2026 PKU Workshop on Optimization Theory and Methods

July 23-25, 2026

Peking University, Beijing, China

http://conference.bicmr.pku.edu.cn/meeting/index?id=133

2026 PKU Workshop on Optimization Theory and Methods

July 23-25, 2026

Peking University, Beijing, China

http://conference.bicmr.pku.edu.cn/meeting/index?id=133

Information for Participants

Conference Hotel for Invited Speakers

Hotel:	Beijing Friendship Hotel
Address:	No. 1 Zhongguancun South Street, Haidian District, Beijing
Phone:	+86-10-68498888
E-mail:	rd@BJfriendshiphotel.com
Website:	http://www.BJfriendshiphotel.com

A shuttle bus will be arranged between the hotel and PKU roughly at the following time:

Hotel to PKU:	08:20 Yibin Building, 08:25 Guibin Building
PKU to Hotel:	12:30
Hotel to PKU:	13:40 Yibin Building, 13:45 Guibin Building
PKU to Hotel:	TBA

Self-Travel Information

The following routes apply in both directions between PKU and Beijing Friendship Hotel.

- Subway: Take Line 4 between East Gate of Peking University Station and Renmin University Station, then walk to the hotel/conference hall. The total travel time is approximately 25-30 minutes.
- Taxi: Take a taxi directly between PKU and the hotel. The total travel time is approximately 15-25 minutes, depending on traffic.
- Bus: Take Bus 305/320 between Zhongguanyuan Station and Chinese Academy of Agricultural Sciences, then walk to the hotel/conference hall. The total travel time is approximately 30-40 minutes, depending on traffic.

Conference Venue

July 23–24:	Wenyuan Hall, Zhihua Building, Peking University
July 25:	Jiayibing Building, Beijing International Center for Mathematical Research, Peking University
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 6.77 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

- Ms. Congcong Zhao: zhaocongcong0613@bicmr.pku.edu.cn
- Dr. Jiang Hu: jianghu@tsinghua.edu.cn

Sponsors

Beijing International Center for Mathematical Research, Peking University

The 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)

Organizing Committee

Yuhong Dai, Chinese Academy of Sciences

Jiang Hu, Tsinghua University

Jong-Shi Pang, University of Southern California

Zaiwen Wen, Peking University

Kun Yuan, Peking University

Yaxiang Yuan, Chinese Academy of Sciences

Conference Schedule

Each talk will be 40 minutes, consisting of 35 minutes for the presentation and 5 minutes for questions.

July 23, Thursday

Venue: Wenyuan Hall, Zhihua Building, Peking University

09:00-10:20 Session 1

09:00-09:40 **Jong-Shi Pang**, Solution of Binary-Constrained Quadratic-Defined Optimization Problems by Progressive Integer Programming
Chair: Xin Liu

09:40-10:20 **Radu Ioan Boţ**, Accelerating Diagonal Methods for Bilevel Optimization: Unified Convergence via Continuous-Time Dynamics
Chair: Qing Ling

10:20-10:40 Coffee Break

10:40-12:00 Session 2

10:40-11:20 **Defeng Sun**, The Aubin Property for Generalized Equations over C^2-cone Reducible Sets
Chair: Yongjin Liu

11:20-12:00 **Kim-Chuan Toh**, A Low-rank Augmented Lagrangian Method for Polyhedral–SDP and Moment-SOS Relaxations of Polynomial Optimization
Chair: Jiaojiao Zhang

12:00-14:20 Lunch

14:20-15:40 Session 3

14:20-15:00 **Ting Kei Pong**, Error Bound for Conic Feasibility Problems: Case Study on Perspective Cones of Some Nonnegative Legendre Functions
Chair: Jianlin Jiang

15:00-15:40 **Johannes O. Royset**, Failure of uniform laws of large numbers for subdifferentials and beyond
Chair: Chao Ding

15:40-16:00 Coffee Break

16:00-18:00 Session 4

16:00-16:40 **Zaiwen Wen**, Mathematical Formalization and Theorem Proving
Chair: Ziyan Luo

16:40-17:20 **Xudong Li**, Semismooth Newton Methods Beyond Strong Jacobian Regularity
Chair: Shaohua Pan

17:20-18:00 **Niao He**, Three Paths to Minima Selection
Chair: Jinyan Fan

July 24, Friday

Venue: Wenyuan Hall, Zhihua Building, Peking University

09:00-10:20 Session 1

09:00-09:40 **Zhi-Quan Luo**, Principled Policy Optimization and Adaptive Sampling for Reinforcing Large Language Models
Chair: Lingchen Kong

09:40-10:20 **Yuhong Dai**, Distributed Recursion Revisited
Chair: Liwei Zhang

10:20-10:40 Coffee Break

10:40-12:00 Session 2

10:40-11:20 **Deren Han**, Randomized conjugate gradient least squares
Chair: Bo Jiang (NJNU)

11:20-12:00 **Jiang Hu**, Rethinking Manifold Augmented Lagrangian Methods with Constant Penalty Parameters
Chair: Xinwei Liu

12:00-14:20 Lunch

14:20-15:40 Session 3

14:20-15:00 **Chenglong Bao**, Interactions Between Dynamical Systems and Numerical Algorithm Design
Chair: Qingna Li

15:00-15:40 **Andre Milzarek**, A Normal Map-based Proximal Stochastic Gradient Method: Convergence and Identification Properties
Chair: Yibin Xiao

15:40-16:00 Coffee Break

16:00-18:00 Session 4

16:00-16:40 **Junyi Liu**, An Augmented Lagrangian Decomposition Method for Two-stage Stochastic Programs with Nonlinearly Bi-parameterized Recourse
Chair: Xiao Wang

16:40-17:20 **Ya-Feng Liu**, One-Bit Precoding in Massive MIMO: Algorithm Design and Asymptotic Performance Analysis
Chair: Junfeng Yang

17:20-18:00 **Kun Yuan**, Accelerating LLM Pre-Training through Flat-Direction Dynamics Enhancement
Chair: Zheng Peng

July 25, Saturday

Venue: Jiayibing Building, Beijing International Center for Mathematical Research, Peking University

09:00-10:20 Session 1

09:00-09:40 **Yinyu Ye**, The Simple Strategy-Iteration Method is Strongly Polynomial for the Turn-Based Deterministic Forward Game
Chair: Wenxun Xing

09:40-10:20 **Shuzhong Zhang**, New Results on Solving Smooth Non-monotone VI Problems
Chair: Liping Zhang

10:20-10:40 Coffee Break

10:40-12:00 Session 2

10:40-11:20 **Zhaosong Lu**, First-order methods for nonconvex–nonconcave minimax optimization under a local Kurdyka–Łojasiewicz condition
Chair: Bo Jiang (SHUFE)

11:20-12:00 **Yao Xie**, New Optimization Formulations in Statistical Learning: Operators, Transport, and Structure
Chair: Zi Xu

12:00-14:20 Lunch

14:20-15:40 Session 3

14:20-15:00 **Guoyin Li**, Generalized Metric Subregularity and Superlinear Convergence Analysis for High-order Regularized Newton's Method
Chair: Yong Xia

15:00-15:40 **Yancheng Yuan**, A Perturbed DCA for Computing d-Stationary Points of Nonsmooth DC Programs
Chair: Cong Sun

Abstracts

Solution of Binary-Constrained Quadratic-Defined Optimization Problems by Progressive Integer Programming

Jong-Shi Pang
University of Southern California

Inspired by the effectiveness of the Progressive Integer Programming (PIP) approach for improving the solution of indefinite quadratic programs (QPs), via their equivalent linear programming with linear complementarity (LPCC) formulation (joint work with Xinyao Zhang and Shaoning Han), we propose the same strategy for solving many classes of binary-constrained quadratic-defined optimization problems, including their game-theoretic extensions. We show that all these problems can be formulated as a LPCC, via the intermediate step of an equivalent formulation as a binary-constrained quadratic program with linear complementarity constraints (QPCC). The LPCC formulation facilitates the application of state-of-the-art IP solvers as the workhorse in a progressive solution procedure for these nonconvex optimization problems that involve both continuous and binary variables.

Accelerating Diagonal Methods for Bilevel Optimization: Unified Convergence via Continuous-Time Dynamics

Radu Ioan Boţ

Faculty of Mathematics, University of Vienna
radu.bot@univie.ac.at

We analyze fast diagonal methods for simple bilevel programs. Guided by the analysis of the corresponding continuous-time dynamics, we provide a unified convergence analysis under general geometric conditions, including Hölderian growth and the Attouch-Czarnecki condition. Our results yield explicit convergence rates and guarantee weak convergence to a solution of the bilevel problem. In particular, we improve and extend recent results on accelerated schemes, offering novel insights into the trade-offs between geometry, regularization decay, and algorithmic design. Numerical experiments illustrate the advantages of more flexible methods and support our theoretical findings.

The Aubin Property for Generalized Equations over C^2-cone Reducible Sets

Defeng Sun

The Hong Kong Polytechnic University

In this talk, we address a fundamental and long-standing question in variational analysis by proving the equivalence between the Aubin property and strong regularity for generalized equations over C^2-cone reducible sets. We also construct a counterexample showing that the C^2-cone reducibility assumption cannot be dropped.

A Low-rank Augmented Lagrangian Method for Polyhedral–SDP and Moment-SOS Relaxations of Polynomial Optimization

Kim-Chuan Toh
National University of Singapore

Polynomial optimization problems (POPs) admit geometric convex conic reformulations (Kim–Kojima–Toh, SIOPT 30, 2020), though they remain NP-hard. We show that several well-known relaxations can be unified under a polyhedral–SDP framework obtained by approximating the intractable cone with intersections of polyhedral cones and the PSD cone. While yielding tight lower bounds, these relaxations become costly as variable dimensions and number of constraints scale as $\Omega(n^{2\tau})$ with the relaxation order τ.

We introduce RiNNAL-POP, a low-rank augmented Lagrangian method (ALM) to solve these large-scale polyhedral–SDP relaxations. For $\Omega(n^{2\tau})$ nonnegativity and consistency constraints, we design a projection scheme whose cost is linear in the number of variables. We further expose facial structure in the relaxation and restrict the matrix variable to affine subspaces tied to exposed faces of the PSD cone, eliminating many linear constraints and enabling efficient factorized ALM subproblems. Each ALM iteration also applies a single projected-gradient step on the original matrix variable to adapt rank and escape spurious local minima when needed. We also extend the framework to moment-SOS relaxations. Experiments on various benchmarks show that RiNNAL-POP is robust and efficient for large-scale polyhedral–SDP relaxations.

[Based on joint work with Di Hou and Tianyun Tang]

Error Bound for Conic Feasibility Problems: Case Study on Perspective Cones of Some Nonnegative Legendre Functions

Ting Kei Pong
Hong Kong Polytechnic University

Error bounds play a central role in the study of conic optimization problems, including the analysis of convergence rates for numerous algorithms. Curiously, those error bounds are often Hölderian with exponent 1/2. In this talk, we try to explain the prevalence of the 1/2 exponent via analyzing the perspective cone constructed from a nonnegative Legendre function on the real line. Our analysis is based on the facial reduction technique and the computation of one-step facial residual functions (1-FRFs). Under appropriate assumptions on the Legendre function, we show that 1-FRFs can be taken to be Hölderian of exponent 1/2 almost everywhere with respect to the two-dimensional Hausdorff measure.

This is a joint work with Xiaozhou Wang and Bruno F. Lourenço.

Failure of uniform laws of large numbers for subdifferentials and beyond

Johannes O. Royset
University of Southern California

We provide counterexamples showing that uniform laws of large numbers do not hold for subdifferentials under natural assumptions. Our constructions are univariate random Lipschitz functions and bivariate random convex functions with two smooth pieces. Consequently, they resolve the questions posed by Shapiro and Xu [J. Math. Anal. Appl., 325 (2007), 1390–1399] in the negative. They also demonstrate the failure of certain graphical and pointwise laws for subdifferentials, revealing fundamental barriers to the consistency of sample-average approximation and subdifferential approximation.

Mathematical Formalization and Theorem Proving

Zaiwen Wen
Peking University

This talk explores a pathway for intelligent mathematical reasoning that transforms mathematical research from natural-language reasoning into a computable, verifiable, and reusable knowledge system. A brief introduction is first given to mathematical formalization, where definitions, theorems, and proofs are represented in formal languages such as Lean, so that every step of reasoning can be checked by machines, thereby enhancing proof reliability, knowledge reuse, and collaborative scalability. Several recent advances in mathematical formalization and automated theorem proving are then presented: for large-scale Lean libraries, premise-selection methods based on expression-tree structures and graph neural networks are developed to improve the efficiency of theorem retrieval and proof search; M2F is proposed as an automatic formalization framework that converts mathematical textbooks and papers into compilable, auditable, and reusable Lean projects, together with ReasBook as a formalized mathematical knowledge base; FaithSieve is introduced to leverage formal evidence and semantic alignment for guiding mathematical proof reviewing and error localization; OptProver is developed to enhance the domain-specific capability of models in formal theorem proving through optimization-domain data and preference learning; and CAM-Bench is constructed as a benchmark for formalized applied mathematics, covering optimization, numerical linear algebra, and numerical analysis. Finally, we introduce ReasLab, an intelligent mathematical reasoning engine, and demonstrate its integrated capabilities in theorem proving and scientific writing.

Semismooth Newton Methods Beyond Strong Jacobian Regularity

Xudong Li
Fudan University

The fast local convergence of Newton-type methods is usually established under strong regularity assumptions, which often fail in degenerate problems. This talk presents two mechanisms for designing semismooth Newton methods beyond classical generalized Jacobian regularity. For nonlinear semidefinite programming, weak second-order conditions, weak strict Robinson constraint qualification, a single nonsingular linearization, and a correction step yield local superlinear or quadratic convergence. For polyhedral projection with linear equality constraints, a lifted equivalent set and the selection of an extreme-point representative lead to nonsingular Newton matrices. These results show that fast Newton convergence in degenerate settings can be derived from refined, problem-dependent variational geometry rather than regularity of the entire generalized Jacobian.

Three Paths to Minima Selection

Niao He

ETH Zurich

Classical optimization is often built around a simple goal: find a minimizer. Most existing theories emphasize convergence rates towards the minima, implicitly treating all solutions as equivalent once optimality is achieved. Modern machine learning applications tell a different story. Distinct minimizers with identical objective values can differ dramatically in properties that matter in practice, including generalization performance, robustness, and even broader societal implications. In overparameterized models, especially in deep learning, a striking phenomenon emerges: even after training error reaches zero, test performance continues to improve, indicating that optimization dynamics keep evolving within the set of global minima. This raises a fundamental and largely under-explored question: which minima do optimization algorithms implicitly select? More ambitiously, can we actively steer optimization dynamics toward desirable minima? In this talk, I will present recent results that shed light on both implicit and active minima selection, highlighting the roles played by optimization dynamics (first- and zeroth-order methods), stochastic noise, and the geometry of the solution landscape.

Principled Policy Optimization and Adaptive Sampling for Reinforcing Large Language Models

Zhi-Quan Luo
The Chinese University of Hong Kong, Shenzhen

Reinforcement learning has become a key framework for training large language models (LLMs), but scaling it to models with hundreds of billions of parameters poses substantial computational challenges. This talk presents two methods for addressing these challenges. First, I will introduce ReMax, a policy-gradient method that removes the auxiliary value network used in PPO-style methods. By using the reward of the greedy response as a control variate, ReMax reduces gradient variance in theory, cuts memory usage by about 50% in practice, and leads to faster training at LLM scale. This line of work later influenced related value-free methods, including GRPO, which was used in DeepSeek-R1. Second, I will introduce Knapsack RL, which formulates sampling-budget allocation as a knapsack-style optimization problem. Instead of distributing computation uniformly across prompts, it reallocates rollouts according to estimated learning progress, increasing the fraction of nonzero policy gradients by 20–40% at no additional computational cost. Taken together, these results provide a principled optimization framework for reinforcing LLMs.

Distributed Recursion Revisited

Yuhong Dai

Chinese Academy of Sciences

The distributed recursion (DR) algorithm is an effective method for solving the pooling problem that arises in many applications. It is based on the well-known P-formulation of the pooling problem, which involves the flow and quality variables; and it can be seen as a variant of the successive linear programming (SLP) algorithm, where the linear programming (LP) approximation problem can be transformed from the LP approximation problem derived by using the first-order Taylor series expansion technique. In this talk, we first propose a new nonlinear programming (NLP) formulation for the pooling problem involving only the flow variables, and show that the DR algorithm can be seen as a direct application of the SLP algorithm to the newly proposed formulation. With this new useful theoretical insight, we then develop a new variant of DR algorithm, called penalty DR (PDR) algorithm based on the proposed formulation. The proposed PDR algorithm is a penalty algorithm where violations of the (linearized) nonlinear constraints are penalized in the objective function of the LP approximation problem with the penalty terms increasing when the constraint violations tend to be large. Compared with the LP approximation problem in the classic DR algorithm, the LP approximation problem in the proposed PDR algorithm can return a solution with a better objective value, which makes it more suitable for finding high-quality solutions for the pooling problem. Numerical experiments on benchmark and randomly constructed instances show that the proposed PDR algorithm is more effective than the classic SLP and DR algorithms in terms of finding a better solution for the pooling problem.

Randomized conjugate gradient least squares

Deren Han

Beihang University

We develop a novel randomized conjugate gradient least squares (RCGLS) method for solving least-squares problems, in which iterative sketching is employed at each step to reduce the dimension and hence the computational cost. In particular, we propose a new perspective on the classical CGLS method, where the next descent direction is determined via a constraint correction problem associated with the gradient. Based on this insight, we replace the gradient with a randomized coordinate gradient that naturally satisfies the variance reduction property, leading directly to the proposed RCGLS method. We prove that RCGLS converges linearly in expectation, with a better convergence bound compared to the randomized coordinate descent method. Furthermore, we investigate an implementation of the method that avoids full-dimensional vector operations, which are the major bottleneck of vanilla RCGLS for sparse matrices and render it impractical. We also show how to apply the RCGLS method to solve the ridge regression problem, yielding a lightweight, parallelizable, and accelerated method for such problems. Numerical experiments are provided to confirm our results.

Rethinking Manifold Augmented Lagrangian Methods with Constant Penalty Parameters

Jiang Hu
Tsinghua University

We study nonsmooth composite optimization over a compact embedded Riemannian submanifold, where a smooth function is combined with a convex Lipschitz term composed with a linear map. Existing non-asymptotic manifold augmented Lagrangian analyses typically use penalty or smoothing parameters that vary with the target accuracy or the iteration counter. This leaves open whether the best known primal-update complexity can be achieved with a genuinely constant penalty parameter. We answer this question affirmatively by proposing a dually regularized manifold augmented Lagrangian method that keeps the penalty parameter fixed and uses a vanishing dual regularization only to make the multiplier computation well posed. The fixed penalty keeps the smooth primal augmented-Lagrangian model differentiated at each iteration uniformly conditioned, with a smoothness constant independent of the target accuracy and the iteration counter. We prove that the method reaches an ϵ-KKT point within $\mathcal{O}(\epsilon^{-3})$ primal Riemannian-gradient updates. For general proximable convex Lipschitz terms, gradient ascent for the multiplier subproblems gives total multiplier-oracle complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$; for separable structured terms whose scalar proximal equations are monotone and cheaply evaluable, including the ℓ_1 norm, a coordinate-wise bisection solver yields $\widetilde{\mathcal{O}}(m\epsilon^{-3})$ scalar bisection steps. We also analyze a single-loop variant with one multiplier gradient-ascent step per iteration and establish an $\mathcal{O}(\epsilon^{-4})$ KKT complexity bound.

Interactions Between Dynamical Systems and Numerical Algorithm Design

Chenglong Bao
Tsinghua University

Numerical algorithms and dynamical systems are closely related, yet are often studied from different perspectives. In this talk, I explore their interaction in the context of scientific computing and optimization. First, we consider phase field crystal models, which are typically formulated as gradient flow. We show how ideas from numerical optimization can be used to design faster algorithms that go beyond standard gradient flow schemes. Second, we discuss the dynamical systems viewpoint of optimization, where accelerated methods can be interpreted as discretizations of suitable ODEs, with connections to modern machine learning architectures.

A Normal Map-based Proximal Stochastic Gradient Method: Convergence and Identification Properties

Andre Milzarek
The Chinese University of Hong Kong, Shenzhen

The proximal stochastic gradient method (prox-SGD) is one of the state-of-the-art approaches for stochastic composite-type problems. In contrast to its deterministic counterpart, prox-SGD has been found to have difficulties with the correct identification of underlying substructures (such as supports, low rank patterns, or active constraints) and it does not possess a finite-time manifold identification property. Existing solutions rely on convexity assumptions or on the additional usage of variance reduction techniques. In this talk, we address these limitations and present a simple variant of prox-SGD based on Robinson ' s normal map. The proposed normal map-based proximal stochastic gradient method (norm-SGD) is shown to converge globally, i.e., accumulation points of the generated iterates correspond to stationary points almost surely. In addition, we establish complexity bounds for norm-SGD that match the known results for prox-SGD and we prove that norm-SGD can almost surely identify active manifolds in finite-time in a general nonconvex setting. Our derivations are built on almost sure iterate convergence guarantees and utilize analysis techniques based on the Kurdyka-Lojasiewicz (KL) inequality. Finally, we discuss how KL-based strategies can be leveraged more broadly in the asymptotic convergence analysis of stochastic and inexact optimization algorithms.

An Augmented Lagrangian Decomposition Method for Two-stage Stochastic Programs with Nonlinearly Bi-parameterized Recourse

Junyi Liu
Tsinghua University

In this work, we consider a class of two-stage stochastic programs in which both the objective and the constraints of the second-stage problem depend nonlinearly on the first-stage decisions. Such a bi-parametrization structure leads to a nonconvex recourse function without Clarke regularity, posing significant challenges in the computation. By lifting the bi-parameterization and duplicating the first-stage variables within the second-stage constraints, we reformulate the original problem as the minimization of composite convex-concave functions subject to non-anticipativity constraints. We develop a two-loop algorithm in which a sequence of augmented Lagrangian subproblems are constructed in the outer-loop and each subproblem is solved by a blockwise decomposition algorithm in the inner loop with finite termination criteria via convex programming. We establish convergence guarantees under both deterministic and stochastic settings. Numerical experiments on a facility location problem demonstrate the effectiveness and computational scalability with respect to the number of scenarios.

One-Bit Precoding in Massive MIMO: Algorithm Design and Asymptotic Performance Analysis

Ya-Feng Liu

Beijing University of Posts and Telecommunications

One-bit precoding is a promising approach to enhancing hardware efficiency in massive MIMO systems and has attracted significant research interests in recent years. However, the inherent one-bit constraint on the transmit signals poses great challenges for both precoding design and performance analysis. This talk presents recent advancements in one-bit precoding. The first part focuses on precoding design for massive MIMO systems, introducing a new negative ℓ_1 penalty approach. This method is based on an exact penalty model that penalizes the one-bit constraint into the objective using a negative ℓ_1-norm term. Compared to existing approaches, our approach achieves a superior trade-off between computational complexity and symbol error rate (SER) performance. Additionally, we will explore one-bit precoding design in massive MIMO integrated sensing and communication (ISAC)—a key use case in 6G systems. The second part of the talk shifts to the asymptotic SER performance analysis of a broad class of linear-quantized precoding schemes. Unlike conventional Bussgang decomposition-based analyzes, our analytical framework is grounded in random matrix theory (RMT), which is more rigorous and can be extended to more general cases. Furthermore, we will present recent findings on the asymptotic SER analysis of a widely studied nonlinear precoding scheme.

Accelerating LLM Pre-Training through Flat-Direction Dynamics Enhancement

Kun Yuan
Peking University

Pre-training Large Language Models requires immense computational resources, making optimizer efficiency essential. The optimization landscape is highly anisotropic, with loss reduction driven predominantly by progress along flat directions. While matrix-based optimizers such as Muon and SOAP leverage fine-grained curvature information to outperform AdamW, their updates tend toward isotropy—relatively conservative along flat directions yet potentially aggressive along sharp ones. To address this limitation, we first establish a unified Riemannian Ordinary Differential Equation (ODE) framework that elucidates how common adaptive algorithms operate synergistically: the preconditioner induces a Riemannian geometry that mitigates ill-conditioning, while momentum serves as a Riemannian damping term that promotes convergence. Guided by these insights, we propose LITE, a generalized acceleration strategy that enhances training dynamics by applying larger Hessian damping coefficients and learning rates along flat trajectories. Extensive experiments demonstrate that LITE significantly accelerates both Muon and SOAP across diverse architectures (Dense, MoE), parameter scales (130M–1.3B), datasets (C4, Pile), and learning-rate schedules (cosine, warmup-stable-decay). Theoretical analysis confirms that LITE facilitates faster convergence along flat directions in anisotropic landscapes, providing a principled approach to efficient LLM pre-training.

The Simple Strategy-Iteration Method is Strongly Polynomial for the Turn-Based Deterministic Forward Game

Yinyu Ye

Shanghai Jiao Tong University
Stanford University

We give a proof, for the first time, that the classic Simple Strategy-Iteration Method solves the Turn-Based Deterministic Forward Game (TBDFG) in strongly polynomial time of the state-action number. More precisely, the algorithm terminates after at most a polynomial number of the Dantzig-Simplex-Pivot steps where each step switches one action in a state. TBDFG, in which no directed action-cycle contains actions controlled by both players, includes popular two-person board games such as Go and Chess and is a subclass of the Turn-Based Deterministic (zero-sum) Game (TBDG) that is in the class of NP∩co-NP. If the forward graph-structure is known and available to explore, we propose an MDP-Based backward propagation algorithm that solves a sequence of deterministic-MDP subproblems and further reduces the total time complexity.

New Results on Solving Smooth Non-monotone VI Problems

Shuzhong Zhang
University of Minnesota

In this talk, we shall present our recent results on solving smooth but non-monotone variational inequality (VI) problems. Our approach is based on a gap-function reformulation of the VI, a homotopy continuation framework, and a study of proximal gradient steps. The overall convergence relies on the error bound properties pertaining to the gap function in question. One such convergence-enabling error bound condition is guaranteed by the assumption that the original mapping and the constraint set are semi-algebraic, and the set is convex and compact. Since the semi-algebraic functions/sets are dense in the spaces of continuous functions/sets, this implies that our method provides an alternative constructive proof for the existence of VI solution if its mapping is continuous and its constraint set is convex and compact in the n-dimensional Euclidean space, which is also a new constructive proof for the famous Brouwer's fixed point theorem.

This is joint work with Lei Zhao.

First-order methods for nonconvex–nonconcave minimax optimization under a local Kurdyka–Łojasiewicz condition

Zhaosong Lu
University of Minnesota

We study a class of nonconvex–nonconcave minimax problems in which the inner maximization problem satisfies a local Kurdyka–Łojasiewicz (KL) condition that may vary with the outer minimization variable. In contrast to the global KL or PL conditions commonly assumed in the literature—which are significantly stronger and often too restrictive in practice—this local KL condition accommodates a broader range of practical scenarios. However, it also introduces new analytical challenges. In particular, as an optimization algorithm approaches a stationary point, the region over which the KL condition holds may shrink, leading to a more intricate and potentially ill-conditioned landscape. To address this challenge, we show that the associated maximal function is locally generalized Hölder smooth. Leveraging this key property, we develop an inexact proximal gradient method for solving the minimax problem, where the inexact gradient of the maximal function is computed by applying a proximal gradient or sequential convex programming method to a KL-structured subproblem. Under mild assumptions, we establish complexity guarantees for computing an approximate stationary point of the minimax problem. This is joint work with Xiangyuan Wang (University of Minnesota).

New Optimization Formulations in Statistical Learning: Operators, Transport, and Structure

Yao Xie

Georgia Institute of Technology

Many problems in modern statistical learning are not fully captured by classical empirical risk minimization. In some settings, the main challenge is not only how to optimize a given loss, but how to choose the right optimization formulation in the first place. This talk discusses several examples where statistical structure leads to optimization problems beyond standard convex loss minimization. I will first discuss operator-based formulations for statistical estimation. In generalized regression models, especially with non-canonical, nonlinear, or nonsmooth links, likelihood minimization may not provide the most natural or stable computational route. Variational inequality formulations offer an alternative by defining estimators through statistically meaningful estimating operators. This viewpoint allows one to separate statistical identification from the existence of a convenient scalar loss, and leads to first-order algorithms with statistical and computational guarantees. I will then turn to optimization over probability distributions. In robust learning, generative modeling, and decision-making under distribution shift, the decision variable may itself be a distribution rather than a finite-dimensional parameter. Transport maps, Wasserstein proximal schemes, and flow-based representations provide computational ways to formulate and solve such distributional optimization problems. Finally, I will discuss the role of structure, such as locality, low rank, index structure, and particle representations, in making high-dimensional statistical optimization tractable. The main message is that new statistical learning problems often require new optimization formulations, where the central objects may be operators, distributions, or structured transport maps rather than only scalar losses.

Generalized Metric Subregularity and Superlinear Convergence Analysis for High-order Regularized Newton's Method

Guoyin Li
The University of New South Wales

In this talk, we introduce a generalized metric subregularity condition that extends the classical notion of metric subregularity and is weaker than the error bound condition used in recent quadratic convergence analyses of cubic regularized Newton methods. We then establish an abstract extended convergence framework that enables one to derive superlinear convergence towards a specific target set (such as the second-order stationary points), under the introduced generalized metric subregularity condition or KL property. We also provide verifiable sufficient conditions under the strict saddle point property and show that they hold easily for several important nonconvex optimization problems, including the Hadamard parameterized compressed sensing model, best rank-one matrix approximation, and generalized phase retrieval.

As an application, we analyze a high-order regularized Newton method with momentum. For the Hadamard parameterized compressed sensing problem, we prove global convergence together with quadratic convergence rate to a local minimizer under a strict complementarity condition, and establish an $\mathcal{O}(1/k^2)$ convergence rate to a local minimizer in the absence of the strict complementarity condition.

This is a joint work with B.S. Mordukhovich and J. Zhu.

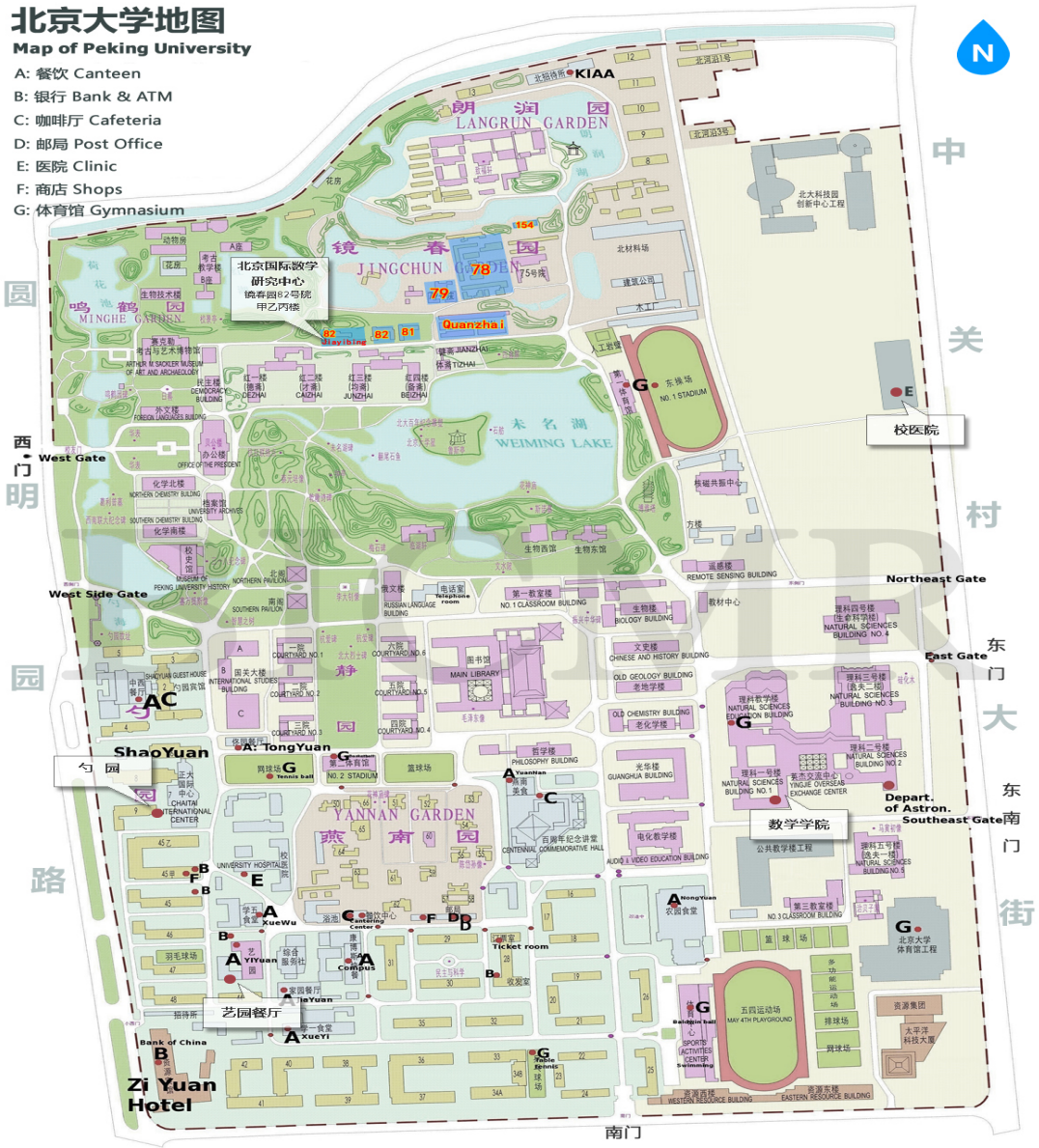
A Perturbed DCA for Computing d-Stationary Points of Nonsmooth DC Programs

Yancheng Yuan

The Hong Kong Polytechnic University

This paper introduces an efficient perturbed difference-of-convex algorithm (pDCA) for computing d-stationary points of an important class of structured nonsmooth difference-of-convex problems. Compared to the principal algorithms introduced in [J.-S. Pang, M. Razaviyayn, and A. Alvarado, Math. Oper. Res. 42(1):95–118 (2017)], which may require solving several subproblems for a one-step update, pDCA solves a perturbed subproblem. Therefore, the computational cost of pDCA for one-step update is comparable to the widely used difference-of-convex algorithm (DCA) introduced in [D. T. Pham and H. A. Le Thi, Acta Math. Vietnam. 22(1):289–355 (1997)] for computing a critical point. Importantly, under practical assumptions, we prove that every accumulation point of the sequence generated by pDCA is a d-stationary point almost surely. Numerical experiment results on several important examples of nonsmooth DC programs demonstrate the efficiency of pDCA for computing d-stationary points.

Map of Peking University



The organizing committee wishes you a pleasant stay in BICMR!