Speakers

1 Ahmet Alacaoglu

2 Krishna Balasubramanian

3 Kabir Chandrasekher

4 Robert Freund

5 Michael Gastner

6 Geovani Grapiglia

7 Serge Gratton

8 Patrice Koehl

9 Donghwan Kim

10 Guanghui Lan

11 Ching-pei Lee

12 Yin Tat Lee

13 Zhaosong Lu

14 Laura Palagi

15 Peter Richtarik

16 Clément W. Royer

17 Katya Scheinberg

18 Yan Shuo Tan

19 Philippe Toint

20 Stephen Wright

21 Stephen Wright

22 Yancheng Yuan

23 Junyu Zhang

24 Yangjing Zhang

Abstracts

Workshop 1: Fast Optimization Algorithms in the Big Data Era

(5–9 Dec 2022)

1 Ahmet Alacaoglu

University of Wisconsin–Madison, USA
Randomized first-order algorithms for min-max problems

Abstract

In this talk, we focus on two classes of randomized first order algorithms for solving min-max problems in the finite sum form. When the coupling between primal and dual variables is bilinear, we use random coordinate updates with first order primal-dual algorithms and derive complexity results improving the deterministic baselines. We also showcase the adaptivity of these algorithms to problem structures such as sparsity, both in theory and practice. When the coupling between primal and dual variables is nonbilinear, we present a variance reduced algorithm that comes with improved theoretical and practical properties compared to baselines such as the extragradient algorithm.

2 Krishna Balasubramanian

University of California at Davis, USA
High-dimensional Inference with Stochastic Approximation Algorithms

Abstract

Stochastic Gradient Descent (SGD) is widely used in modern data science. Existing analyses of SGD have predominantly focused on the fixed-dimensional

setting. In order to perform high-dimensional statistical inference with such algorithms, it is important to study the dynamics of SGD under high-dimensional scalings. In this talk, I will discuss high-dimensional limit theorems and bounds for the online least-squares SGD iterates for solving over-parameterized linear regression. First, focusing on the asymptotic setting (i.e., when both the dimensionality and iterations tend to infinity), I will first present the mean-field limit (in the form of infinite-dimensional ODEs) and fluctuations (in the form of infinite-dimensional SDEs) for the online least-squares SGD iterates. A direct consequence of the result is obtaining explicit forms and related fluctuations for the mean-squared error. Next, focusing on the non-asymptotic setting, I will discuss Berry-Esseen bounds for linear functionals of the online least-squares SGD iterates. This result holds in particular as long as the dimensionality grows at least polynomially in terms of the number of iterations (or equivalently the number of observations used), and could be used to obtain prediction confidence intervals or entry-wise estimation confidence intervals.

3 Kabir Chandrasekher

Stanford University, USA

Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization

Abstract

We consider the problem of estimating the factors of a rank-1 matrix with i.i.d. Gaussian, rank-1 measurements that are nonlinearly transformed and corrupted by noise. Considering two prototypical choices for the nonlinearity, we study the convergence properties of a natural alternating update rule for this nonconvex optimization problem starting from a random initialization. We show sharp convergence guarantees for a sample-split version of the algorithm by deriving a deterministic recursion that is accurate even in high-dimensional problems. Notably, while the infinite-sample population update is uninformative and suggests exact recovery in a single step, the algorithm—and our deterministic prediction—converges geometrically fast from a random initialization. Our sharp, non-asymptotic analysis also exposes several other fine-grained properties of this problem, including how the nonlinearity and noise level affect convergence behavior. On a technical

level, our results are enabled by showing that the empirical error recursion can be predicted by our deterministic sequence within fluctuations of the order $n^{-1/2}$ when each iteration is run with n observations. Our technique leverages leave-one-out tools originating in the literature on high-dimensional M-estimation and provides an avenue for sharply analyzing higher-order iterative algorithms from a random initialization in other high-dimensional optimization problems with random data.

4 Robert Freund

MIT Sloan School of Management, USA
Recent Advances in First-Order Methods for Linear Programming

Abstract

Linear programming (LP) is an essential optimization problem, either by itself or as a critical subproblem of a mixed integer program. In the last several years, as the scale of data and LPs has increased enormously, first-order methods (FOMs) have been developed for LP with the goal of solving huge-scale applications for which classic algorithms fail (namely simplex and interior-point methods). In this talk we present some new theoretical and computational results for FOMs for LP. We introduce new approaches to analyzing convergence guarantees that are more consistent with observed practical performance. We identify key bottlenecks of some existing methods, and we present new methods that resolve these bottlenecks and improve on current algorithms. This is joint work with Zikai Xiong, an MIT OR Center doctoral student.

5 Michael Gastner

Yale-NUS College, Singapore
Accelerating the calculation of optimally smooth pycnophylactic interpolations

Abstract

Many quantitative geospatial data are collected and aggregated in discrete enumeration units (e.g. countries or provinces). Smooth pycnophylactic interpolation aims to find a smooth, non-negative function such that the area

integral over each enumeration unit is equal to the aggregated data. Conventionally, smooth pycnophylactic interpolation is achieved by a cellular automaton algorithm that converts a piecewise constant function into an approximately smooth function defined on a grid of coordinates on a geographic equal-area map projection. Although many software packages include implementations of this algorithm, its convergence is slow. A faster method is to construct a density-equalising map projection in which areas of enumeration units are proportional to the aggregated data. A pycnophylactic interpolation can be efficiently obtained from the Jacobian of this map projection, but solutions do not necessarily achieve optimal smoothness. In this talk, I present ongoing work that aims to further accelerate the density-equalising algorithm without sacrificing the near-optimality achieved by the conventional cellular-automaton method.

6 Geovani Grapiglia

UC Louvain, Belgium
Derivative-Free Optimization Methods based on Finite-Differences

Abstract

In many practical optimization problems, the gradients of the functions involved are not readily available. Examples include the tuning of algorithmic parameters, model calibration, and also the design of black-box attacks on Deep Neural Networks. These problems can be addressed with Derivative-Free Optimization (DFO) methods, which are methods that rely only on function evaluations. Very often, the evaluation of the objective functions is computationally expensive. Therefore, one of the main concerns in DFO is the development of methods with a low worst-case complexity in terms of function evaluations. In this talk, I will discuss some recent worst-case complexity results obtained for DFO methods based on finite-difference gradient approximations.

7 Serge Gratton

ENSEEIHT and Université de Toulouse, France
Complexity and performance for two classes of noise-tolerant first-order algorithms

Abstract

Two classes of algorithms for optimization in the presence of noise are presented, that do not require the evaluation of the objective function. The first generalizes the well- known Adagrad method. Its complexity is then analyzed as a function of its parameters, and it is shown that some methods of the class enjoy a better asymptotic convergence rate than previously known. A second class of algorithms is then derived whose complexity is at least as good as that of the first class. Initial numerical experiments on finite-sum problems arising from deep-learning applications suggest that that methods of the second class often outperform those of the first.

8 Patrice Koehl

University of California, Davis, USA

Light speed computation of exact solutions to generic and to degenerate assignment problems

Abstract

The linear assignment problem is a fundamental problem in combinatorial optimization with a wide range of applications, from operational research to data sciences. It consists of assigning "agents" to "tasks" on a one-to-one basis, while minimizing the total cost associated with the assignment. While many exact algorithms have been developed to identify such an optimal assignment, most of these methods are computationally prohibitive for large size problems. In this talk, I will describe a novel approach to solving the assignment problem using techniques adapted from statistical physics. In particular I will derive a strongly concave effective free energy function that captures the constraints of the assignment problem at a finite temperature. This free energy decreases monotonically as a function of beta, the inverse of temperature, to the optimal assignment cost, providing a robust framework for temperature annealing. For large enough beta values the exact solution to the generic assignment problem can be derived using a simple round-off to the nearest integer of the elements of the computed assignment matrix. I will also describe a provably convergent method to handle degenerate assignment problems. Finally, I will describe computer implementations of this framework that are optimized for parallel architectures, one based on CPU,

the other based on GPU. These implementations enable solving large assignment problems (of the orders of a few 10000s) in computing clock times of the orders of minutes.

9 Donghwan Kim

Korea Advanced Institute of Science Technology, Korea
Semi-Anchored Multi-Step Gradient Method for Nonconvex-Nonconcave Minimax Optimization

Abstract

Minimax problems, such as generative adversarial network, adversarial training, and fair training, are widely solved by a multi-step gradient descent ascent (MGDA) method in practice. However, its convergence guarantee is limited. In this work, inspired by the primal-dual hybrid gradient method, we propose a new semi-anchoring (SA) technique for the MGDA method. This makes the MGDA method find a stationary point of a structured nonconvex-nonconcave composite minimax problem The resulting method, named SA-MGDA, is built upon the Bregman proximal point method.

10 Guanghui Lan

Georgia Institute of Technology, USA
Policy Optimization over General State and Action Spaces

Abstract

Reinforcement learning (RL) problems over general state and action spaces are notoriously challenging. In contrast to the tableau setting, one cannot enumerate all the states and then iteratively update the policies for each state. This prevents the application of many well-studied RL methods especially those with provable convergence guarantees. In this talk, we first present a substantial generalization of the recently developed policy mirror descent method to deal with general state and action spaces. We introduce new approaches to incorporate function approximation into this method, so that we do not need to use explicit policy parameterization at all. Moreover, we present a novel policy dual averaging method for which possibly simpler

function approximation techniques can be applied. We establish linear convergence rate to global optimality or sublinear convergence to stationarity for these methods applied to solve different classes of RL problems under exact policy evaluation. We then define proper notions of the approximation errors for policy evaluation and investigate their impact on the convergence of these methods applied to general-state RL problems with either finite-action or continuous-action spaces. To the best of our knowledge, the development of these algorithmic frameworks as well as their convergence analysis appear to be new in the literature.

11 Ching-pei Lee

Academia Sinica, Taipei

Solution Structure Utilization for Efficient Optimization and Large-scale Machine Learning

Abstract

Regularized optimization that adds a regularizer to the objective function for minimization is widely used in numerous applications in machine learning and signal processing to induce desired structures in an optimal solution. In this talk, I will first describe how to find such a structure in an approximate solution, without obtaining the optimal solution that is usually only the limit point of an iterative algorithm, in different scenarios and discuss why this matters.

I will then show how we can utilize such an optimal structure to devise more efficient optimization algorithms. Examples of our algorithms designed for specific regularizers used in various tasks in machine learning and signal processing to either identify such optimal structures for better performance or utilize them to accelerate the optimization will be discussed, including training structured neural network models, matrix completion, federated learning, and the best subset selection problem.

This is joint work with Yu-Sheng Li, Wei-Lin Chiang, Zih-Syuan Huang, Jan Harold Alcantara, Ling Liang, Tianyun Tang, Kim-Chuan Toh.

12 Yin Tat Lee

University of Washington, USA
From Robustness to Efficiency

Abstract

Traditionally, optimization methods like (stochastic) gradient descent take time at least linear to the number of parameters. For problems with many parameters, each step is too expensive.

In this talk, I will discuss how designing robust algorithms can lead to faster algorithms. In particular, I will explain the robust interior point method. Its robustness allows us to speed up its iteration cost via data structures. This results in theoretically faster algorithms for many important problems such as linear programming, semidefinite programming, and the maximum flow problem.

13 Zhaosong Lu

University of Minnesota, USA
First-order methods for convex optimization and monotone inclusions under local Lipschitz conditions

Abstract

In this talk, we first consider convex optimization whose smooth components have a locally Lipschitz continuous gradient and propose a first-order method for finding an epsilon-KKT solution. We then consider monotone inclusion in which the point-valued operator is locally Lipschitz continuous and propose a primal-dual extrapolation method for finding an epsilon-residual solution. All the proposed methods are parameter free with a verifiable termination criterion and also enjoy a nearly optimal complexity. This is joint work with Sanyou Mei (University of Minnesota).

14 Laura Palagi

Sapienza University of Rome, Italy
Convergence under Lipschitz smoothness of ease-controlled Random Reshuffling gradient Algorithms

Abstract

We consider minimizing the average of a very large number of smooth and possibly non-convex functions. This optimization problem has deserved much attention in the past years due to the many applications in different fields, the most challenging being training Machine Learning models. Widely used approaches for solving this problem are mini-batch (online) gradient methods which, at each iteration, update the decision vector moving along the gradient of a mini-batch selection of the component functions. Depending on the selection rule for the mini-batch different methods have been defined. We consider the Incremental Gradient (IG) and the Random reshuffling (RR) methods which proceed in cycles, picking a fixed or uniformly random order (permutation), respectively, and processing the component functions according to this order. Convergence properties of these schemes have been proved under different assumptions, usually quite strong. We aim to define ease-controlled modifications of the IG/RR schemes, which require a light additional computational effort and can be proved to converge under very weak and standard assumptions. In particular, we define two algorithmic schemes in which the IG/RR iteration is controlled by using a watchdog rule and a derivative-free line search that activates only sporadically to guarantee convergence. The two schemes differ in the watchdog and the line search, which are performed using either a monotonic or a non-monotonic rule. The two schemes also allow controlling the updating of the learning rate used in the main IG/RR iteration, avoiding the use of preset rules that may drive it to zero too fast, thus overcoming another tricky aspect in implementing online methods, which is the updating rule of the stepsize. We proved convergence under the lonely assumption of Lipschitz continuity of the gradients of the component functions. We performed an extensive computational test using different Deep Neural Architectures and a benchmark of varying size datasets. We compare our implementation with both full batch gradient methods and online standard implementation of IG/RR methods, proving that the computational effort is comparable with the corresponding online methods and that the control on the learning rate may allow faster decrease.

15 Peter Richtarik

King Abdullah University of Science and Technology, Saudi Arabia
ProxSkip: Local gradient steps provably lead to communication acceleration

Abstract

In this talk I will introduce ProxSkip [1] - a surprisingly simple and provably efficient method for minimizing the sum of a smooth (f) and an expensive nonsmooth proximable (ψ) function. The canonical approach to solving such problems is via the proximal gradient descent (ProxGD) algorithm, which is based on the evaluation of the gradient of f and the prox operator of ψ in each iteration. In this work we are specifically interested in the regime in which the evaluation of prox is costly relative to the evaluation of the gradient, which is the case in many applications. ProxSkip allows for the expensive prox operator to be skipped in most iterations: while its iteration complexity is $O(\kappa log 1/\epsilon)$, where κ is the condition number of f, the number of prox evaluations is $O(\sqrt{\kappa} log 1/\epsilon)$ only. Our main motivation comes from federated learning, where evaluation of the gradient operator corresponds to taking a local GD step independently on all devices, and evaluation of prox corresponds to (expensive) communication in the form of gradient averaging. In this context, ProxSkip offers an effective acceleration of communication complexity. Unlike other local gradient-type methods, such as FedAvg [2], SCAFFOLD [3], S-Local-GD [4] and FedLin [5], whose theoretical communication complexity is worse than, or at best matching, that of vanilla GD in the heterogeneous data regime, we obtain a provable and large improvement without any heterogeneity-bounding assumptions.

Time permitting, I will mention several subsequent extensions, generalizations and improvements [6, 7, 8].

References

[1] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich and Peter Richtárik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! 39th International Conference on Machine Learning (ICML 2022)

[2] Brendan McMahan, Eider Moore, Daniel Ramage and Blaise Agüera y Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, https://arxiv.org/abs/1602.05629v3, 2016

[3] Sai Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, Ananda Suresh, SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning, ICML 2020

[4] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik, Local SGD: unified theory and new efficient methods, NeurIPS 2020

[5] Aritra Mitra and Rayana Jaafar and George J. Pappas and Hamed Hassani, Linear Convergence in Federated Learning: Tackling Client Heterogeneity and Sparse Gradients, NeurIPS 2021

[6] Abdurakhmon Sadiev, Dmitry Kovalev and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with inexact prox, https://arxiv.org/abs/2207.03957, 2022

[7] Grigory Malinovsky, Kai Yi and Peter Richtárik. Variance reduced ProxSkip: Algorithm, theory and application to federated learning, https://arxiv.org/abs/2207.04338, 2022

[8] Laurent Condat and Peter Richtárik. RandProx: Primal-dual optimization algorithms with randomized proximal updates, https://arxiv.org/abs/2207.12891, 2022

16 Clément W. Royer

Université Paris Dauphine-PSL, France

Optimization methods for highly nonconvex data science tasks

Abstract

Nonconvex optimization formulations arise not only in deep learning, but also in data analysis and robust statistics. Although such problems typically come with a great deal of structure, certain formulations are quite challenging from an algorithmic perspective, as they induce complex optimization landscapes and numerous critical points. Designing fast algorithms for such problems is thus a challenging endeavor. On one hand, theoretical guarantees such as

complexity results (or convergence rates), that characterize how fast a given method can be in the worst case, often fail at representing the behavior of an algorithm over more specific instances. On the other hand, general purpose nonlinear optimization schemes may prove quite efficient on such problems, often without convergence rate guarantees or principled ways of exploiting the problems' structure.

In this talk, we investigate two highly nonconvex optimization problems from data science. For both problems, our goal is to propose a method that is not also theoretically fast in a complexity sense but also practical for the task at hand. The first part of the talk will discuss nonconvex formulations of robust linear regression. On such instances, we observe that the theoretically fastest methods are outperformed by standard schemes such as nonlinear conjugate gradient. We then propose an algorithm that is both equipped with complexity guarantees and of practical interest for this task. In the second part of the talk, we consider a nonconvex optimization problem resulting from the training of a discretized neural ordinary differential equation model. We leverage recent results from the landscape of shallow networks to understand that of our problem, thereby illustrating its highly nonconvex nature. Given such a landscape, we consider several optimization schemes with complexity guarantees, and explore both their convergence rates and their ability to avoid spurious critical points in practice.

17 Katya Scheinberg

Cornell University, USA
Stochastic oracles and how to define them

Abstract

Continuous optimization is a mature field, which has recently undergone major expansion and change. One of the key new directions is the development of methods that do not require exact information about the objective function.

Nevertheless, the majority of these methods, from stochastic gradient descent to "zero-th order" methods use some kind of approximate first order information.

We will introduce a general definition of a stochastic oracle and show how this definition applies in a variety of familiar settings, including simple

stochastic gradient via sampling, traditional and randomized finite difference methods and more. We will overview several stochastic methods and how the general definition extends to the oracles used by these methods.

18 Yan Shuo Tan

National University of Singapore, Singapore

A Mixing Time Lower Bound for a Simplified Version of Bayesian Additive Regression Trees (BART)

Abstract

Bayesian Additive Regression Trees (BART) is a popular Bayesian nonparametric regression algorithm. The posterior is a distribution over sums of decision trees, and predictions are made by averaging approximate samples from the posterior. The combination of strong predictive performance and the ability to provide uncertainty measures has led BART to be commonly used in the social sciences, biostatistics, and causal inference. BART uses Markov Chain Monte Carlo (MCMC) to obtain approximate posterior samples over a parameterized space of sums of trees, but it has often been observed that the chains are slow to mix. We provide the first lower bound on the mixing time for a simplified version of BART in which we reduce the sum to a single tree and use a subset of the possible moves for the MCMC proposal distribution. Our lower bound for the mixing time grows exponentially with the number of data points. Inspired by this new connection between the mixing time and the number of data points, we perform rigorous simulations on BART. We show qualitatively that BART's mixing time increases with the number of data points. The slow mixing time of the simplified BART suggests a large variation between different runs of the simplified BART algorithm. A similar large variation is known for BART in the literature. This large variation could result in a lack of stability in the models, predictions, and posterior intervals obtained from the BART MCMC samples. Our lower bound and simulations suggest the need for improvements to the MCMC sampler.

19 Philippe Toint

Université de Namur, Belgium
Objective-Function-Free Optimization, Part II:
Complexity of Adaptive Regularization and Numerical Experiments

Abstract

An adaptive regularization algorithm for unconstrained nonconvex optimization is presented in which the objective function is never evaluated, but only derivatives are used. This algorithm belongs to the class of adaptive regularization methods, for which optimal worst-case complexity results are known for the standard framework where the objective function is evaluated. It is shown in this paper that these excellent complexity bounds are also valid for the new algorithm, despite the fact that significantly less information is used. In particular, it is shown that, if derivatives of degree one to p are used, the algorithm will find a ϵ_1-approximate first-order minimizer in at most $(\epsilon_1^{-(p+1)/p})$ iterations, and an (ϵ_1, ϵ_2)-approximate second-order minimizer in at most $(\max[\epsilon^{-(p+1)/p}, \epsilon_2^{-(p+1)/(p-1)}])$ iterations. As a special case, the new algorithm using first and second derivatives, when applied to functions with Lipschitz continuous Hessian, will find an iterate x_k at which the gradient's norm is less than ϵ_1 in at most $(\epsilon_1^{-3/2})$ iterations.

Numerical experiments will finally be presented showing the surprisingly good performance and robustness of "flexible" variants of the new methods (Adagrad-like and adaptive regularization).

20 Stephen Wright

University of Wisconsin–Madison, USA
Optimization in theory and practice

Abstract

Complexity analysis in optimization seeks upper bounds on the amount of work required to find approximate solutions of problems in a given class with a given algorithm, and also lower bounds, usually in the form of a worst-case example from a given problem class. The relationship between theoretical complexity bounds and practical performance of algorithms on "typical" problems varies widely across problem and algorithm classes, and relative

interest among researchers between the theoretical and practical aspects of algorithm design and analysis has waxed and waned over the years. This talk surveys complexity analysis and its relationship to practice in optimization, with an emphasis on linear programming and convex and nonconvex nonlinear optimization, providing historical (and cultural) perspectives on research in these areas.

21 Stephen Wright

University of Wisconsin–Madison, USA
Primal-dual optimization methods for robust machine learning

Abstract

We consider a convex-concave primal-dual optimization framework in which the coupling between primal and dual variables is bilinear. This framework admits linearly constrained optimization together with a variety of interesting problems in machine learning, including (linear) empirical risk minimization with various regularization terms. It also includes a formulation that we term "generalized linear programming" (GLP) in which regularization terms and constraints are added to the traditional linear programming formulation, provided they admit efficient prox operations. Problems from differentially robust optimization (DRO), using either f-divergence metrics or Wasserstein metrics, can be formulated as GLPs.

We describe algorithms for our framework that take prox-gradient steps alternately in the primal and dual variables, but incorporate such additional features as coordinate descent, variance reduction, dual averaging, importance sampling, and iterate averaging. Our methods can also exploit sparsity in the matrix that couples primal and dual variables. Our methods match or improve on the best known worst-case complexity bounds in various settings. Computational experiments indicate that our methods also have good practical performance.

The talk represents joint work with Ahmet Alacaoglu, Jelena Diakonikolas, Chaobing Song, Eric Lin, and Volkan Cevher

22 Yancheng Yuan

The Hong Kong Polytechnic University, Hong Kong

An Efficient HPR Algorithm for the Wasserstein Barycenter Problem with $O(\mathrm{Dim(P)}/\varepsilon)$ Computational Complexity

Abstract

In this talk, we will introduce and analyze an efficient Halpern-Peaceman-Rachford (HPR) algorithm for solving the Wasserstein barycenter problem (WBP) with fixed supports. While the Peaceman-Rachford (PR) splitting method itself may not be convergent for solving the WBP, the HPR algorithm can achieve an $O(1/\varepsilon)$ non-ergodic iteration complexity with respect to the Karush-Kuhn-Tucker (KKT) residual. More interestingly, we propose an efficient procedure with linear time computational complexity to solve the linear systems involved in the subproblems of the HPR algorithm. As a consequence, the HPR algorithm enjoys an $O(\mathrm{Dim(P)}/\varepsilon)$ non-ergodic computational complexity in terms of flops for obtaining an ε-optimal solution measured by the KKT residual for the WBP, where Dim(P) is the dimension of the variable of the WBP. This is better than the best-known complexity bound for the WBP. Moreover, the extensive numerical results on both the synthetic and real data sets demonstrate the superior performance of the HPR algorithm for solving the large-scale WBP.

23 Junyu Zhang

National University of Singapore, Singapore

A Unified Primal-Dual Algorithm Framework for Inequality Constrained Problems

Abstract

In this paper, we propose a unified primal-dual algorithm framework based on the augmented Lagrangian function for composite convex problems with conic inequality constraints. The new framework is highly versatile. First, it not only covers many existing algorithms such as PDHG, Chambolle-Pock (CP), GDA, OGDA and linearized ALM, but also guides us to design a new efficient algorithm called Simi-OGDA (SOGDA). Second, it enables us to study the role of the augmented penalty term in the convergence analysis.

Interestingly, a properly selected penalty not only improves the numerical performance of the above methods, but also theoretically enables the convergence of algorithms like PDHG and SOGDA. Under properly designed step sizes and penalty term, our unified framework preserves the $\mathcal{O}(1/N)$ ergodic convergence while not requiring any prior knowledge about the magnitude of the optimal Lagrangian multiplier. Linear convergence rate for affine equality constrained problem is also obtained given appropriate conditions. Finally, numerical experiments on linear programming, ℓ_1 minimization problem, and multi-block basis pursuit problem demonstrate the efficiency of our methods.

24 Yangjing Zhang

Chinese Academy of Sciences, China

On Efficient and Scalable Computation of the Nonparametric Maximum Likelihood Estimator in Mixture Models

Abstract

The nonparametric maximum likelihood estimation (NPMLE) is a classic and important method to estimate the mixture models from finite observations. The discretization of the infinite dimensional probability measure with a fixed support in the NPMLE leads to a finite dimensional convex optimization problem. Although can be solved by off-the-shelf interior point based solvers, the algorithm does not scale well with the number of grid points (denoted as m) and the number of observed data points (denoted as n). In this talk, leveraging the observation that the solution of the finite dimensional NPMLE is usually sparse, we propose an efficient and scalable semismooth Newton based augmented Lagrangian method (ALM). By carefully exploring the structure of the ALM subproblem, we show that the computational cost of the generalized Hessian (second order information) is independent of the number of grid points. Hence, compared with the recent work Kim et al. (2020,JCGS) that using the active set based sequential quadratic programming to solve the NPMLE whose Hessian evaluation scales quadratically in m, our proposed algorithm is in particular suitable for the multivariate mixture models where a large number of grid points is needed in the multi-dimensional space. Extensive numerical experiments are conducted to show the effectiveness of our approach. In particular, we are able to get highly accurate solutions for the multivariate synthetic data with n=100,000 and m=10,000

within one minute. For the astronomy data set Gaia TGAS that contains 1.4 million dereddened stars with 10,000 grid points in the 2-dimensional space, our algorithm successfully solve it in about seven hours.