor of mathematics, distinguished professor, dean of the Graduate School of Mathematics, the founding director of the Institute of Mathematics for Industry, executive vice president of Kyushu University, and vice president and professor at Tokyo University of Science. He is now also a professor emeritus of Kyushu University. He specializes in representation theory, number theory, and mathematical physics.

Digital Longitudinal Monitoring of Fiber-optic Link Using Coherent Receiver

Takeo Sasai

Abstract

In fiber-optic communication systems, it is crucial for operators to accurately monitor various physical parameters along optical links to fully leverage the potential transmission capacity and conduct fault analysis. Digital longitudinal monitoring (DLM) has been intensively studied for its capability of monitoring various physical parameters, such as optical power, distributed along the fiber-longitudinal direction by solely processing signals received at coherent receivers. In this article, the fundamentals and recent advances in DLM are reviewed, including working principles, spatial resolution, and key experiments demonstrating its feasibility for use in operations.

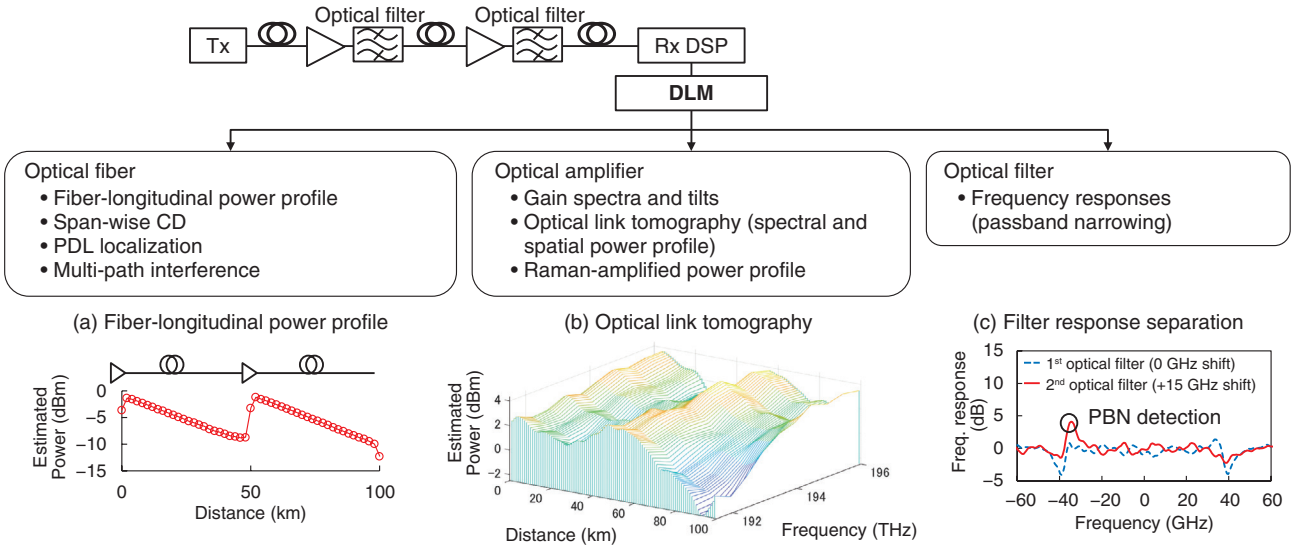
Keywords: digital longitudinal monitoring, digital coherent receiver, fiber nonlinearity

1.Introduction

Optical networks are becoming increasingly complex due to trends such as disaggregation, dynamic provisioning, and ultra-wideband transmission. To fully leverage the potential capacity and maintain these advanced networks efficiently, it is crucial for operators to monitor the physical parameters of the entire link, including optical power and locations of loss anomalies.

Digital longitudinal monitoring (DLM), which has been studied intensively, estimates various physical link parameters distributed in the *fiber-longitudinal* direction solely by processing signals received at a digital coherent receiver (**Fig. 1**). Demonstrated monitored parameters include the longitudinal optical power profile [1–7], span-wise chromatic dispersion (CD) map [2], amplifiers' gain tilt [2, 8], optical filter detuning [2], polarization dependent loss (PDL) [9–11], multi-path interference [6], and spectral and spatial power monitoring called optical link tomography over C [2], C+L [8], and S+C+L [12] bands. DLM enables the localization of multiple anomalies over multi-span links without the need for dedicated hardware devices such as optical time domain reflectometers (OTDRs) and optical spectrum analyzers, thus reducing installation costs of these devices. Longitudinal power monitoring (LPM) is of particular importance since optical power determines the generalized signal-to-noise ratio (SNR) and its distributed measurement allows the localization of loss anomaly, both of which facilitate network management and control. Various demonstrations of LPM have showcased its capabilities, including a precise LPM closely matching the OTDR [4], demonstrations over 10,000 km [5], LPM using commercial transponders [7], and field experiments [13].

The primary challenge with DLM is that it relies on the fiber nonlinearity, and high fiber launch power is necessary to achieve sufficient accuracy, which causes a quality of transmission (QoT) degradation in adjacent wavelength division multiplexing (WDM) channels due to excessive nonlinear interference (NLI). Most demonstrations used optical power far higher than the optimal operational point. We recently demonstrated LPM under system optimal launch power and WDM conditions with sufficient accuracy to locate several loss anomalies in field-deployed fibers [13]. In this demonstration, we also showed four-dimensional optical link tomography, which



DSP: digital signal processing
PBN: passband narrowing
Rx: receiver
Tx: transmitter

Fig. 1. Concept of DLM and monitored physical parameters.

visualizes optical power not only in the distance direction but also in the time, frequency, and polarization, allowing for the localization of multiple QoT degradation causes such as PDL, spectral tilt, and time-varying power anomaly (e.g., fiber bending loss).

In this article, the fundamentals and recent advancements in DLM are reviewed, with a particular focus on LPM, including the localization principle, an inherent limitation on spatial resolution (SR), algorithms, and several key demonstrations of DLM.

2. Working principle

LPM estimates the fiber-longitudinal optical power $P(z)$ from received waveforms by extracting the nonlinear phase shift $\gamma'(z) = \gamma(z)P(z)$ that the signals experienced during the fiber transmission, where $\gamma(z)$ is the fiber nonlinear coefficient at position z. The key mechanism for the localization of the optical power is the interaction between fiber nonlinearity and CD in optical fibers [3]. To elucidate the localization principle, let us consider the regular perturbation model for the fiber nonlinear propagation. In the first-order regular perturbation, the additive NLI $j\gamma'(z)|\boldsymbol{A}(z)|^2 \boldsymbol{A}(z)$ is excited at each position on fibers, which is dependent on the original signal waveform $\boldsymbol{A}(z)$ (see **Fig. 2**). Such local NLIs propagate to the receiver, undergoing the remaining CD $\widehat{D}_{zL}$ from z to the link end L, and evolve as $\gamma'(z)\boldsymbol{g}(z)$, where

$$\boldsymbol{g}(z) = j\widehat{D}_{zL}[|\boldsymbol{A}(z)|^2 \boldsymbol{A}(z)]. \quad (1)$$

The total NLI at the receiver is the accumulation of the received local NLIs and represented as

$$\boldsymbol{A}_1(L) = \int_0^L \gamma'(z)\boldsymbol{g}(z)dz, \quad (2)$$

which shows that $\{\boldsymbol{g}(z)\}_z$ form a basis of the total NLI. Two of these basis vectors $\boldsymbol{g}(z)$ and $\boldsymbol{g}(z + \Delta z)$, separated by a distance Δz, are linearly independent in the presence of sufficient CD, allowing the corresponding $\gamma'(z)$ to be extracted at the receiver. The qualitative understanding is that sufficient CD alters the original signal waveforms during the propagation, and the excited NLIs at different locations are thus unique and distinguishable upon reception.

3. Spatial resolution

One straightforward approach to extract the expansion coefficient $\gamma'(z_k)$ at position z_k ($k \in [0, K-1]$) is to take the inner product of $\boldsymbol{A}_1(L)$ and the corresponding basis vector $\boldsymbol{g}_k = \boldsymbol{g}(z_k)$. However, the basis $\{\boldsymbol{g}(z)\}_z$

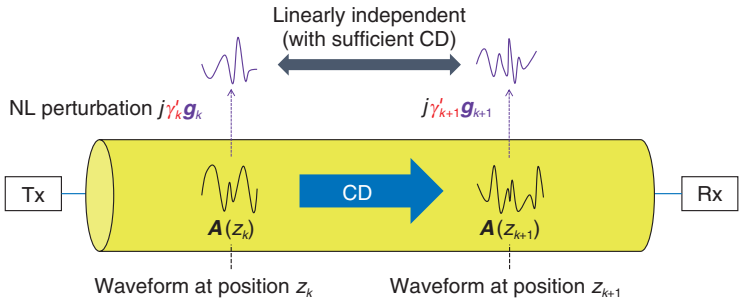


Fig. 2. Perturbation model of fiber nonlinear propagation. NLIs from positions z_k and z_{k+1} are linearly independent with sufficient CD, allowing the estimation of $\gamma'_k = \gamma'(z_k)$.

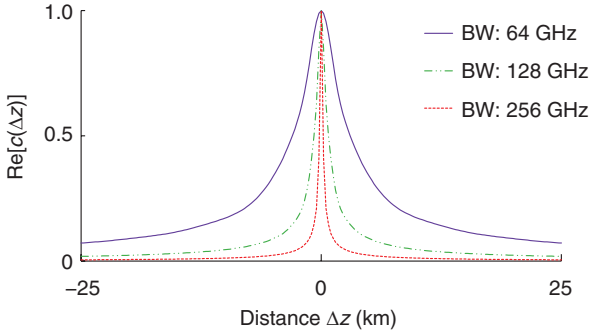


Fig. 3. SCF with various BW. β_2 = –20.5 ps²/km is assumed.

is not orthogonal: the resulting inner product $\boldsymbol{g}_k^{\dagger}\boldsymbol{A}_1(L)$ involves not only $\gamma'(z_k)$ but also those at neighboring positions. In fact, it has been shown [3] that the expectation of the inner product of two vectors $\boldsymbol{g}(z)^{\dagger}\boldsymbol{g}(z+\Delta z)$ is expressed under assumptions of stationary Gaussian signal and constant CD β_2 over the link with negligible high-order dispersions as

$$c(\Delta z) \propto \frac{1}{\sqrt{1 + 2j\left(\frac{\Delta z}{z_{CD}}\right) + 3\left(\frac{\Delta z}{z_{CD}}\right)^2}}$$

$$\left(z_{\mathrm{CD}} \simeq \frac{0.288}{|\beta_2|\mathrm{BW}^2} \text{ for Nyquist signals}\right), \qquad (3)$$

which is called the spatial correlation function (SCF) or spatial response function [3, 14]. Here, BW is the bandwidth of the signal. **Figure 3** shows the SCF for various BWs. The SCF has a 'width' with long tails, suggesting that the estimated $\gamma'(z_k)$ values contain contributions from neighboring positions. This means that there is an inherent uncertainty in determining the position of loss events, limiting the SR of LPM. By defining the SR as the full width at half maximum of the SCF, obtain

$$\mathrm{SR} \simeq \frac{0.507}{|\beta_2|\mathrm{BW}^2}, \qquad (4)$$

indicating that the SR is enhanced with an increased CD and BW [3].

4. Methods

The simple inner-product approach described above is called the correlation method (CM) [1, 3, 6]. However, due to the non-orthogonality shown in the SCF, the entire output of CM $\mathbf{G}^{\dagger}\boldsymbol{A}_1 = [\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_{K-1}]^{\dagger}\boldsymbol{A}_1$ is expressed as the convolution of the true power profile and SCF [3], which indicates the sensitivity of the CM is limited as shown in **Fig. 4** (blue). Another approach is the least squares (LS) $(\mathbf{G}^{\dagger}\mathbf{G})^{-1}\mathbf{G}^{\dagger}\boldsymbol{A}_1$ [4], which minimizes $\|\mathbf{G}\boldsymbol{\gamma}' - \boldsymbol{A}_1\|^2$. LS naturally deconvolves the convolution effects in CM by $(\mathbf{G}^{\dagger}\mathbf{G})^{-1}$, achieving precise LPM, as shown in red. However, the simple LS suffers from instability

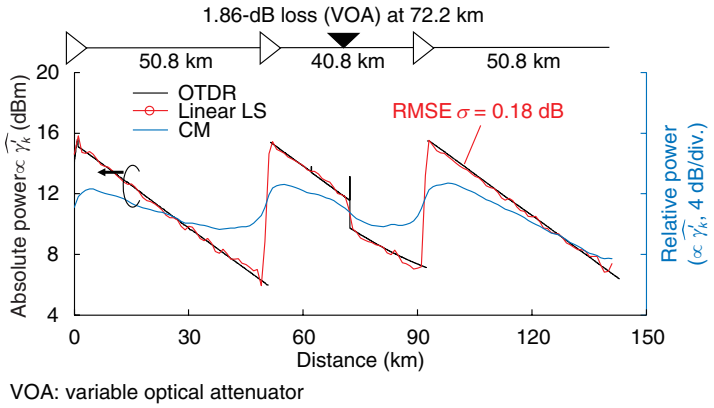


Fig. 4. Experimental results of LPM with LS (red) and CM (blue) with 1.86-dB attenuation inserted at 72.2 km. RMSE from OTDR was 0.18 dB.

related to the ill-posedness of LPM, as pointed out in a previous study [4]. The penalized LS was therefore proposed [7] as

$$\widehat{\boldsymbol{\gamma}}' = (\mathbf{G}^{\dagger}\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^{\dagger}\boldsymbol{A}_1, \quad (5)$$

where λ is a regularization parameter and **I** is the identity matrix. This method generalizes CM and LS as it approaches CM for $\lambda \to \infty$, while it becomes LS for $\lambda = 0$.

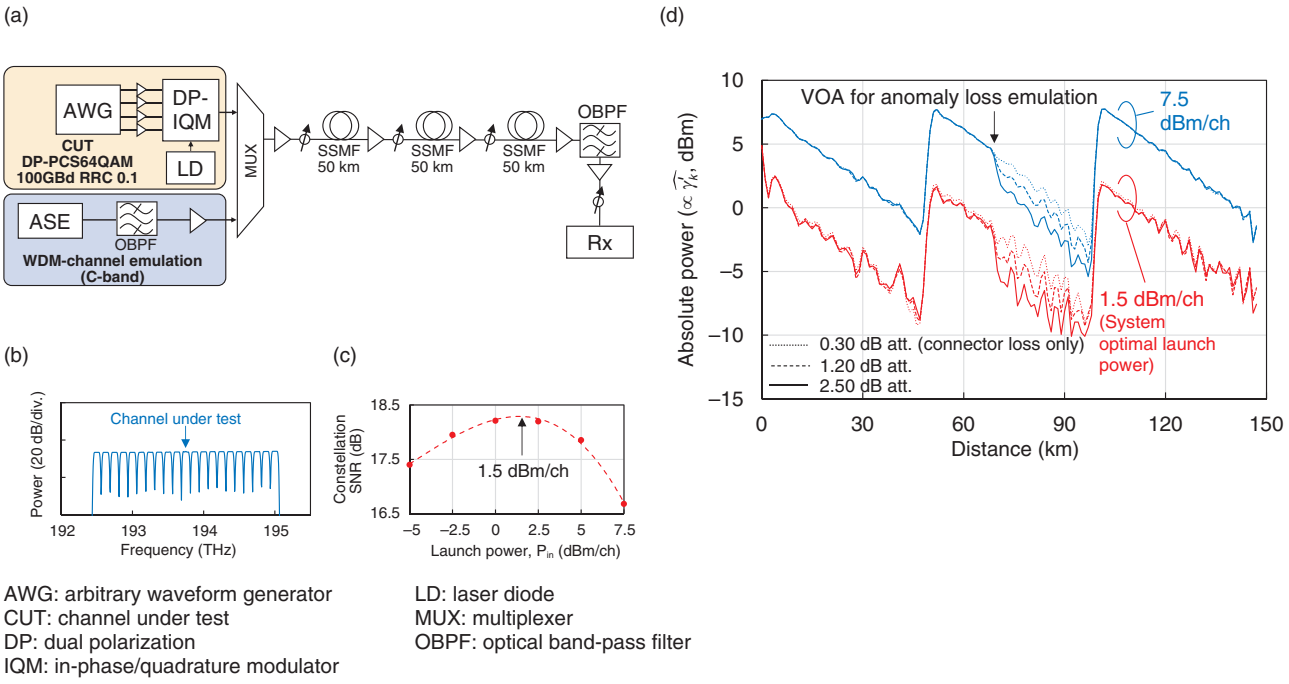
Although most LPM demonstrations have used self-channel interference, cross-channel interference (XCI) or cross phase modulation can also be used to localize power events [15–17]. Although XCI-based methods require an access to two channels and their timing synchronization, they achieve high SR due to a large walk-off between two channels.

5. Experimental demonstrations

Figure 4 shows an experimental demonstration of LPM using the LS estimation [4], which achieved precise estimation. This experiment used probabilistically constellation shaped (PCS) 64 quadrature amplitude modulation (QAM) with a roll-off factor of 0.1 and symbol rate of 100 GBd. The link under test was a 142.4-km 3-span standard single-mode fiber (SSMF) link with a 1.86-dB attenuation inserted at 72.2 km. The fiber launch power was set to 15 dBm/ch. While the CM (blue) reflects the overall power trend, it fails to align with the OTDR and is less sensitive to the loss anomaly due to the convolution effect. On the other hand, the LS demonstrated a closer match with the OTDR, having a root mean square error (RMSE) of 0.18 dB and maximum absolute error of 0.57 dB. **Figure 5** shows the LPM experiment under the system optimal launch power and WDM conditions [4]. The WDM channels were loaded from an amplified spontaneous emission (ASE) source, shaped using an optical filter, with the channel under test set at the center of the WDM channels (Figs. 5(a)(b)). The optimal power was around 1.5 dBm/ch (Fig. 5(c)). As shown in Fig. 5(d), LPM showed superior performance with high power (blue). However, the estimated power profiles at 1.5 dBm/ch are still clearly visible, enough to locate a loss anomaly. The RMSE from the OTDR prior to the loss event was $\sigma = 0.20$ dB, and we set the detection threshold of $4\sigma = 0.80$ dB. Since an inserted loss of 1.20 dB exceeded the threshold, LPM successfully detected the 1.20-dB loss anomaly and can potentially localize a 0.80-dB loss. These results indicate the feasibility of LPM for use in system operations.

6. Summary

In this article, the fundamentals and recent advancements in LPM are reviewed. Recent intensive efforts have led to significant progress, such as a precise LPM that closely matches the OTDR, the feasibility demonstration at operational launch power, and adapting LPM for use with commercial transponders. To achieve more reliable performance for deployment, future research should include (i) improving noise and distortion robustness for enhanced accuracy at operational optical power levels, (ii) developing lightweight algorithms, and (iii) enhancing functionality for monitoring a wider range of link parameters.



AWG: arbitrary waveform generator
CUT: channel under test
DP: dual polarization
IQM: in-phase/quadrature modulator
LD: laser diode
MUX: multiplexer
OBPF: optical band-pass filter

Fig. 5. (a) Experimental setup for WDM transmission. (b) Transmitted WDM spectra. (c) Constellation SNR as a function of fiber launch power. System optimal launch power was approximately 1.5 dBm/ch. (d) Experimental results of LPM under WDM conditions with various attenuation levels are also shown.

References

[1] T. Tanimura, S. Yoshida, K. Tajima, S. Oda, and T. Hoshida, "Fiber-longitudinal Anomaly Position Identification over Multi-span Transmission Link Out of Receiver-end Signals," J. Lightw. Technol., Vol. 38, No. 9, pp. 2726–2733, 2020. https://doi.org/10.1109/JLT.2020.2984270

[2] T. Sasai, M. Nakamura, E. Yamazaki, S. Yamamoto, H. Nishizawa, and Y. Kisaka, "Digital Longitudinal Monitoring of Optical Fiber Communication Link," J. Lightw. Technol., Vol. 40, No. 8, pp. 2390–2408, 2022. https://doi.org/10.1109/JLT.2021.3139167

[3] T. Sasai, E. Yamazaki, and Y. Kisaka, "Performance Limit of Fiber-longitudinal Power Profile Estimation Methods," J. Lightw. Technol., Vol. 41, No. 11, pp. 3278–3289, 2023. https://doi.org/10.1109/JLT.2023.3234534

[4] T. Sasai, M. Takahashi, M. Nakamura, E. Yamazaki, and Y. Kisaka, "Linear Least Squares Estimation of Fiber-longitudinal Optical Power Profile," J. Lightw. Technol., Vol. 42, No. 6, pp. 1955–1965, 2024. https://doi.org/10.1109/JLT.2023.3327760

[5] A. May, F. Boitier, A. C. Meseguer, J. U. Esparza, P. Plantady, A. Calsat, and P. Layec, "Longitudinal Power Monitoring over a Deployed 10,000-km Link for Submarine Systems," The 46th Optical Fiber Communications Conference and Exhibition (OFC 2023), Tu2G.3, San Diego, CA, USA, 2023. https://doi.org/10.1364/OFC.2023.Tu2G.3

[6] C. Hahn, J. Chang, and Z. Jiang, "Localization of Reflection Induced Multi-path-interference over Multi-span Transmission Link by Receiver-side Digital Signal Processing," The 45th Optical Fiber Communications Conference and Exhibition (OFC 2022), Th1C.3, San Diego, CA, USA, 2022. https://doi.org/10.1364/OFC.2022.Th1C.3

[7] T. Sasai, Y. Sone, E. Yamazaki, M. Nakamura, and Y. Kisaka, "A Generalized Method for Fiber-longitudinal Power Profile Estimation," The 49th European Conference on Optical Communications (ECOC 2023), Tu.A.2.6., Glasgow, UK, 2023. https://doi.org/10.1049/icp.2023.2471

[8] M. Sena, P. Hazarika, C. Santos, B. Correia, R. Emmerich, B. Shariati, A. Napoli, V. Curri, W. Forysiak, C. Schubert, J. K. Fischer, and R. Freund, "Advanced DSP-based Monitoring for Spatially Resolved and Wavelength-dependent Amplifier Gain Estimation and Fault Location in C+L-band Systems," J. Lightw. Technol., Vol. 41, No. 3, pp. 989–998, 2023. https://doi.org/10.1109/JLT.2022.3208209

[9] M. Eto, K. Tajima, S. Yoshida, S. Oda, and T. Hoshida, "Location-resolved PDL Monitoring with Rx-side Digital Signal Processing in Multi-span Optical Transmission System," OFC 2022, Th1C.2, San Diego, CA, USA, 2022. https://doi.org/10.1364/OFC.2022.Th1C.2

[10] M. Takahashi, T. Sasai, E. Yamazaki, and Y. Kisaka, "Experimental Demonstration of Monitoring PDL Value and Location Using DSP-based Longitudinal Power Estimation with Linear Least Squares," ECOC 2023, P14, Glasgow, UK, 2023. https://doi.org/10.1049/icp.2023.2057

[11] L. Andrenacci, G. Bosco, and D. Pilori, "PDL Localization and Estimation through Linear Least Squares-based Longitudinal Power Monitoring," IEEE Photon. Technol. Lett., Vol. 35, No. 24, pp. 1431–1434, 2023. https://doi.org/10.1109/LPT.2023.3331110

[12] R. Kaneko, T. Sasai, F. Hamaoka, M. Nakamura, and E. Yamazaki, "Fiber Longitudinal Monitoring of Inter-band-SRS-induced Power Transition in S+C+L WDM Transmission," The 47th Optical Fiber Communications Conference and Exhibition (OFC 2024), W1B.4, San Diego, CA, USA, 2024. https://doi.org/10.1364/OFC.2024.W1B.4

[13] T. Sasai, G. Borraccini, Y.-K. Huang, H. Nishizawa, Z. Wang, T. Chen, Y. Sone, T. Matsumura, M. Nakamura, E. Yamazaki, and Y. Kisaka, "4D Optical Link Tomography: First Field Demonstration of Autonomous Transponder Capable of Distance, Time, Frequency, and Polarization Resolved Monitoring," OFC 2024, Th4B.7, San Diego,

CA, USA, 2024. https://doi.org/10.1364/OFC.2024.Th4B.7
[14] C. Hahn and Z. Jiang, "On the Spatial Resolution of Location-resolved Performance Monitoring by Correlation Method," OFC 2023, W1H.3, San Diego, CA, USA, 2023. https://doi.org/10.1364/OFC.2023.W1H.2
[15] R. Hui, C. Laperle, and M. O'Sullivan, "Measurement of Total and Longitudinal Nonlinear Phase Shift as Well as Longitudinal Dispersion for a Fiber-optic Link Using a Digital Coherent Transceiver," J. Lightw. Technol., Vol. 40, No. 21, pp. 7020–7029, 2022. https://doi.org/10.1109/JLT.2022.3198549
[16] P. Serena, C. Lasagni, A. Bononi, F. Boitier, A. May, P. Ramantanis, and M. Lonardi, "Locating Fiber Loss Anomalies with a Receiver-side Monitoring Algorithm Exploiting Cross-phase Modulation," OFC 2023, W1H.3, San Diego, CA, USA, 2023. https://doi.org/10.1364/OFC.2023.W1H.3
[17] I. Kim, O. Vassilieva, R. Shinzaki, M. Eto, S. Oda, and P. Palacharla, "Multi-channel Longitudinal Power Profile Estimation," ECOC 2023, Tu.A.2.4, Glasgow, UK, 2023. https://doi.org/10.1049/icp.2023.2217

Takeo Sasai
Associate Distinguished Researcher, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in electrical engineering from the University of Tokyo in 2015 and 2017. In 2017, he joined NTT Network Innovation Laboratories. His research interests include fiber-optic link modeling and digital signal processing in optical fiber communication. He is a member of the IEEE, the Optica, and Institute of Electronics, Information and Communication Engineers (IEICE) of Japan. He was the recipient of the IEICE Communications Society Optical Communication Systems Young Researchers Award in 2021 and IEICE Young Researcher Award in 2022.

Collaborative Business Navigation Platform That Comprehensively Supports Work of Operators

Hiroki Koike, Hideaki Tanaka, Hidetaka Koya, Fumihiro Yokose, Hajime Nakajima, and Haruo Oishi

Abstract

NTT Access Network Service Systems Laboratories has long been developing technologies that contribute to more-efficient operations on personal computers and other information devices. We have implemented a collaborative business navigation platform as a new technology to continue this trend. Using this platform makes it possible to easily develop digital transformation (DX) tools suited to diverse work environments and operator roles. This article discusses the functions of the collaborative business navigation platform and the use cases of DX tools by using this platform.

Keywords: DX promotion, automatic operation, user interface

1. Adapting digital transformation tools for diverse work processes and system applications

Digital transformation (DX) has been attracting attention, and DX tools that run on personal computers (PCs), such as robotic process automation (RPA)[*1] and digital adoption platform (DAP)[*2], are becoming widely used. However, it is difficult to improve work efficiency in many fields by using conventional DX tools. For example, the telecommunications business provides a variety of services with a complex combination of physical devices and logical functions. Operators working in this business are required to be accurate and efficient when handling vast amounts of information in real time. Carrying out such work involves many factors, one of the most important of which is the work system and applications. Each function has its own processes, and operations proceed through a combination of various work systems and applications (e.g., schedulers, email, and reception systems). Since the combination and use of these work systems and applications differ as the work changes, the operations have their own unique processes; however, they are not the best work processes[*3] for each individual task. Even in such cases, the operator needs to integrate each process for each task and carry out the work (many tasks) as a single work process. However, in complex work processes, conventional DX tools have not been able to provide sufficient support due to the challenges described in the following section.

2. Challenges facing traditional DX tools

Two typical examples of situations that could not be improved with conventional DX tools are shown in **Fig. 1**. In the first example shown in Fig. 1(a), providing a user interface (UI) suitable for a variety of work environments has proved difficult. If a work system and applications are commonly used for both field (i.e., outdoor) work and office automation (OA) work, the UI suitable for each type of work will differ. Implementing a UI tailored to such diverse work

*1 RPA: A software technology that automates user operations.
*2 DAP: A software technology that supports the implementation of systems and tools.
*3 Work process: A way of executing operations related to work. It can also be paraphrased as a process, method, procedure, etc.

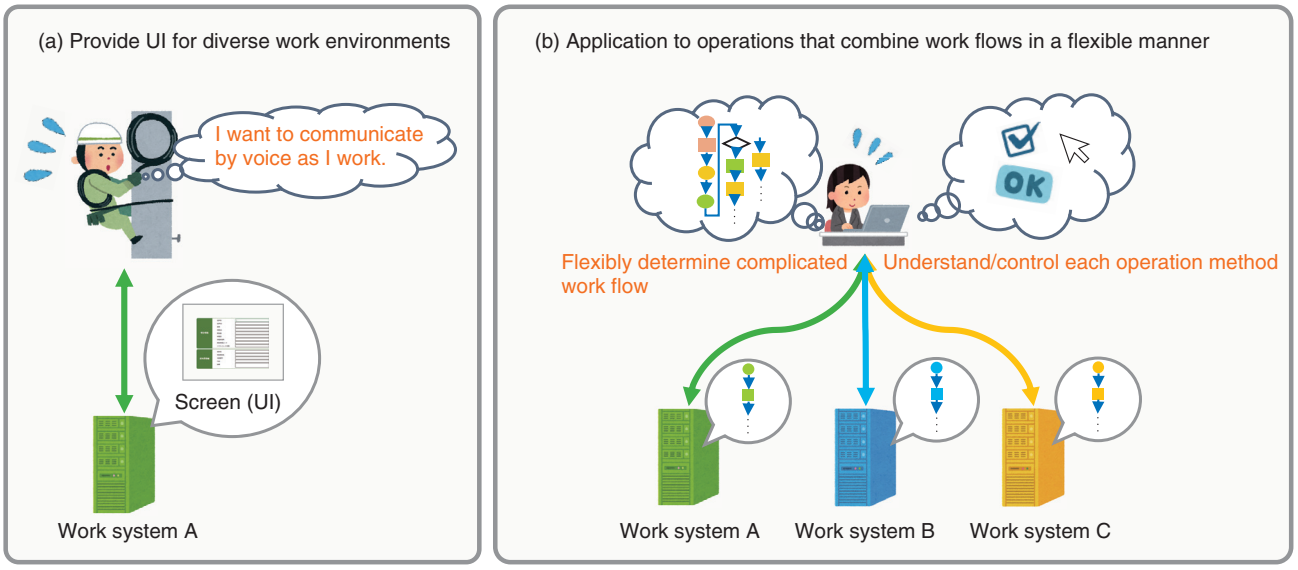


Fig. 1. Challenges facing traditional DX tools.

environments into the work systems and applications by using current DX tools can lead to high costs.

As shown in Fig. 1(b), it is difficult to flexibly handle complex and sophisticated work flows. Work processes involve diverse and multiple operators, work systems, and applications. Work efficiency can be improved by breaking down work processes into smaller parts and using DX tools to handle situations in which work can be streamlined (e.g., by automation); however, the number and types of situations that require improvement vary in accordance with the type of work, and it is inefficient and impractical to develop DX tools for each work type individually from scratch. To address the above challenges, the NTT Network Innovation Center developed the collaborative business navigation platform that is based on technology developed by NTT Access Network Service Systems Laboratories.

3. Technical overview of the collaborative business navigation platform

The collaborative business navigation platform is software that runs in the background of the local environment on Windows 10 and 11. The platform consolidates and integrates two major functions commonly required for DX tools: (i) a function for managing and controlling screen elements (operation targets[*4]), such as buttons and text boxes, registered in the DX tool, and (ii) a function for monitoring the operation status (work status[*5]) during work processes such as operator operations and screen changes resulting from those operations. By centralizing the above functions and eliminating the need to develop each DX tool separately, most development tasks can be focused solely on preparing UIs, work processes, etc. (i.e., upper-level applications) that are tailored to the operator's work environment and role. Thus, it becomes easy to develop DX tools that are tailored to the operator's work environment and role.

To use DX tools developed using the collaborative business navigation platform, the upper-level application and platform must be connected. As shown in **Fig. 2**, a DX tool can be developed by preparing arbitrary upper-level applications (such as chatbots and voice-recognition applications) that fit the operator's environment and role. Two functions included on the platform become available: (1) a management-and-control function for operation targets of work systems and applications (operation-target-abstraction function) and (2) a monitoring function for work status (work-status-abstraction function).

3.1 Operation-target-abstraction function

Conventionally, when developing a DX tool with

*4 Operation target: A screen element on the work-system application that is registered in the operation-target database.

*5 Work status: Operations carried out by the operator during the work process (starting the system, updating files, setting values, etc.).

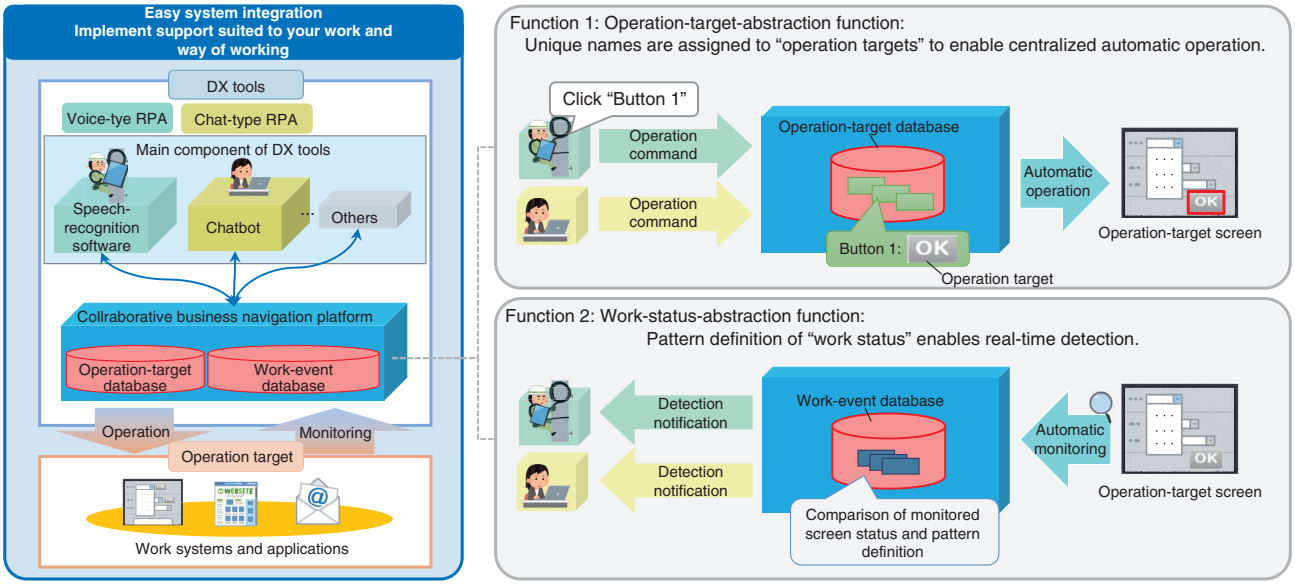


Fig. 2. Technical overview of collaborative business navigation platform.

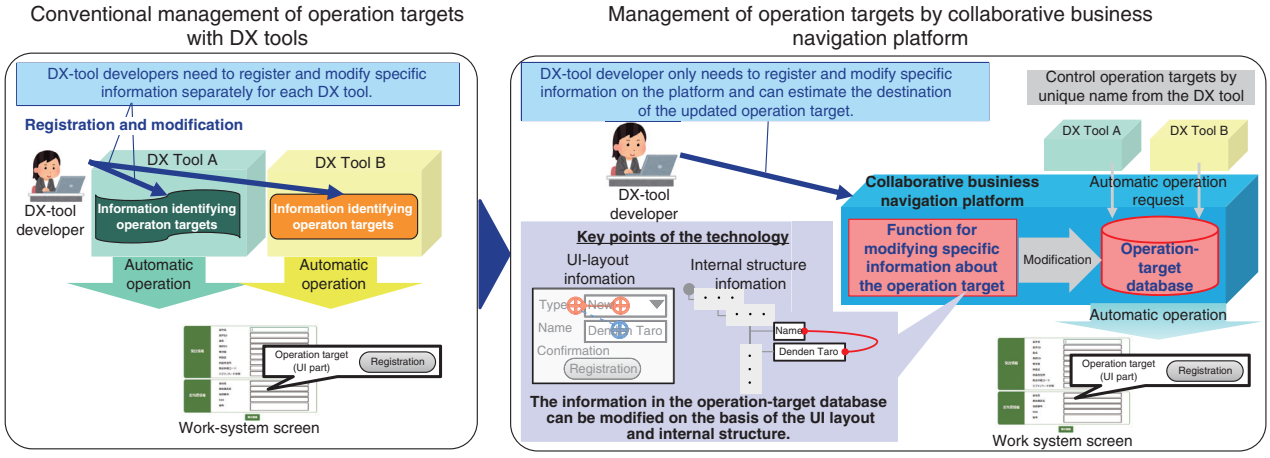


Fig. 3. Operation-target-abstraction function.

an automatic-operation function, as shown in **Fig. 3**, the developer had to manage the registration and modification of information identifying the operation target for each DX tool separately. By using this platform, the developer can assign a unique, easy-to-read name to information that identifies the operation target and can centrally manage the corresponding information in the operation-target database on the platform. It is also possible to automatically create an operation-target database by extracting candidates for operation targets from the screens of work systems and applications [1]. When updating a work system or application, it is possible to estimate the destination of changes in operation targets in accordance with the UI layout of the screen and the internal structure and enable the modification of information registered in the operation-target database.

3.2 Work-status-abstraction function

The ideal way to ensure efficient work is to take early action in response to ever-changing work conditions. However, work processes become complex and

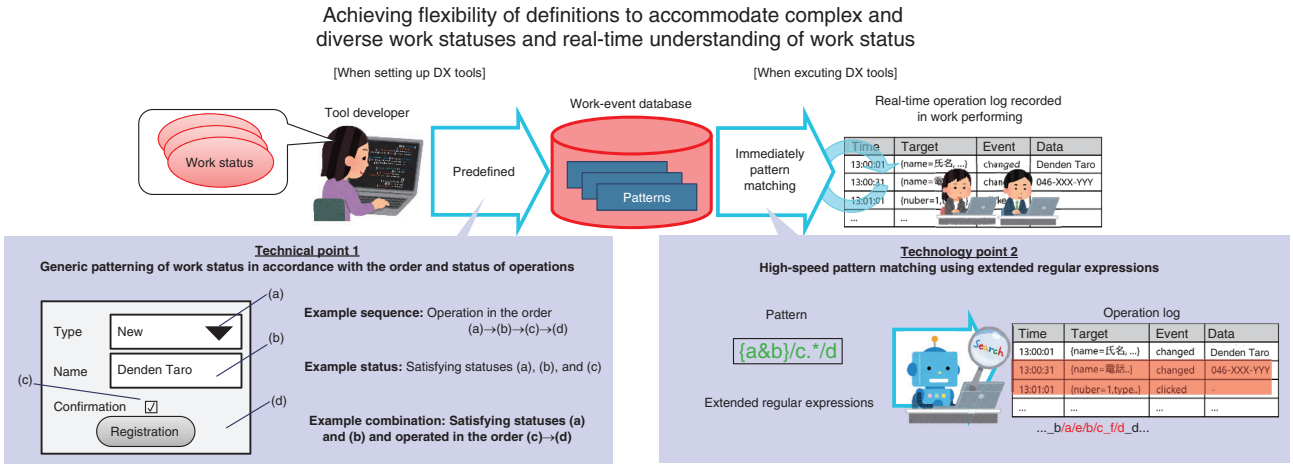


Fig. 4. Work-status-abstraction function.

diverse. To determine the work status from conventional DX tools accurately and in real time, it was necessary to implement individual mechanisms. Thus, it was difficult to use each mechanism easily, and operators needed to determine the status. To maximize efficiency, we developed a mechanism that enables real-time understanding of the status of work using DX tools.

The platform determines the work status by monitoring changes in the operation targets of work systems and applications displayed on the screen. By defining the work status to be detected as an "event" in advance and registering it in the work-event database on the platform, the work status can be determined in real time when a DX tool is executed. As shown in **Fig. 4**, events are defined as patterns (monitoring conditions) that combine the "operation order" logged from PC terminals, etc. and the "status" such as screen content. The work-status-abstraction function is defined by using the extended regular expression*6, and its flexibility allows for the expression of complex and diverse work situations. When a DX tool is executed, the work status is determined by pattern matching of events and operation logs recorded by monitoring the screen. A regular expression engine can be adapted to speed up the pattern matching and provide a real-time understanding of the work status [2].

4. Developer's tool that supports visual configuration

For the operation-target-abstraction function, the setup screen shown in **Fig. 5** is used to obtain information that can identify the operation target of the work system and application. Multiple methods (cursor position, click, focus, etc.) can be used in accordance with particular complex work environments. By giving the operation target a unique name, it can then be registered in the operation-target database.

The work-status-abstraction function has a setup screen, as shown in **Fig. 6**. The basic monitoring conditions (basic conditions) are set by selecting the "operation status" to be monitored (existence and display of the target, value settings, position changes, etc.) for the screen elements registered in the operation-target database. If the developer/user wants to monitor a series of operations rather than a single operation, they can set a compound of conditions (e.g., "basic condition 1 followed by basic condition 2" or "basic condition 2 followed by basic condition 1") by considering the detection order of multiple basic conditions. By defining basic and compound conditions as events, it is then possible to register events in the work-event database.

By using functions based on NTT Access Network Service Systems Laboratories' technology, as described above, the costs required to develop and manage DX tools can be greatly reduced, and a wide range of operating situations can be determined in real time, making it possible to easily develop DX tools tailored to various work environments and roles.

*6 Regular expression: A method of expressing variations of strings that follow certain rules in a single string. A technique used primarily in pattern matching.

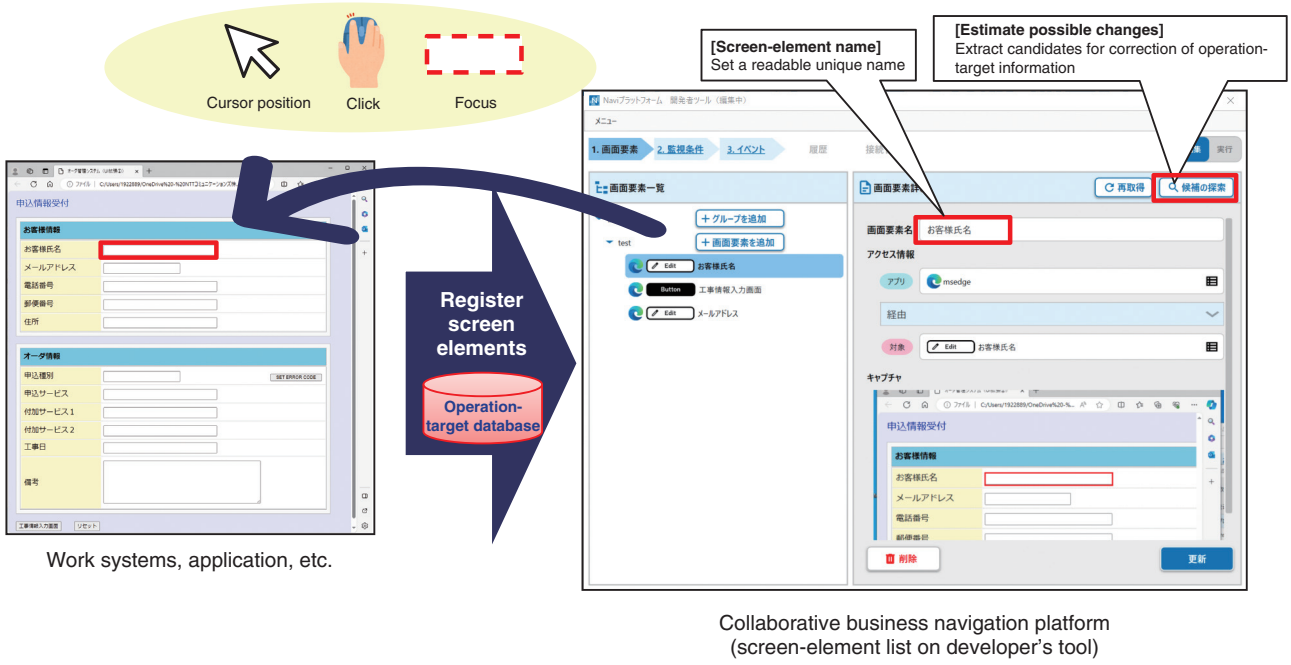


Fig. 5. Operation-target-abstraction function on developer's tool.

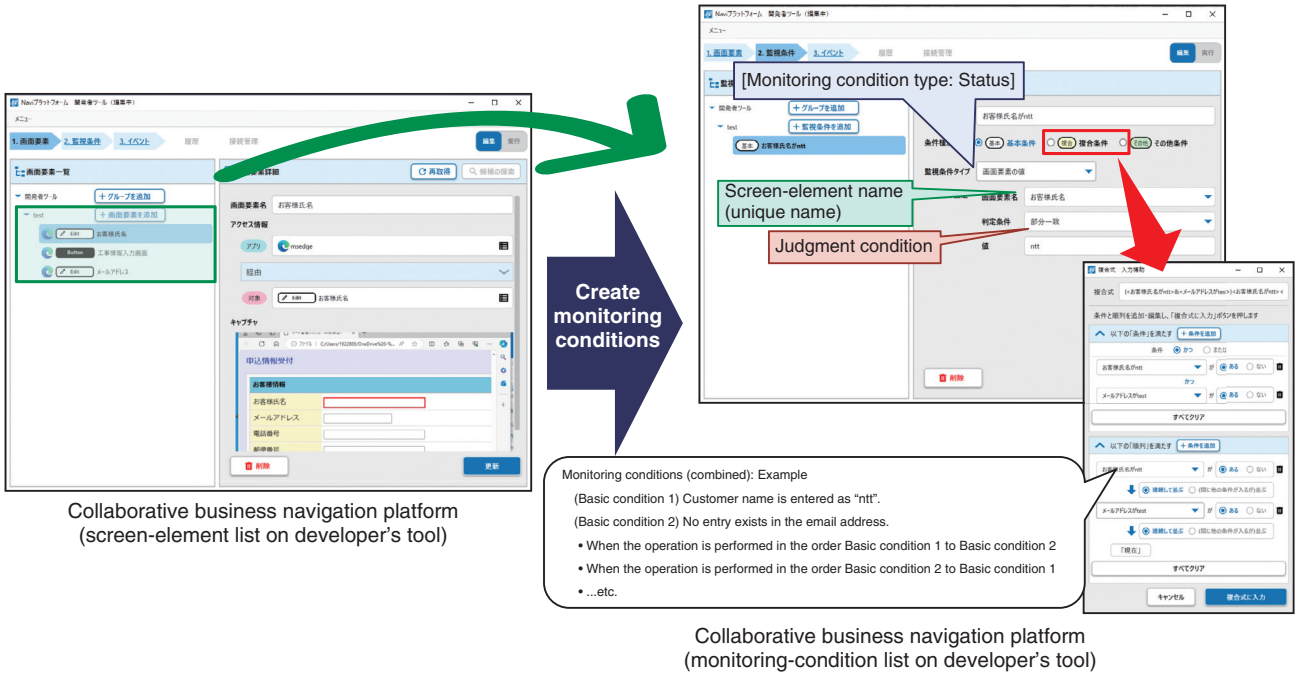


Fig. 6. Work-status-abstraction function on developer's tool.

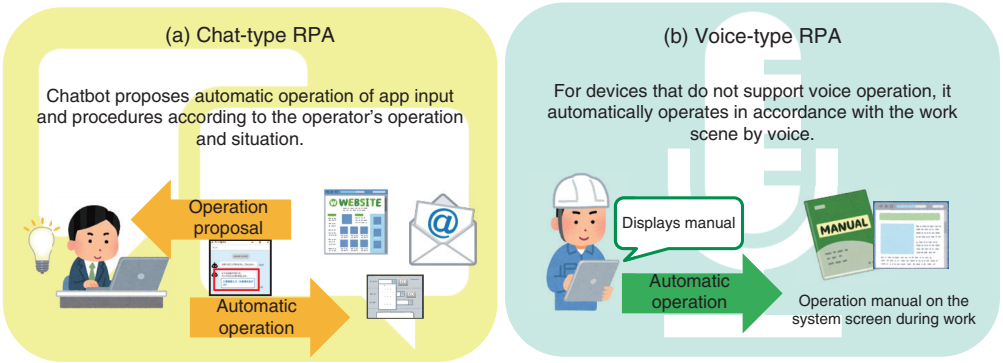


Fig. 7. DX-tool usage scenarios using collaborative business navigation platform.

5. Example usage of DX tools using the collaborative business navigation platform

5.1 Proposal-based operation support combined with chatbots

This example shows a DX tool that operators use for carrying out OA work. As the number of operations (scenarios) automated by RPA increases (due to consolidation and diversification of work), it is becoming cumbersome for operators to understand which scenarios to execute for each work status, and it is inefficient for them to execute scenarios. To achieve high operational efficiency independent of the operator's skills and experiences, a mechanism is required that can automatically propose the most-appropriate scenario in accordance with the work status.

Combining a chatbot with the platform enables the operator to receive operation suggestions from the chatbot in accordance with the operator's work status detected in real time by the work-status-abstraction function. The operator can proceed with automation scenarios while interacting with the chatbot. While carrying out the job interactively, the operator can acquire knowledge on their own, so it becomes possible to achieve high and stable job quality regardless of years of experience or work experience (**Fig. 7(a)**).

5.2 Voice-based operation support combined with voice UI

This example shows a DX tool used by operators in the field. When operators mainly work outdoors, they often operate work systems and applications on mobile devices while handling other tools. If they operate the device for long periods, especially in bad weather, the device might malfunction. Therefore, it is necessary to provide a DX tool with an appropriate UI that can address the above issues, rather than the traditional UI using touch, keyboard, and mouse operation.

Combining speech recognition and text-to-speech software with the collaborative business navigation platform enables the operator to operate work systems and applications by voice command without seeing the screen (**Fig. 7(b)**). Making it possible to call up manuals and input information into work systems and applications by voice leads to improved job efficiency in work environments in which it is difficult to operate devices using a mouse, keyboard, or touch UI.

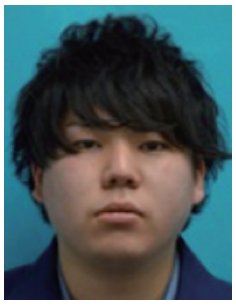
6. Future work

The collaborative business navigation platform makes it possible to provide operators with DX tools suited to their various work environments and roles simply by changing the upper-level applications. This platform is scheduled to be introduced within the NTT Group in 2024. We then plan to promote its general commercialization.

References

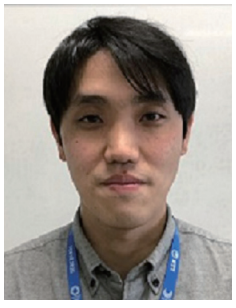
[1] H. Koya, M. Komiyama, A. Kataoka, and H. Oishi, "Proposal and Evaluation of a Target Extraction and Mapping Method for Automatic System Operation," IEICE Technical Report, Vol. 121, No. 13, ICM2021-8, pp. 41–46, 2021 (in Japanese).

[2] H. Koya, A. Kataoka, and H. Oishi, "Proposal and Evaluation of a Query Representation and Retrieval Method for Operation Logs," IEICE Technical Report, Vol. 121, No. 399, ICM2021-47, pp. 29–34, 2022 (in Japanese).



Hiroki Koike
Researcher, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received a B.E and M.E. in civil and environmental engineering from Waseda University, Tokyo, in 2018 and 2020. He joined NTT EAST in 2020 and was engaged in communications construction. He joined NTT Access Network Service Systems Laboratories in 2022 and has been researching business-operator psychology and navigation technology for business operation. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Hideaki Tanaka
Researcher, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received a B.E and M.E. in machine learning from Doshisha University, Kyoto, in 2012 and 2014. He joined NTT WEST in 2014 and was engaged in network operation. He joined NTT Access Network Service Systems Laboratories in 2021 and has been researching a remote operation assistant system and navigation technology for business operation.



Hidetaka Koya
Senior Research Engineer, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received an M.E. in computational intelligence and systems science from Tokyo Institute of Technology, Kanagawa, in 2011 and joined NTT Access Network Service Systems Laboratories the same year. He has been engaged in the research and development of business process automation and navigation technology. He is a member of IEICE.

Fumihiro Yokose
Senior Research Engineer, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received an M.E. in electrical engineering from the University of Electro-Communications, Tokyo, in 2007 and joined NTT Access Network Service Systems Laboratories the same year. He has been engaged in the research and development of navigation technology for business operation.

Hajime Nakajima
Senior Research Engineer, Supervisor, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received an M.E. in information engineering from the University of Tsukuba, Ibaraki, in 2005 and joined NTT Access Network Service Systems Laboratories the same year. He has been engaged in the research and development of operation support systems. He is a member of IEICE.

Haruo Oishi
Senior Research Engineer, Supervisor, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received a B.E and M.E. in electrical engineering from Tokyo Institute of Technology in 1995 and 1997 and received a Ph.D. in computer science and communications engineering from Waseda University, Tokyo, in 2022. He joined NTT in 1997. Since then, he has engaged in the research and development on educational technology, mobile PC middleware technology, business Ethernet service management system development, network failure estimation technology, and business navigation technology. He is currently with NTT Access Network Service Systems Laboratories as the leader of the Access Operation Project Collaboration Operation Technology Group. He is a senior member of IEICE and the chair of the IEICE Technical Committee on Information and Communication Management (ICM).

Work-improvement-support Technology That Supports Wide-ranging Implementation and Application of Digital Transformation Measures

Shinji Ogawa, Ryou Uchida, Misa Fukai, Ikuko Takagi, Masanobu Sakamoto, Haruo Oishi, Kimio Tsuchikawa, and Taisuke Wakasugi

Abstract

To make operations of an organization digital, i.e., digital transformation (DX), companies and local governments are promoting on-site efforts to improve business efficiency using digital technology and the implementation of measures by DX-promotion departments. To further enhance the effectiveness of DX promotion, it is necessary to create a DX cycle in which workers in the field and DX-promotion departments cooperate to efficiently and accurately implement effective DX measures over a wide area. The work-improvement-support technology developed by NTT to meet this necessity is introduced in this article.

Keywords: DX promotion, business analysis, operations

1. Importance of implementing DX measures in collaboration with workers in the field and DX-promotion departments

The use of digital technology is an effective means to achieve business transformation that can promote both operational efficiency and new-value creation in response to rapidly changing social and economic conditions. Companies and local governments are promoting organization-wide digital transformation (DX) of operations, such as field-led improvement of work efficiency through digital technology (field-led DX) and implementation of digital technology by DX-promotion departments.

Field-led DX has made progress in terms of automating and streamlining work procedures through the use of tools that automate desk work such as robotic process automation (RPA). However, field-led DX is often designed and implemented in a manner that is optimized for individual field operations; therefore, it is difficult for workers in other fields to determine whether the conditions for using field-led DX are compatible with their work. Currently, the only way to compare DX measures and field work is to observe and compare each manually, which requires many operations simply to select applicable examples from a variety of DX measures implemented at other fields. It will be important to improve business practices and procedures flexibly with a sense of urgency to meet the needs of society. This issue is faced by DX-promotion departments, and smooth rollout of DX measures is an essential element of fundamental

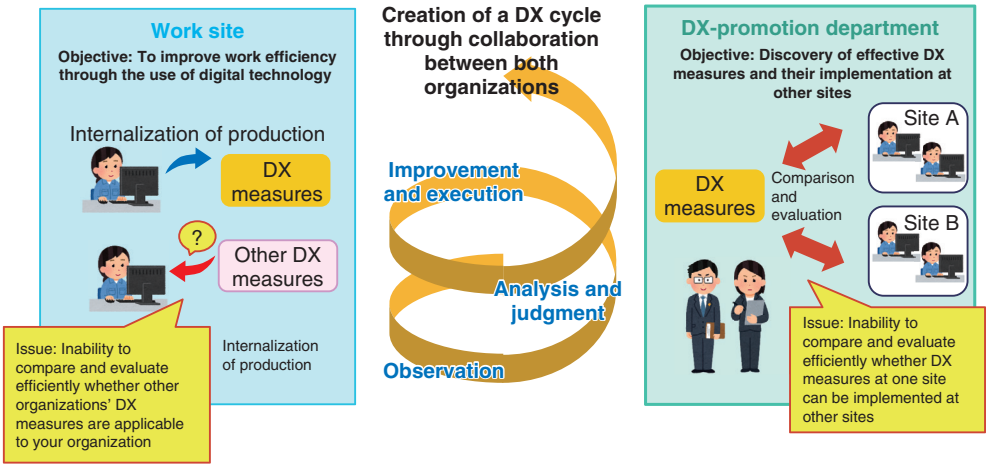


Fig. 1. DX cycle of collaboration between workers in the field and DX-promotion department.

business improvements that will have an impact throughout the company. Accordingly, to further increase the effectiveness of DX promotion, it is necessary to create a DX cycle in which workers in the field and the DX-promotion departments cooperate in a manner that efficiently and accurately implements effective DX measures in a wide range of areas (**Fig. 1**).

2. Overview of work-improvement-support technology

To address the above issues, NTT Network Innovation Center developed work-improvement-support technology, which is based on technology developed at NTT Access Network Service Systems Laboratories, that can be used immediately in the field and in DX-promotion departments. This technology uses historical information on operations carried out on personal computers (PCs) (operation logs) to quantitatively compare and visualize DX measures and the work procedures at the site that is considering applying the measures. The technology provides the following two tools that can be used immediately in the field or in DX-promotion departments.

- Operation-log-acquisition tool: A tool that records the operational history of the screen and graphical user interface (GUI) of a PC in real time as an operation log.
- Operation-log-analysis tool: A tool that reads data from the obtained operation logs, simplifies the logs as work procedures, quantitatively compares similarities in the work procedures, and visualizes the comparison results in an easily understood format.

Using these tools makes it possible to quantitatively compare DX measures and on-site work to which those measures are considered being applied by using operation-log data accumulated in real time from daily work. These tools thus make it easier to objectively determine the suitability of DX measures on the basis of the actual state of work operations.

3. Operation-log-acquisition tool

The operation-log-acquisition tool records operation history of the screen and GUI on PCs in real time as an operation log. The tool is installed in a PC in the Windows environment and records operation logs in file format according to operations by workers and automation tools such as RPA (see **Table 1** for operating environment). The tool detects user input operations and changes in application and window statuses on the terminal screen and captures the operations carried out by the user as operation events. This information can also be recorded and saved as log files (in text format) at an arbitrary location. The operation screen can be recorded and saved as a captured image upon detection of an operation event.

4. Operation-log-analysis tool

The operation-log-analysis tool compares the similarity of work procedures from the data of the obtained operation logs and visualizes them in an easy-to-understand format. It has the following three

Table 1. Operating environment of the operation-log-acquisition tool.

Item	Requirement
OS	Windows 10/11
CPU	Intel Core i3 2.9 GHz or more
Memory	4 GB or more
Applications that can be acquired	Microsoft Edge*1,3, Google Chrome*1,3, Firefox*1,3, Microsoft Excel*2,3, Windows OS*3 *1: Record changes to contents of GUI *2: Record changes to contents of cell *3: Window state changes, copy/paste operations, etc.

CPU: central processing unit
OS: operating system

Fig. 2. Screenshot of visualization screen of similar areas.

functions.

(1) Automatic extraction and visualization of similarities from work procedures

The similarity between work procedures is quantitatively evaluated from the two types of operation-log data, and the evaluation results are used to visualize similar and dissimilar points of procedures (see "Similarity assessment based on operation co-occurrence" below for details of the evaluation technique). The screenshot in **Fig. 2** is an example of analyzing and visualizing the similarities between the two types of operation-log data. For example, when the operation-log data concerning a DX measure are used as the comparison standard, and the operation-log data concerning the work at the site where the application of the DX measure is being considered are used as the comparison target, an operation range similar to that of the DX measure is extracted. This range is highlighted with a red border, so it is possible to see at a glance the similar points of work procedures. Similarly, dissimilar points can also be extracted by specifying the settings.

(2) Detailed comparison of similar work procedures

The similar operation ranges of each operation extracted by function (1) are arranged side-by-side, and the similarities between them are compared in detail. The screen shot in **Fig. 3** is an example of visualizing similar operations side-by-side. For example, if the background color of the operation logs to be compared is blue or green, the user can see at a glance that the operations are similar; if it is red, the operations are different (see **Table 2**). Each operation is converted into an explanation expressed in natural language that can be read and understood directly by the user (see "Generation of explanatory information about operation" below for details). This function makes it possible to understand the operations that can implement DX measures at the operational level.

(3) Generation of material for manuals from operation logs

By combining the text data of the operation history

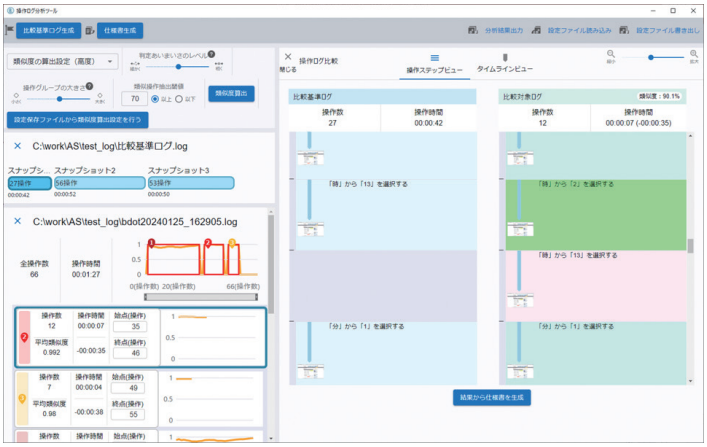


Fig. 3. Detailed screenshot of comparison screen of similar areas.

Table 2. Judging whether operations are similar or dissimilar.

Background color of the log to be compared	Explanation	Determination of identity of operation	Examples of operation logs
Blue ( )	When the operated screen and GUI parts are the same and the input data are the same	Same	• Comparison standard: Input value “1” is selected from the label name “minute” • Comparison target: Input value “1” is selected from the label name “minute”
Green ( )	When the operated screen and GUI parts are the same but the input data are different	Different	• Comparison standard: Input value “13” is selected from the label name “hour” • Comparison target: Input value “2” is selected from the label name “hour”
Red ( )	When the operated screen and GUI parts are different	Example of operation log	• Comparison standard: Input value “13” is selected from the label name “hour” • Comparison target: Input value “1” is selected from the label name “minute”

acquired as an operation log with screen-capture data, it is possible to generate a document file combining the work procedure and screen of the corresponding work target as material for the manual. This function makes it possible to identify specific work procedures related to DX measures and on-site work and can be used as an aid for creating manuals used in implementing DX measures.

5. Features of NTT technology

5.1 Similarity assessment based on operational co-occurrence

Operation logs often contain fluctuations due to, for example, rework or interruptions by other operations, which depend on the situation during the operation. Therefore, NTT Access Network Service Systems Laboratories developed a technology for evaluating the similarity of operations by using operational co-occurrence, which means that operations involved in similar work show mutual co-occurrence. This technology absorbs fluctuations in operations, correctly evaluates the similarity of operations, and exactly matches the range of similarity if they are similar, even if the operations are not exactly the same.

The similarity-assessment technology based on operational co-occurrence sets one operation log as a reference log and extracts parts that are similar to the reference log from another operation log containing the operation target (target log).

This technology involves the following three steps (**Fig. 4**).

(1) Information on operation events located closely together and near the reference log is

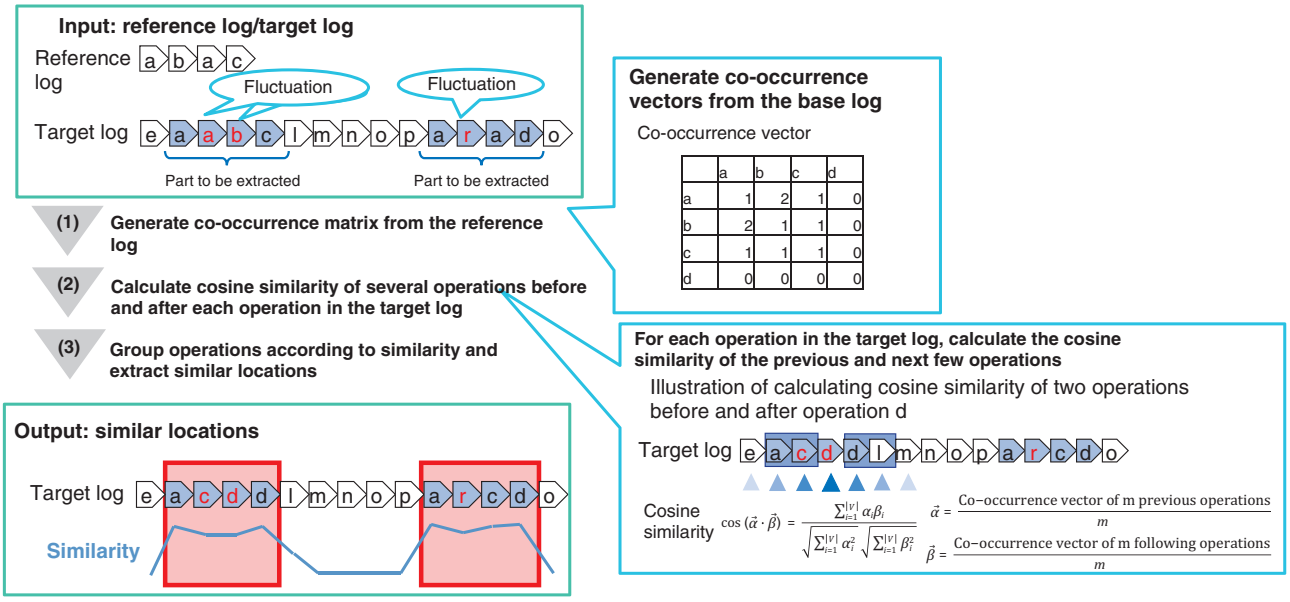


Fig. 4. Extraction of location similarity by operation co-occurrence.

stored as a co-occurrence matrix.

(2) The cosine similarity of the previous and next few operations from a given operation in the target log is calculated using the information on the co-occurrence vector of the reference log created in step (1). This similarity is calculated for all operations in the target log. The key point of this technology is that it uses the co-occurrence vector and similarity between several operations before and after the target operation in the target log. By appropriately rounding off the sequential order in the operation log, it is possible to calculate a similarity close to the semantic similarity recognized by people.

(3) The operations are grouped in accordance with their similarity and extracted as matching and mismatching regions.

5.2 Generation of explanatory information about operations

The operation log records information that indicates operation history, such as the time the operation was carried out and GUI-component information, but it is in machine language, which is unsuitable for viewing and understanding by people. Therefore, NTT Access Network Service Systems Laboratories developed a technology that can be used as explanatory information by mapping labels to GUI components. The technology takes advantage of the fact that labels that are easy for people to understand are expressed next to the GUI components. This makes it possible to generate text explaining operation procedures automatically from operation logs in a manner that makes it possible to visualize the operation without much effort. Logs can be obtained not only for operations carried out by people but also for operations executed with automation tools such as RPA, and explanatory information can be generated to facilitate understanding of the execution content of automated operation scenarios.

This technology involves the following three steps (**Fig. 5**).

(1) On the basis of the format recorded in the acquired operation log, rules are used to determine the appropriate sentence structures for the explanatory text and the parts of speech that make up the text.

(2) Two elements, parent-child relationships in the HTML and positional relationships on the display screen, are focused on, and the correspondence with nouns (labels) that describe the GUI components at the operation location is inferred from information around the GUI components on the screen.

(3) Explanatory text based on the sentence pattern determined in step (1) and the label obtained in step (2) is generated.

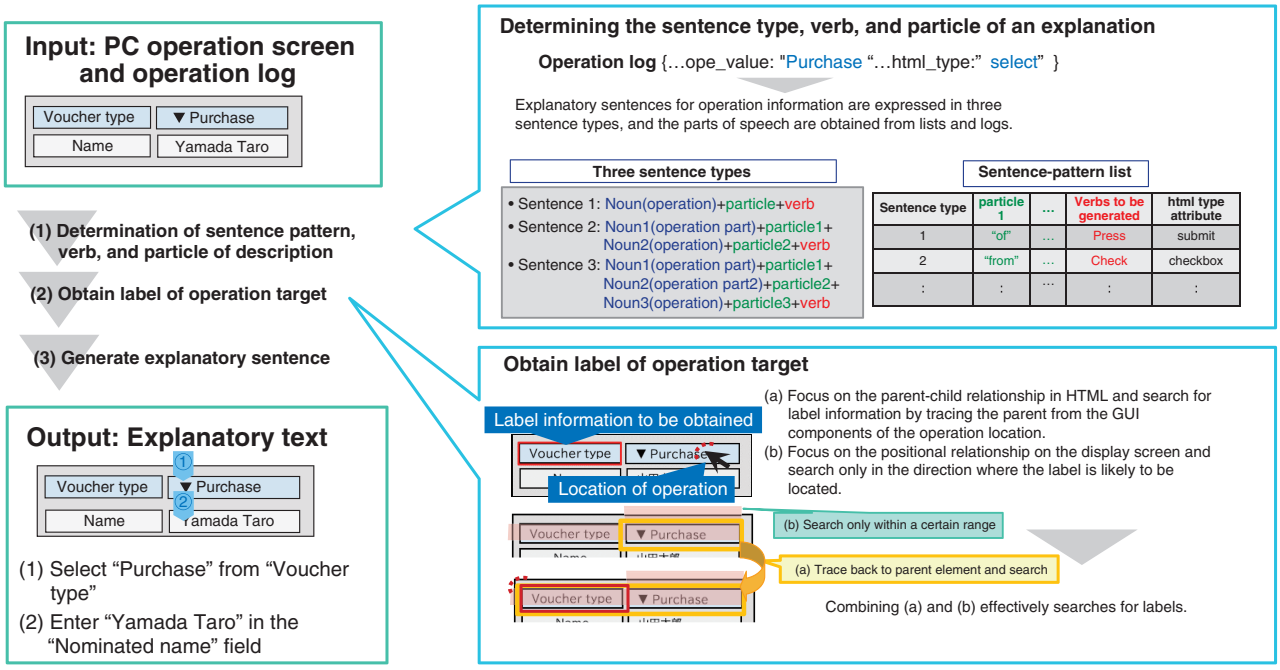


Fig. 5. Extraction of explanatory information describing operation.

6. Use cases

Three example situations in which the two above tools can be used are given below.

(1) Creation of DX cycle through collaboration between the workers in the field and DX-promotion departments

DX-promotion departments can use these tools to quantitatively compare the work involved in field-led DX measures and work involved in the field that is considering implementing the measures, and they can analyze and determine whether the measures are compatible. The tools thus enable efficient and accurate implementation of DX measures in line with on-site operations.

(2) Expanding the scope of automation by evaluating the applicability of RPA scenarios

By comparing work automated by tools such as RPA as DX measures with manual work, it is possible to identify the scope of the RPA scenario. It thus becomes possible to identify similar areas as candidates for RPA implementation and consider further improvements in business efficiency for the areas that differ.

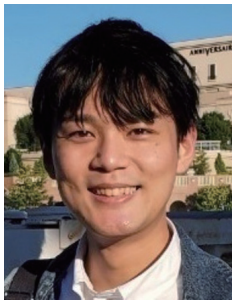
(3) Assistance in improving work skills through comparison with more-efficient employees

By using these tools to compare differences in work procedures executed by employees with high work efficiency, we can identify effective work carried out by certain employees that is not being carried out by other employees and use it as a teaching tool to improve the skills of the latter employees. Using this teaching tool will help create new DX measures that will improve overall operational efficiency.

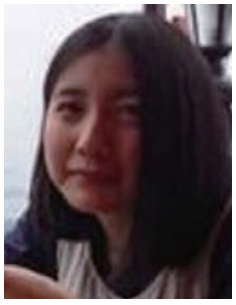
7. Future developments

Work-improvement-support technology for creating a DX cycle in which workers in the field and the DX-promotion departments cooperate to efficiently and accurately implement effective DX measures was introduced. This technology has just completed product development and can now be used for various DX-promotion measures. We plan to promote application of this technology to actual business operations and commercialization in the general market.

Shinji Ogawa
Senior Research Engineer, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.



Ryou Uchida
Researcher, Navigation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received a B.E. and M.E. from Chuo University, Tokyo, in 2017 and 2019. He joined NTT EAST in 2019 and was engaged in communications construction. He joined NTT Access Network Service Systems Laboratories in 2021 and has been researching business operation analysis methods. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).



Misa Fukai
Researcher, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.

Ikuko Takagi
Senior Research Engineer, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.

Masanobu Sakamoto
Senior Research Engineer, Intelligent Service Architecture Design Group, Human Insight Laboratory Project, NTT Human Informatics Laboratories.

Haruo Oishi
Senior Research Engineer, Supervisor, Collaboration Operation Technology Group, Access Network Operation Project, NTT Access Network Service Systems Laboratories.
He received a B.E and M.E. in electrical engineering from Tokyo Institute of Technology in 1995 and 1997 and received a Ph.D. in computer science and communications engineering from Waseda University, Tokyo, in 2022. He joined NTT in 1997. Since then, he has engaged in the research and development (R&D) on educational technology, mobile PC middleware technology, business Ethernet service management system development, network failure estimation technology, and business navigation technology. He is currently with NTT Access Network Service Systems Laboratories as the leader of the Access Operation Project Collaboration Operation Technology Group. He is a senior member of IEICE and the chair of the IEICE Technical Committee on Information and Communication Management (ICM).

Kimio Tsuchikawa
General Manager, Planning Department, NTT Information Network Laboratory Group.
He received a B.E. and M.E. in applied physics from Nagoya University, Aichi, in 2000 and 2002. He joined NTT in 2002 and engaged in R&D of navigation technology for business operation. He is a member of IEICE.

Taisuke Wakasugi
Section Manager, Business Promotion Department, NTT FIELDTECHNO CORPORATION.
He received an M.E. in materials science from Japan Advanced Institute of Science and Technology, Ishikawa, in 1998. He joined NTT the same year and engaged in R&D on the IPv6 domain name system server and navigation technology for business operation.

ITU-T SG16 Meeting Report

Jiro Nagao

Abstract

The International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Study Group 16 (SG16) meeting was held in Rennes, France, in April 2024 with online participation (hybrid). NTT Group made a proposal to Question 8 to initiate a new work item on "first-person transfer immersive live experience," which enables users to share (feel) the sensation of the performer at a remote event site (e.g. the piano player in a music concert). This article mainly reports on the discussion and result of Question 8.

Keywords: ITU-T, first-person transfer immersive live experience (FT-ILE), FEEL TECH

1. ITU-T SG16 meeting, April 2024, Rennes, France

The International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Study Group 16 (SG16) meeting was held 15–26 April 2024 in Rennes, France with remote participation (hybrid). The author participated in the meeting online as an associate rapporteur for Question 8 (Q8) of Working Party 3 (WP3). **Tables 1** and **2** show the management team of Q8 and the list of Contributions to Q8, respectively. Among the 11 Contributions, 6 were additions and revisions to the existing draft Recommendations, and 5 were proposals to initiate new work items. The discussion and results are described below with emphasis on the proposal from NTT Group to initiate a new work item on first-person transfer immersive live experience (FT-ILE).

Table 1. Management team of ITU-T SG16 Q8.

Position	Name	Affiliation (Country)
Rapporteur	Hideo Imanaka	NICT (Japan)
Associate Rapporteur	Hoerim Choi	KT (Korea (Rep. of))
Associate Rapporteur	Jiro Nagao	NTT (Japan)

2. Discussion summary and results of ITU-T SG16 Q8 meeting

2.1 Proposal of draft Recommendation H.ILE-FT (SG16-C400)

A proposal to initiate a new draft Recommendation (H.ILE-FT: An architectural framework for FT-ILE) was submitted to internationally standardize NTT DOCOMO's FEEL TECH technology [1]. The Contribution proposed to describe the requirements, functional components, and architectural framework of FT-ILE in the draft Recommendation. "First-person ILE" is first proposed in the Contribution. This is a new type of ILE in which the audience can experience first-person sensation. At a piano concert, for example, a remote user can experience vision, sound, haptic sensation, etc. of the actual piano player. In the conventional ILE, a remote user can experience the piano concert as if they were among the audience at the concert venue. This can be called third-person ILE. **Figures 1** and **2** show the conventional third-person ILE and proposed first-person ILE, respectively.

In the example of transmission of haptic sensation of a pianist, a remote user can experience the pianist's haptic sensation at the tip of their fingers. The sensation can differ from person to person, affected by their sensitivity to stimulation, the size of their hands, etc. Therefore, data and processing to adjust such differences are necessary. First-person ILE uses data that are more closely related to the pianist and the

Table 2. List of Contributions to ITU-T SG16 Q8.

Contribution	Summary	Source
SG16-C401-R1	Editorial revision of H.IIS-FA	NICT
SG16-C610	Editorial revision and proposal for consent of H.IIS-FA	China Telecom
SG16-C452	Addition to H.ILE-QR Clause 7	China Telecom
SG16-C527	Addition to H.ILE-QR	China Telecom, China Unicom
SG16-C454	Addition to requirements of H.ILE-AMR	China Telecom
SG16-C609	Addition to H.ILE-AMR Clauses	OKI
SG16-C400	Proposal of new draft Recommendation H.ILE-FT	NTT DOCOMO
SG16-C419	Proposal of new draft Recommendation H.ILE-3DIT	NICT
SG16-C477	Proposal of new draft Recommendation H.ILE-ER	China Telecom, China Unicom
SG16-C587-R1	Proposal of new draft Recommendation F.ARSArch	China Telecom, China Unicom, ICT-CAS
SG16-C592-R1	Proposal of new draft Recommendation H.3D-INR	China Telecom, MIIT

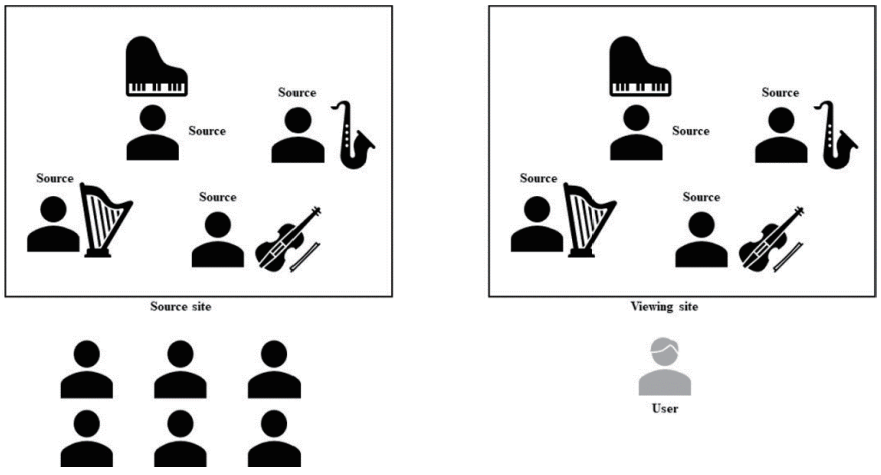


Fig. 1. Conventional ILE (third-person ILE).

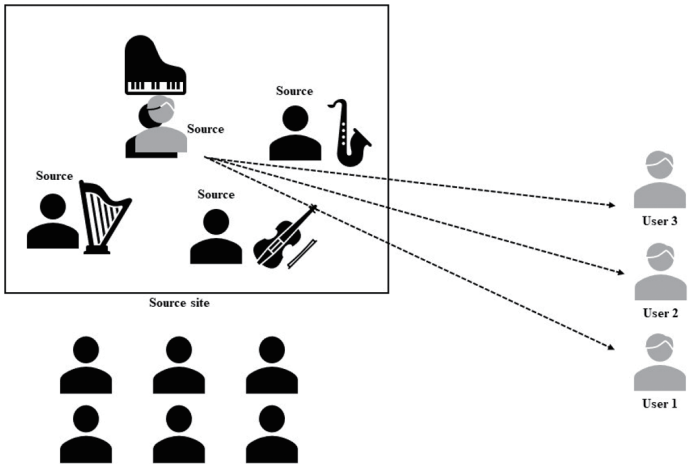


Fig. 2. Proposed ILE (first-person ILE).

users than the conventional third-person ILE, necessitating a different architecture. With those considerations as a background, a new draft Recommendation was proposed.

The proposal was discussed at the meeting, and the initiation was agreed with comments. Support of the initiation was offered from KT (formerly, Korea Telecom), National Institute of Information and Communications Technology (NICT), and China Telecom during the meeting in addition to the source of the Contribution, NTT DOCOMO and NTT. Ms. Nishio (NTT DOCOMO) was appointed as an editor along with one from China Telecom.

2.2 Proposal of H.ILE-3DIT (SG16-C419)

Initiation of a new draft Recommendation H.ILE-3DIT (Functional requirements and frameworks of three-dimensional (3D) model-based immersive telepresence system) was proposed by NICT. It proposes to clarify the functional requirements and frameworks of a 3D model-based immersive telepresence system in which a 3D model of each remote participant is constructed from image information and displayed in real time at appropriate locations with appropriate posture in a shared 3D space. This is in line with an ILE service scenario of a remote meeting described in ITU-T Recommendation H.430.3 (ILE service scenario). The initiation was agreed.

2.3 Other proposals and discussion results

OKI and NICT contributed proposals to revise the existing draft Recommendations H.IIS-FA (Functional architecture of interactive immersive services system) and H.ILE-AMR (Framework of ILE using multiple autonomous multimedia-enhanced mobile robots), respectively. These were agreed after discussion. China Telecom and others proposed initiation of three new draft Recommendations. Two (SG16-C587-R1, SG16-C592-R1) were agreed with revisions to their titles (H.ILE-AR, H.ILE-3DINR). The other was related to Q26 of SG16. After consultation with Q26, it was agreed to continue collaboration with Q26, and the initiation was postponed. H.IIS-FA was consented. Other revision proposals to the existing draft Recommendations were discussed and agreed.

3. Conclusions

New draft Recommendation H.ILE-FT, which aims for international standardization of NTT DOCOMO's FEEL TECH technology, was initiated. NTT will collaborate with NTT DOCOMO to enrich the draft Recommendation. The current study period of ITU-T is in its final year. This means the World Telecommunication Standardization Assembly (WTSA), ITU-T's highest decision-making body, will be held (this time in India in October 2024). It will be the first WTSA for the current Director of the ITU Telecommunication Standardization Bureau, Mr. Onoe, a former NTT executive. NTT will collaborate with ITU-T further for a successful WTSA.

Reference

[1] Press release issued by NTT DOCOMO, "DOCOMO Announces World's First Technology that Utilizes Human-Augmentation Platform for Sharing Haptic Information Between People," Jan. 25, 2023. https://www.docomo.ne.jp/english/info/media_center/pr/2023/0125_00.html

Jiro Nagao
Senior Research Engineer, Digital Twin Computing Laboratory, NTT Human Informatics Laboratories.

He received a Ph.D. in information science from Nagoya University, Aichi, in 2007 and joined NTT the same year. From 2007 to 2011, he was engaged in research and development of image-processing and content-distribution technology. From 2012 to 2017, he worked for NTT Communications, serving as the technical leader of commercial video-streaming services. From 2017 to 2021, he was engaged in research and development of immersive media and presentation technology at NTT Service Evolution Laboratories. From 2022 to July 2024, his mission was to promote standardization activities for global deployment of NTT research and development (R&D) technologies and services at Standardization Office, R&D Planning Department, NTT Corp. He has been engaged in R&D at NTT Human Informatics Laboratories since August 2024. He is currently an associate rapporteur of ITU-T Study Group 16 Question 8 (Immersive Live Experience) since 2022.

External Awards

Achievement Award
Winners: Shinji Matsuo, NTT Device Technology Laboratories; Koji Takeda, NTT Device Technology Laboratories; Takuro Fujii, NTT Device Technology Laboratories
Date: June 6, 2024
Organization: The Institute of Electronics, Information and Communication Engineers (IEICE)

For pioneering research on membrane lasers on Si.

Outstanding Catalyst - Beyond Telco
Winner: Autonomous Network Hyperloops Phase V Team (Antel, Celfocus, Chunghwa Telecom, Cognizant, ESRI, Futurewei, Infosim, Intersec, MTN, NTT, Orange, Telecom Italia, UBiqube, and Verizon)
Date: June 20, 2024
Organization: TMF

For "Autonomous Networks Hyperloops - Phase V - Virtual Command Center as a Service."

The 35th Radio Achievement Award, Minister of Internal Affairs and Communications Award
Winner: Doohwan Lee, NTT Network Innovation Laboratories
Date: June 25, 2024
Organization: Association of Radio Industries and Businesses (ARIB)

For research and development of OAM-MIMO wireless multiplexing technology.

Best Demo Award
Winners: Hiroki Baba, NTT Network Service Systems Laboratories; Shiku Hirai, NTT Network Service Systems Laboratories; Kentarou Hayashi, NTT Network Service Systems Laboratories; Tomonori Takeda, NTT Network Service Systems Laboratories
Date: June 27, 2024
Organization: The 10th IEEE International Conference on Network Softwarization (IEEE Netsoft 2024)

For "In-network Computing Architecture for Service Acceleration for 6G Networks."
Published as: H. Baba, S. Hirai, K. Hayashi, and T. Takeda, "In-network Computing Architecture for Service Acceleration for 6G Networks," IEEE Netsoft 2024, St. Louis, MO, USA, June 2024.

Papers Published in Technical Journals and Conference Proceedings

Distribution of Control during Bimanual Movement and Stabilization
A. Takagi and M. Kashino
Scientific Reports, Vol. 14, 16506, July 2024.

In two-handed actions like baseball batting, the brain can allocate the control to each arm in an infinite number of ways. According to hemispheric specialization theory, the dominant hemisphere is adept at ballistic control, while the non-dominant hemisphere is specialized at postural stabilization, so the brain should divide the control between the arms according to their respective specialization. Here, we tested this prediction by examining how the brain shares the control between the dominant and non-dominant arms during bimanual reaching and postural stabilization. Participants reached with both hands, which were tied together by a stiff virtual spring, to a target surrounded by an unstable repulsive force field. If the brain exploits each hemisphere's specialization, then the dominant arm should be responsible for acceleration early in the movement, and the non-dominant arm will be the prime actor at the end when holding steady against the force field. The power grasp force, which signifies the postural stability of each arm, peaked at movement termination but was equally large in both arms. Furthermore, the brain predominantly used the arm that could use the stronger flexor muscles to mainly accelerate the movement. These results point to the brain flexibly allocating the control to each arm according to the task goal without adhering to a strict specialization scheme.

Efficient Fiber-inspection and Certification Method for Optical-circuit-switched Datacenter Networks
K. Anazawa, T. Inoue, T. Mano, H. Nishizawa, and E. Oki
Journal of Optical Communications and Networking, Vol. 16, No. 8, pp. 788–799, July 2024.

Datacenter networks (DCNs) consisting of optical circuit switches (OCSs) have been considered as a promising solution to dramatically improve their transmission capacity, energy efficiency, and

communication latency. To scale optical-circuit-switched DCNs (OCS DCNs), hierarchical OCSs with tens of thousands of optical fibers need to be installed, and they should be inspected before starting datacenter operations. Since traditional DCNs consist of electrical-packet switches (EPSs), the condition and cabling of fibers can be inspected easily by probing neighboring EPSs. However, OCS networks cannot be inspected in the same manner because OCSs cannot transmit and receive probe signals. Thus, we have had to attach and detach a light source and power meter (LSPM) to every switch for probing all the fibers, which takes weeks. This paper proposes an efficient method for inspecting and certifying fibers in an entire DCN without repeating LSPM reattachment. Our method is based on (1) theories on quickly estimating the fiber condition on the basis of the intensity of received probe signals, (2) the maximum allowable loss of each fiber derived from the transceiver budget used in operations, and (3) an algorithm that reduces the number of probes needed. The results from an extensive numerical evaluation indicate that our method inspected a DCN with 18,432 fibers in at most a day, whereas a baseline method involving repeated LSPM reattachment would take more than a week. We also confirmed that our method never produced false negatives and false positives under practical network conditions.