2025 PKU Workshop on Optimization Theory and Methods

June 24-26, 2025

Beijing, China

http://conference.bicmr.pku.edu.cn/meeting/index?id=125

2025 PKU Workshop on Optimization Theory and Methods

June 24-26, 2025

Beijing, China

http://conference.bicmr.pku.edu.cn/meeting/index?id=125

Information for Participants

Accommodation Information for Invited Guests

Hotel:	"Zhong Guan Xin Yuan" Global Village, Building 1 or 9 (中关新园1号楼或9号楼)
Address:	No. 126 Zhongguancun North Street, Haidian District 北京市海淀区中关村北大街126号
Dates:	By default, the hotel room is reserved from June 23rd (check in) to June 26th (check-out). Please let us know if you have a different arrival-departure schedule.
Website:	http://pkugv.pku.edu.cn
Tel:	+86-10-62752288

Hotel:	The Lakeview Hotel Beijing (北京北大博雅国际酒店)
Address:	No. 127 Zhongguancun North Street, Haidian District 北京市海淀区中关村北大街127号
Dates:	By default, the hotel room is reserved from June 23rd (check in) to June 27th (check-out). Please let us know if you have a different arrival-departure schedule.
Website:	https://www.thelakeviewbeijing.cn
Tel:	+86-20-86009099

Travel Information

You can choose one of the following ways to the hotels

- Take a taxi directly to the location mentioned above.
- Take Subway Line 4 to East Gate of Peking University Station.

Conference Venue

Venue:	Wenyuan Hall, Zhihua Building, Peking University 北京大学数学科学学院智华楼文远堂
Map:	PKU campus map
Arrival:	You can enter the campus through the southeast gate and walk to the venue.

Meals

- Breakfasts will be complimentary at the hotel.
- Lunches and dinners are provided by the workshop. Please let us know if you have any dietary restrictions or preferences.

Currency

Chinese currency is RMB. The current rate is about 7.20 RMB for 1 US dollar. The exchange of foreign currency can be done at the airport or the conference hotel. Please keep the receipt of the exchange so that you can change back to your own currency if you have RMB left before you leave China. Please notice that some additional processing fee will be charged if you exchange currency in China.

Parking at PKU Campus

If you plan to drive to PKU, please send us your license plate number; otherwise, your car cannot enter the PKU campus.

Contact Information

If you need any help, please feel free to contact

- Mrs.Congcong Zhao: zhaocongcong0613@bicmr.pku.edu.cn

Sponsors

Beijing International Center for Mathematical Research, Peking University

Center for Machine Learning Research, Peking University

School of Mathematical Sciences, Peking University

National Natural Science Foundation of China

Great Bay University (Great Bay Institute for Advanced Study)

MindOpt Team, Alibaba DAMO Academy

Organizing Committee

Guanghui Lan (Georgia Institute of Technology)

Zaiwen Wen (Peking University)

Wotao Yin (Alibaba DAMO Academy)

Kun Yuan (Peking University)

Yaxiang Yuan (Chinese Academy of Sciences)

Conference Schedule

Each talk is 30 minutes.

June 24th, Tuesday

08:50-09:00 Opening Ceremony: TBD

09:00-10:00 Session 1

- **09:00-09:30** **Jongshi Pang**, Heaviside Composite Optimization and Complementarity Constraints by a Progressive Integer Programming Method
  Chair: Yin Zhang
- **09:30-10:00** **Yao Xie**, Generative Models for Distributionally Robust Optimization (DRO): Iterative Algorithms in Continuous Probability spaces
  Chair: Lingchen Kong

10:00-10:30 Group Photo and Tea Break

10:30-12:00 Session 2

- **10:30-11:00** **Defeng Sun**, Accelerating Preconditioned ADMM for Solving Linear Programming and Composite Convex Quadratic Programming
  Chair: Yafeng Liu
- **11:00-11:30** **Jiawang Nie**, A Characterization for Tight Sparse Relaxations
  Chair: Yongjin Liu
- **11:30-12:00** **Benjamin Grimmer**, Beyond Minimax Optimality: A Subgame Perfect Gradient Method
  Chair: Yiju Wang

12:00-14:00 Lunch

14:00-15:00 Session 3

- **14:00-14:30** **Guanghui Lan**, Projected Gradient Methods for Nonconvex and Stochastic Optimization with Auto-Conditioned Stepsizes
  Chair: Wenxun Xing
- **14:30-15:00** **Liwei Zhang**, The Rate of Convergence of Augmented Lagrangian Method for Minimax Optimization Problems with Semidefinite Constraints
  Chair: Jiaxin Xie

15:00-15:30 Tea Break

15:30-17:00 Session 4

- **15:30-16:00** **Weijie Su**, PolarGrad: A Class of Matrix-Gradient Optimizers from a Unifying Preconditioning Perspective
  Chair: Jianlin Jiang
- **16:00-16:30** **Kun Yuan**, A Memory Efficient Randomized Subspace Optimization Method for Training Large Language Models
  Chair: Jinbao Jian

16:30-17:00 **Tiande Guo**, Generative Artificial Intelligence - Models and Algorithms
Chair: Chen Ling

17:00-20:00 Banquet

June 25th, Wednesday

09:00-10:00 Session 1

09:00-09:30 **Ernest Ryu**, A Duality Correspondence Between Optimization Algorithms
Chair: Haoyang Liu

09:30-10:00 **Jelena Diakonikolas**, On Efficient Solvers for Fixed-Point Equations: Classical Results and New Twists
Chair: Changjun Yu

10:00-10:30 Tea Break

10:30-12:00 Session 2

10:30-11:00 **Deren Han**, Connecting Randomized Iterative Methods with Krylov Subspaces
Chair: Shaohua Pan

11:00-11:30 **Chao Ding**, Adaptive Regularized Newton-CG for Nonconvex Optimization: Optimal Global Complexity and Quadratic Local Convergence
Chair: Yitian Xu

11:30-12:00 **Chenglong Bao**, Global Optimization for A Class of Nonconvex Problems
Chair: Xiaojiao Tong

12:00-14:00 Lunch

14:00-15:00 To Alibaba's Beijing Headquarters by Bus

15:00-15:30 Tour Around Alibaba's Beijing Headquarters

15:30-17:30 Session 3

15:30-16:00 **Zaiwen Wen**, Advancing Mathematical Formalization: Tools and Techniques for Lean
Chair: Liping Wang

16:00-16:30 **Wotao Yin**, Automated Problem Decomposition and Solution for Extended Monotropic Optimization
Chair: Minru Bai

16:30-17:30 **Discussion: How Industry-Academia Collaborations Create Impact?**

17:30-20:30 Dinner

20:30-21:30 Return to Peking University

June 26th, Thursday

09:00-10:00 Session 1

09:00-09:30 **Laurent Condat**, Communication-efficient Distributed Optimization Algorithms
Chair: Ziyan Luo

09:30-10:00 **Yong Xia**, Split-Merge: A Difference-based Approach for Dominant Eigenvalue Problem
Chair: Nanjing Huang

10:00-10:30 Tea Break

10:30-12:00 Session 2

10:30-11:00 **Congying Han**, Purity Law for Neural Routing Problem Solvers with Enhanced Generalizability
Chair: Shenglong Zhou

11:00-11:30 **Junyu Zhang**, A Few Issues on Stochastic Mirror Descent: Optimality Measure and Approach to Acceleration
Chair: Qi Ye

11:30-12:00 **Jiashuo Jiang**, Achieving Instance-dependent Guarantees for (Constrained) Markov Decision Process
Chair: Wenbao Ai

12:00-14:00 Lunch

Abstracts

Global Optimization for A Class of Nonconvex Problems

Chenglong Bao
Tsinghua University

Solving nonconvex optimization problems is a fundamental challenge with wide applications across many fields. Theoretically, identifying classes of nonconvex problems that admit global optimization is an important topic. In this talk, we propose an adaptive stochastic gradient descent algorithm for minimizing nearly convex functions. The algorithm features an adaptive noise injection mechanism based on the current function values and can be theoretically shown to converge to the global minimum at a linear rate. Additionally, I will introduce a zeroth-order method that achieve global convergence even when the objective function is nondifferentiable.

Communication-efficient Distributed Optimization Algorithms

Laurent Condat
King Abdullah University of Science and Technology

In distributed optimization and machine learning, a large number of machines perform computations in parallel and communicate back and forth with a server. In particular, in federated learning, the distributed training process is run on personal devices such as mobile phones. In this context, communication, that can be slow, costly and unreliable, forms the main bottleneck. To reduce communication, two strategies are popular: 1) local training, that consists in communicating less frequently; 2) compression. Also, a robust algorithm should allow for partial participation. I will present several randomized algorithms we developed recently, with proved convergence guarantees and accelerated complexity. Our most recent paper "LoCoDL: Communication-Efficient Distributed Learning with Local Training and Compression," has been presented at the International Conference on Learning Representations (ICLR) 2025, as a Spotlight.

On Efficient Solvers for Fixed-Point Equations: Classical Results and New Twists

Jelena Diakonikolas
University of Wisconsin-Madison

Fixed-point operator equations–where one seeks solutions to $T(x) = x$ for an operator T mapping a vector space to itself–form a fundamental class of problems with broad applications in optimization theory, game theory, economics, and, more recently, reinforcement learning. In this talk, I will begin by reviewing classical algorithmic results for solving such equations under oracle access to the operator T. I will then highlight key gaps in the literature, particularly in settings where T may be expansive–though in a controlled sense–or where access to T is limited to stochastic queries. Finally, I will present recent results that address some of these challenges and conclude with open questions and potential directions for future work.

Adaptive Regularized Newton-CG for Nonconvex Optimization: Optimal Global Complexity and Quadratic Local Convergence

Chao Ding

Chinese Academy of Sciences

Finding an ϵ-stationary point of a nonconvex function with a Lipschitz continuous Hessian is a central problem in optimization. However, existing regularized Newton methods face a fundamental trade-off: methods achieving optimal $O(\epsilon^{-\frac{3}{2}})$ global complexity typically exhibit only linear local convergence. In contrast, those with quadratic local rates often have suboptimal global performance. To bridge this long-standing gap, in this talk, we propose a new class of regularizers constructed using the conjugate gradient approach with a negative curvature monitor to solve the regularized Newton equation. The proposed algorithm is adaptive, requiring no prior knowledge of the Hessian Lipschitz constant, and achieves a global complexity of $O(\epsilon^{-\frac{3}{2}})$ in terms of the second-order oracle calls, and $\tilde{O}(\epsilon^{-\frac{7}{4}})$ for Hessian-vector products, respectively. When the iterates converge to a point where the Hessian is positive definite, the method exhibits quadratic local convergence. Preliminary numerical results, including training the physics-informed neural networks, illustrate the competitiveness of our algorithm.

Beyond Minimax Optimality: A Subgame Perfect Gradient Method

Benjamin Grimmer
Johns Hopkins University

This talk will take up the task of designing the provably best possible gradient method for smooth convex optimization. Methods with big-O optimal worst-case guarantees were (famously) discovered in the 80s by Nesterov. Methods with exactly minimax optimal worst-case guarantees were developed in the last decade. As the main new result, we will present a "subgame perfect" method that is not only optimal against a worst-case problem instance but also optimally leverages all gradient information revealed at each step. This corresponds to being dynamically minimax optimal, or in game theory terms, provides us with a subgame perfect strategy for optimization. Besides attaining this high standard (beyond minimax optimality), our subgame-perfect gradient method is also very fast, able to compete with the quasi-second-order method BFGS.

Generative Artificial intelligence - Models and Algorithms

Tiande Guo
University of Chinese Academy of Sciences

With the continuous development of deep learning and neural network technologies, generative artificial intelligence has achieved significant breakthroughs and demonstrated great potential in multiple application scenarios. This talk first constructs a unified mathematical framework for generative methods in artificial intelligence and systematically introduces its core technologies, including variational autoencoder (VAE), Generative Adversarial Networks (GAN), diffusion Model and Flow-based model. Meanwhile, it analyzes the advantages and limitations of different methods in various tasks. Furthermore, this talk presents the application prospects of generative methods in artificial intelligence in fields such as mathematics, physics, life sciences, medicine, computer science and engineering. Finally, summarizes the key challenges currently faced by generative methods in artificial intelligence and focuses on discussing their future development directions in mathematics and intelligent optimization research.

Purity Law for Neural Routing Problem Solvers with Enhanced Generalizability

Congying Han
University of Chinese Academy of Sciences

Achieving generalization in neural approaches across different scales and distributions remains a significant challenge for routing problems. A key obstacle is that neural networks often fail to learn robust principles for identifying universal patterns and deriving optimal solutions from diverse instances. In this presentation, we first uncover Purity Law, a fundamental structural principle for optimal solutions of routing problems, defining that edge prevalence grows exponentially with the sparsity of surrounding vertices. Statistically validated across diverse instances, Purity Law reveals a consistent bias toward local sparsity in global optima. Building on this insight, we propose Purity Policy Optimization (PUPO), a novel training paradigm that explicitly aligns characteristics of neural solutions with Purity Law during the solution construction process to enhance generalization. Extensive experiments demonstrate that PUPO can be seamlessly integrated with popular neural solvers, significantly enhancing their generalization performance without incurring additional computational overhead during inference.

Connecting Randomized Iterative Methods with Krylov Subspaces

Deren Han
Beihang University

Randomized iterative methods, such as the randomized Kaczmarz method, have gained significant attention for solving large-scale linear systems due to their simplicity and efficiency. Meanwhile, Krylov subspace methods have emerged as a powerful class of algorithms, known for their robust theoretical foundations and rapid convergence properties. Despite the individual successes of these two paradigms, their underlying connection has remained largely unexplored. In this talk, we develop a unified framework that bridges randomized iterative methods and Krylov subspace techniques, supported by both rigorous theoretical analysis and practical implementation. The core idea is to formulate each iteration as an adaptively weighted linear combination of the sketched normal vector and previous iterates, with the weights optimally determined via a projection-based mechanism. This formulation not only reveals how subspace techniques can enhance the efficiency of randomized iterative methods, but also enables the design of a new class of iterative-sketching-based Krylov subspace algorithms. We prove that our method converges linearly in expectation and validate our findings with numerical experiments.

Achieving Instance-dependent Guarantees for (Constrained) Markov Decision Process

Jiashuo Jiang
The Hong Kong University of Science and Technology

We consider the reinforcement learning problem for the constrained Markov decision process (CMDP), which plays a central role in satisfying safety or resource constraints in sequential learning and decision-making. In this problem, we are given finite resources and a MDP with unknown transition probabilities. At each stage, we take an action, collecting a reward and consuming some resources, all assumed to be unknown and need to be learned over time. In this work, we take the first step towards deriving optimal instance-dependent guarantees for the CMDP problems. We derive a logarithmic regret bound, which translates into a $O(1/(\Delta * \epsilon) * \log^2(1/\epsilon))$ sample complexity bound, with Δ being a problem-dependent parameter, yet independent of ϵ. Our sample complexity bound improves upon the state-of-art $O(1/\epsilon^2)$ sample complexity for CMDP problems established in the previous literature, in terms of the dependency on ϵ. To achieve this advance, we develop a new framework for analysing CMDP problems. To be specific, our algorithm operates in the primal space and we resolve the primal LP for the CMDP problem at each period in an online manner, with adaptive remaining resource capacities. The key elements of our algorithm are: i) a characterization of the instance hardness via LP basis, ii) an eliminating procedure that identifies one optimal basis of the primal LP, and; iii) a resolving procedure that is adaptive to the remaining resources and sticks to the characterized optimal basis. We further explore the extensions to RL problems with function approximation, developing new algorithms and achieving new instance-dependent guarantees. This is based on joint work with Prof. Yinyu Ye.

Projected Gradient Methods for Nonconvex and Stochastic Optimization with Auto-Conditioned Stepsizes

Guanghui Lan
Georgia Institute of Technology

We propose a new class of projected gradient (PG) methods for minimizing smooth, possibly nonconvex functions over convex compact sets. We first present a sharp complexity analysis of the standard PG method, matching the best known results for approximate stationarity. Building on this, we introduce an auto-conditioned PG (AC-PG) variant that adapts stepsizes using first-order information, eliminating the need for Lipschitz constants or line search, while retaining optimal complexity. We extend these ideas to stochastic settings, developing stochastic PG (SPG) and variance-reduced SPG (VR-SPG) methods, along with adaptive stepsize strategies, and establish new convergence guarantees under different oracle models.

A Characterization for Tight Sparse Relaxations

Jiawang Nie
University of California, San Diego

We study the sparse Moment-SOS hierarchy of relaxations for solving sparse polynomial optimization problems. We show that this sparse hierarchy is tight if and only if the objective can be written as a sum of sparse nonnegative polynomials, each of which belongs to the sum of the ideal and quadratic module generated by the corresponding sparse constraints. Based on this characterization, we give several sufficient conditions for the sparse Moment-SOS hierarchy to be tight. In particular, we show that this sparse hierarchy is tight under some assumptions such as convexity, optimality conditions or finiteness of constraining sets.

Heaviside Composite Optimization and Complementarity Constraints by a Progressive Integer Programming Method

Jongshi Pang
the University of Southern California

A univariate Heaviside function is the (discontinuous) indicator function of an (open or closed) interval. By its name, a Heaviside composite function is the composition of a Heaviside function with a multivariate function that may be nonconvex and nondifferentiable. Complementarity constraints are well known for their modeling breadth in many optimization problems, ranging from an indefinite quadratic program to bilevel optimization to structured sparsity selection. In principle, these two features: Heaviside functions and complementarity constraints, can be formulated by integer variables; yet the resulting optimization problems present great challenges for state-of-the-art integer programming methods beyond small-to-medium sizes, per solution by the well-known Gurobi solver. This talk presents an elementary yet very effective way, termed the progressive integer programming (PIP) method, to exploit the solver's capability and push it beyond what it can normally handle. Extensive computational results demonstrate PIP's great promise for solving many classes of nonconvex optimization problems that involve Heaviside composite functions and complementarity constraints, and beyond.

A Duality Correspondence Between Optimization Algorithms

Ernest Ryu
University of California, Los Angeles

Optimization theory is classically built upon a rich theory of dualities between spaces, sets, functions, and problems. In this talk, we present a different kind of duality correspondence, one between first-order optimization *algorithms*. We name this phenomenon H-duality and show how accelerated methods of one kind are H-duals of accelerated methods of another kind.

PolarGrad: A Class of Matrix-Gradient Optimizers from a Unifying Preconditioning Perspective

Weijie Su
University of Pennsylvania

The ever-growing scale of deep learning models and datasets underscores the critical importance of efficient optimization methods. While preconditioned gradient methods such as Adam and AdamW are the de facto optimizers for training neural networks and large language models, structure-aware preconditioned optimizers like Shampoo and Muon, which utilize the matrix structure of gradients, have demonstrated promising evidence of faster convergence. In this talk, we introduce a unifying framework for analyzing “matrix-aware” preconditioned methods, which not only sheds light on the effectiveness of Muon and related optimizers but also leads to a class of new structure-aware preconditioned methods. A key contribution of this framework is its precise distinction between preconditioning strategies that treat neural network weights as vectors (addressing curvature anisotropy) versus those that consider their matrix structure (addressing gradient anisotropy). This perspective provides new insights into several empirical phenomena in language model pre-training, including Adam’s training instabilities, Muon’s accelerated convergence, and the necessity of learning rate warmup for Adam. Building upon this framework, we introduce PolarGrad, a new class of preconditioned optimization methods based on the polar decomposition of matrix-valued gradients. As a special instance, PolarGrad includes Muon with updates scaled by the nuclear norm of the gradients. We provide numerical implementations of these methods, leveraging efficient numerical polar decomposition algorithms for enhanced convergence. Our extensive evaluations across diverse matrix optimization problems and language model pre-training tasks demonstrate that PolarGrad outperforms both Adam and Muon.

Accelerating Preconditioned ADMM for Solving Linear Programming and Composite Convex Quadratic Programming

Defeng Sun
The Hong Kong Polytechnic University

We aim to employ an accelerated preconditioned alternating direction method of multipliers (pADMM), whose proximal terms are convex quadratic functions, to solve linearly constrained convex optimization problems. To achieve this, we first reformulate the pADMM into a form of proximal point method (PPM) with a positive semidefinite preconditioner which can be degenerate due to the lack of strong convexity of the proximal terms in the pADMM. Then we accelerate the pADMM by accelerating the reformulated degenerate PPM (dPPM). Specifically, we first propose an accelerated dPPM by integrating the Halpern iteration and the fast Krasnosel'ski-Mann iteration into it, achieving asymptotic and non–asymptotic convergence rates. Subsequently, building upon the accelerated dPPM, we develop an accelerated pADMM algorithm that exhibits both asymptotic and non-asymptotic nonergodic convergence rates concerning the Karush-Kuhn-Tucker residual and the primal objective function value gap. Extensive numerical experiments on large-scale linear programming and convex composite quadratic programming benchmark datasets, conducted using a GPU, demonstrate the substantial advantages of our Halpern–Peaceman-Rachford (HPR) method—a special case of the Halpern–accelerated pADMM framework over state-of-the-art solvers, including the award-winning PDLP, as well as PDQP, SCS, CuClarabel, and Gurobi, in achieving high–accuracy solutions. [This talk is based on joint work with Kaihuang Chen, Yancheng Yuan, Guojun Zhang, and Xinyuan Zhao.]

Advancing Mathematical Formalization: Tools and Techniques for Lean

Zaiwen Wen
Peking University

This talk explores cutting-edge tools and techniques for mathematical formalization in Lean, beginning with optlib, a Lean library for mathematical optimization (github.com/optsuite/optlib), and ReasLab, an online collaborative IDE for Lean (alpha.reaslab.io). We then present two methodological advances: (1) tree-based premise selection, which uses the core representation of Lean for automate proof premise discovery, and (2) a framework for translating informal proofs into formal proofs via a chain of states, systematically transforming human-written arguments into machine-verifiable proofs. Together, we aim to accelerate formalization workflows and broaden the accessibility of mathematical proof.

Split-Merge: A Difference-based Approach for Dominant Eigenvalue Problem

Yong Xia
Beihang University

The computation of the dominant eigenvector of symmetric positive semidefinite matrices is a cornerstone operation in numerous optimization-driven applications. Traditional methods, typically based on the Quotient formulation, often suffer from challenges related to computational efficiency and reliance on prior spectral knowledge. In this work, we leverage the alternative Difference formulation to reinterpret the classical power method as a first-order optimization algorithm. This perspective allows for a novel convergence analysis and facilitates the development of accelerated variants with larger step-sizes, achieving faster convergence without additional computational cost. Building on this insight, we introduce a generalized family of Difference-based methods, with the power method as a special case. Within this family, we propose Split-Merge, an algorithm that attains accelerated convergence without requiring spectral knowledge and operates solely via matrix-vector products. Extensive experiments on both synthetic and real-world datasets demonstrate that Split-Merge consistently outperforms state-of-the-art methods in both efficiency and scalability. In particular, it achieves more than a $10\times$ speedup over the classical power method, underscoring its practical effectiveness for large-scale problems.

Generative models for Distributionally Robust Optimization (DRO): Iterative algorithms in continuous probability spaces

Yao Xie

Georgia Institute of Technology

We consider a minimax problem motivated by distributionally robust optimization (DRO) when the worst-case distribution is continuous, leading to significant computational challenges due to the infinite-dimensional nature of the optimization problem. Recent research has explored learning the worst-case distribution using neural network-based generative models to address these computational challenges but lacks algorithmic convergence guarantees. This paper bridges this theoretical gap by presenting an iterative algorithm to solve such a minimax problem, achieving global convergence under mild assumptions and leveraging technical tools from vector space minimax optimization and convex analysis in the space of continuous probability densities. In particular, leveraging Brenier's theorem, we represent the worst-case distribution as a transport map applied to a continuous reference measure and reformulate the regularized discrepancy-based DRO as a minimax problem in the Wasserstein space. Furthermore, we demonstrate that the worst-case distribution can be efficiently computed using a modified Jordan-Kinderlehrer-Otto (JKO) scheme with sufficiently large regularization parameters for commonly used discrepancy functions, linked to the radius of the ambiguity set. Additionally, we derive the global convergence rate and quantify the total number of subgradient and inexact modified JKO iterations required to obtain approximate stationary points. These results are potentially applicable to nonconvex and nonsmooth scenarios, with broad relevance to modern machine learning applications.

Automated Problem Decomposition and Solution for Extended Monotropic Optimization

Wotao Yin
Alibaba DAMO Academy

Optimization practitioners encounter substantial hurdles when transforming optimization problems and applying appropriate algorithms, mainly when these problems comprise diverse components. Some elements may be ideally suited for proximal operators, while some others lend themselves to parallel computing. Practitioners must manually determine how to integrate these approaches effectively, demanding deep expertise and considerable time.

We introduce an automated system PDMO.jl that optimizes extended monotropic optimization problems, which encompasses conic programs and many other nonlinear programs, allowing experts to concentrate on strategic tasks like problem formulation and result interpretation. PDMO.jl analyzes problem structures, seamlessly applies appropriate techniques to each component, and harnesses recent advancements in automated parameter selection and acceleration techniques.

This is joint work with: Kaizhao Sun, Baihao Wu, and Kun Yuan

A Memory Efficient Randomized Subspace Optimization Method for Training Large Language Models

Kun Yuan
Peking University

The memory challenges associated with training Large Language Models (LLMs) have become a critical concern, particularly when using the Adam optimizer. To address this issue, numerous memory-efficient techniques have been proposed, with GaLore standing out as a notable example designed to reduce the memory footprint of optimizer states. However, these approaches do not alleviate the memory burden imposed by activations, rendering them unsuitable for scenarios involving long context sequences or large mini-batches. Moreover, their convergence properties are still not well-understood in the literature. In this work, we introduce a Randomized Subspace Optimization framework for pre-training and fine-tuning LLMs. Our approach decomposes the high-dimensional training problem into a series of lower-dimensional subproblems. At each iteration, a random subspace is selected, and the parameters within that subspace are optimized. This structured reduction in dimensionality allows our method to simultaneously reduce memory usage for both activations and optimizer states. We establish comprehensive convergence guarantees and derive rates for various scenarios, accommodating different optimization strategies to solve the subproblems. Extensive experiments validate the superior memory and communication efficiency of our method, achieving performance comparable to GaLore and Adam.

A Few Issues on Stochastic Mirror Descent: Optimality Measure and Approach to Acceleration

Junyu Zhang
National University of Singapore

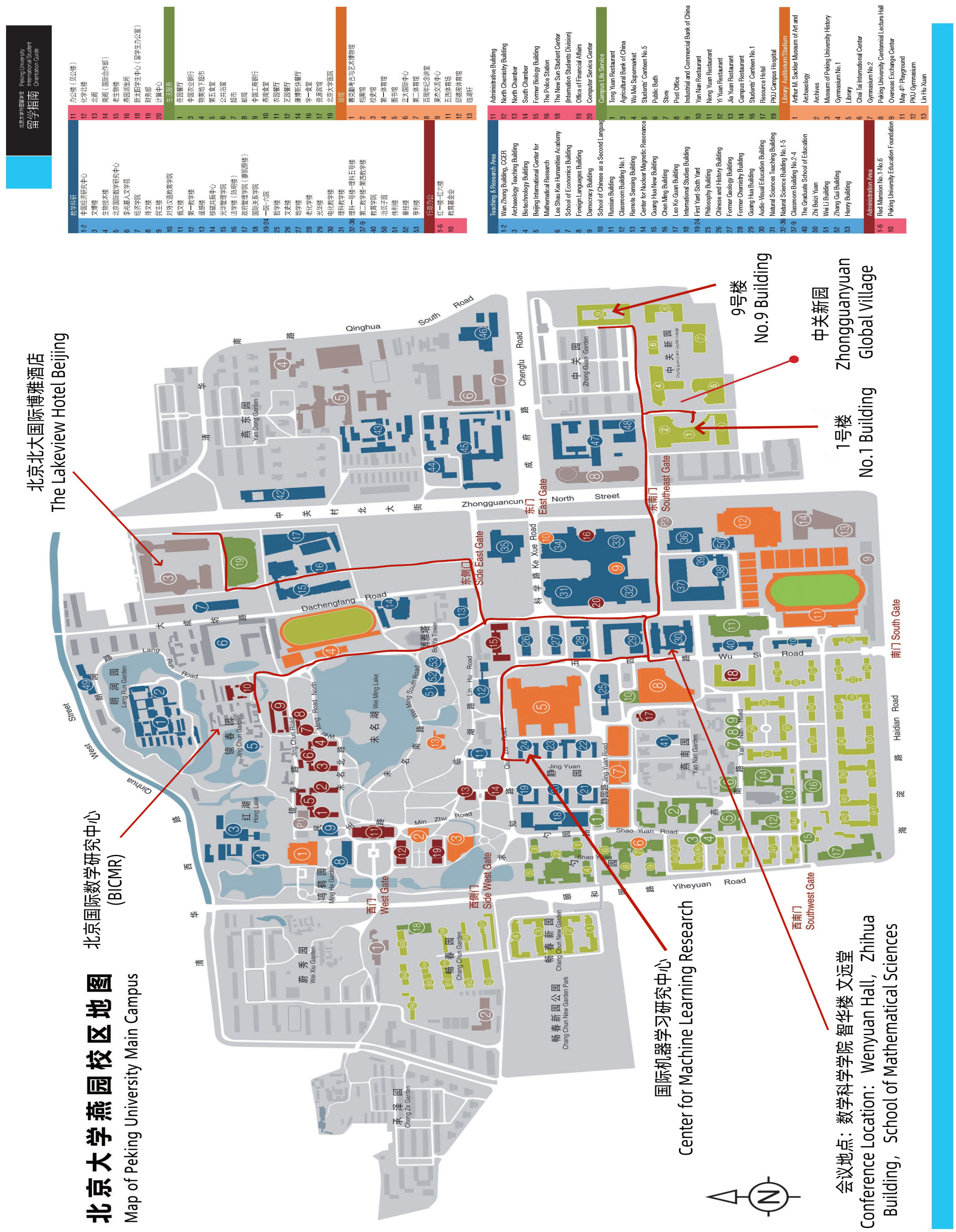
Mirror descent (MD) and its stochastic variants are well-understood first order methods in convex optimization under appropriate smoothness conditions and have been extended to general relative smooth and nonconvex problems in recent years. Despite being widely studied, there are still many issues in the existing research. In this talk, we will first reveal the inappropriateness of the currently widely adopted optimality measures, which casts doubts about the validity of current complexity results, for both deterministic and stochastic MD. Second, due to the absence of Lipschitz continuity, most existing results on nonconvex stochastic mirror descent (SMD) do not allow popular acceleration techniques such as random shuffling and variance reduction, etc. To fix these issues, we propose a novel regularity condition that is widely satisfied by most of the popular kernels. Combined with the dual space perspective of mirror descent, a new optimality measure has been proposed and a Lipschitz-like gradient bound has been established. We elaborate how these concepts can help the analysis SMD by establishing new and improved complexity results for SMD with variance reduction.

The Rate of Convergence of Augmented Lagrangian Method for Minimax Optimization Problems with Semidefinite Constraints

Liwei Zhang
Dalian University of Technology

The augmented Lagrangian method for solving a class of minimax optimization problems with semidefinite constraints is investigated. Firstly, we propose the second-order optimality conditions for the minimax optimization problem. Under the second-order sufficiency optimality conditions, we prove that the sequence of multiplier vectors generated by the augmented Lagrangian method has at least Q-linear convergence if the sequence of penalty parameters is bounded and the convergence rate is superlinear if the sequence of penalty parameters is increasing to infinity. Finally, we use the second-order derivative of the value function to estimate the rate of convergence of the sequence of multiplier vectors.

Map of Peking University



The organizing committee wishes you a pleasant stay in Peking University!