Speakers

1 **Waheed Bajwa**

2 **Chao Ding**

3 **Ethan Xingyuan Fang**

4 **Niao He**

5 **Michael Hintermuller**

6 **Nhat Ho**

7 **Lek-Heng Lim**

8 **Meixia Lin**

9 **Yurii Nesterov**

10 **Yurii Nesterov**

11 **Viet Anh Nguyen**

12 **Jong-Shi Pang**

13 **Jong-Shi Pang**

14 **Houduo Qi**

15 **Anthony Man-Cho So**

16 **Yong Sheng Soh**

17 **Mahdi Soltanolkotabi**

18 **Akiko Takeda**

19 **Christos Thrampoulidis**

20 **Xin Tong**

21 Shuoguang Yang

22 Antonios Varvitsiotis

23 Man-Chung Yue

24 Anru Zhang

Abstracts

Workshop 2: Structured Optimization Models in High-Dimensional Data Analysis

(12–16 Dec 2022)

1 Waheed Bajwa

Rutgers University, USA

Learning Mixtures of Separable Dictionaries for High-Dimensional Tensor Data

Abstract

Data-driven feature representations comprise one of the first and most important steps within a data analysis pipeline. Dictionary learning, which involves using training data to obtain an overcomplete matrix that sparsifies the unseen data, has emerged as one of the most powerful data-driven feature representation methods during the last decade-and-a-half. When utilized for high-dimensional tensor (aka, multiway) data, however, conventional dictionary learning suffers from high sample complexity and computational overhead. We address this challenge in the talk by proposing a new model for dictionary learning for high-dimensional tensor data that corresponds to learning a mixture of separable dictionaries. The proposed model better captures the richness of tensor data by generalizing the separable dictionary learning model to one that offers an improved tradeoff between bias and variance. In the talk, we explore two different structured optimization approaches for learning a mixture of separable dictionaries and also derive sufficient conditions for local identifiability of the underlying dictionary in each case. Moreover, we discuss computational algorithms that can be used to solve the problem of learning a mixture of separable dictionaries in both batch and online settings.

This talk is based on a joint work with Mohsen Ghassemi, Zahra Shakeri, and Anand Sarwate.

2 Chao Ding

Chinese Academy of Sciences, China
On the convergence analysis of augmented Lagrangian method for matrix optimization

Abstract

The augmented Lagrangian method (ALM) has gained tremendous popularity for its elegant theory and impressive numerical performance since it was proposed by Hestenes and Powell in 1969. It has been widely used in numerous efficient solvers to improve numerical performance to solve many problems. In this talk, we will introduce some new convergence results on the matrix optimization problems, including nonlinear semidefinite programming and nonsmooth optimization on Riemannian manifold.

3 Ethan Xingyuan Fang

Penn State University, USA
Implicit Regularization of Bregman Proximal Point Algorithm and Mirror Descent on Separable Data

Abstract

Bregman proximal point algorithm (BPPA), as one of the centerpieces in the optimization toolbox, has been witnessing emerging applications. With simple and easy to implement update rule, the algorithm bears several compelling intuitions for empirical successes, yet rigorous justifications are still largely unexplored. We study the computational properties of BPPA through classification tasks with separable data, and demonstrate provable algorithmic regularization effects associated with BPPA. We show that BPPA attains non-trivial margin, which closely depends on the condition number of the distance generating function inducing the Bregman divergence. We further demonstrate that the dependence on the condition number is tight for a class of problems, thus showing the importance of divergence in affecting the quality of the obtained solutions. In addition, we extend our findings to mirror

descent (MD), for which we establish similar connections between the margin and Bregman divergence. We demonstrate through a concrete example, and show BPPA/MD converges in direction to the maximal margin solution with respect to the Mahalanobis distance. Our theoretical findings are among the first to demonstrate the benign learning properties BPPA/MD, and also provide corroborations for a careful choice of divergence in the algorithmic design. Joint work with Yan Li, Caleb Ju, and Tuo Zhao

4 Niao He

ETH Zurich, Switzerland

Nonconvex Min-Max Optimization: fundamental limits, acceleration, and adaptivity

Abstract

Min-max optimization plays a critical role in emerging machine learning applications from training GANs to robust reinforcement learning, and from adversarial training to fairness. In this talk, we discuss some recent results on min-max optimization algorithms with a special focus on the modern nonconvex regime, including their fundamental limits, acceleration, and adaptivity. We introduce the first accelerated algorithms that achieve near-optimal complexity bounds as well as a family of adaptive algorithms with parameter-free adaptation under various problem settings.

5 Michael Hintermuller

Weierstrass Institute, Germany

A descent algorithm for the optimal control of ReLU neural network informed PDEs based on approximate directional derivatives

Abstract

We introduce a numerical solver for a class of optimal control problems with learning-informed semilinear partial differential equations (PDEs). The latter have constituents that are in principle unknown and are approximated by nonsmooth ReLU neural networks. We argue that a direct smoothing of the ReLU network with the aim to make use of classical solvers can have

significant disadvantages. This motivates us to devise a numerical algorithm that treats directly the nonsmooth optimal control problem, by employing a descent algorithm inspired by a bundle-free method. Several numerical examples are provided and the efficiency of the algorithm is shown.

6 Nhat Ho

The University of Texas at Austin, USA
Instability, Computational Efficiency and Statistical Accuracy

Abstract

Many statistical estimators are defined as the fixed point of a data-dependent operator, with estimators based on minimizing a cost function being an important special case. The limiting performance of such estimators depends on the properties of the population-level operator in the idealized limit of infinitely many samples. We develop a general framework that yields bounds on statistical accuracy based on the interplay between the deterministic convergence rate of the algorithm at the population level, and its degree of (in)stability when applied to an empirical object based on n samples. Using this framework, we analyze both stable forms of gradient descent and some higher-order and unstable algorithms, including Newton's method and its cubic-regularized variant, as well as the EM algorithm. We provide applications of our general results to several concrete classes of singular statistical models, including Gaussian mixture estimation, single-index models, and informative non-response models. We exhibit cases in which an unstable algorithm can achieve the same statistical accuracy as a stable algorithm in exponentially fewer steps—namely, with the number of iterations being reduced from polynomial to logarithmic in sample size n.

7 Lek-Heng Lim

The University of Chicago, USA

Abstract

8 Meixia Lin

Singapore University of Technology and Design, Singapore
Determinantal point processes for sampling minibatches in SGD

Abstract

In this work, we contribute an orthogonal polynomial-based determinantal point process paradigm for performing minibatch sampling in SGD. Our approach leverages the specific data distribution at hand, which endows it with greater sensitivity and power over existing data-agnostic methods. We substantiate our method via a detailed theoretical analysis of its convergence properties, interweaving between the discrete data set and the underlying continuous domain. In particular, we show how specific DPPs and a string of controlled approximations can lead to gradient estimators with a variance that decays faster with the batchsize than under uniform sampling. Coupled with existing finite-time guarantees for SGD on convex objectives, this entails that, for a large enough batchsize and a fixed budget of item-level gradients to evaluate, DPP minibatches lead to a smaller bound on the mean square approximation error than uniform minibatches. Moreover, our estimators are amenable to a recent algorithm that directly samples linear statistics of DPPs (i.e., the gradient estimator) without sampling the underlying DPP (i.e., the minibatch), thereby reducing computational overhead.

9 Yurii Nesterov

UCLouvain, Belgium
Set-Limited Functions and Polynomial-Time Interior-Point Methods

Abstract

In this talk, we revisit some elements of the theory of self-concordant functions. We replace the notion of self-concordant barrier by a new notion of set-limited function, which forms a wider class. We show that the proper set-limited functions ensure polynomial time complexity of the corresponding path-following method (PFM). Our new PFM follows a deviated path, which connects an arbitrary feasible point with the solution of the problem. We present some applications of our approach to the problems of unconstrained optimization, for which it ensures a global linear rate of convergence even in for nonsmooth objective function.

10 Yurii Nesterov

UCLouvain, Belgium
New perspectives for higher-order methods in Convex Optimization

Abstract

In the recent years, the most important developments in Optimization were related to clarification of abilities of the higher-order methods. These schemes have potentially much higher rate of convergence as compared to the lower-order methods. However, the possibility of their implementation in the form of practically efficient algorithms was questionable during decades. In this talk, we discuss different possibilities for advancing in this direction, which avoid all standard fears on tensor methods (memory requirements, complexity of computing the tensor components, etc.). Moreover, in this way we get the new second-order methods with memory, which converge provably faster than the conventional upper limits provided by the Complexity Theory.

11 Viet Anh Nguyen

The Chinese University of Hong Kong, China
Fair Principal Component Analysis under Optimal Transport Perturbations

Abstract

Principal component analysis (PCA) is a simple yet useful dimensionality reduction technique in modern machine learning pipelines. In consequential domains such as college admission, healthcare and credit approval, it is imperative to take into account emerging criteria such as the fairness and the robustness of the learned projection. In this paper, we measure fairness using the equality of the reconstruction errors criterion. We study three aspects of the fair PCA problem under this criterion: (1) we delineate an optimal transport statistical test to detect if a given projection matrix is unfair, (2) we provide a uncertainty quantification tool that determines the maximal amount of possible unfairness of a projection matrix, and (3) we study how the Riemannian optimization framework can be used to find a fair projection matrix that is distributionally robust against perturbations of the empirical measures.

12 Jong-Shi Pang

University of Southern California, USA
Nonconvex Stochastic Programs: Deterministic Constraints

Abstract

Since its early days, the field of stochastic programming has benefitted from the advances of convex programming, particularly large-scale linear programming. A major drawback of this approach is that for ease of computations and analysis, the models are of the convex kind and are formulated at the expense of simplifications but lacking generality and faithfulness to their source applications. A simple case in point is the classical two-stage linear stochastic program with recourse where the first-stage decision variable appears linearly only in the constraints of the second-stage linear program, resulting in the recourse function being convex and piecewise linear. Starting with the linearly bi-parameterized two-stage stochastic program with recourse, the speaker and his co-authors have begun a rigorous study of nonconvex (and typically nondifferentiable) stochastic programs of various kinds that arise from diverse sources. Illustrated with motivating applications, this general talk presents some selected results of our research published in several papers addressing problems in this exciting domain of modern optimization, where the combination of uncertainty, nonconvexity, and nondifferentiability constitutes the key analytical and computational challenges, in addition to the common issue of sampling of the randomness. As the first part of this vast topic, the presentation is restricted to the case where the constraints are deterministic and the randomness occurs only in the objective function of the optimization problem. Models, theory, and algorithms are sketchily covered that together are the central elements of a new chapter of stochastic optimization where the surface has barely been scratched.

The bulk of the presented materials is drawn from joint work with Drs. Ying Cui (University of Minnesota), Junyi Liu (Tsinghua University), and Suvrajeet Sen (University of Southern California).

13 Jong-Shi Pang

University of Southern California, USA
Nonconvex Stochastic Programs: Chance Constraints

Abstract

Chance-constrained programs (CCPs) constitute a difficult class of stochastic programs (SPs) due to its possible nondifferentiability and nonconvexity even with simple linear random functionals. Existing approaches for solving the CCPs mainly deal with convex random functionals within the probability function. This work considers two generalizations of the class of chance constraints commonly studied in the literature; one generalization involves probabilities of disjunctive nonconvex functional events and the other generalization involves mixed-signed affine combinations of the resulting probabilities; together, we coin the term affine chance constraint (ACC) system for these generalized chance constraints. The treatment of such an ACC system involves the fusion of several individually known ideas: (a) parameterized upper and lower approximations of the indicator function in the expectation formulation of probability; (b) external (i.e., fixed) versus internal (i.e., sequential) sampling-based approximation of the expectation operator; (c) constraint penalization as relaxations of feasibility; and (d) convexification of nonconvexity and nondifferentiability via surrogation. These ideas lead to several algorithmic strategies with various degrees of practicality and computational efforts for the nonconvex ACC-SP. In an external sampling scheme, a given sample batch (presumably large) is applied to a penalty formulation of a fixed-accuracy approximation of the chance constraints of the problem via their expectation formulation. This results in a sample average approximation scheme, whose almost-sure convergence under a directional derivative condition to a Clarke stationary solution of the expectation constrained-SP as the sample sizes tend to infinity is established. In contrast, sequential sampling, along with surrogation leads to a sequential convex programming based algorithm whose asymptotic convergence for fixed- and diminishing-accuracy approximations of the indicator function can be established under prescribed increments of the sample sizes.

This work is joint with Drs. Ying Cui (University of Minnesota) and Junyi Liu (Tsinghua University).

14 Houduo Qi

University of Southampton, UK

Global and Local Convergence-Rate Analysis of an Inexact Newton Augmented Lagrangian Method for Zero-One Composite Optimization

Abstract

Zero-One Composite Optimization (0/1-COP) is a prototype of nonsmooth, nonconvex optimization problems and it has attracted much attention recently. Augmented Lagrangian Method (ALM) has stood out as a leading methodology for such problems. The main purpose of this paper is to extend the classical theory of ALM from smooth problems to 0/1-COP. We propose, for the first time, second-order optimality conditions for 0/1-COP. In particular, under a second-order sufficient condition (SOSC), we prove Q-linear convergence rate of the proposed ALM. In order to identify the subspace used in SOSC, we employ the proximal operator of the 0/1-loss function, leading to an active-set identification technique. Built around this identification process, we design practical stopping criteria for any algorithm to be used for the subproblem of ALM. We justify that Newton's method is an ideal candidate for the subproblem and it enjoys both global and quadratic convergence. Those considerations result in an inexact Newton ALM (iNALM) for 0/1-COP. The method of iNALM is unique in the sense that it is active-set based, it is inexact (hence more practical), and SOSC instead of widely assumed Kurdyka-Lojasiewicz (KŁ) properties plays an important in its R-linear convergence analysis. The numerical results on both simulated and real datasets show the fast running speed and high accuracy of iNALM when compared with several leading solvers. This is joint work with Penghe Zhang and Naihua Xiu.

15 Anthony Man-Cho So

The Chinese University of Hong Kong, China

On the Complexity of Approximate Stationarity Concepts in Non-Smooth Optimization

Abstract

Non-smooth non-convex optimizaton problems pose many challenges to the definition and computation of stationarity concepts. Although the field of variational analysis has over the years developed various stationarity concepts for Lipschitz functions and provided many beautiful theoretical tools for studying them, the computational complexity of these concepts remains largely open. In this talk, we discuss the complexity of finding (approximate)

stationary points of certain sub-classes of Lipschitz functions. On the negative side, we show that under a standard first-order oracle framework, no algorithm that can find a near-approximate stationary (NAS) point of any Clarke regular function has dimension-free finite-time complexity. On the positive side, we show that with a standard first-order oracle, there is an algorithm with dimension-free finite-time complexity for computing a Goldstein approximate stationary (GAS) point of a Lipschitz function. If time permits, we also discuss how such an algorithm can be used to compute NAS points of certain Clarke irregular Lipschitz functions that arise in machine learning applications.

The talk is based on joint work with Lai Tian.

16 Yong Sheng Soh

National University of Singapore, Singapore

Optimal Regularizers for Data via Shape Regression

Abstract

Regularization techniques are frequently deployed in the solution of ill-posed inverse problems. These take the form of penalty functions that are appended to the objective, and whose role is to encourage certain structure in solutions – prominent examples include the L1 norm for inducing sparsity, and the total variation norm for inducing smoothness.

The process of choosing an appropriate regularizer is a delicate process which requires deep domain expertise about the application domain. More recently, data-driven techniques in which regularizers learned directly from clean data are increasingly used in practice, and often out-perform carefully handcrafted ones.

In this talk, we try to understand what an optimal choice of regularizer for a specific application domain looks like. The key idea is to transform the optimization problem, which is stated over the space of functions, into a shape regression instance. Many statements about optimality seamlessly translate to extremal geometric problems, which have been widely studied. At the end, we briefly discuss what our analyses teach us in relation to the dictionary learning problem, as well as learned regularizers specified as neural networks.

17 Mahdi Soltanolkotabi

University of Southern California, USA
Demystifying Feature learning via gradient descent with applications to medical image reconstruction

Abstract

In this talk I will discuss the challenges and opportunities for using deep learning in medical image reconstruction. Contemporary techniques in this field rely on convolutional architectures that are limited by the spatial invariance of their filters and have difficulty modeling long-range dependencies. To remedy this, I will discuss our work on designing new transformer-based architectures called HUMUS-Net that lead to state of the art performance and do not suffer from these limitations. A key component in the success of the above approach is a unique feature learning capability of unrolled neural networks trained based on end-to-end training. In the second part of the talk I will demystify this feature learning capability of neural networks in this context as well as more broadly for other problems. Our result is based on an intriguing spectral bias phenomena for gradient descent, that puts the iterations on a particular trajectory towards solutions that learn good features that generalize well. Notably this analysis overcomes a major theoretical bottleneck in the existing literature and goes beyond the "lazy" training regime which requires unrealistic hyperparameter choices (e.g. very small step sizes, large initialization or wide models).

18 Akiko Takeda

The University of Tokyo, Japan
Generalized Levenberg–Marquardt method with oracle complexity bound and local quadratic convergence

Abstract

Nonconvex optimization problems of minimizing the sum of a possibly nonsmooth convex function and a smooth composite function arise naturally in various contemporary applications such as machine learning. The generalized Levenberg– Marquardt (LM) method, also known as the prox-linear method,

has been developed for such problems. In this talk, we propose a new generalized LM method with three theoretical guarantees: iteration complexity bound, oracle complexity bound, and local convergence under a Holderian growth condition. The method iteratively solves strongly convex subproblems with a damping term, and these theoretical guarantees are achieved by the update rule of the damping parameter and the inexact solution method for subproblems. This is a joint work with Naoki Marumo and Takayuki Okuno.

19 Christos Thrampoulidis

University of British Columbia, Canada
Finding Structures in Large Models: Imbalance Trouble

Abstract

What are the unique structural properties of models learned deep nets? Is there an implicit bias towards solutions of a certain geometry? How does this vary across training instances, architectures, and data? Towards answering these questions, the recently discovered Neural Collapse phenomenon formalizes simple geometric properties of the learned embeddings and of the classifiers, which appear to be "cross-situational invariant" across architectures and different balanced classification datasets.

But what happens when classes are imbalanced? Is there a (ideally equally simple) description of the geometry that is invariant across class-imbalanced datasets? By characterizing the global optima of an unconstrained-features model, we formalize a new geometry that remains invariant across different imbalance levels. Importantly, it, too, has a simple description despite the asymmetries imposed by data imbalances on the geometric properties of different classes.

Overall, we show that it is possible to extend the scope of the neural-collapse phenomenon to a richer class of geometric structures. We also motivate further investigations into the impact of class imbalances on the implicit bias of first-order methods and into the potential connections between such geometry structures and generalization.

20 Xin Tong

National University of Singapore, Singapore
Sampling with constraints using variational methods

Abstract

Sampling-based inference and learning techniques, especially Bayesian inference, provide an essential approach to handling uncertainty in machine learning (ML). As these techniques are increasingly used in daily life, it becomes essential to safeguard the ML systems with various trustworthy-related constraints, such as fairness, safety, interpretability. We propose a family of constrained sampling algorithms which generalize Langevin Dynamics (LD) and Stein Variational Gradient Descent (SVGD) to incorporate a moment constraint or a level set specified by a general nonlinear function. By exploiting the gradient flow structure of LD and SVGD, we derive algorithms for handling constraints, including a primal-dual gradient approach and the constraint controlled gradient descent approach. We investigate the continuous-time mean-field limit of these algorithms and show that they have $O(1/t)$ convergence under mild conditions.

21 Shuoguang Yang

Hong Kong University of Science and Technology, China
Decentralized Gossip-Based Stochastic Bilevel Optimization over Communication Networks

Abstract

Bilevel optimization have gained growing interests, with numerous applications found in meta learning, minimax games, reinforcement learning, and nested composition optimization. This paper studies the problem of distributed bilevel optimization over a network where agents can only communicate with neighbors, including examples from multi-task, multi-agent learning and federated learning. In this paper, we propose a gossip-based distributed bilevel learning algorithm that allows networked agents to solve both the inner and outer optimization problems in a single timescale and share information via network propagation. We show that our algorithm enjoys the $O(1/K\epsilon^2)$ per-agent sample complexity for general nonconvex bilevel

optimization and $O(1/K\epsilon)$ for strongly convex objective, achieving a speedup that scales linearly with the network size. The sample complexities are optimal in both ϵ and K. We test our algorithm on the examples of hyperparameter tuning and decentralized reinforcement learning. Simulated experiments confirmed that our algorithm achieves the state-of-the-art training efficiency and test accuracy.

22 Antonios Varvitsiotis

Singapore University of Technology and Design, Singapore
TBA

Abstract

Understanding how populations of infinitesimal anonymous agents learn to adapt their behavior over time as a result of repeated strategic interactions is a fundamental problem in science and engineering. In this work we develop a theoretical and algorithmic framework for identifying the dynamics that govern agent behavior using only samples from a short-run of a single system trajectory. In contrast to the modern data-driven paradigm for model discovery, our framework is applicable to settings where observational data is scarce. Our computational approach uses sum-of-squares optimization to solve a side-information assisted polynomial regression problem, where we compensate for the absence of data by incorporating side-information constraints modeling a wide range of agent behaviors. Our experimental results demonstrate that the dynamics recovered by our method are indistinguishable from the ground truth dynamics with respect to important benchmarks, including the equilibrium selection problem in congestion games and the exact identification of chaotic dynamics. Based on joint work with Georgios Piliouras and Joseph Sakos.

23 Man-Chung Yue

The University of Hong Kong, China
Nonlinear Covariance Shrinkage Estimator via Distributionally Robust Optimization

Abstract

The covariance matrix of a random vector is an important object in probability theory and statistics that finds a wide range of applications. A standard covariance matrix estimator is the sample covariance matrix, which performs poorly in high-dimensional regime. One way to improve the performance is to apply a linear shrinkage operation on the sample covariance matrix. In this talk, we develop a new family of nonlinear covariance shrinkage estimators based on the concept of distributionally robust optimization and provide a unified theoretical analysis. Not only do the proposed nonlinear shrinkage covariance matrix estimators enjoy many desirable properties, but they are also efficiently computable. We then conclude the talk by presenting some numerical experiments on the proposed estimators using both synthetic and real data.

24 Anru Zhang

Duke University, USA
Tensor Learning in 2020s: Methodology, Theory, and Applications

Abstract

The analysis of tensor data, i.e., arrays with multiple directions, has become an active research topic in the era of big data. Datasets in the form of tensors arise from a wide range of scientific applications. Tensor methods also provide unique perspectives to many high-dimensional problems, where the observations are not necessarily tensors. Problems in high-dimensional tensors generally possess distinct characteristics that pose great challenges to the data science community.

In this talk, we discuss several recent advances in tensor learning and their applications in genomics, computational imaging, and electronic health records. We also illustrate how we develop statistically optimal methods and computationally efficient algorithms that interact with the modern theories of computation and non-convex optimization.