March 20, 2026, 8:00 AM - 9:30 AM

FA01

Grand Ballroom

Optimization and Computational Economics

Invited Session

Optimization in Data Science

Chair: Chuwen Zhang, University of Chicago Booth School of Business, Chicago, IL, United States

1 - The Second-Order Tonnement: Decentralized Interior-Point Methods for Market Equilibrium

Chuwen Zhang, University of Chicago Booth School of Business, Chicago, IL, United States, Chang He, Bo Jiang, Yinyu Ye

The Walras tâtonnement process and Smale's process are two classical approaches to compute market equilibria in exchange economies such as the Fisher market in the last century. While the tâtonnement process can be seen as a first-order method where only customers’ best-demand-response information is needed, Smale's process is a second-order Newton method and it is less realistic due to its reliance on additional customer-demand derivative information and computation-expensive Newton steps. At the beginning of 2000, the convex optimization such as interior-point methods were developed for computing Fisher-market equilibria and they were proved to be polynomial-time. However, they are also second-order Newton algorithms that need full customer utility-demand function information. In this work, for the first time, we develop second-order polynomial-time tâtonnement methods that only need consumers' best-demand-response information. To address the expensive Newton steps, we also introduce an explicitly Newton-step approximation with high-probability guarantees and a scaling matrix that optimally minimizes the Newton-system condition number, both of which again rely solely on best-demand-responses. Preliminary tests are presented to justify the capability of the proposed methods for solving large-scale problems. Extensions of our approach are also discussed.

2 - Tight Non-asymptotic Local Convergence and Acceleration of Sinkhorn's Algorithm

Zhaonan Qu, Columbia University, New York, NY, United States, Wenzhi Gao

Sinkhorn's algorithm is a simple yet powerful method that has found wide applications, most notably in computational optimal transport. More recently it has featured prominently in the training of large language models such as DeepSeek. Despite the many previous works that study its convergence behavior, a tight non-asymptotic local convergence analysis remains absent. In this paper, we solve this open problem. We show that the local convergence rate Sinkhorn algorithm is $O(1/\lambda_2 \log(1/\varepsilon))$, where λ_2 is the second largest eigenvalue of a matrix associated with the original problem data. Extensive numerical experiments validate the tightness of our result. Our analysis also shows an implicit connection between the Sinkhorn algorithm and the preconditioned gradient descent, and with this connection, we show that an accelerated $O(1/\sqrt{\lambda_2} \log(1/\varepsilon))$ local convergence is achievable through acceleration.

3 - Accelerated Price Adjustment for Fisher Markets with Exact Recovery of Competitive Equilibrium

Chonghe Jiang, Massachusetts Institute of Technology, Cambridge, MA, United States, He Chen, Anthony So

The canonical price-adjustment process, t\^atonnement, typically fails to converge to the exact competitive equilibrium (CE) and requires a high iteration complexity of $\tilde{\cO}(1/\epsilon)$ to compute ϵ-

CE prices in widely studied linear and quasi-linear Fisher markets. This paper proposes refined price-adjustment processes to overcome these limitations. By formulating the task of finding CE of a (quasi-)linear Fisher market as a strongly convex nonsmooth minimization problem, we develop a novel accelerated price-adjustment method (APM) that finds an ϵ-CE price in $\tilde{\cO}(1/\sqrt{\epsilon})$ lightweight iterations, which significantly improves upon the iteration complexities of t\^atonnement methods. Furthermore, through our new formulation, we construct a recovery oracle that maps approximate CE prices to exact CE prices at a low computational cost. By coupling this recovery oracle with APM, we obtain an adaptive price-adjustment method whose iterates converge to CE prices in finite steps. To the best of our knowledge, this is the first convergence guarantee to exact CE for price-adjustment methods in linear and quasi-linear Fisher markets. Our developments pave the way for efficient lightweight computation of CE prices. We also present numerical results to demonstrate the fast convergence of the proposed methods and the efficient recovery of CE prices.

4 - Competitive Equilibrium for Chores: from Dual Eisenberg-Gale to a Fast, Greedy, LP-based Algorithm

Tianlong Nan, Columbia University, New York, NY, United States, Bhaskar Ray Chaudhury, Christian Kroer, Ruta Mehta

We study the computation of competitive equilibrium for Fisher markets with n agents and m divisible chores. Competitive equilibria for chores are known to correspond to the nonzero KKT points of a program that minimizes the product of agent disutilities, which is a non convex program whose zero points foil iterative optimization methods. We introduce a dual like analogue of this program, and show that a simple modification to our “dual” program avoids such zero points, while retaining the correspondence between KKT points and competitive equilibria. This allows, for the first time ever, application of iterative optimization methods over a convex region for computing competitive equilibria for chores. We next introduce a greedy Frank-Wolfe algorithm for optimization over our program and show a new state-of-the-art convergence rate to competitive equilibrium. Moreover, our method is significantly simpler than prior methods: each iteration of our method only requires solving a simple linear program. We show through numerical experiments that our method is computationally efficient: it easily solves every instance we tried in a few seconds, including instances with hundreds of agents and up to 1000 chores, usually in 10-30 iterations, is simple to implement, and has no numerical issues.

FA02

Georgian Room

Power Systems Planning: Optimization Modeling and Software Advancements

Contributed Session

Computational Optimization

Chair: Rachael Alfant, Sandia National Laboratories, Albuquerque, NM, United States

1 - Selecting Key Days for Electricity Grid Capacity Expansion Planning Considering Resilience to Compound Hazards

Rachael Alfant, Sandia National Laboratories, Albuquerque, NM, United States, Nicole Jackson, Kyle Skolfield, Casey Doyle, Thom Edwards

In electricity grid capacity expansion planning, representative days are used to model daily and hourly operating conditions for long-term investment planning. Common approaches to representative day selection rely on clustering algorithms (e.g., k-means). However, these methods may fail to capture operating conditions on “extreme days”, which are the days when peak loads occur. Failure to model these conditions can negatively impact system resilience. In particular, there is a need to effectively capture the operational and structural impact to the grid of compound hazards, as they may be more severe than individual hazards in isolation. We present a probabilistic approach for selecting extreme days for capacity expansion planning that explicitly considers the spatio-temporal relationships among compound hazards.

2 - Multifidelity Progressive Hedging and Bundling Strategies for Diverse Stochastic Programming Problems

Zachary Kilwein, Sandia National Laboratories, Albuquerque, NM, United States, Rachael Alfant, Kyle Skolfield, William Hart

In several scientific and engineering fields, well-defined models exist on a spectrum of fidelity with an inherent zero sum trade-off between accuracy and problem complexity. In many stochastic programming (SP) applications, computational tractability is a major challenge that in practice limits either the space of exploration, the fidelity of model used, or both. Decomposition methods such as progressive hedging can reduce CPU cost through parallelization but still requires choosing a single model fidelity. In this work, we detail our framework for creating multifidelity representations of stochastic problems and novel bundling strategies of scenarios that can accelerate search and convergence in large-scale SPs, while leveraging both low and high fidelity model representations. We demonstrate our methods on a diverse set of SP applications, including scheduling, grid operations and planning, and chemical process control in order to explore the effects of problem linearity, discrete/continuous variables, bundling strategies, and algorithmic hyperparameters.

3 - Alternative Solutions for Energy Applications without Recourse Assumptions

Matthew Viens, University of Wisconsin-Madison/Sandia National Labs, Madison, WI, United States, William Hart, Michael Ferris

Optimization problems can have multiple alternative solutions that achieve an optimal or near-optimal result. When applied to the power systems optimization context, our models can be stochastic programs with many scenarios and lack guarantees about (relatively) complete recourse. Such features can present a tension between standard stochastic programming decomposition strategies and needing to maintain feasibility. We extend our previous work on alternative solution generation in Benders Decomposition to address these concerns while preserving our enumeration and projection guarantees. We demonstrate how these extended tools allow alternative solution generation on a range of power systems optimization problems.

4 - Enhancing Stochastic Power Planning through Structured Nonlinear Global Decomposition

Georgia Stinchfield, Carnegie Mellon University, Pittsburgh, PA, United States, Matthew Viens, William Hart, Carl Laird

Two-stage stochastic power planning problems are central to modeling uncertainty in generation capacity expansion and investment decisions. These problems are often large-scale and potentially nonconvex, motivating the use of decomposition-based global optimization methods. In this work, we introduce a priority branch-and-bound (B&B) framework for decomposable nonlinear optimization, implemented in our open-source Python package, SNoGloDe (Structured Nonlinear Global Decomposition). The algorithm exploits the natural scenario-based separability of stochastic formulations: each realization of uncertain parameters defines a scenario subproblem, and solving these subproblems separately (ideally in parallel) provides a relaxation that serves as a valid bound for the original coupled problem.

Unlike conventional B&B algorithms that minimize node evaluations to close the optimality gap, our approach identifies diverse, high-quality alternative solutions within the search tree. In the context of power systems planning, such alternative solutions offer operators valuable insights into trade-offs among cost, reliability, and resilience under uncertainty.

SNoGloDe provides a flexible software framework for embedding these algorithms, integrating parallel computation, global optimization solvers, and customizable branching logic. This enables systematic exploration of the feasible space and supports the generation of multiple near-optimal solutions. Our results demonstrate that structured global decomposition can both accelerate solution times and enhance interpretability in stochastic power planning optimization.

FA03

Plaza I

Efficient Second-Order Methods

Invited Session

Nonlinear Optimization

Chair: Chang He, Shanghai University of Finance and Economics, Shanghai, N/A

1 - Accelerating Nonconvex Optimization via Online Learning

Ruichen Jiang, University of Texas at Austin, Austin, TX, United States, Aryan Mokhtari, Francisco Patitucci

A fundamental problem in optimization is finding an ϵ-first-order stationary point of a smooth nonconvex function using only gradient information. The best-known gradient complexity for this task, assuming both the gradient and Hessian of the objective function are Lipschitz continuous, is $O(\epsilon^{-1.75})$. In this talk, I will present a new method with a gradient complexity of $O(d^{0.25}\epsilon^{-1.625})$, where d is the problem dimension, yielding an improvement when $d = O(\epsilon^{-0.5})$.

Our key idea is to design a quasi-Newton method that operates through two levels of online learning. Specifically, we first reformulate the task of finding a stationary point for a nonconvex problem as minimizing the regret in an online convex optimization problem, where the loss is determined by the gradient of the objective function. Then, we introduce a novel optimistic quasi-Newton method to solve this online learning problem, with the Hessian approximation update itself framed as an online learning problem in the space of matrices. Beyond improving the known complexity bounds, our result provides the first guarantee that quasi-Newton methods can outperform gradient descent-type methods in nonconvex optimization.

2 - Accelerating Trust-Region Methods: An Attempt to Balance Global and Local Efficiency

Yuntian Jiang, Shanghai University of Finance and Economics, Shanghai, China, People's Republic of, Chuwen Zhang, Bo Jiang, Yinyu Ye

Historically speaking, it is hard to balance the global and local efficiency of second-order optimization algorithms. For instance, the classical Newton's method possesses excellent local convergence but lacks global guarantees, often exhibiting divergence when the starting point is far from the optimal solution~\cite{more1982newton,dennis1996numerical}. In contrast, accelerated second-order methods offer strong global convergence guarantees, yet they tend to converge with slower local rate~\cite{carmon2022optimal,chen2022accelerating,jiang2020unified}. Existing second-order methods struggle to balance global and local performance, leaving open the question of how much we can globally accelerate the second-order methods while maintaining excellent local convergence guarantee. In this paper, we tackle this challenge by proposing for the first time the accelerated trust-region-type methods, and leveraging their unique primal-dual information.Our primary technical contribution is Accelerating with Local Detection, which utilizes the Lagrange multiplier to detect local regions and achieves a global complexity of $\tilde{O}(\epsilon^{-1/3})$, while maintaining quadratic local convergence. We further explore the trade-off when pushing the global convergence to the limit. In particular, we propose the Accelerated Trust-Region Extragradient Method that has a global near-optimal rate of $\tilde{O}

(\epsilon^{-2/7})$ but loses the quadratic local convergence. This reveals a phase transition in accelerated trust-region type methods: the excellent local convergence can be maintained when achieving a moderate global acceleration but becomes invalid when pursuing the extreme global efficiency. Numerical experiments further confirm the results indicated by our convergence analysis.

3 - Sequential Quadratic Optimization for Solving Stochastic Optimization Problems with Stochastic Nonlinear Inequality and Equality Constraints

Yang Zeng, Arizona State University, Tempe, AZ, United States

The objective of this study is to develop an algorithm for addressing general smooth nonlinear optimization problems in which both the objective function and the constraints—including equality and inequality constraints—are defined through expectations. Within this framework, the exact values and derivatives of the objective and constraint functions are assumed to be inaccessible. The proposed algorithm incorporates an adaptive step-size scheme and updates the iterates solely on the basis of estimated objective gradients, constraint values, and constraint derivatives. Analysis of the algorithm's asymptotic convergence properties is conducted. Under mild and standard assumptions, it is established that the method achieves convergence in expectation. The result of the numerical experiment also illustrates that our proposed algorithm outperforms compared with the traditional methods.

4 - A Single-Loop Second-Order Method for Finding Local Minimum in Bilevel Optimization

Jiawen Bi, University of Minnesota, Minneapolis, MN, United States, Jiaxiang Li, Mingyi Hong, Shuzhong Zhang

Bilevel optimization has become a central tool for modern machine learning and engineering design, yet practical methods still struggle when the upper level landscape is nonconvex and common algorithms could be trapped in saddle points. In this paper, we study nonconvex–strongly-convex bilevel optimization and propose a Single-Loop Cubic-Regularized Newton algorithm (SLCRN) that that leverages carefully designed inexpensive follower updates to obtain accurate upper level information. Under standard smoothness and strong convexity of the lower level, SLCRN achieves deterministic convergence to an ε-SOSP in O$(\varepsilon^{-3/2})$ oracle calls, matching the optimal second order rate for nonconvex optimization while requiring only two tunable parameters.
We further present counterexamples showing that several natural modifications could not achieve the same rate of convergence, or even destabilize the iterates, underscoring the necessity of our design choices. Empirically, on controlled test function featuring multiple strict saddles and a unique global minimizer, the proposed method consistently escapes saddles more efficiently, which verifies our theoretical analysis.

FA04

Plaza II

Derivative-Free Optimization: Theory, Algorithms, and Applications

Invited Session

Nonlinear Optimization

Chair: Kwassi Joseph Dzahini, Argonne National Laboratory, Lemont, IL, United States

1 - CLARSTA: A Random Subspace Trust-region Algorithm for Convex-constrained Derivative-free Optimization

Yiwen Chen, University of British Columbia, Kelowna, BC, Canada, Warren Hare, Amy Wiebe

Model-based derivative-free optimization (DFO) methods are one major class of DFO methods widely used in practice. However, these methods are known to struggle in high dimensions. Recent research tackles this issue by searching for decreases in low-dimensional subspaces sampled randomly in each iteration. This talk proposes a random subspace trust-region algorithm for general convex-constrained DFO problems. We provide a new subspace sampling technique that guarantees that the subspace preserves the first-order criticality measure by a certain percentage with a fixed probability. We define a new class of models that can be constructed using only feasible points in the subspace. Based on these new theoretical results, an almost-sure global convergence and a worst-case complexity analysis of our algorithm are presented. Numerical experiments demonstrate the strong performance of our algorithm in high dimensions.

2 - Approximation in Model-Based Derivative-Free Optimization

Pengcheng Xie, Lawrence Berkeley National Laboratory, Berkeley, CA, United States

This talk will discuss traditional and modern approximation techniques for (expensive) derivative-free optimization, in which the approximation serves as a tool for identifying the optimality of the black-box objective, rather than merely approximating the black box itself. Both static (offline) strategies and dynamic (online) strategies will be discussed. Such approximation ideas can also be naturally extended to first-order and higher-order methods.

3 - Stratified Adaptive Sampling for Derivative-free Trust Region Optimization: Complexity and Applications

Giovanni Amici, North Carolina State University, Raleigh, NC, United States, Sara Shashaani, Pranav Jain

There is emerging evidence that trust-region (TR) algorithms are very effective at solving derivative-free nonconvex stochastic optimization problems in which the objective function is a Monte Carlo (MC) estimate. A recent strand of methodologies adaptively adjusts the sample size of the MC estimates by keeping the estimation error below a measure of stationarity induced from the TR radius. In this work we explore stratified adaptive sampling strategies to equip the TR framework with accurate estimates of the objective function, thus optimizing the required number of MC calls to reach a given ϵ-accuracy of the solution. We prove a reduced sample complexity and explore inexpensive implementations in high dimension.

4 - Bayesian Optimization with Multi-Modal Surrogate Models

Ian Taylor, National Renewable Energy Laboratory, Westminster, CO, United States, Juliane Mueller, Julie Bessac

Multi-modal learning, a machine learning task in which multiple modalities and sources of data contribute to an objective, is becoming increasingly important as data collection and simulation capabilities advance. Non-linear global optimization is one objective which stands to benefit from the availability of auxiliary information in the form of multi-modal data. We leverage recently developed multi-modal surrogate models in a novel multi-modal Bayesian optimization framework. Our framework guides the acquisition of new observations both of the expensive objective function and the cheaper multi-modal data. We demonstrate empirically the improved optimization performance of multi-modal Bayesian optimization compared to uni-modal alternatives.

5 - A Derivative-Free Solver for Multi-Region Stochastic Equilibrium Problems under Climate Change Risks

Yunsoo Ha, Cornell University, Ithaca, NY, United States, Linda Nozick

This study presents a derivative-free solver for multi-region stochastic equilibrium problems motivated by insurance market dynamics under climate change risks. The proposed framework models insurers competing across multiple regions and transferring a portion of their catastrophe risk to reinsurers, all under uncertain hurricane losses. Since the underlying equilibrium conditions are simulation-based and lack analytic gradients, a derivative-free approach is developed to efficiently search for equilibria. Numerical experiments using coastal North Carolina data demonstrate that the proposed method reliably computes equilibrium outcomes under climate-induced uncertainty within a reasonable computational budget.

FA05

Plaza III

Recent Advances in Optimization under Uncertainty: Methodologies and Applications

Invited Session

Optimization under Uncertainty

Chair: Man Yiu Tsang, Texas Tech University, Lubbock, TX, United States

1 - Scalable Finite Adaptability via Polyhedral Partition and Learning

Zolykha Rezaei, Texas Tech University, Lubbock, TX, United States, Ningji Wei, Eojin Han

In many practical scenarios—especially those involving sequential decision making—professionals prefer decisions that are adaptable, interpretable, and robust against uncertainties. In this research, we explore the concept of finite adaptability to construct robust optimization policies under uncertainty. More specifically, we address a general setting where uncertainty affects both the objective function and the constraints within the model. Our aim is to construct policies by partitioning the uncertainty set and assigning a contingent decision to each segment.

To solve the resulting robust optimization problem, we propose both exact and approximation methods. The exact methods include Benders' decomposition and dualization techniques for linear models. To enhance scalability and effective partition learning, we also develop approximation methods based on machine learning techniques such as support vector machines (SVMs) and decision trees.

2 - Bayesian Learning for Two-Stage Stochastic Programming

Hamed Rahimian, Clemson University, Clemson, SC, United States, Amin Khademi, Seyed Mahbid Abtahi

Despite significant theoretical and computational advances in two-stage stochastic programming (SP), most existing research assumes that the uncertain parameters (or their probability distributions) are known a priori or belong to a predefined set. In particular, these studies typically assume that the distribution of random variables is non-adaptive to decisions. While such assumptions simplify mathematical formulation and computation, they seldom hold in practice. In this talk, we present a novel Bayesian learning framework for two-stage SP, in which the underlying distribution of random parameters is adaptive and can be progressively learned from the observed effects of decisions. We demonstrate this paradigm through a capacitated facility location problem under demand uncertainty. Furthermore, we propose a solution methodology for the resulting model and provide a theoretical analysis on the sensitivity of the optimal learning policy with respect to sampling costs and prior information.

3 - Model Uncertainty-Aware Residuals-Based Sample Average Approximation

Man Yiu Tsang, Texas Tech University, Lubbock, TX, United States, Karmel Shehadeh, Guzin Bayraksan

We study a contextual stochastic optimization problem, where uncertain parameters and covariates are observed jointly, and the goal is to choose decisions minimizing expected cost given new covariates. Existing regression-based approaches assume that the prediction model fully captures the relationship between uncertainty, covariates, and decisions—yet this relationship is fundamentally unknown, a challenge known as model uncertainty. To address this, we propose a model uncertainty–aware residuals-based SAA framework (MUR-SAA), formulated as a distributionally robust optimization problem. By treating the regression function as a random element and optimizing against the worst-case expected cost over all plausible distributions of the unknown regression function, MUR-SAA produces solutions that are robust to model misspecification. We establish, for the first time, efficient mechanisms for quantifying the stability of general residual-based stochastic programs and our MUR-SAA model. Moreover, we identify conditions under which our MUR-SAA model enjoys desirable asymptotic convergence and finite-sample properties. We derive equivalent tractable reformulations of MUR-SAA and develop a decomposition algorithm to solve them. We illustrate our theoretical results and demonstrate the advantages of our MUR-SAA model using a contextual portfolio optimization problem.

4 - Data-Driven Contextual Optimization with Gaussian Mixtures: Flow-Based Generalization, Robust Models, and Multistage Extensions

YoungChul Yoon, University of Illinois Urbana-Champaign, Champaign, IL, United States, Grani Adiwena Hanasusanto, Yijie Wang

Contextual optimization enhances decision quality by leveraging side information to improve predictions of uncertain parameters. However, existing approaches face significant challenges when dealing with multimodal or mixtures of distributions. The inherent complexity of such structures often precludes an explicit functional relationship between the contextual information and the uncertain parameters, limiting the direct applicability of parametric models. Conversely, while non-parametric models offer greater representational flexibility, they are plagued by the "curse of dimensionality," leading to unsatisfactory performance in high-dimensional problems. To address these challenges, this paper proposes a novel contextual optimization framework based on Gaussian Mixture Models (GMMs). This model naturally bridges the gap between parametric and nonparametric approaches, inheriting the favorable sample complexity of parametric models while retaining the expressiveness of non-parametric schemes. By employing normalizing flows, we further relax the GM assumption and extend our framework to arbitrary distributions. Finally, inspired by the structural properties of GMMs, we design a novel GMM-based solution scheme for multistage stochastic optimization problems with Markovian uncertainty. This method exhibits significantly better sample complexity compared to traditional approaches, offering a powerful methodology for solving long-horizon, high-dimensional multistage problems. We demonstrate the effectiveness of our framework through extensive numerical experiments on a series of operations management problems. The results show that our proposed approach consistently outperforms state-of-the-art methods, underscoring its practical value for complex decision-making problems under uncertainty.

FA06

Director's and Lounge

Machine Learning for Global Optimization

Invited Session

Global Optimization

Chair: Rohit Kannan, Virginia Tech, Blacksburg, VA, United States

Co-Chair: Harsha Nagarajan, Los Alamos National Laboratory, Los Alamos, NM, United States

1 - Self-Supervised Learning to Tune Low-Fidelity Optimization Models

Can Li, Purdue University, West Lafayette, IN, United States, Gonzalo Constante

Recent advances in machine learning have enabled data-driven proxies that approximate optimization models using historical or synthetic data. These models offer rapid, near-optimal solutions, making them attractive for time-sensitive applications. However, they often lack explicit representations of system constraints, resulting in infeasible or unreliable decisions in safety-critical contexts.

Tuned low-fidelity optimization models have emerged as a promising alternative by combining historical information with known system dynamics and constraints. While more interpretable and reliable, their

training typically requires solving expensive bilevel optimization problems to identify optimal tuning parameters.

This work introduces a self-supervised learning framework for tuning low-fidelity optimization models using differentiable optimization layers. The approach integrates three components: a learning model that predicts tunable parameters, a differentiable optimization layer mapping these parameters to approximate decisions, and a high-fidelity model that implicitly evaluates solution quality. Applied to the AC Optimal Power Flow problem, with the DC-OPF model as its low-fidelity counterpart, the proposed method achieves superior data efficiency, training speed, and solution accuracy compared to conventional supervised approaches.

2 - Machine Learning-Guided Undercover for Mixed-Integer Nonlinear Programs

Dhruva Sundararajan, Virginia Tech, Blacksburg, VA, United States, Jude Sanborn, Gavin Esperanza, Rohit Kannan

We design a machine learning (ML)-guided enhancement of the Undercover primal heuristic for mixed-integer nonlinear programs (MINLPs). Undercover identifies a subset of variables to fix and assigns them values, thereby restricting the MINLP to a mixed-integer linear program (MILP) that can be solved using off-the-shelf solvers to potentially obtain a feasible solution. While Undercover performs well on mixed-integer quadratically constrained quadratic programs (MIQCQPs), its effectiveness on general MINLPs is limited by the difficulty of choosing good fixing values. To address this challenge, we propose a graph convolutional network (GCN) that predicts high-quality fixing values for selected variables. The model is pretrained using best-known feasible solutions and further optimized through an end-to-end loss function by leveraging LP sensitivity theory. We integrate our ML-guided Undercover heuristic within the global solver SCIP and evaluate it on benchmark instances from QPLIB and MINLPLIB. Computational results demonstrate that the proposed approach substantially improves the performance of the default Undercover implementation in SCIP.

3 - Strict Constraint Satisfaction in Neural Networks

Faruque Hasan, Texas A&M University, College Station, TX, United States, Rahul Golder, Bimol Nath Roy

Physics-informed neural networks (PINNs) do not always guarantee strict constraint satisfaction. This is problematic in engineering systems where minor violations of governing laws can degrade the reliability and consistency of model predictions. To overcome these challenges, we introduce KKT-Hardnet (https://github.com/SOULS-TAMU/kkt-hardnet), a neural network architecture that enforces linear and nonlinear equality and inequality constraints to machine precision. It leverages a differentiable projection onto the feasible region by solving Karush-Kuhn-Tucker (KKT) conditions of a distance minimization problem. Our major contributions are twofold (1) we provided a unique strategy to include the inequality constraints by considering the complementary conditions through Fischer-Burmeister reformulation to ensure nonnegativity in corresponding Lagrange multipliers, (2) we developed the framework for hard nonlinear constraint satisfaction for both equality and inequality. Compared to multilayer perceptrons (MLPs) and PINNs, KKT-Hardnet achieves strict constraint satisfaction. It also circumvents the need to balance data and physics residuals in PINN training. This enables the integration of domain knowledge into machine learning towards reliable hybrid modeling of complex systems. Moreover, our framework allows us to train neural networks to learn and predict solutions to nonlinear optimization problems, offering a new avenue for integrating machine learning with optimization.

4 - Decision-Focused Learning for Underground Pumped Hydroelectric Storage Day-ahead Scheduling

Honghui Zheng, Johns Hopkins University, Baltimore, MD, United States, Pietro Favaro, Jan Drgona, Yury Dvorkin

Underground pumped hydroelectric storage (UPHES) systems are essential for grid-scale renewable energy integration, yet their optimal day-ahead scheduling presents significant computational challenges due to nonlinear turbine characteristics, discrete operational modes, and complex hydraulic dynamics. Traditional mixed-integer nonlinear programming (MINLP) formulations are computationally prohibitive for real-time operations. This paper introduces a decision-focused learning (DFL) framework that achieves near-optimal UPHES scheduling with dramatically reduced computational requirements. The proposed methodology employs neural networks to predict penalty weights that guide recursive linearization, transforming the intractable MINLP into a sequence of convex quadratic programs via differentiable optimization layers. Validation on twenty representative Belgian electricity market scenarios demonstrates that the DFL approach achieves 1.1% higher ex-post profit than mixed-integer quadratic programming (MIQP) benchmarks while reducing solution time by three orders of magnitude (1.24s versus 1205.79s). Furthermore, the framework exhibits superior robustness under initialization noise, maintaining stable performance across perturbation levels from 10% to 80%, while conventional MIQP methods degrade by 12.3%-22.5%. These results establish DFL as an enabling technology for real-time UPHES scheduling in volatile electricity markets, bridging the computational-accuracy gap that has limited operational deployment of optimization-based control strategies.

FA07

Congress Room

Recent Advances in Sequential Decision Making

Invited Session

Discrete Optimization

Chair: Alfredo Torrico, University of Southern California, Los Angeles, CA, United States

Co-Chair: Sebastian Perez-Salazar, Rice University

1 - Improved Approximations for Free-Order Prophet Matching with Patience Constraints

Calum MacRury, Columbia University, New York, NY, United States

We study a bipartite online stochastic matching problem where the set of feasible edges is known in advance, but the weight of each edge is drawn from a known distribution and revealed only upon querying. The algorithm may adaptively decide which edges to query in an order of its choosing, subject to \emph{patience constraints} that bound the number of queries per vertex. Upon querying an edge, it must immediately and irrevocably decide whether to include it in the matching.

Computing the optimal algorithm is known to be NP-hard, even when one side consists of a single vertex. On the positive side, Pollner, Roghani, Saberi, and Wajc (EC 2022) give a polynomial-time algorithm with an approximation ratio of \$0.426\$ when vertices on one side have unlimited patience, which drops to \$0.395\$ when both sides have bounded patience. Their result is based on rounding a linear program that cannot be rounded within a factor better than \$0.544\$ in the worst case. In this work, we significantly improve these guarantees to \$1 - 1/e\$ and \$0.58\$, respectively, for the one-sided and two-sided patience settings. A key ingredient in our work is a new configuration LP that allows for more effective rounding.

Our approach is based on designing \emph{threshold policies}, where each queried edge is accepted if its realized weight exceeds a chosen threshold. This framework also extends our results to more general objectives beyond maximizing expected weight, including \emph{revenue} in the sequential pricing problem.

2 - Dynamic Interval Scheduling with Random Start and End Times

Rui Gong, Georgia Institute of Technology, Atlanta, GA, United States, Alejandro Toriello

The interval scheduling problem is a cornerstone of operations research and computer science; its deterministic form admits Θ(n)-time solutions. We introduce a new stochastic variant in which the number of tasks, job weights, available time slots, and the start–end time distributions of each task are known in advance, but a task's realized start and end times are revealed only after it is committed. The objective is to sequentially schedule non-overlapping tasks to maximize expected total weight. We study two formulations: (i) conflicts are determined with respect to the underlying distributions, and (ii) a task is dropped whenever any time in its support is already occupied. Both formulations are NP-hard. We develop linear and semidefinite programming relaxations of the resulting dynamic program that encode a stochastic analog of the core interval-scheduling constraint—each slot can be occupied by at most one task. We analyze these relaxations theoretically and computationally, and we design dynamic heuristics.

3 - Adaptive Two-sided Assortment Optimization: Revenue Maximization

Mohammadreza Ahmadnejadsaein, Cornell University, New York, NY, United States, Omar El Housni

We study adaptive two-sided assortment optimization for revenue maximization in choice-based matching platforms. The platform has two sides of agents: an initiating side and a responding side. The decision-maker sequentially selects initiating agents, shows each an assortment of responding agents, and observes their choices. After processing all initiating agents, the responding agents are shown assortments and make selections. A match occurs when two agents mutually select each other, generating pair-dependent revenue. Choices follow Multinomial Logit (MNL) models. This setting generalizes prior work on maximizing the number of matches under submodular demand assumptions, which do not hold in our revenue-maximization context.

Our main contribution is the design of polynomial-time approximation algorithms with constant-factor guarantees. For general pairwise revenues, we develop a randomized algorithm achieving a $(\frac{1}{2}-\epsilon)$-approximation in expectation for any $\epsilon > 0$. The algorithm provides guarantees under various agent arrival settings, including fixed order, simultaneous processing, and adaptive selection. When revenues are equal for any given responding-side agent, the guarantee improves to $(1 - \frac{1}{e} - \epsilon)$. In structural settings where responding agents share a common revenue-based ranking, we design a simpler adaptive deterministic algorithm achieving a $\frac{1}{2}$-approximation. Our approach leverages novel linear programming relaxations, correlation gap arguments, and structural properties of the revenue functions.

4 - Prophets in Parallel: Dynamic Cut Minimization in Series-Parallel Graphs

William Frendreiss, Georgia Institute of Technology, Atlanta, GA, United States, Alejandro Toriello

We introduce a sequential version of the minimum s-t cut problem, defined by a directed graph with source s and sink t, and nonnegative random variables for each arc representing arc weights. Starting with $S = \{s\}$, we observe weight realizations for outgoing arcs from S, and choose to either stop, or to add a node to S, with the objective of minimizing the expected weight of the chosen cut. When the graph is a directed s-t path, the model reduces to the classical prophet inequality problem in minimization form.

We formulate the problem as a dynamic program (DP) and show that computing the expected cost of an optimal policy is #P-hard even on simple series-parallel graphs. We leverage a linear programming (LP) formulation of the DP to construct a separable cost-to-go function approximation (CFA), yielding a lower bound. We design a heuristic policy based on the CFA and provide a guarantee on its worst-case performance that depends on graph topology but is independent of arc weight distributions.

FA08

Committee Room

Applications of Mixed-Integer Linear Programming

Invited Session

Emerging Applications of Optimization

Chair: Norbert Trautmann, University of Bern, Department of Business Administration, Bern, 03012, Switzerland

Co-Chair: Vanessa Sawkmie, University of Wisconsin Madison, Madison, WI, United States

1 - Evaluating Redistricting Reform Proposals with MIP-based Heuristics

Ian Ludden, Rose-Hulman Institute of Technology, Terre Haute, IN, United States, Mason Wiedow

As political redistricting in the United States continues to suffer from partisan gerrymandering, reform proposals proliferate. Several reform ideas fall into the category of two-player redistricting games. Inspired by the famous "I cut, you choose" protocol for fair division of a cake between two agents, these proposals turn drawing a state's districts into a two-player (i.e., two-party) extensive form game. Throughout such a game, each party controls some of the final voting district boundaries. The hope is that Nash equilibria yield fairer district plans than those proffered by either party. To find analytical solutions of Nash equilibria, proposers of redistricting games often relax geographic constraints, switching from a balanced, connected graph partitioning problem to one of rearranging voters arbitrarily into bins as if red and blue marbles.

To what degree will fairness results from this geography-free setting translate to reality? In this talk, we show that several redistricting games have pathological examples (despite strong restrictions) in which the problem's geography gives one player an outsized advantage. We then apply MIP-based heuristics to simulate redistricting games for more realistic instances. Applying standard partisan fairness measures, we compare our resulting district plans to those found by other heuristics. We conclude with policy implications and possible extensions.

2 - Bi-Parameterized Two-Stage Stochastic Min-Max and Min-Min Mixed Integer Programs: Applications for Decision-Dependent Uncertainty

Manish Bansal, Virginia Tech, Blacksburg, VA, United States, Sumin Kang

We introduce two-stage stochastic min-max and min-min integer programs with bi-parameterized recourse (BTSPs), where the first-stage decisions affect both the objective function and the feasible region of the second-stage problem. We present solution approaches for solving these programs. Interestingly, distributionally robust optimization problems with decision-dependent and non-relatively complete ambiguity sets, and stochastic programs with decision-dependent probability distribution can be reformulated as BTSPs. The results of our computational experiments show that solving these reformulations using our proposed approaches is effective and efficient in comparison to state-of-the-art approaches.

3 - Identifying Pareto Efficient Integer Assignments in Biobjective Mixed-Integer Optimization

Philip de Castro, Clemson University, Clemson, SC, United States, Margaret Wiecek

In biobjective mixed-integer optimization, the Pareto set is, in general, a disconnected set which may be parameterized by the value of the integer variables. Identifying the integer values corresponding to the Pareto set is a difficult problem, but has important applications such as radiation therapy planning. In this talk, we present an algorithm which can exactly identify the integer values of the Pareto set, even when the continuous relaxation is nonlinear. We discuss the theory, describe the algorithm, and demonstrate its performance on test problems from the literature.

4 - Asset Deployment for Anti-Submarine Search Operations: Stochastic Optimization Frameworks

Vanessa Sawkmie, University of Wisconsin-Madison, Madison, WI, United States, Jeff Linderoth

Detection of undersea targets rely on successful encounters with active search assets, each with distinct operational capabilities. We present stochastic integer programming frameworks for optimizing search asset deployment to maximize detection likelihood under uncertainty. Our initial two-stage model strategically allocates assets across search patterns, defined by regions and time intervals, over a finite planning horizon, accounting for stochastic target motion and detection outcomes. We extend this framework to a three-stage model to incorporate intelligence received from pre-positioned fixed bottom sensors, where observed sensor detections update posterior target motion distributions, and asset deployments are optimized against these updated distributions. Computational experiments demonstrate solution scalability and the operational value of stochastic optimization for Anti-submarine search operations.

5 - Asset Deployment for Anti-Submarine Search Operations: Stochastic Optimization Frameworks

FA09

Cabinet Room

Linear and Conic Optimization I

Invited Session

Linear and Conic Optimization

Chair: David Papp, North Carolina State University, Raleigh, NC, United States

Co-Chair: Anita Varga, North Carolina State University, Raleigh, NC, United States

1 - Facial Reduction for Semidefinite Relaxations of Combinatorial Optimization Problems

Hao Hu, Clemson University, Clemson, SC, United States, Mingming Xu, Boshi Yang

In this talk, we present new findings on facial reduction for semidefinite relaxations of combinatorial optimization problems. In semidefinite programming (SDP), Slater's condition is crucial for both theoretical convergence guarantees and the practical performance of optimization algorithms. When Slater's condition fails, facial reduction can restore it through a finite sequence of reformulations. However, these reformulations often involve solving auxiliary optimization problems that can be as challenging as the original. Recent research has therefore focused on developing more efficient strategies for performing facial reduction.

In our work, we specifically consider SDP problems that arise as relaxations of combinatorial optimization problems. This perspective enables us to exploit the underlying combinatorial structure, allowing the development of novel and highly efficient facial reduction techniques. We establish theoretical results demonstrating the effectiveness of our approach. Numerical experiments further show that applying our specialized facial reduction method significantly improves both the speed and accuracy of solving SDP problems.

2 - On Computing Search Directions for Primal-Dual Interior Point Method for Non-Symmetric Conic Optimization

Utkarsh Detha, MOSEK ApS, Copenhagen, Denmark

Primal-Dual interior point methods for symmetric conic optimization primarily rely on the computation of three search directions, namely the affine scaling direction (predictor), the centering (or corrector) direction and lastly a higher-order correction (such as Mehrotra's correction). Computation of these directions is closely tied to the choice of a scaling matrix. The Nesterov-Todd scaling is widely accepted as the best choice for the symmetric conic case. However, the choice isn't as clear for the case of non-symmetric conic optimization. One approach is based on using a multi-secant generalization of a quasi-Newton type update to the Hessian to ensure it satisfies two secant equations. The effectiveness of this approach is demonstrated in the case of 3-D non-symmetric cones in the MOSEK solver. In this talk, we discuss our experience with extending this choice to the case of higher-dimensional non-symmetric cones. We will discuss the exploitation of sparsity in solving the normal equations and/or the augmented system and also how to handle proximity to the central path in a reasonable manner. Lastly, we will also discuss the computation and impact of the higher-order correction.

3 - Dual Certificates of Primal Cone Membership–Part I

David Papp, North Carolina State University, Raleigh, NC, United States, Anita Varga

Sums-of-squares (SOS) and other popular convex relaxations of polynomial minimization problems are often solved using off-the-shelf numerical methods which do not immediately yield rigorously provable lower bounds on the global minimum. For instance, the moment-SOS hierarchy yields rigorous upper (instead of lower) bounds on the sought lower bound, while the numerical solution of the SOS SDP (the dual of the moment-SOS SDP) yields a rigorous nonnegativity certificate of a slightly different polynomial than the one at hand, due to the inexactly satisfied equality constraints in the SDP.

More generally, for conic programs in standard form, numerically computed primal "feasible" solutions usually aren't feasible, and do not provide rigorous upper bounds on the optimal value.

The first talk of this two-part series presents a theoretical framework (termed "dual certificates") for obtaining rigorous upper bounds from suitable dual feasible solutions, complementing the standard duality theory in which the same dual solutions yield lower bounds. When applied to tractable subcones of nonnegative polynomials commonly encountered in polynomial optimization (SOS, SONC, SAGE and SDSOS polynomials, etc.), dual certificates enable the computation of rigorous nonnegativity certificates using inexact, numerical methods.

4 - Dual Certificates of Primal Cone Membership-Part II

Anita Varga, North Carolina State University, Raleigh, NC, United States, David Papp

Dual membership certificates are easily verifiable, rigorous certificates of membership in a cone represented by vectors in the dual cone. This construction complements the classical interpretation of the dual cone, in which such vectors serve as certificates of non-membership. An important feature of the approach is that exact certificates can be computed using standard numerical methods.

The second talk of this two-part series is concerned with the computational aspects of dual certificates in optimization problems. First, we show that dual certificates can be obtained entirely numerically, for example by using primal-dual interior-point algorithms (IPAs). We then show that, using dual certificates, we can recover primal feasible solutions from suitable dual feasible ones. This enables the computation of both rigorously certifiable lower and upper bounds from the same dual strictly feasible solution. The gap between these bounds tends to zero as the dual solution approaches optimality.

Dual certificates, in turn, can be used to enhance existing IPAs: one may, e.g., "retrofit" dual methods to return exact feasible solutions, or certify the progress of IPAs. We present numerical examples for the theoretical results.

FA10

Caucus Room

Primal-dual Perspectives on Continuous Optimization Algorithms

Invited Session

Optimization in Data Science

Chair: Jiaming Liang, University of Rochester, ATLANTA, United States

Co-Chair: Renbo Zhao, University of Iowa, Iowa City, IA, United States

1 - Primal-Dual Proximal Bundle and Conditional Gradient Methods for Convex Problems

Jiaming Liang, University of Rochester, Rochester, NY, United States

This paper studies the primal-dual convergence and iteration-complexity of proximal bundle methods for solving nonsmooth problems with convex structures. More specifically, we develop a family of primal-dual proximal bundle methods for solving convex nonsmooth composite optimization problems and establish the iteration-complexity in terms of a primal-dual gap. We also propose a class of proximal bundle methods for solving convex-concave nonsmooth composite saddle-point problems and establish the iteration-complexity to find an approximate saddle-point. This paper places special emphasis on the primal-dual perspective of the proximal bundle method. In particular, we discover an interesting duality between the conditional gradient method and the cutting-plane scheme used within the proximal bundle method. Leveraging this duality, we further develop novel variants of both the conditional gradient method and the cutting-plane scheme. Additionally, we report numerical experiments to demonstrate the effectiveness and efficiency of the proposed proximal bundle methods in comparison with the subgradient method for solving a regularized matrix game.

2 - Improved Analysis of Restart Accelerated Gradient and Augmented Lagrangian Methods via Inexact Proximal Point Frameworks

Matthew Burns, University of Rochester, Rochester, NY, United States, Jiaming Liang

This work proposes and applies an inexact proximal point (IPP) framework for convex optimization with relative error criteria. Convergence rates for the proposed Lower Model Oracle (LMO) and Accelerated LMO (ALMO) frameworks are provided, with the accelerated variant attaining optimal rates for first-order convex and strongly convex optimization. We demonstrate that the (A)LMO framework can be applied in both primal and dual contexts, with an emphasis on `two-loop' algorithms. From the primal perspective, we apply the ALMO framework to prove convergence of a doubly accelerated restarted Nesterov acceleration method for composite (strongly) convex optimization. From the dual perspective, we analyze the inexact augmented Lagrangian method (iALM) for linearly constrained composite convex optimization. The LMO framework is used to provide nonergodic guarantees matching existing rates for iALM. The ALMO framework is then applied to propose an accelerated iALM scheme that attains $\mathcal{O}(\varepsilon^{-1})$ complexity for linearly constrained convex composite optimization. Numerical experiments are provided to support the theoretical analysis and demonstrate the performance of the primal and dual methods.

3 - A Large-Scale Benchmarking of Deterministic and Stochastic Derivative-Free Optimization Algorithms

Donghyun Oh, Georgia Institute of Technology, Atlanta, GA, United States, Nikolaos Sahinidis

We present a systematic benchmarking study of state-of-the-art derivative-free optimization algorithms on a broad suite of test problems. We compare a number of global and local solvers, including customary stochastic methods and modern deterministic methods like the recently proposed branch-and-model algorithm. We evaluate optimization performance using several metrics such as scalability and sensitivity to initialization. The comparison demonstrates the strengths and limitations of each algorithm across diverse problem classes.

4 - Convergence Rate and Iteration Complexity of Trust-Region Method for Set-Valued Maps

Suprova Ghosh, Indian Institute of Technology (BHU) , Varanasi, India

In this paper, we present convergence-rate and iteration-complexity analysis of a recently proposed trust-region method for set optimization problems (TRM-SOP), where the objective is a set-valued map generated by finitely many vector-valued functions \citep{Ghosh et al.2025b}. We establish that the method achieves a sublinear convergence rate under mild assumptions, and that the number of iterations required to reach a K-critical point with accuracy ϵ is bounded by \[O\!\left(\sqrt{m}\,\Sigma\,\epsilon^{-2} \right), \] where m denotes the number of components in each vector-valued function and Σ is the optimality gap. Notably, this bound is dimension-free with respect to the decision space and closes existing theoretical gaps in the performance guarantees of trust-region algorithms for set optimization. These results provide the first iteration-complexity theory and convergence-rate analysis for trust-region methods in the set-valued context.

FA11

Charter Room

Advances in Districting

Invited Session

Network Optimization

Chair: Xiaocong Zhen, Oklahoma State University, 802 West Highpoint Drive, Stillwater, OK, 74075, United States

1 - Local Markov Chains for Redistricting with Network Partitions

Daryl DeFord, Vassar College, Poughkeepsie, NY, United States

Electoral districts can be modeled as partitions of a network representing the underlying geographic units, which allows for analysis using discrete optimization and sampling methods. Spanning tree approaches have become popular for generating large samples of comparison plans in this setting, which can introduce a strong preference for partitions with many spanning trees in the target distribution. In this talk I will discuss recent applications and extensions of these techniques, including for court cases, support of line-drawing commissions, and evaluating nonpartisan justifications for proposed plans. Along the way I will present related open problems and proposals based on the cycle basis walk and classic network partitioning methods including spectral clustering and modularity.

2 - Sampling Tree-weighted Partitions without Sampling Trees

Jamie Tucker-Foltz, Yale SOM, New Haven, CT, United States, Sarah Cannon, Wesley Pegden

We give a new algorithm for sampling tree-weighted partitions of a large class of planar graphs. Formally, the tree-weighted distribution on k-partitions of a graph weights k-partitions proportional to the product of the number of spanning trees of each partition class. Recent work on problems in computational redistricting analysis has driven special interest in the conditional distribution where all partition classes have the same size (balanced partitions). One class of Markov chains in wide use aims to sample from balanced tree-weighted k-partitions using a sampler for balanced tree-weighted 2-partitions. Previous implementations of this 2-partition sampler would draw a random spanning tree and check whether it contains an edge whose removal produces a balanced 2-component forest; if it does, this 2-partition is accepted, otherwise the algorithm rejects and repeats. In practice, this is a significant computational bottleneck.
We show that in fact it is possible to sample from the balanced tree-weighted 2-partition distribution directly, without first sampling a spanning tree; the acceptance and rejection rates are the same as in previous samplers. We prove that on a wide class of planar graphs encompassing network structures typically arising from the geographic data used in computational redistricting, our algorithm takes expected linear time O(n). Notably, this is asymptotically faster than the best known method to generate random trees, which is $O(n \log^2 n)$ for approximate sampling and $O(n^{1 + \log \log \log n / \log n})$ for exact sampling

3 - Some Advances in Sampling Graph Partitions

Gregory Herschlag, Duke University, Durham, NC, United States, jonathan mattingly

When auditing a redistricting plan, one often wants to compare what *would* have happened under non-partisan plans. In accessing this counterfactual, one first places a probability distribution on the space of redistricting plans and then uses Monte Carlo sampling to integrate the space and determine features of interest (e.g., how many party members does any given map elect?)

Specifically, a class of methods called Recombination (or Recom) has become a popular Monte Carlo method to use in practice, however, it is only (empirically) capable of efficiently sampling from certain classes of measures. In this talk we will use spectral methods to understand these barriers to mixing on an extended class of measures. I will conclude with some discussion on a survey of alternative methods that extend the class of measures that we can efficiently sample and avoid the barriers found in Recom.

4 - Fair Districts v. Geography: A False Dichotomy?

Xiaocong Zhen, Oklahoma State University, Stillwater, OK, United States, Austin Buchanan

In recent months, politicians across the spectrum have complained about maps drawn by their political opponents that slice-and-dice communities to gerrymander seats for their party. One proposed reform is to explicitly require mapmakers to seek proportionality: the idea that if a party earns X% of the votes, then they should win roughly X% of the seats. However, some have expressed concern that achieving proportionality will require non-compact districts that shred counties. In this talk, we propose optimization methods to test whether this is true. Specifically, we ask whether it is possible to draw maps that achieve partisan proportionality and minimize the extent to which counties are divided across districts. For many states, we find that the answer is yes, i.e., there exist maps that achieve ideal performance on both objectives, simultaneously. Further, the generated maps are more compact than those currently in use. We accomplish this with a combination of heuristic and exact integer programming methods (e.g., Benders decomposition). Our results are practically relevant for proposed anti-gerrymandering legislation and inform independent redistricting commissions and special masters who must delicately balance competing criteria.

March 20, 2026, 10:00 AM - 11:00 AM

Welcome & Plenary 1

Grand Ballroom

Sampling on Graphs

Plenary

Plenary

Chair: Mathieu Dahan, Georgia Institute of Technology, Atlanta, GA, United States

1 - Sampling on Graphs

Rekha Thomas, University of Washington, Seattle, WA, United States

Is it possible to compute the average value of a function on a graph by knowing the function values at a small set of vertices? If so, how small can this subset be and how does one find it? What sorts of functions can we average? The story draws on a range of mathematics from analysis to geometry to discrete math.

In this talk I will introduce graphical designs which allow us to answer these questions. When the graphs have special structure, these designs become highly structured, connecting them to well-known and classical objects in mathematics and computer science. For instance, graphical designs in graphs of hypercubes are orthogonal arrays which are related to linear codes.

Joint work with Catherine Babecki, Zawad Chowdhury and Stefan Steinerberger

March 20, 2026, 11:00 AM - 11:45 AM

Plenary 2

Grand Ballroom

Distributionally Robust – Risk Averse Multistage Stochastic Optimization

Plenary

Plenary

Chair: Vidya Muthukumar, ISyE Georgia Tech, Atlanta, GA, United States

1 - Distributionally Robust – Risk Averse Multistage Stochastic Optimization

Alexander Shapiro, ISyE Georgia Tech, Atlanta, GA, United States

In this talk we discuss several approaches to the formulation of distributionally robust counterparts of Markov Decision Processes, where the transition kernels are not specified exactly but rather are assumed to be elements of the corresponding ambiguity sets. The intent is to clarify some connections between the game and static formulations of distributionally robust MDPs and explain the role of rectangularity associated with ambiguity sets in determining these connections.

March 20, 2026, 12:45 PM - 2:15 PM

FB01

Grand Ballroom

Theoretical Foundations of Reinforcement Learning: Algorithms and Convergence Guarantees

Invited Session

Optimization in Data Science

Chair: Zaiwei Chen, Purdue University, West Lafayette, IN, United States

1 - Sobolev-Prox Algorithm for Continuous Time Reinforcement Learning

Wenlong Mou, University of Toronto, Toronto, ON, Canada

Model-free function approximation is the workhorse of modern reinforcement learning (RL). By solving projected Bellman fixed-point equations, these methods can learn the value functions and optimal policies efficiently without requiring a model of the environment. While function approximation is known to be hard for general discrete-time RL problems, recent work discovered that RL in continuous-time control problems exhibits remarkable properties that enable the design of provably efficient algorithms.

In this talk, we introduce recent advances in optimization algorithms for continuous-time RL with function approximation. We present a new class of Sobolev-prox algorithms that leverage the function space structures of the the Bellman operator from a controlled diffusion process. We prove non-asymptotic convergence guarantees for general function approximations. We will also discuss applications to fine-tuning of diffusion-based generative models.

2 - Natural Policy Gradient: All Roads Lead Back to Bellman

Zaiwei Chen, Purdue University, West Lafayette, IN, United States

Policy gradient (PG) and natural policy gradient (NPG) methods have become central tools in modern reinforcement learning. In contrast to classical dynamic programming methods such as value iteration and policy iteration, these gradient-based approaches are often viewed as departing from Bellman-operator-based algorithm design and analysis. As a consequence, existing convergence analyses of NPG tend to be considerably more involved than those for traditional policy iteration and typically yield weaker guarantees.

In this talk, we present a new viewpoint that reconnects natural policy gradient with Bellman operators. We show that NPG can be formulated as a doubly smooth policy iteration method. This reformulation allows us to directly leverage the contraction and monotonicity properties of the Bellman operator, leading to the strongest known convergence rate guarantees for NPG through a significantly simpler proof framework. We believe this perspective has the potential to become a standard, textbook-level understanding of natural policy gradient.

3 - Sample Complexity of Distributionally Robust Average Reward MDP

Yuepeng Yang, Yale University, New Haven, CT, United States

Robustness has been a key challenge in reinforcement learning, as the performance gap between the optimal policy and the estimated policy can be quite large in the presence of model uncertainty. In this work we study the average reward Markov Decision Process under the distributionally robust setting, where the transition probability is subject to some uncertainty set.

Our goal is to understand the sample complexity in this distributionally robust setting. To handle the average reward setting, we employ a reduction to its discounted counterpart. Leveraging prior understanding of the discounted setting, we are able to establish a sharp sample complexity result for robust average reward MDP with uncertainty set induced by total variation distance. We also complement this with a matching information-theoretical lower bound that highlights the distinction between the robust average reward setting and both its non-robust and discounted robust counterparts.

4 - A Broader View of Thompson Sampling

Yanlin Qu, Columbia Business School, New York, NY, United States

Thompson Sampling is one of the most widely used and studied bandit algorithms, known for its simple structure, low regret performance, and solid theoretical guarantees. Yet, in stark contrast to most other families of bandit algorithms, the exact mechanism through which posterior sampling (as introduced by Thompson) is able to "properly" balance exploration and exploitation, remains a mystery. In this talk we show that the core insight to address this question stems from recasting Thompson Sampling as an online optimization algorithm. To distill this, a key conceptual tool is introduced, which we refer to as "faithful" stationarization of the regret formulation. Essentially, the finite horizon dynamic optimization problem is converted into a stationary counterpart which "closely resembles" the original objective (in contrast, the classical infinite horizon discounted formulation, that leads to the Gittins index, alters the problem and objective in too significant a manner). The newly crafted time invariant objective can be studied using Bellman's principle which leads to a time invariant optimal policy. When viewed through this lens, Thompson Sampling admits a simple online optimization form that mimics the structure of the Bellman-optimal policy, and where greediness is regularized by a measure of residual uncertainty based on point-biserial correlation. This answers the question of how Thompson Sampling balances exploration-exploitation, and moreover, provides a principled framework to study and further improve Thompson's original idea.

FB02

Georgian Room

Advances in Linear Algebra for Optimization

Invited Session

Computational Optimization

Chair: Robert Parker, Los Alamos National Laboratory, Mason, OH, United States

1 - Exploiting Block-triangular Submatrices in KKT Systems

Robert Parker, Los Alamos National Laboratory, Los Alamos, NM, United States

The computational bottleneck in many nonlinear optimization algorithms is the factorization of the symmetric, indefinite KKT matrix. For some problems, this factorization can be accelerated using a Schur complement decomposition to exploit submatrices that are known to be nonsingular. In this work, we develop a Schur complement decomposition that exploits block-triangular submatrices, which can be factorized and solved with their own structure-exploiting algorithms. We demonstrate that this method accelerates solution of KKT systems from optimization problems with neural networks as constraints.

2 - Advancements in Re::Solve, a Linear Solver for Nonlinear Problems

Shaked Regev, Oak Ridge National Laboratory, Oak Ridge, TN, United States, Nicholson Koukpaizan, Slaven Peles

Re::Solve is a library of linear solvers designed for solving series of similar systems such as those arising in optimization or dynamic simulations. It provides a unified interface for built-in and vendor-specific solvers that run on AMD and NVIDIA GPUs, as well as conventional CPUs. Through this interface, the developer can reuse elements of the problem that repeat themselves such as the sparsity structure of matrices, pivoting and ordering sequences, and data allocation. Our recent release includes:
1. An extension of our transpose class to work with asymmetric matrices (including rectangular) without duplicating memory allocation. This is crucial for certain specialized solvers.
2. Added diagonal scaling kernels for matrices on the left or right and for vectors. This allows efficient matrix-matrix multiplication when one of the matrices is diagonal and more general norm measurements, such as weighting different parts of the residual differently.
3. Improved efficiency of the multi-vector class, which improves the speed and memory footprint of randomized methods such as RandFGMRES.
4. A smooth interface for the user who provides the matrix data in CSR format to use KLU for the first solve, even though it is built for CSC format, and then moves the rest of the calculations to the GPU without explicit CSC to CSR conversions. This allows the user to use tools built for different formats, on different machines, with only one data transfer to the GPU and no explicit transpose operations.

We also detail our plans for future releases.

3 - MadIPM.jl – GPU Solver for (Batch) Linear Programming

Michael Klamkin, Georgia Institute of Technology, Atlanta, GA, United States, Alexis Montoison

MadIPM.jl is a GPU-accelerated interior-point solver for large-scale linear programming.
It is built on top of MadNLP and implements the Mehrotra predictor-corrector method. Currently, MadIPM.jl efficiently solves individual LP problems on the GPU using the NVIDIA cuDSS solver for linear systems and high-performance GPU kernels developed in MadNLPGPU with KernelAbstractions.jl.
These kernels handle critical operations such as assembling KKT systems, performing line search, and other core computations in interior-point methods.
We are actively working on extending MadIPM.jl to support batch LPs, enabling multiple linear programs to be solved in parallel on the GPU.

4 - CVXPYlayers v1.0: GPU-accelerated Differentiable Optimization

Parth Nobel, Stanford University, Palo Alto, CA, United States

We present the recent development of GPU-accelerated optimization solvers and optimization differentiators. Optimization layers in neural networks—nonlinear layers where the parameters of a convex optimization problem are learned by differentiating a function of the optimal solution—were developed at the end of the 2010s. At that time, their real-world applications were limited by the performance of CPU-bound optimization solvers and differentiators. GPU-acceleration of convex optimization problems had been solved only for special cases like dense QPs and cVar-constrained optimization. In the last two years, general conic interior point methods began to achieve speedups on GPUs, thanks to advancements in sparse matrix factorization. Coupled with new approaches to argmin differentiating, CVXPYlayers---a modeling language for differentiable optimization---developed support for GPU-acceleration, unlocking new performance opportunities for optimization layers in neural networks. We present recent advancements in the space and discuss applications unlocked by the newly achieved high-performance solvers.

FB03

Plaza I

Advanced Methods for Large-scale Nonlinear and Stochastic Optimization I

Invited Session

Nonlinear Optimization

Chair: Qi Wang, University of Michigan, Ann Arbor, MI, United States

1 - Two-Step Negative Curvature Methods for Nonconvex Optimization with Second-Order Convergence Results

Wanping Dong, University of Michigan, Ann Arbor, MI, United States, Albert Berahas, Raghu Bollapragada

We consider a negative curvature method for continuous nonlinear nonconstrained optimization problems under a stochastic setting where the function values, gradients, and Hessian (products) are available only through inexact probabilistic oracles. Our goal is to develop algorithms that have high probabilistic second-order convergence and affordable complexity so that they can be used for large-scale problems. We introduce general conditions on the probabilistic oracles and propose a method that uses both negative curvature and gradient directions. We derive a high probability tail bound on the iteration complexity of the algorithm and show improvements compared to our previous negative curvature method. A practical variant is implemented to illustrate the power of the proposed algorithm.

2 - Generalization of Silver Stepsize Schedule to Stochastic Optimization

Baoyu Zhou, Arizona State University, Tempe, AZ, United States, Luwei Bai, Yang Zeng

A two-step stepsize schedule is introduced for stochastic gradient methods applied to smooth, strongly convex optimization problems. We consider the setting where only stochastic gradient approximations, which are unbiased, of bounded variance, and supported on a finite set, are accessible. When the variance bound is relatively smaller than a ratio of the initial optimality gap, our proposed stepsize schedule achieves faster convergence performance compared to the well-regarded constant stepsize 2/(M+m), where m and M are parameters describing strong convexity and Lipschitz gradient continuity, respectively. Our stepsize schedule can be viewed as a generalization of the well-known two-step silver stepsize schedule from deterministic setting to stochastic optimization.

3 - Smoothing-Enabled Randomized Stochastic Gradient Schemes for Solving Nonconvex Nonsmooth Potential Games

Zhuoyu Xiao, University of Michigan, Ann Arbor, MI, United States

The state of the art in solving nonconvex nonsmooth games under uncertainty remains in its infancy. Many existing studies primarily rely on suitable growth conditions or local convexity-like property. However, these conditions are restrictive and sometimes difficult to justify, motivating the development of alternative algorithms that are not reliant on them. In this work, we focus on the computation of equilibria of stochastic noncooperative games characterized by a potential function. We first consider the nonconvex smooth setting and propose the randomized stochastic gradient (RSG) scheme with $\mathcal{O}(\epsilon^{-4})$

complexity guarantee. Based on this result, we further propose smoothing-enabled RSG methods to address stochastic potential games afflicted by nonconvexity and nonsmoothness. In addition, the biased RSG variant is discussed and we demonstrate its efficiency through a class of stochastic hierarchical games, where the lower-level exact solution information is unavailable. Promising numerics are provided to support our theoretical analysis.

4 - A Gradient Sampling Algorithm for Noisy Nonsmooth Optimization

Lara Zebiane, Lehigh University, Bethlehem, PA, United States, Frank E. Curtis, Albert Berahas

We present an algorithm for solving nonconvex nonsmooth optimization problems when function and generalized gradient evaluations are subject to errors. In particular, we consider settings when these evaluations are subject to so-called deterministic noise, such as from bounded computational errors due to numerical simulations. Our proposed algorithm is of the gradient sampling variety, a powerful methodology for nonconvex nonsmooth optimization. We present theoretical convergence guarantees for our method as well as the results of numerical experiments when solving challenging test problems.

FB04

Plaza II

Recent Advances in Two-Stage Stochastic Programming

Invited Session

Optimization under Uncertainty

Chair: Akshita Gupta, Purdue University, West Lafayette, IN, United States

Co-Chair: Susan Hunter, Purdue University, West Lafayette, IN, United States

1 - Scalarization Based Approach to Two-Stage Robust Multiobjective Optimization

Juho Roponen, University of Jyvaskyla, Jyvaskyla, Finland, Babooshka Shavazipour, Margaret Wiecek

Multistage robust multiobjective optimization addresses multistage decision-making problems under multiple conflicting objectives and uncertainty that results from the volatility of the environment. Although robust multiobjective optimization has been extensively studied, difficulties and complexities of solving multistage robust multiobjective optimization problems remain underexplored.
We propose a two-stage robust multiobjective optimization problem (TSRMOP) with uncertainty in general locations. Relying on policy optimization and scalarization of the second-stage problem, we develop a max-min property and reformulate the TSRMOP into a single-stage, single-objective problem whose optimal solutions are two-stage robust ε-properly efficient solutions to the original TSRMOP. By employing an achievement scalarizing function, the proposed approach is readily applicable to interactive decision-support. We apply this solution approach to a real-life forest management problem that is solved interactively. By iteratively updating preferences and solving the scalarized problem, we find a two-stage robust ε-properly efficient solution that is the most preferred by the decision maker. A comparison of the computed two-stage solutions against the solutions computed with a single-stage robust multiobjective model on the same data reveals the advantage of two-stage modelling.

2 - Solving Two-Stage Robust Biobjective Linear Programs with Parametric Benders' Decomposition

Margaret Wiecek, Clemson University, Clemson, SC, United States, Herve Kerivin, Rakhi Goswami

Multistage multiobjective optimization addresses decision problems under conflict and uncertainty allowing for decision making in stages to suit changing environments or currently available information. A two-stage robust biobjective linear program (TSRBOLP) models two-stage decision processes with strategic decisions being made first, and operational decisions being made at later time after more information about the decision problem becomes known. The goal is to compute the first-stage feasible solutions that are Pareto-efficient with respect to the first and second-stage objectives and account for the worst-case uncertainty scenario.

The TSRBOLP is studied under discrete or continuous uncertainty in the right-hand-side of the constraints and with one or two second-stage objectives. The weighted-sum scalarization transforms the TSRBOLP into a single-objective optimization problem (SOP) whose optimal solutions are Pareto-efficient to the TSRBOLP. The SOP assumes the form of a parametric linear program with bilinear terms in the constraints. A parametric Benders' decomposition algorithm is designed to compute the exact or approximate parametric optimal solutions to the SOP. The algorithm relies on an adaptation of Benders' decomposition into a parametric optimization environment. The implementation of the algorithm makes use of MATLAB Multiparametric Toolbox and the Biconvex Approximate Simplex Method (Pangia, 2024). Numerical examples are included.

3 - Stochastic Dual Dynamic Programming for Decision Making Under Conflict and Uncertainty

Benjamin Hamlin, Clemson University, Clemson, SC, United States, Cheng Guo, Margaret Wiecek

We introduce a Stochastic Dual Dynamic Programming method for multiobjective or single-objective-parametric multistage stochastic programming problems. A parametric approach improves on existing methods by eliminating the need for the Simplex method and thereby allowing for more than two objectives. Commercial solvers are used to efficiently perform forward and backward passes. We demonstrate the algorithm on a triobjective Brazilian hydrothermal problem and explore decreasing solving time through cut management and neural networks.

4 - Properties of Two-Stage Stochastic Multi-Objective Linear Programs

Akshita Gupta, Purdue University, West Lafayette, IN, United States

We study a two-stage stochastic multi-objective linear program (TSSMOLP) under risk neutral formulation, which is a natural generalization of a two-stage stochastic linear program (TSSLP). The proposed TSSMOLP formulation allows modelers to specify multiple objectives in each stage. The second-stage recourse decision is modeled by a random multi-objective linear program (MOLP) whose solution maps to a random second-stage nondominated set. The TSSMOLP then comprises the objective, which is the Minkowski sum of a linear term plus the expected value of second-stage nondominated set, and the constraints, which are linear. We prove properties of TSSMOLPs, which are stochastic set optimization problems, and the multifunctions that arise therein. To aid theoretical analysis, we propose nondominance-equivalent reformulations to the original problem and show that (i) the global Pareto set of a TSSMOLP is cone-convex, (ii) solving a parameterized TSSLP produces a weakly efficient point for the TSSMOLP, and (iii) a TSSMOLP in which the uncertainty is governed by an atomic probability measure with finite support can be reformulated as an MOLP. Leveraging the above properties we demonstrate that solving a TSSMOLP under a sample average approximation framework is equivalent to solving a large scale MOLP, and validate our results through a numerical example.

FB05

Plaza III

Interplay of Optimization and Machine Learning

Invited Session

Optimization under Uncertainty

Chair: Hongcheng Liu, University of Florida; HCL Optimal, LLC., Gainesville, FL, United States

1 - Estimation of the Conditional Mean of a Distribution via Piecewise-Linear Convex Optimization

Anton Malandii, Stony Brook University, Stony Brook, NY, United States

Least-squares regression is typically formulated as a quadratic program. This talk presents a novel approach for reducing it to a piecewise linear convex minimization problem within the Risk Quadrangle (RQ) Framework. Evidently, this problem can be reduced to linear programming. Crucially, this is not a heuristic step: the linearized formulation is statistically justified and shown to be equivalent to quantile regression, estimating the specific quantile corresponding to the mean. As a result, it inherits many benefits of quantile-based approaches, such as effective handling of heteroskedastic data and robustness to outliers.

We conduct extensive numerical experiments to justify our theoretical claims and highlight the computational benefits of our method. Our results highlight the versatility of the RQ Framework in unifying and extending classical regression methods, opening the door to broader applications and deeper theoretical understanding.

2 - Fast Presolving Framework For Sparse Ridge Regression: Screening-Based Cutting Plane Generation and Selection

Haozhe Tan, National University of Singapore, Singapore, Singapore, Guanyi Wang

Screening is a commonly utilized presolving technique for mixed binary nonlinear programming (MBNLP) optimization. It aims to identify whether one or multiple binaries take a specific value in an optimal solution of the original MINLP in the presolving step, before exact implementation of Branch-and-Bound (BnB) algorithms. This paper studies sparse ridge regression (SRR) problem as a typical case of MBNLP, and proposes a novel prosolving framework based on screening methods for SRR. Compared with existing (safe) variable screening and optimality-based cut generation algorithms for domain/range reduction in the presolving step, we show that the proposed framework offers superior screening capabilities with a tiny increase in computational cost, leading to a significant reduction in total running time. Extensive numerical experiments further validate these theoretical findings and demonstrate superior computational performance over state-of-the-art methods. Specifically, the proposed framework shows particular advantages in challenging optimization scenarios, including (i) difficult real-world instances, and (ii) synthetic simulations with high dimensionality, low ridge regularization parameters, and challenging modeling assumptions.

3 - Recursive Inequality Bounding Techniques for Optimization Algorithms

Renyuan Li, NUS, Singapore, Singapore, Yutong Dai, Guanyi Wang, Frank E. Curtis, Daniel Robinson

This work presents a formal framework for analyzing iterative optimization algorithms. We develop a general method for deriving bounds on sequences defined by recursive inequalities, based on constructing a majorizing system of equalities. This approach provides a foundation for approximating the divergence rates of Lipschitz functions. We illustrate its effectiveness by applying it to establish convergence bounds for various momentum-based methods, including SPSTORM algorithm and Nesterov's acceleration.

4 - Relatively Robust Multiproduct Pricing

Ru Zhang, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Thomas Weber

This study presents a relatively robust pricing framework for multiproduct monopolies under distribution-free demand uncertainty characterized by Frobenius-norm constraints. Compared to extant models based on interval uncertainty, our approach employs ambiguity sets that align naturally with standard multivariate estimation techniques. The optimal robust price vector, obtained as the solution to a quasi-concave programming problem, maximizes the worst-case ratio of achieved profit to the ex-post optimal profit under full information. A key structural result is that the performance index is strictly quasi-concave in price, allowing the unique optimum to be characterized by a first-order condition. An embedded fixed-point problem can be solved efficiently, and closed-form expressions are derived for symmetric linear, semi-log, and log-log demand systems with associated parameter uncertainty. In general nonlinear settings, the model remains applicable through local linearization and regression-based confidence regions. To evaluate the approach, we conduct both extensive simulation experiments and empirical analyses using large-scale datasets from a major online retailer. The results demonstrate the superior performance of the relatively robust pricing method compared to alternative robustness criteria such as maximin payoff, minimax regret, and Laplacian uncertainty.

FB06

Director's and Lounge

Advances in Solution Methods for MINLPs

Invited Session

Global Optimization

Chair: Danial Davarnia, Purdue University, Ames, IA, United States

1 - Sparse Cuts for the Positive Semidefinite Cone

Jeff Linderoth, University of Wisconsin-Madison, Madison, WI, United States, Oktay Gunluk, Paul Junger, Andrea Lodi, James R. Luedtke

We consider optimization problems containing nonconvex quadratic functions for which semidefinite programming (SDP) relaxations often yield strong bounds. We investigate linear inequalities that outer approximate the positive semidefinite cone and are sparse in the sense that they are supported only on the variables corresponding to products of variables that are present in quadratic functions. We show that these sparse linear inequalities yield an LP relaxation that gives the same bound as the SDP relaxation. We demonstrate how to identify these inequalities via a separation procedure that involves solving a structured ``projection'' SDP. In a computational study, we find that the sparse LP relaxations defined by these inequalities can significantly accelerate branch-and-bound methods for globally solving nonconvex optimization problems.

2 - Generalizing Double McCormick Volume Formulas: A Computational Exploration

Lillian Makhoul, University of Colorado Denver, Denver, CO, United States, Emily Speakman

By iteratively applying the standard McCormick inequalities, we obtain a so-called double-McCormick relaxation of the graph of the trilinear monomial over a specified box domain. The specific relaxation obtained depends on which bilinear term we chose to convexify first, resulting in a choice of three distinct double-McCormick relaxations, in addition to the convex hull. In previous work, a triangulation method was used to obtain volume formulas for each of these relaxations, assuming the box domain is contained in the nonnegative orthant. A consequence of this analysis is a ranking of the quality of the three double-McCormick relaxations by volume, meaning that given any box domain in the nonnegative orthant, we know a-priori which double-McCormick relaxation is tightest, and exactly how much "tightness" we are losing with respect to the convex hull. In the case of a general box domain, we have performed a complete case analysis for the convex hull polytope and have obtained a volume formula for each case. Volume formulas for the double-McCormick relaxations in this general setting are yet to be found. Thus, as a first step towards determining a ranking of the relaxations, we computationally explore the quality of the objective function bounds the relaxations produced.

3 - McCormick Envelopes for Mixed-integer PDE-constrained Optimization

Sven Leyffer, Argonne National Laboratory, Lemont, IL, United States, Paul Manns

McCormick envelopes are a standard tool for deriving convex relaxations of optimization problems that involve polynomial terms. McCormick relaxations provide lower bounds, for example, in branch-and-bound procedures for mixed-integer nonlinear programs but have not gained much attention in PDE-constrained optimization so far. This lack of attention may be due to the distributed nature of such problems, which on the one hand leads to infinitely many linear constraints (generally state constraints that may be difficult to handle) in addition to the state equation for a pointwise formulation of the McCormick envelopes and renders bound-tightening procedures that successively improve the resulting convex relaxations computationally intractable. We analyze McCormick envelopes for problems governed by a semilinear PDE involving a bilinearity and integrality constraints. We approximate the nonlinearity by averaging the nonconvex terms over the cells of a partition of the computational domain of the PDE. This approach yields convex relaxations that underestimate the original problem up to an a priori error estimate that depends on the mesh size of the discretization. These new approximate McCormick relaxations can be improved by means of an optimization-based bound-tightening procedure. We show that their minimizers converge to minimizers to a limit problem with a pointwise formulation of the McCormick envelopes when driving the mesh size to zero. We provide a computational example, for which we certify all of our imposed assumptions. The results point to both the potential of the methodology and the gaps in the research that need to be closed.

4 - An Aggregation-Based Convexification Framework for Chance-Constrained Programs

Danial Davarnia, Purdue University, West Lafayette, IN, United States, Hamed Rahimian

This work presents a new convexification framework that reformulates the single-row relaxations of chance-constrained programs as bilinear sets over a simplex in a lifted space. Using an aggregation-based procedure, we derive facet-defining inequalities in the original space of variables, unifying and extending known families of valid inequality while introducing new ones that capture larger portions of the convex hull. Main contributions include a general aggregation-based convexification method for bilinear sets over a simplex, a bilinear reformulation of mixing sets with knapsack constraints, and a unified description of existing and new inequality families. Computational results show that the proposed inequalities recover over 90% of convex hull facets, yielding stronger relaxations and improved efficiency.

FB07

Congress Room

Recent Advancements in Discrete Optimization with Applications

Invited Session

Discrete Optimization

Chair: Beste Basciftci, Tippie College of Business, University of Iowa, Iowa City, IA, United States

Co-Chair: Shunyu Yao, University of Iowa, Iowa City, IA, United States

1 - Influence Spread under Affective Polarization

Shunyu Yao, Tippie College of Business, University of Iowa, Iowa City, IA, United States, Beste Basciftci, Buddhika Nettasinghe

In this talk, we consider the problem of influence spread under affective polarization. Departing from the classic diffusion models that assume purely positive reinforcement among nodes, we study a diffusion process that depends on both the positive reinforcement from in-group members and the negative influence from out-group members. We define two optimization problems—Influence Maximization under Affective Polarization (IMAX-AP) and Influence Minimization under Affective Polarization (IMIN-AP), and formulate the diffusion process as an integer linear programming (ILP) model to characterize how individuals' opinions evolve over time through antagonistic interactions. We further establish complexity and inapproximability results for these problems and discuss solution approaches to address them.

2 - Fast Neural Value Function Surrogates for Nested Mixed-Integer Programs

Justin Dumouchelle, University of Calgary, Calgary, AB, Canada, Rahul Mihir Patel, Esther Julien, Merve Bodur, Jannis Kurtz, Elias B. Khalil

Mathematical optimization is an invaluable framework for modeling and solving decision-making problems with many successes in single-level deterministic problems (e.g., mixed-integer linear or nonlinear optimization). However, many real-world problems require accounting for uncertainty or the reaction of another agent. Paradigms such as stochastic programming, bilevel optimization, and robust optimization can model these situations but are much slower to solve than their deterministic counterparts, especially when discrete decisions must be made. This work demonstrates how a single framework based on value function approximation and optimizing over trained neural networks can be adapted to all three domains. Empirically, we find solutions of similar, and in some cases significantly better, quality than state-of-the-art algorithms in each field, often within a fraction of the running time. The datasets and three frameworks, Neur2SP (NeurIPS'22), Neur2RO (ICLR'24), and Neur2BiLO (NeurIPS'24), are open-sourced for further research.

3 - Integrated Layout Planning for Large-Scale Onshore Wind Farms

Yu Yang, University of Florida, Gainesville, FL, United States

As the world accelerates its shift toward clean energy, onshore wind, responsible for over 90% of global wind power, plays a pivotal role. However, large-scale planning of onshore wind farms remains a high-stakes, high-complexity problem that has received surprisingly less attention than its offshore counterpart. Motivated by collaborative projects with a leading electrical engineering company in China, we study a fully integrated layout design problem that jointly optimizes turbine placement, cable routing, and access road construction, three critical decisions that drive both capital investment and energy efficiency. We formulate the problem as a mixed-integer quadratically constrained program (MIQCP) that incorporates a quadratic wake effect model, accurately capturing the cumulative wake interactions. To solve this problem at scale, we design a specialized B&C algorithm that combines a dynamic linearization technique to effectively handle nonlinear constraints and novel cutting planes to significantly tighten the formulation. Our method not only outperforms commercial solvers by a wide margin but also unlocks substantial economic value, reducing infrastructure costs by around 20%. This work offers new insights into sustainable infrastructure design and demonstrates the power of integrated optimization in addressing real-world energy challenges.

4 - Transit Network Design with Two-Level Demand Uncertainties: A Machine Learning and Contextual Stochastic Optimization Framework

Hongzhao Guan, Georiga Institute of Technology, Atlanta, GA, United States, Beste Basciftci, Peter Van Hentenryck

Transit Network Design is a well-studied problem in the field of transportation, typically addressed by solving optimization models under fixed demand assumptions. Considering the limitations of these assumptions, this paper proposes a new framework, namely the Two-Level Rider Choice Transit Network Design (2LRC-TND), that leverages machine learning and contextual stochastic optimization (CSO) to incorporate two layers of demand uncertainties into the network design process. The first level identifies travelers who rely on public transit (core demand), while the second level captures the conditional adoption behavior of those who do not (latent demand), based on the availability and quality of transit services. To capture these two types of uncertainties, 2LRC-TND relies on two travel mode choice models, that use multiple machine learning models. To design a network, 2LRC-TND integrates the resulting choice models into a CSO, formulated using two approaches: Mixed-Integer Programming (MIP) and Constraint Programming (CP).

2LRC-TND is evaluated through a case study involving over 6,600 travel arcs and 38,000 trips in the Atlanta metropolitan area. Under different budget and ridership scenarios, the computational results demonstrate the effectiveness of the 2LRC-TND in designing transit networks that account for demand uncertainties and contextual information, offering a more realistic alternative to fixed-demand models. Additionally, under different experimental scenarios, not all MIP and CP models can be solved to optimality. This work also compares the effectiveness of MIP and CP in addressing the problem and discusses the ongoing development of customized solution algorithms.

FB08

Committee Room

Optimization for Carbon Accounting and Effective Grid Decarbonization

Invited Session

Emerging Applications of Optimization

Chair: Luc Cote, Cambridge, MA, United States

1 - Locational Marginal Emissions for Carbon-Aware Data Center Operations in Large-Scale Power Grids

Luc Cote, Massachusetts Institute of Technology, Cambridge, MA, United States

Carbon accounting methods for electricity consumption face challenges regarding physical deliverability, double counting, additionality, and impact magnitude. Locational Marginal Emissions (LMEs) show potential to address many of these key issues. However, their use in a large-scale power grids remains understudied. We analyze the properties of LMEs from a data center's perspective in a 1493-bus Western Interconnection over one year of hourly operation. We find that LME characteristics create three distinct regions: the hydropower-dominated Pacific Northwest, with low and stable LMEs; the coal-heavy Intermountain West, containing often high LMEs; and the Sunbelt, where mixed generation leads to variable LMEs correlated with solar output. This characterization provides analytical guidance for data center emission reduction. In particular, LME-guided emission reduction interventions through data center temporal-spatial load shifting, siting, and renewable procurement display over 85% accuracy with respect to actual emission reduction. Moreover, large-scale, nodal grid simulation is shown to be critical to accurate evaluation.

2 - Pricing Voluntary Renewable Contracts for an Optimal Decarbonization Pathway

Bolun Xu, Columbia University, New York, NY, United States, Zhiyuan Fan, Tianyi Lin

This talk presents a framework to analyze the optimal strategy for a vertically integrated power utility to price voluntary renewable energy and allocate the resulting revenue between the utility and renewable generators in its portfolio or contracts. We model voluntary renewable demand as households' and firms' willingness to pay to differentiate from the grid-average mix that includes fossil generation. By characterizing the Karush–Kuhn–Tucker (KKT) conditions of the utility's problem—augmented to reflect standard regulatory feasibility constraints—we show that the binding conditions align with distinct phases of the energy transition. The analysis yields policy-relevant guidance for tariff/program design under renewable integration. Two central insights emerge. First, the system cannot attain 100\% renewable penetration relying solely on voluntary demand for renewable power. Second, as renewable penetration deepens, three phases arise with different optimal pricing and revenue-assignment regimes: (i) spontaneous expansion, where willingness to pay suffices without revenue support; (ii) revenue-sharing–supported expansion, where optimal internal transfers from REC/tariff revenues sustain further build-out; and (iii) eventual stagnation, where voluntary demand is exhausted. These phases map to systematic adjustments in optimal REC/tariff pricing and revenue allocation as penetration grows, offering a theoretical foundation for utility program and tariff design in the renewable transition.

3 - Greening the Grid: Carbon-Aware Electricity Market Design

Line Roald, University of Wisconsin - Madison, Madison, WI, United States, Wenqian Jiang

As a growing number of electricity consumers, such as data centers and hydrogen companies, become carbon-sensitive, incorporating carbon emissions into their price-sensitive decision-making is crucial. This presentation discusses how electricity markets could be designed to account for consumer preferences for low-carbon energy. We show how such markets allow consumers to balance price and carbon considerations, and enable effective decarbonization of the power system.

FB09

Cabinet Room

Linear and Conic Optimization II

Invited Session

Linear and Conic Optimization

Chair: Jacob M. Aguirre, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Renato Monteiro, Georgia Institute of Technology

1 - When Does Primal/dual IPM Beat Primal-dual IPM in Linear Optimization

Wenzhi Gao, Stanford University, Stanford, CA, United States, Huikang Liu, Yinyu Ye, Madeleine Udell

The primal–dual interior point method (IPM) is widely regarded as the most efficient and robust IPM variant for linear optimization. In this talk, we present scenarios where pure primal and pure dual IPMs can outperform the standard primal–dual IPM. For linear programming, we show that the enhanced numerical stability of the pure primal IPM can lead to significant speedups over a primal–dual solver, particularly near convergence. For semidefinite programming, we extend the dual-scaling IPM with the self-dual embedding technique, demonstrating that the dual-scaling approach can better exploit problem sparsity while maintaining robustness in detecting infeasibility. We demonstrate the efficiency of our proposed approaches through extensive experiments on benchmark datasets.

2 - A Hybrid Sparse Augmented Lagrangian Method for Large-Scale Linear Programming

Jacob M. Aguirre, Georgia Institute of Technology, Atlanta, GA, United States, Renato D.C. Monteiro, Anton J. Kleywegt

We present a hybrid low-rank augmented Lagrangian method for large-scale linear programs. Each outer iteration solves a convex quadratic subproblem with an Enhanced Frank–Wolfe routine that alternates a support-restricted accelerated composite gradient refinement with a global Frank–Wolfe step, warm-started from the previous outer iterate. The subproblem is the augmented Lagrangian, with full multiplier updates and a doubling schedule for β_t until a joint residual-and-gap test terminates. The theory gives an $O(1/\bar\varepsilon)$ inner rate and a logarithmic bound on outer iterations via a potential $\Phi(\varepsilon_p, \varepsilon_g)$ (where ε_p is the primal-feasibility tolerance and ε_g is the primal-dual gap tolerance). For general LPs over the nonnegative orthant, an adaptive-radius Phase-1/Phase-2 scheme doubles the radius until the complementarity surrogate is small, guaranteeing finite-phase convergence and explicit bounds on total Frank–Wolfe steps. Empirically, the method is competitive with the simplex method, the interior-point method, PDLP, and other first order methods on synthetic and benchmark datasets, with both CPU and GPU implementations compared.

3 - SDP Approach to Quadratic k-Vertex-Disjoint Paths Problem

Mingming Xu, Clemson University, Clemson, SC, United States, Hao Hu

We consider the quadratic k-vertex–disjoint paths problem (Q-VDP). Given a directed graph G and k source–target pairs, the objective is to find k mutually vertex–disjoint paths that minimize a quadratic cost function. Such quadratic objectives naturally arise when interactions among paths matter. The problem is NP-hard when k is part of the input. In this work, we formulate Q-VDP as an integer quadratic program and derive semidefinite-programming (SDP) relaxations. We also develop a tailored Alternating Direction Method of Multipliers (ADMM) to solve these relaxations, yielding a fast algorithm that recovers exact solutions in practice. Numerical results demonstrate that our method achieves greater computational efficiency than the commercial solver Gurobi.

4 - Local Second-Order Limit Dynamics of the Alternating Direction Method of Multipliers for Semidefinite Programming

Shucheng Kang, Harvard University, Boston, MA, United States

The alternating direction method of multipliers (ADMM) is widely used for solving large-scale semidefinite programs (SDPs), yet on degenerate instances it often enters prolonged slow-convergence regions where the Karush--Kuhn--Tucker (KKT) residuals nearly stall. To explain and predict the fine-grained dynamical behavior inside these regions, we develop a local second-order limit dynamics framework for ADMM near an arbitrary KKT point. Assuming the existence of a strictly complementary primal--dual solution pair, we derive a second-order local expansion of the ADMM dynamics by leveraging a refined and simplified variational characterization of the (parabolic) second-order directional derivative of the PSD projection operator. This expansion reveals a closed convex cone of directions along which the local first-order update vanishes, and it induces a second-order limit map that governs the persistent drift after transient effects are filtered out. We characterize fundamental properties of this mapping, including its kernel, range, and continuity. A primal--dual decoupling further yields a clean scaling law for the effect of the penalty parameter in ADMM. We connect these properties to second-order dynamical features of ADMM, including fixed points, almost-invariant sets, and microscopic phases. Three empirical phenomena in slow-convergence regions are then explained or predicted: (1) angles between consecutive iterate differences are small yet nonzero, except for sparse spikes; (2) primal and dual infeasibilities are insensitive to penalty-parameter updates; (3) iterates can be transiently trapped in a low-dimensional subspace for an extended period. Extensive numerical experiments on the Mittelmann dataset corroborate our theoretical predictions.

FB10

Caucus Room

Data-Driven Algorithm Design and Analysis for Parametric Optimization

Invited Session

Computational Optimization

Chair: Rajiv Sambharya, University of Pennsylvania, Philadelphia, PA, United States

1 - Exact Verification of First-Order Methods via Mixed-Integer Linear Programming

Bartolomeo Stellato, Princeton University, Princeton, NJ, United States, Vinit Ranjan, Jisun Park, Stefano Gualandi, Andrea Lodi

We present exact mixed-integer linear programming formulations for verifying the performance of first-order methods for parametric quadratic optimization. We formulate the verification problem as a mixed-integer linear program where the objective is to maximize the infinity norm of the scaled fixed-point residual after a given number of iterations, corresponding to violations of the KKT conditions. Our approach captures a wide range of gradient, projection, proximal iterations through affine or piecewise affine constraints. We derive tight polyhedral convex hull formulations of the constraints representing the algorithm iterations. To improve the scalability, we develop a custom bound tightening technique combining interval propagation, operator theory, and optimization-based bound tightening. Numerical examples, including linear and quadratic programs from network optimization, sparse coding using Lasso, and optimal control, show that our method provides several orders of magnitude reductions in the worst-case fixed-point residuals, closely matching the true worst-case performance.

2 - Verification of Sequential Convex Programming for Parametric Non-convex Optimization

Rajiv Sambharya, University of Pennsylvania, Philadelphia, PA, United States, Nikolai Matni, George Pappas

We introduce a verification framework to exactly verify the worst-case performance of sequential convex programming (SCP) algorithms for parametric non-convex optimization. The verification problem is formulated as an optimization problem that maximizes a performance metric (e.g., the suboptimality after a given number of iterations) over parameters constrained to be in a parameter set and iterate sequences consistent with the SCP update rules. Our framework is general, extending the notion of SCP to include both conventional variants such as trust-region, convex-concave, and prox-linear methods, and algorithms that combine convex subproblems with rounding steps, as in relaxing and rounding schemes. Unlike existing analyses that may only provide local guarantees under limited conditions, our framework delivers global worst-case guarantees—quantifying how well an SCP algorithm performs across all problem instances in the specified family. Applications in control, signal processing, and operations research demonstrate that our framework provides, for the first time, global worst-case guarantees for SCP algorithms in the parametric setting.

3 - Why Recurrent Models Are a Natural Fit for Learning to Optimize?

Jialin Liu, University of Central Florida, Orlando, FL, United States, Lisang Ding, Stanley Osher, Wotao Yin

When designing AI to solve optimization problems, which model class is superior: end-to-end feedforward networks or recurrent fixed-point learners? We argue the answer lies in the problem's intrinsic structure. For many canonical optimization tasks, the target "solution map"—the function from problem coefficients to the optimal solution—is only locally, not globally, Lipschitz. This is due to discrete shifts in the solution structure, such as basis or active-set changes. This structural fact presents a challenge for feedforward models but creates a natural advantage for recurrent architectures. We show that a simple, regular update operator, when iterated, can elegantly represent these complex, piecewise-regular solution maps. Our theory makes this precise, proving that the expressive power of such recurrent models grows with the number of test-time iterations, ultimately matching the exact class of locally Lipschitz mappings. This finding yields a clear architectural criterion for the field of Learning-to-Optimize: when a problem's solution map exhibits this local-but-not-global regularity, recurrent models are the superior choice. This allows practitioners to scale solution accuracy simply by increasing test-time compute, without enlarging the model. We validate this principle across a diverse set of optimization problems.

4 - Deep Learning-Aided Large-Scale Distributed Optimization

Augustinos Saravanos, Massachusetts Institute of Technology, Cambridge, MA, United States, Alex Oshin, Hunter Kuperman, Arshiya Taj Abdul, Vincent Pacelli, Evangelos Theodorou

As modern decision-making problems continue to grow in scale and structural complexity, there is a pressing need for optimization frameworks that remain both computationally efficient and reliable. Classical model-based distributed optimization is a powerful paradigm for large-scale decision-making, yet it typically requires extensive tuning, faces computational and communication limitations, and has limited ability to exploit data or generalize across problem instances. On the other hand, black-box deep neural networks have witnessed widespread success in various domains, but they often lack interpretability and performance guarantees. To leverage their complementary strengths, we introduce a new class of deep learning-aided distributed architectures for large-scale convex optimization. The proposed frameworks rely on unrolling the iterations of distributed optimizers as layers of a deep learning network, and then incorporating learnable components. Extensive experiments demonstrate substantial improvements over traditional optimization algorithms in wall-clock time across diverse applications including optimal control, linear regression and transportation networks. Critically, they exhibit strong generalization: trained on small problem instances, they scale to solve significantly larger ones using the same learned policies. We also provide certifiable performance guarantees on the outputs, ensuring high-quality solutions for unseen problem instances.

FB11

Charter Room

Recent Developments in Network Optimization and Applications

Invited Session

Network Optimization

Chair: Zolykha Rezaei, Texas Tech University, Lubbock, TX, United States

1 - Stochastic Bilevel Optimization for the Network Design of Multimodal Transit Systems with Heterogeneous Rider Preferences under Uncertain Travel Times and Demand

Yiling Zhang, University of Minnesota, Minneapolis, MN, United States, Suri Liu, Beste Basciftci, Wenyuan Wang

Designing efficient and user-friendly multimodal transit networks is critical for modern urban mobility. We study a novel stochastic multimodal transit network design problem that integrates fixed-route services with on-demand shuttles, explicitly accounting for heterogeneous rider preferences, uncertain travel times and passenger demand. The hierarchical decision-making process is modeled using a two-stage stochastic bilevel optimization problem, where the transit agency (leader) determines the network design, and riders (followers) select their preferred routes based on realized traffic conditions. The model inherits the complexity of a nonconvex bilevel structure with stochastic programming, posing significant computational challenges.
To address this, we first develop an equivalent single-level mixed integer linear programming (MILP) reformulation by introducing a response search algorithm that efficiently enumerates critical follower route choices. To further enhance scalability, we propose a decomposition method that combines a relaxed formulation with a subset of follower responses and iteratively strengthens it with valid cutting planes. Computational experiments on instances derived from a public transit network in Dalian, China, demonstrate the efficiency and effectiveness of our approaches, achieving significant speedups compared to existing single-level reformulations. Additionally, a comprehensive case study on the Ann Arbor/Ypsilanti region in Michigan highlights practical benefits, including cost savings and more convenient route suggestions, demonstrating the value of the proposed stochastic bilevel model over deterministic or single-level counterparts.

2 - A Robust Bilevel Interdiction Problem with Applications in Human Trafficking Disruption

Daniel Lopes da Silva, Florida Polytechnic University, Lakeland, FL, United States, Thomas Sharkey, Yongjia Song

We investigate a class of robust bilevel interdiction problems motivated by applications in human trafficking disruption. In these problems, the follower, who we assume to be rational, will solve a linear program where each variable has an upper bound. The leader, who interdicts the linear program by removing variables, minimizes the total level of activity across a special set of variables in the optimal solution of the follower's problem. The problem includes data uncertainty as the leader does not know the follower's operational cost vector, but only that it belongs to a given uncertainty set. The follower has complete knowledge about its own parameters. We present a column-and-constraint generation (C&CG) method to solve the optimistic version of these problems. We discuss the difficulties in solving the standard subproblem from the C&CG method and propose a method to solve the subproblem by exploiting the structure of the uncertainty set. Computational experiments with synthetic but realistic domestic sex trafficking networks demonstrate that the alternative approach to solving the subproblem consistently outperforms the standard approach in terms of running time and numerical stability. We then solve various realistic instances of sex trafficking networks through the C&CG method with the alternative approach and discuss insights obtained.

3 - Commit or Defer Shortest Path Interdiction with Incomplete Information

Juan Borrero, University of South Florida, Tampa, FL, United States, Shuai Shao, Oleg Prokopyev

We study a sequential shortest path interdiction problem where the interdiction resources are not renewed once they are used. Assuming the interdictor does not know the cost vector of the network and only has access to an uncertainty set, we consider finding policies that are worst-case optimal. We show such policies are intractable to compute, and provide set of conditions that can be checked in real-time on whether the interdictor should commit or defer.

4 - A Bilevel Stochastic Optimization Framework for Strategic Deployment of Survivor Geolocation Systems and Rescue Teams in Post-Disaster Operations

Jose Walteros, University at Buffalo, Buffalo, NY, United States, Angelo Soto-Vergel, Diana Ramirez-Rios, Juan Borrero

The prompt collection of accurate information about potential survivor locations in the aftermath of a disaster is critical for the success of search and rescue operations.
Traditional methods rely on teams combing through affected areas for signs of survivors. However, limitations on data availability, resources, and personnel can severely
restrict the effectiveness of these efforts. Recent technological advancements have enabled the strategic deployment of remote sensors in affected areas, providing an
effective tool for identifying signs of survivors before rescue missions are launched. The data collected by these sensors can significantly enhance the situational
awareness of rescue planners, allowing them to guide rescue teams toward regions with potentially high survivor densities, thereby improving the likelihood of successful
outcomes. This presentation introduces a bilevel stochastic optimization model with recourse that addresses the joint problem of placing sensor probes and deploying
rescue teams. In the first level, the model selects candidate zones for deploying a set of geolocated probes equipped with various sensors (e.g., heat, sound, motion),
subject to a predefined budget. These decisions rely solely on a-priori probabilities of the survivors' locations, inferred from pre-disaster indicators, hazard exposure risks,
and socio-demographic markers. These factors serve as a general indicator of the potential presence of survivors and their level of endangerment. In the second level, the
model determines the areas the teams will traverse, using the partial information collected from the deployed probes. The overarching objective of the model is to
maximize the expected number of survivors that can be rescued.

March 20, 2026, 2:45 PM - 4:15 PM

FC01

Grand Ballroom

Iterative Optimization in the Average Case

Invited Session

Optimization in Data Science

Chair: Ashwin Pananjady, ISyE Georgia Tech, Atlanta, GA, United States

1 - State Evolution Beyond First-order Methods

Kabir Verchand, University of Southern California (USC), Los Angeles, CA, United States

We consider the dynamics of iterative optimization algorithms when applied to instances with high-dimensional, random data. When the algorithm of choice is a first-order method, it is known that the dynamics of the method are well approximated by a low-dimensional deterministic recursion known as state evolution. In this talk, we move beyond first-order methods and develop a rigorous state evolution for a far larger set of algorithms. The main technical tool is a variant of Bolthausen’s conditioning method based on a sequential variant of Gordon’s Gaussian comparison inequality.

2 - Asymptotic Theory of Iterated Empirical Risk Minimization, with Applications to Active Learning

Yue Lu, Harvard University, Cambridge, MA, United States, Hugo Cui

We study iterated empirical risk minimization (ERM) procedures in which multiple ERMs are trained on the same dataset, with predictions from earlier stages explicitly entering the loss functions of later ones. Such procedures arise naturally in active learning and reweighting algorithms, but induce strong statistical dependencies that place them beyond the scope of classical asymptotic ERM theory. For linear models trained with a broad class of convex losses on Gaussian mixture data, we derive a sharp and fully asymptotic characterization of the test error in the high-dimensional regime where the sample size and ambient dimension scale proportionally. Our analysis builds on a nested leave-one-out approach that allows us to explicitly control the dependencies induced by data reuse and prediction-dependent losses, yielding closed-form predictions for the performance of the ERM estimators. We apply this framework to pool-based active learning, lifting oracle and sample-splitting assumptions pervasive in existing analyses. The resulting theory reveals a previously unexplained tradeoff in labeling budget allocation across stages, as well as a selection-driven double-descent phenomenon that occurs at fixed model size and fixed total sample count.

3 - Algorithm Inference via Nonconvex Gradient Descent

Qiyang Han, Rutgers University, Piscataway, NJ, United States

Conventional statistical inference methods are typically developed for models simple enough to admit tractable estimators through carefully designed iterative algorithms. In contrast, modern deep learning models are enormously complex, yet are trained by simple gradient-descent-type algorithms, often without any provable guarantee of algorithmic convergence to global/local optima.

Can we reconcile classical inference principles with these highly complicated, modern learning paradigms? In this talk, we will present a new inference framework addressing this question, by showing that valid statistical inference can be performed along the entire gradient descent trajectory, iteration by iteration, without requiring convexity of the loss landscape or convergence of the algorithm.

To illustrate this concept, we begin with a single-index (one-layer neural network) regression model and demonstrate how gradient descent iterates can be "debiased", at each iteration, to yield valid confidence intervals for the underlying signal and consistent estimates of generalization errors. We then extend this paradigm to the much more challenging setting of learning with general multi-layer neural networks in their full complexity, where the loss landscape can be arbitrarily complex. Crucially, the proposed method remains valid without requiring either algorithmic convergence or oracle knowledge of the unknowns, and may therefore inform practical decisions such as early stopping and hyperparameter tuning.

The key technical ingredient underlying this new inference paradigm is a recent entrywise dynamical mean-field theory for a broad class of first-order algorithms developed by the speaker.

This talk is partially based on joint with Masaaki Imaizumi and Xiaocong Xu.

4 - A New Approach to Polychromatic Computed Tomography

Ashwin Pananjady, ISyE Georgia Tech, Atlanta, GA, United States

I will present a new class of methods for single material recovery polychromatic computed tomography that is based on formulating the signal recovery problem using a careful monotone variational inequality. The new class of methods is fast, accurate and flexible, and can produce reconstructions that are comparable in quality to state-of-the-art algorithms while reducing radiation doses by two orders of magnitude. I will also present a class of random data assumptions that can serve as useful guides when comparing different iterative algorithms in this space of problems.

Based on joint work with Mengqi Lou, Sara Fridovich-Keil, Namhoon Kim, Amir Pourmorteza and Kabir

Verchand.

FC02

Georgian Room

Recent Advances in Stochastic and Large-Scale Optimization

Invited Session

Optimization in Data Science

Chair: Chuan He, Linköping University, Linköping, Sweden

Co-Chair: Hanyang Li, UC Berkeley, Albany, CA, United States

1 - Some Unified Theory for Variance Reduced Prox-Linear Methods

Yue Wu, Johns Hopkins University, Baltimore, MD, United States, Benjamin Grimmer

We consider the nonconvex, nonsmooth problem of minimizing a composite objective of the form f(g(x)) + h(x) where the inner mapping g is a smooth finite summation or expectation amenable to variance reduction. In such settings, prox-linear methods can enjoy variance-reduced speed-ups despite the existence of nonsmoothness. In this talk, we will present a unified convergence theory applicable to a wide range of common variance-reduced vector and Jacobian constructions. Our theory (i) only requires operator norm bounds on Jacobians (whereas prior works used potentially much larger Frobenius norms), (ii) provides state-of-the-art high probability guarantees, and (iii) allows inexactness in proximal computations.

2 - Nested Stochastic Algorithm for Generalized Sinkhorn distance-Regularized Distributionally Robust Optimization

Yufeng Yang, Texas A&M University, College Station, TX, United States, Yi Zhou, Zhaosong Lu

Distributionally robust optimization (DRO) is a powerful technique to train robust models against data distribution shift. This paper aims to solve regularized nonconvex DRO problems, where the uncertainty set is modeled by a so-called generalized Sinkhorn distance and the loss function is nonconvex and possibly unbounded. Such a distance allows to model uncertainty of distributions with different probability supports and divergence functions. For this class of regularized DRO problems, we derive a novel dual formulation taking the form of nested stochastic optimization, where the dual variable depends on the data sample. To solve the dual problem, we provide theoretical evidence to design a nested stochastic gradient descent (SGD) algorithm, which leverages stochastic approximation to estimate the nested stochastic gradients. We study the convergence rate of nested SGD and establish polynomial iteration and sample complexities that are independent of the data size and parameter dimension, indicating its potential for solving large-scale DRO problems. We conduct numerical experiments to demonstrate the efficiency and robustness of the proposed algorithm.

3 - A Bregman ADMM for Bethe Variational Problem

Tianyun Tang, The University of Chicago, Chicago, IL, United States, Yuehaw Khoo, Kim-Chuan Toh

In this work, we propose a novel Bregman ADMM with nonlinear dual update to solve the Bethe variational problem (BVP), a key optimization formulation in graphical models and statistical physics. Our algorithm provides rigorous convergence guarantees, even if the objective function of BVP is non-convex and non-Lipschitz continuous on the boundary. A central result of our analysis is proving that the entries in local minima of BVP are strictly positive, effectively resolving non-smoothness issues caused by zero entries. Beyond theoretical guarantees, the algorithm possesses high level of separability and parallelizability to achieve highly efficient subproblem computation. Our Bregman ADMM can be easily extended to solve the quantum Bethe variational problem. Numerical experiments are conducted to validate the effectiveness and robustness of the proposed method.

4 - Subgradient Regularization: A Descent-Oriented Subgradient Method for Nonsmooth Optimization

Hanyang Li, UC Berkeley, Berkeley, CA, United States, Ying Cui

In nonsmooth optimization, a negative subgradient is not necessarily a descent direction, making the design of convergent descent methods based on zeroth-order and first-order information a challenging task. The well-studied bundle methods and gradient sampling algorithms construct descent directions by aggregating subgradients at nearby points in seemingly different ways, and are often complicated or lack deterministic guarantees. In this work, we identify a unifying principle behind these approaches, and develop a general framework of descent methods under the abstract principle that provably converge to stationary points. Within this framework, we introduce a simple yet effective technique, called subgradient regularization, to generate stable descent directions for a broad class of nonsmooth marginal functions, including finite maxima or minima of smooth functions. When applied to the composition of a convex function with a smooth map, the method naturally recovers the prox-linear method and, as a byproduct, provides a new dual interpretation of this classical algorithm. Numerical experiments demonstrate the effectiveness of our methods on several challenging classes of nonsmooth optimization problems, including the minimization of Nesterov's nonsmooth Chebyshev-Rosenbrock function.

FC03

Plaza I

Advanced Methods for Large-scale Nonlinear and Stochastic Optimization II

Invited Session

Nonlinear Optimization

Chair: Qi Wang, University of Michigan, Ann Arbor, MI, United States

1 - A Hybrid Newton-CG-MR Method with Worst Case Complexity Guarantees for Nonconvex Optimization

Michael O'Neill, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, Chengze Xie

Inexact Newton methods have a long history of popularity in large scale, nonlinear optimization. We propose a new Newton type algorithm which uses a combination of
the conjugate gradient (CG) and minimum residual (MR) solvers for linear systems to generate search directions. The new method outperforms both Newton-CG and
Newton-MR methods on a wide range of test problems while maintaining optimal worst-case complexity guarantees for nonconvex optimization.

2 - Fair Supervised Learning Through Constraints on Smooth Nonconvex Unfairness-Measure Surrogates

Zahra Khatti, Lehigh University, Bethlehem, PA, United States, Daniel Robinson, Frank E. Curtis

A new strategy for fair supervised machine learning (ML) is proposed. Its advantages compared to others are as follows. (a) We introduce a new smooth nonconvex surrogate to approximate the Heaviside functions involved in discontinuous unfairness measures. The surrogate is a tight approximation that ensures the trained prediction models are fair, as opposed to other (e.g., convex) surrogates that can fail to lead to fair prediction models. (b) Rather than rely on regularizers (that lead to optimization problems that are difficult to solve) and corresponding regularization parameters (that can be expensive to tune), we propose a strategy that employs hard constraints so that specific tolerances for unfairness can be enforced. (c) Our strategy readily allows for constraints on multiple (potentially conflicting) unfairness measures at the same time. Multiple measures can be considered with a regularization approach, but at the cost of having even more difficult training problems and further expense for tuning. By contrast, through hard constraints, our strategy leads to training problems that can be solved tractably through minimal tuning.

3 - Projected Gradient Methods for Nonconvex and Stochastic Optimization: New Complexities and

Auto-Conditioned Stepsizes

Tianjiao Li, Massachusetts Institute of Technology, Cambridge, MA, United States, Guanghui Lan, Yangyang Xu

We present a novel class of projected gradient (PG) methods for minimizing a smooth but not necessarily convex function over a convex compact set. We first provide a novel analysis of the "vanilla" PG method, achieving the best-known iteration complexity for finding an approximate stationary point of the problem. We then develop an "auto-conditioned" projected gradient (AC-PG) variant that achieves the same iteration complexity without requiring the input of the Lipschitz constant of the gradient or any line search procedure. The key idea is to estimate the Lipschitz constant using first-order information gathered from the previous iterations, and to show that the error caused by underestimating the Lipschitz constant can be properly controlled. We then generalize the PG methods to the stochastic setting, by proposing a stochastic projected gradient (SPG) method and a variance-reduced stochastic gradient (VR-SPG) method, achieving new complexity bounds in different oracle settings. We also present auto-conditioned stepsize policies for both stochastic PG methods and establish comparable convergence guarantees.

4 - A Sequential Cubic Optimization Method with Second-Order Complexity Guarantees for Equality Constrained Optimization

Nikolaos Dimou, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, Michael ONeill

We develop a new method for equality constrained optimization problems based on a sequential cubic programming framework. Each iteration utilizes a step decomposition based on the Jacobian of the constraints into a normal and a tangential component, the latter of which is found by solving a subproblem involving cubic regularization. The method incorporates second order correction steps as necessary to ensure global convergence to second order stationary points as well as local quadratic convergence. In addition, we show that the algorithm is the first to obtain worst case complexity guarantees on the order of $\epsilon_g^{-3/2}$ for the gradient of the Lagrangian, ϵ_H^{-3} in terms of second order stationarity, and ϵ_c^{-1} in terms of the constraint violation. These complexity guarantees are the best known results of any method for this class of problems.

FC04

Plaza II

Advances in Continuous Optimization Algorithms

Invited Session

Nonlinear Optimization

Chair: Gabriel Provencher Langlois, University of Illinois Urbana-Champaign, Champaign, IL, United States

Co-Chair: Akwum Onwunta, Lehigh University, Bethlehem, PA, United States

1 - Exact Solutions of Differential Inclusions Arising from Linear and Certain Quadratic Programs

Gabriel Provencher Langlois, University of Illinois Urbana-Champaign, Champaign, IL, United States, Jerome Darbon

In this talk, I will show that differential inclusions arising from linear and certain quadratic programs are integrable, that is, their solutions can be computed explicitly. The analysis naturally leads to simple algorithms that compute solutions exactly, numerically up to machine precision, and are very fast. I will illustrate these algorithms on several numerical examples of practical interest in statistics and scientific computing.

2 - Low-Rank Gradient Flow with Multilevel Acceleration - A First Order Algorithm for Non-Convex Optimization

Andrei Draganescu, University of Maryland, Baltimore County, Baltimore, MD, United States

We introduce a novel class of first-order methods for unconstrained optimization, called low-rank gradient flows (LRGFs). The idea behind these methods is to construct at every optimization step a low-rank quadratic surrogate for the cost function, followed by an analytic solve for the gradient flow on the surrogate model; the optimization step concludes with a line search on the curve representing the gradient flow. It is shown that the above steps are condensed in a very simple formula for the gradient flow, at a cost per step that is comparable to that of a nonlinear conjugate gradient algorithm. The fact that the line search is conducted along a curve distinguishes LRGF from other first order optimization methods, where the line search is conducted along a search direction, that is, a straight line. This may also help LRGF better navigate the geometry of the cost function, allowing the method to avoid local minima more often than other first-order methods, as shown by numerical experiments. For higher dimensional problems the convergence can be accelerated using a multilevel strategy based on reduced order models.

3 - Natural Gradient Descent in the Context of PDE-Based Optimization

Levon Nurbekyan, Emory University, Atlanta, GA, United States

I will discuss the natural gradient descent (NGD) optimization technique in the context of PDE-based optimization problems. In particular, I will discuss the impact of the norm used for generating the "information matrix" in connection with the regularity theory of the PDE under consideration. I will also discuss efficient computational methods for performing NGD using a streamlined least squares formulation and numerical linear algebra techniques.

4 - Existence and Computation of Solutions to Non-Monotone Variational Inequalities

Sina Arefizadeh, Arizona State University, Tempe, AZ, United States, Angelia Nedich

In this talk, we present sufficient conditions for the existence of solutions to non-monotone Variational Inequalities (VIs) based on inverse mapping theory and degree theory. In contrast to the common approach in the current literature that studies the sum of the Jacobian of the VI mapping and its transpose, our conditions obtained through the Inverse Function Theorem rely on the product of the Jacobian and its transpose, which naturally appears in the derivative of the residual norm. This perspective opens up broad possibilities for investigating non-monotone variational inequalities, both in establishing the existence of solution results and in developing solution methods. We have obtained several sufficient conditions for the existence of solutions, including a sufficient condition for the existence of a Minty solution. Furthermore, we have shown that the Korpelevich and Popov methods converge to a solution of a non-monotone VI, provided that a Minty solution exists.

FC05

Plaza III

Optimization and Learning in Sequential Decision Making

Invited Session

Optimization under Uncertainty

Chair: Negar Soheili, University of Illinois-Chicago, Chicago, IL, United States

Co-Chair: Andre Cire, University of Toronto, Toronto, ON, Canada

1 - Bayesian Risk-sensitive Policy Optimization for MDPs with General Loss Functions

Xiaoshuang Wang, Georgia Tech, Atlanta, GA, United States, Enlu Zhou

Motivated by many application problems, we consider Markov decision processes (MDPs) with a general loss function and unknown parameters. To mitigate the epistemic uncertainty associated with unknown parameters, we take a Bayesian approach to estimate the parameters from data and impose a coherent risk functional (with respect to the Bayesian posterior distribution) on the loss. Since this formulation usually does not satisfy the interchangeability principle, it does not admit Bellman equations and cannot be solved by approaches based on dynamic programming. Therefore, We propose a policy gradient optimization method, leveraging the dual representation of coherent risk measures and extending the envelope theorem to continuous cases. We then show the stationary analysis of the algorithm with a convergence rate of $O(T^{-1/2} + r^{-1/2})$, where T is the number of policy gradient iterations and r is the sample size of the gradient estimator. We further extend our algorithm to an episodic setting, and establish the global convergence of the extended algorithm and provide bounds on the number of iterations needed to achieve an error bound $O(\epsilon)$ in each episode.

2 - Delayed Allocation in Marginalized Flow Models for Weakly Coupled Markov Decision Processes

Andre Cire, University of Toronto, Toronto, ON, Canada, Selva Nadarajah

We introduce a process interpretation of fluid policies derived from Lagrangian relaxation of weakly coupled Markov decision processes (WDPs), revealing that existing approaches follow an "allocate-then-correct" sequence: first allocating marginal probability flows to component actions fluid linear program, then correcting for constraint violations. We identify that these feasibility corrections can be costly, degrading expected policy performance and increasing sample path variability, which we term fluidmodel feasibility risk. We propose delayed allocation, which reverses this process by identifying feasible actions before allocation. This creates a hierarchy of structure-aware fluid models that restrict actions to subsets of provably feasible solutions, operating on convex hulls of feasible action sets rather than linear relaxations. Our hierarchy converges in the limit to exact-feasible models from the literature, but remains computationally tractable at intermediate levels. For a broad class of problems, we establish transferability results showing that approximation guarantees for deterministic constraint systems carry over to our stochastic formulations. We prove that polynomial-time approximation algorithms yield polynomial-time delayed allocation models, even when exact-feasible formulations require exponential representation – a striking complexity difference supporting intermediate models in our hierarchy. Numerical experiments on dynamic assortment and maintenance problems demonstrate that delayed allocation reduces optimality gaps of standard approaches, while also decreasing performance variability. Our models also reduce reliance on frequent reoptimization, broadening applicability to environments where regular reoptimization is costly or impractical.

3 - Multistage Stochastic Dispatch of Mobile Power Sources for Enhanced Grid Resilience under Extreme Weather

Jean-Christophe Raymond-Bertrand, Virginia Tech, Blacksburg, VA, United States, Mithun Goutham, Deepjyoti Deka, Harsha Nagarajan, Russell Bent, Rohit Kannan

This paper presents a multistage stochastic framework to enhance operational resilience of power grids during hurricanes by incorporating uncertainty in the weather forecasts. Unlike deterministic models which rely on a single most-probable trajectory, our model anticipates, adapts, and restores against hurricane-induced damage. Using ensemble forecasts and component fragility curves, a data-driven scenario generator quantifies the likelihood of damage across power-stations and transmission lines. Scenario reduction techniques identify a subset of highest-impact scenarios compatible with stochastic dual dynamic programming. Within each stage, the placement of mobile power sources is optimized to minimize loadshed, constrained by network, ramping, and movement constraints. The dynamic nature of weather forecasting is captured using a rolling-horizon framework, allowing the model to be re-optimized as new forecasts become available. Computational experiments using modified IEEE 5, 39, and 118-bus networks under a Category-5 hurricane scenario demonstrate that our proposed framework significantly reduces the expected loadshed compared to wait-and-see approaches, while remaining tractable on commercially available hardware. This combination of ensemble forecast data, stochastic optimization, rolling-horizon, and adaptive-restorative planning establishes a novel and practical approach to increase real-time resilience of power-grids during extreme weather events.

4 - A Doubly Randomized Fluid Method for MDPs

Negar Soheili, University of Illinois-Chicago, Chicago, IL, United States, Selva Nadarajah, Shakiba Rahnama

Fluid models are often used to simplify large stochastic control problems. Instead of tracking stochastic transitions directly, they replace the transition dynamics with their expected evolution, resulting in a deterministic system that is much easier to analyze and optimize. While fluid models can reveal useful structural properties of good policies, they smooth out the randomness in the system and therefore lose information about the long-run state-action frequencies that are important for designing effective policies in the original stochastic problem.

We introduce a doubly randomized fluid model that keeps the simplicity of fluid approximations while preserving the original model's key stochastic structure. Our starting point is a linear-program representation of long-run behavior, which directly captures how often each state-action pair is used under a policy. This formulation is typically intractable because it involves infinitely many variables and must satisfy a flow-balance condition for every state.

Our method addresses both difficulties using two layers of randomization. We approximate long-run state-action frequencies using a finite set of randomly generated basis functions, and we enforce flow balance through a finite number of randomly selected surrogate conditions. The resulting tractable problem recovers the long-run state-action frequencies that characterize good policies - something classical fluid models cannot do. We show that as we increase the number of basis functions and surrogate constraints, the approximation improves and converges to the behavior of the true stochastic system with epsilon-error guarantees.

FC06

Director's and Lounge

Theoretical Advances in Global Optimization

Invited Session

Global Optimization

Co-Chair: Harsha Nagarajan, Los Alamos National Laboratory, Los Alamos, NM, United States

Co-Chair: Adrian Göß, University of Technology Nuremberg (UTN), Nuremberg, N/A, Germany

1 - New Convexification Techniques for Graphical Decision Models

Chengwenjian Wang, University of Minnesota, Minneapolis, MN, United States, Jean-Philippe Richard, Mohit Tawarmalani

Graphical decision models, such as decision trees and decision diagrams, represent decision processes using graphs whose nodes or edges encode decision variables. Their intuitive structure has made them a popular tool for prediction, modeling human decision-making, and capturing combinatorial structure. Hence, these models have deep connections to optimization, where they are used to formulate complex discrete problems. In this work, we propose new convexification techniques for graphical decision models that yield stronger formulations. Our approach systematically exploits graph structure to derive strong valid inequalities, enhancing both theoretical understanding and computational performance. We further demonstrate that these techniques extend naturally to a broader class of problems, including network design, showing their versatility.

2 - Convex Relaxations of Unit Commitment with AC Power Flow

Xiangxin An, Georgia Institute of Technology, Atlanta, GA, United States, Constance Crozier, Santanu Dey

Motivated by the challenges from nonconvex AC power flow physics and binary unit-commitment decisions, we investigate how convexification strategies—originally developed for AC Optimal Power Flow (ACOPF)—can be extended and unified within a full AC Unit Commitment framework. Our work compares several relaxations, including Second-Order Cone (SOC) relaxations, McCormick relaxations, and outer approximations of SOC constraints. To further strengthen these relaxations, we incorporate tightening techniques such as valid inequalities and a feasibility-based bound-tightening (FBBT) procedure that iteratively refines variable bounds. A key feature of our approach is that FBBT is performed on a relaxed UC model with slackened demand, allowing the resulting bounds to be reused across different demand profiles within a prescribed tolerance. Together, these components provide a more scalable convex relaxations for Unit Commitment with AC Power Flow.

3 - Accelerating Global Parabolic Approximations: A Comparison with Piecewise Linear Relaxations

Adrian Göß, University of Technology Nuremberg (UTN), Nuremberg, Germany

Solving mixed-integer nonlinear programs (MINLPs) typically relies on constructing relaxations that are easier to tackle than the original problem. Recently, global parabolic (PARA) relaxations were introduced, featuring separable quadratic functions — paraboloids — as global under- or overestimators of general nonlinear constraint functions. At its core, this approach computes PARA approximations a-priori for a given tolerance and function domain. The resulting paraboloids are used in the form of a lookup table to construct relaxations by replacing the original constraint.

Although this approach provides great flexibility in approximating multivariate functions, its computational tractability is limited, preventing a direct comparison with established methods. We therefore propose a new method to compute PARA approximations a-priori that is bound to univariate constraint functions but achieves orders-of-magnitude speedups. This improvement enables a computational comparison with piecewise linear (PWL) relaxations on general MINLP instances from the well-established MINLPLib.

In particular, relaxations with varying tolerance are constructed for either technique and then solved. The results show that PWL relaxations are faster to solve when the tolerance is high, whereas for tighter tolerances the advantage shifts toward the PARA approach. This effect is attributed to the large number of auxiliary variables introduced by PWL to identify the active piece, which grows with smaller tolerance, while PARA relaxations do not require additional variables.

4 - Some Theoretical Results on the Complex System with Linear and Conic Complementary Constraints

Jiming Peng, University of Houston, Houston, TX, United States, Wissam AlAli

In this talk, we consider the issue of determining the feasibility of a complex system defined by linear and conic complementary constraints (SL3Cs). SL3C arises frequently in economic analysis and engineering management. Subclasses of SL3Cs include several complex systems that have been well-studied in the literature such as linear complementary problems, affine variational inequalities, systems with linear and binary constraints, systems with linear and second-order conic complementary constraints, and systems with linear and semidefinite complementary constraints. It has been shown that finding a feasible solution to SL3C is NP-hard.

In this talk, we introduce several foundational theories on SL3Cs. First, we introduce the so-called universal relaxation theory (URT), which shows the feasibility issue of a given SL3C can be addressed via solving some linear conic optimization problem with a suitable linear objective function. Based on the new URT, we recast the issue of determining the feasibility of an SL3C as another equivalent bilinear conic optimization problem. Then we explore the constructed bilinear model to derive different sufficient and necessary conditions for the global optimal solution of the underlying bilinear model. We will also discuss how the new theoretical insights can help the design of novel trustworthy and explainable back-propagation methods for SL3Cs.

5 - Asymptotically Tight Lagrangian Dual of Smooth Nonconvex Problem

Jingye Xu, Georgia Institute of Technology, Atlanta, GA, United States

In convex geometry, the Shapley–Folkman Lemma asserts that the nonconvexity of a Minkowski sum of n-dimensional nonconvex sets does not accumulate once the number of summands exceeds the dimension n, and thus the sum becomes approximately convex. Originally published by Starr in the context of quasi-equilibrium in nonconvex market models in economics, the lemma has since found widespread use in optimization, particularly for estimating the duality gap of the Lagrangian dual of separable nonconvex problems.
Given its foundational nature, we pose the following geometric question: \emph{Is it possible for the nonconvexity of the Minkowski sum of n-dimensional nonconvex sets to vanish as the number of summands increases, under some general conditions?} We answer this affirmatively. First, we provide a new and elementary geometric proof of the Shapley–Folkman Lemma based on the facial structure of the convex hull of each set. This leads to an unconditional improvement over the classical error bound derived from the lemma.
Building on this new geometric perspective, we further show that when most of the sets satisfy a certain ``local smoothness'' condition, their Minkowski sum converges directly to a convex set, with a vanishing nonconvexity measure. In optimization, this implies that the Lagrangian dual of block-structured smooth nonconvex problems—with potentially additional sparsity constraints—is asymptotically tight under mild assumptions, which contracts non-vanishing duality gap obtained via classical Shapley-Folkman Lemma.

FC07

Congress Room

Advances in MIP and MINLP

Invited Session

Discrete Optimization

Chair: Robert Hildebrand, Virginia Tech, Blacksburg, VA, United States

1 - Scaled Simple Disjunctive Cuts

Aleksandr Kazachkov, University of Florida, Gainesville, FL, United States, Chen Chen, Connor Johnston, Diego Moran

We introduce a new perspective on computing facet-defining inequalities for the convex hull of a union of polyhedra, where each term in this disjunction is defined by a single linear inequality imposed on a (shared) simple cone. The classic approach to generating these inequalities is a linear program from a reverse polar formulation. We avoid such optimization and instead provide an outline of a tree-based enumeration algorithm through a geometric interpretation of scaling the disjunctive cut principle, which computes inequalities with a simple formula. We provide the first example in the literature of a disjunction whose closed convex hull has an exponential number of facets. We prove a linear upper bound when the number of disjunctive terms is two and show that the reverse polar formulation, even when there are only a linear number of facets, may have an exponential number of basic feasible solutions.

2 - Complexity in Reverse Convex Integer Programming

Robert Hildebrand, Virginia Tech, Blacksburg, VA, United States

We study the complexity of identifying the integer feasibility of reverse convex sets. We present various settings where the complexity can be either NP-Hard or efficiently solvable when the dimension is fixed. Of particular interest is the case of bounded reverse convex constraints with a polyhedral domain.
We introduce a structure, a Boundary Hyperplane Cover, that permits this problem to be solved in polynomial

time in the number of removed sets, hyperplanes, and encoding size provided the dimension is fixed.

3 - Parametric Polyhedra in Mixed-Integer Programming

Diego Moran, Rensselaer Polytechnic Institute, Troy, NY, United States, Gustavo Angulo, Silvia Di Gregorio

We present some old and new results on arbitrary families of parametric polyhedra. First, if the constraint matrix is fixed, in the literature there are structural results for the integer hull and the finiteness of cutting plane closures for varying r.h.s. For instance, recently, Becu et al. proved in "Approximating the Gomory Mixed-Integer Cut Closure Using Historical Data" that the GMI closure of this family is finitely generated, in the sense that there exists a finite list of aggregation weights defining the GMI cuts that give the GMI closure for any polyhedra in the family. We extend this result for other cutting plane closures. Second, if the family of parametric polyhedra is arbitrary but all polyhedra in the family have the same integer hull, they define the same MIP, and we can leverage this information to understand and solve MIPs better. These families have been used to understand theoretical properties of the rank of cutting planes and to obtain better formulations. We present an application of these same-integer-hull families to formulations for the Asymmetric Traveling Salesman Problem.

4 - Cutting Planes for Binarized Integer Programs

Anton Derkach, Cornell University, New York, NY, United States

We consider integer programming problems with bounded general-integer variables belonging to the general class of network flow problems. For those, we computationally investigate the effect on mixed-integer linear programming (MIP) solvers of the different ways of producing extended formulations that replace a bounded general integer variable by a linear combination of a set of auxiliary binary variables linked by additional linear constraints. We show that MILP solvers perform very differently depending on which extended formulations is used and we interpret that different performance through the lens of cutting planes generation. Finally, we discuss a simple family of mixed-integer rounding inequalities that especially benefit from the reformulation, and we show its benefit within different MIP solvers. This provides methodological and practical guidelines for the use of those extended formulations in MIP and, to the best of our knowledge, this is the first extensive computational analysis of the topic.

FC08

Committee Room

From Science to Sustainability: Large-scale Optimization in Action

Invited Session

Emerging Applications of Optimization

Chair: Ryan Cory-Wright, Imperial, London, N/A, United Kingdom

1 - AI Noether -- Bridging the Gap Between Scientific Laws Derived by AI Systems and Canonical Knowledge via Abductive Inference

Ryan Cory-Wright, Imperial, London, United Kingdom, Karan Srivastava, Sanjeeb Dash, Barry Trager, Cristina Cornelio, Lior Horesh

A core goal in modern science is to harness recent advances in AI and computer processing to automate and accelerate the scientific method. Symbolic regression can fit interpretable models to data, but these models often sit outside established theory. Recent systems (e.g., AI Descartes, AI Hilbert) enforce derivability from prior axioms. However, sometimes new data and associated hypotheses derived from data are not consistent with existing theory because the existing theory is incomplete or incorrect. Automating abductive inference to close this gap remains open. We propose a solution: an algebraic geometry-based system that, given an incomplete axiom system and a hypothesis that it cannot explain, automatically generates a minimal set of missing axioms that suffices to derive the axiom, as long as axioms and hypotheses are expressible as polynomial equations. We formally establish necessary and sufficient conditions for the successful retrieval of such axioms. We illustrate the efficacy of our approach by demonstrating its ability to explain Kepler's third law and a few other laws, even when key axioms are absent.

2 - Predictive and Prescriptive Analytics toward Optimizing Wildfire Suppression

Alexandre Jacquillat, MIT Sloan School of Management, Cambridge, MA, United States, Leonard Boussioux, Ryne Reger, Jacob Wachspress

Intense wildfire seasons require critical prioritization decisions to suppress wildfires over a disperse geographic area with limited resources. This paper develops a predictive and prescriptive approach to jointly optimize crew assignments and wildfire suppression. The problem features a combinatorial resource allocation structure with endogenous wildfire demand and non-linear wildfire dynamics. We formulate an integer optimization model in a time-space-rest network representation of crew assignments and a time-state network representation of wildfire dynamics, with linking constraints to ensure consistency between both. We develop a branch-and-price-and-cut algorithm based on: (i) a two-sided column generation scheme that generates fire suppression plans and crew routes iteratively; (ii) a new family of cuts exploiting the knapsack structure of the linking constraints; and (iii) novel branching rules to accommodate non-linear wildfire dynamics. We also develop a data-driven approach based on double machine learning and causal forests to estimate wildfire spread as a function of covariate information and suppression efforts, while controlling for the endogeneity of treatment and outcome variables. Extensive computational experiments show that the optimization algorithm scales to practical and otherwise intractable instances; and that the methodology can enhance suppression effectiveness as compared to several practical baselines.

3 - Optimizing Advanced Air Mobility Operations in a Corridor Network

Felipe Cordera, Massachusetts Institute of Technology, Cambridge, MA, United States, Alexandre Jacquillat, Spencer McDonald

Electric vertical takeoff and landing (eVTOL) vehicles are driving the emergence of Advanced Air Mobility (AAM). AAM systems are expected to rely on corridor networks, where en-route separation and directional flow management are essential to ensure safe and efficient operations. This paper develops a tractable optimization framework for AAM operations that jointly determines vehicle dispatching and routing, four-dimensional flight trajectories, and flow directionality in capacitated corridors. We formulate an integer optimization model in a time–space–lane network that exploits a subpath structure at the flight level. We solve it via column generation to decompose vehicle dispatching and corridor flow decisions in a master problem and flight trajectories in a pricing problem, using a tailored backward label-setting algorithm. Our method scales to realistic instances with up to 50 vertiports, 2,000 corridor conjunctions, and hundreds of trip requests. Results demonstrate that the integrated optimization methodology can provide significant benefits as compared to benchmarks, resulting in higher operating profits and larger demand coverage.

4 - Optimal Batching and In-Building Delivery Routing with Capacitated Residential Parcel Lockers

Ignacio Erazo, Amazon Fulfillment Technologies, Bellevue, WA, United States, Dipayan Banerjee

We optimize the operations of a parcel delivery firm that has access to a residential parcel locker (RPL) in the lobby of a high-rise building. We assume that the number of parcels to be delivered on a particular day exceeds the number of available RPL compartments; parcels that cannot be delivered to the RPL must be delivered via elevator and walking to residents' doorsteps on upper floors of the building. With the objective of minimizing expected makespan, we seek to jointly determine the optimal subset of parcels to deliver to the RPL, determine the driver's travel routes within the building, and partition the non-RPL packages into batches. Under mild assumptions on the building's layout and elevator behavior, we first derive a polynomial-time algorithm for the subproblem of determining the optimal sequence of floor visits given a set of parcels to deliver at residents' doorsteps. We then leverage this algorithm to develop a column generation approach capable of solving realistically sized instances of the full problem to optimality. We also propose straightforward heuristics for the full problem and derive tight approximation ratios. Computational studies provide evidence that both exact and heuristic optimization, when used correctly, provide significant benefits in this setting.

FC09

Cabinet Room

Linear and Conic Optimization III

Invited Session

Linear and Conic Optimization

Chair: Joachim Dahl, Cardinal Operations, Ballerup, Denmark

1 - On Primal-dual Scalings for Exponential Cone Programming

Joachim Dahl, Cardinal Operations, Ballerup, Denmark

Recently exponential cone programming has received increased attention, and has been implemented in both commercial optimization packages and open-source projects. One of proposed algorithms is based on primal-dual scalings developed by Tuncel. In this talk we discuss such scalings and present numerical results to illustrate the impact of different approaches.

2 - Complexity of Chordal Conversion for Sparse Semidefinite Programs with Small Treewidth

Richard Zhang, University of Illinois Urbana-Champaign, Urbana, IL, United States

If a sparse semidefinite program (SDP), specified over n by n matrices and subject to m linear constraints, has an aggregate sparsity graph G with small treewidth, then chordal conversion will sometimes allow an interior-point method to solve the SDP in just O(m+n) time per-iteration, which is a significant speedup over the n^3 time per-iteration for a direct application of the interior-point method. Unfortunately, the speedup is not guaranteed by an O(1) treewidth in G that is independent of m and n, as a diagonal SDP would have treewidth zero but can still necessitate up to n^3 time per-iteration. Instead, we construct an extended aggregate sparsity graph \bar{G} by forcing each constraint matrix A_{i} to be its own clique in G. We prove that a small treewidth in \bar{G} does indeed guarantee that chordal conversion will solve the SDP in O(m+n) time per-iteration, to \epsilon-accuracy in at most O(\sqrt{m+n}\log(1/\epsilon)) iterations. This sufficient condition covers many successful applications of chordal conversion, including the MAX-k-CUT relaxation, the Lovász theta problem, sensor network localization, polynomial optimization, and the AC optimal power flow relaxation, thus allowing theory to match practical experience.

3 - Generalized Cuts and Grothendieck Covers: Algorithmic Performance

Nathan Benedetto Proença, University of Waterloo, Waterloo, ON, Canada, Marcel K. de Carli Silva, Cristiane M. Sato, Levent Tuncel

We have exploited the Antiblocking Duality (as introduced by Fulkerson
in the 1970s) between the maximum cut and (fractional) cut-covering
problems to develop approximation algorithms.
We strengthened and extended this theory as well as the underlying
algorithms to a wide class of generalized cuts and generalized
fractional covers.
This wide class includes both maximizing a convex quadratic form over
the hypercube, as well as all Boolean 2-Constraint Satisfaction
problems.
These algorithms all obtain fractional covers by repeated sampling
from a distribution obtained via semidefinite programming (SDP),
following the framework of the seminal work of Goemans and Williamson.
We explore the practical performance of this family of algorithms,

experimentally studying the behavior of the approximation ratio with respect to the samples taken and the quality of the SDP solution, as well as other aspects of theoretical and practical interest.

4 - ***Cancelled*A Warmstarting Technique for General Conic Optimization in Interior Point Methods**

Yuwen Chen, EPFL, Lausanne, Switzerland, Paul Goulart, Colin Jones

We propose a novel warm-starting method for primal-dual interior point methods that is applicable to general conic optimization problems. We prove that the proposed approach maintains small feasibility residuals and duality gaps that are proportional to the perturbation on problems, while generating an initial point that remains close to the central path. We will also show the computation of the warm start is efficient. Numerical experiments demonstrate that the proposed strategy effectively reduces the number of iterations and overall computational time across a variety of test problems.

FC10

Caucus Room

Computational Optimization at Google

Invited Session

Computational Optimization

Chair: Juan Pablo Vielma, Google, Cambridge, MA, United States

1 - Accelerating Large-scale Network Capacity Planning via Set-cover-enhanced Benders Decomposition.

Eduardo Moreno, Google Research, Paris, France, Babak Moazzez

Modern network capacity planning for large-scale infrastructures, such as Google's network, relies on complex, large-scale Mixed-Integer Programming (MIP) models to determine cost-effective capacity allocations while guaranteeing Quality of Service (QoS). Existing methods often struggle with long solution times due to the size and complexity of these models.

We propose a Benders decomposition framework to significantly enhance the performance of large-scale capacity planning solvers. The core innovation is the dynamic generation of Benders cuts, which replace cumbersome flow-conservation constraints and efficiently enforce feasibility within the capacity allocation master problem.

Furthermore, we show that the feasibility subproblem for cut generation is equivalent to a Set Cover problem. By exploiting this structure, we apply advanced Set Cover heuristics and a novel hierarchical clustering technique to generate multiple violated cuts in a single iteration, dramatically improving the efficiency of the decomposition. This integration provides a new family of valid inequalities that more effectively guide the optimal search. Our preliminary results validate that this approach significantly reduces computational time compared to existing methods, enabling rapid, high-quality network design decisions

2 - Duality and Decomposition in Google's OR-Tools

Juan Pablo Vielma, Google, Cambridge, MA, United States

OR-Tools MathOpt is a software library for algebraic modeling of mathematical optimization problems. MathOpt supports solver-independent modeling of a wide range of optimization problems including continuous and mixed-integer problems with linear or quadratic objectives and various classes of constraints (including linear, quadratic, second order cone and some specialized constraints). In this talk, we describe various aspects of MathOpt's design and implementation with a particular focus on duality and decomposition methods.

3 - Practical Performance Guarantees for Pipelined DNN Inference

Matthew Fahrbach, Google, Atlanta, GA, United States, Aaron Archer, Kuikui Liu, Prakash Prabhu

We optimize pipeline parallelism for deep neural network (DNN) inference by partitioning model graphs into k stages and minimizing the running time of the bottleneck stage, including communication. We give practical and effective algorithms for this NP-hard problem, but our emphasis is on tackling the practitioner's dilemma of deciding when a solution is good enough. To this end, we design novel mixed-integer programming (MIP) relaxations for proving lower bounds. Applying these methods to a diverse testbed of 369 production models, for $k \in \{2,4,8,16,32,64\}$, we empirically show that these lower bounds are strong enough to be useful in practice. Our lower bounds are substantially stronger than standard combinatorial bounds. For example, evaluated via geometric means across a production testbed with k=16 pipeline stages, our MIP formulations raise the lower bound from 0.4598 to 0.9452, expressed as a fraction of the best partition found. In other words, our improved lower bounds close the optimality gap by a factor of 9.855x.

4 - OR-Tools CP-SAT Solver: Achieving State-of-the-art Performance with a Portfolio Solver

Vitor Sessak, Google, Paris, France, Eric Bruneton, Toby Davies, Frederic Didier, Laurent Perron

> This presentation will start with an overview of the Google OR-Tools' **CP-SAT** discrete optimization solver. Classical discrete optimization approaches, such as **Mixed Integer Programming (MIP), SAT formulations, or Constraint Programming (CP)**, are often presented as competing techniques from which a practitioner must choose depending on the specific problem solved. CP-SAT achieves state-of-the-art performance by combining these diverse methodologies into a **hybrid solver**. Crucially, we combine the different approaches so they reinforce each other, resulting in a stronger solver than any single approach alone. In the second part of the talk, I will focus on the latest developments and new heuristics. Finally, I will discuss, from a practical point of view, how a hybrid solver architecture can help or hinder the integration of new heuristics.

FC11

Charter Room

Advances in Power Systems Problems with Discrete Decisions

Invited Session

Network Optimization

Chair: Behnam Jabbari Marand, North Carolina State University, Raleigh, NC, United States

1 - Angle-Based Valid Inequalities for Switchable DC Transmission Networks

Behnam Jabbari Marand, North Carolina State University, Raleigh, NC, United States, Adolfo Escobedo

Switching transmission lines on and off can improve the economic and operational performance of power systems. The Direct-Current Optimal Transmission Switching (DC-OTS) problem provides a formal framework for minimizing generation costs by reconfiguring network topology under a linearized power-flow model. DC-OTS is typically formulated as a mixed-integer linear program with disjunctive constraints encoded via big-M parameters that bound voltage angle differences across non-operational lines. In practice, these bounds are often chosen in an overly conservative and ad hoc manner, under the prevailing belief that computing tight values requires solving the computationally intractable longest path problem. This work overcomes this limitation by developing a novel polyhedral analysis of an angle-based DC-OTS formulation. We construct extended formulations whose projections yield the convex hulls of angle-based relaxations and, from these, derive two families of facet-defining inequalities that tighten angle-difference bounds. We then compare one of these inequalities against a variant of the cycle-based inequalities previously introduced for DC-OTS. Computational experiments on standard benchmark instances show that incorporating these inequalities within a cutting-plane framework substantially improves solver performance, demonstrating their practical effectiveness for solving large-scale DC-OTS problems.

2 - Grid-ECO: Grid-Aware Electric Vehicle Charging Stations Placement Optimizer

Amritanshu Pandey, University of Vermont, Burlington, VT, United States, Bikram Panthee

The rapid adoption of electric vehicles (EVs) requires expanding public EV charging stations, but placing these stations into the existing distribution grid can result in transformer overloading and excessive voltage drop, necessitating expensive upgrades. In this work, we identify the optimal size and placement of EV charging stations (EVCS) by maximizing the number of installed chargers while satisfying grid physics constraints, subject to the available budget, EV charger demand, and charging station spatial distribution. Due to nonlinear AC network constraints and integer decision variables, the underlying optimization problem is a **NP-hard** Mixed-Integer Nonlinear Program (MINLP). Prior works have not addressed the difficulty of MINLP in solving this optimization problem. Standard approaches include MILP relaxation by avoiding or linearizing AC network constraints or nonlinear programming relaxation by preselecting integer decisions. We solve the original MINLP problem by including the three-phase AC network model in the current-voltage (IV) formulation. Next, we reformulate the MINLP problem into an equivalent Mixed-Integer Bilinear Program (MIBLP) that the Gurobi solver has shown to solve, with spatial branch-n-bound to near 0% optimality gap if good initial conditions and bounds on the bilinear state variables are available. We relax MIBLP using McCormick relaxation to obtain good initial conditions. We use a sequential bound tightening approach to generate tight cuts for the original problem. We are the first to demonstrate that we can achieve a near-0 % optimality gap for the underlying MIBLP problem in real-world distribution networks.

3 - Second-Order-Cone Formulations of Power Flow for Topology Optimization

Noah Rhodes, Los Alamos National Laboratory, Los Alamos, NM, United States, Jim Luedkte, Line Roald

Optimization problems that involve topology optimization in scenarios with large scale outages, such as post-disaster restoration or public safety power shutoff planning, are very challenging to solve. Using simple power flow representations such as DC power flow or network flow models results in low quality solutions which requires significantly higher-than-predicted load shed to become AC feasible. Recent work has shown that formulations based on the Second Order Cone (SOC) power flow formulation find very high-quality solutions with low load shed, but the computational burden of these formulations remains a significant challenge. With the aim of reducing computational time while maintaining high solution quality, this work explores formulations which replace the conic constraints with a small number of linear cuts. The goal of this approach is not to find an exact power flow solution, but rather to identify good binary decisions, where the power flow can be resolved after the binary variables are fixed. We find that a simple reformulation of the Second Order Cone Optimal Power Shutoff problem can greatly improve the solution speed, but that a full linearization of the SOC voltage cone equation results in an overestimation of the amount of power that can be delivered to loads.

4 - A Hybrid Decomposition Approach for Stochastic Unit Commitment with Combined-Cycle Generators

Rosemary Barrass, AI Institute for Advances in Optimization, Atlanta, GA, United States, Harsha Nagarajan, Mathieu Tanneau, Russell Bent, Pascal Van Hentenryck

The U.S. power grid is undergoing a major paradigm shift with the increased development of renewable generators, electric vehicles, and data centers. In response to this growing need, the U.S. has ramped up the construction of combined-cycle generators (CCs). CCs are fast-ramping generators that utilize variable configurations of combustion turbines (CTs) and steam turbines (STs) to achieve much higher efficiency than traditional CTs alone. Schedulers must select a valid configuration of the CTs and STs for each time step of a given study horizon. For schedule optimization, this requires the addition of a large number of binary constraints and variables in Unit Commitment (UC) problem formulations. This change of paradigm makes it difficult to commit generators efficiently, especially when considering the system's growing stochasticity. Algorithms for this class of UC problems are almost non-existent due to their computational complexity. We present a novel hybrid Benders' (BD) and Dantzig-Wolfe (DW) decomposition algorithm for stochastic UC problems with CCs. The algorithm exploits the separability of the linear constraints and the integer CC constraints through BD and DW, respectively. Results are presented for varying load instances of the 935-generator FERC test data set, modified to include mode data for CCs. The algorithm demonstrates a significant speed-up in solving the outer problem over traditional BD across all cases. It also finds higher lower bounds on cases with 30 or more scenarios earlier than both BD and Gurobi's branch-and-bound solver. This demonstrates that the proposed algorithm is a scalable approach for solving stochastic UC with CCs.

March 20, 2026, 4:30 PM - 6:00 PM

FD01

Grand Ballroom

Accelerated Methods and Sharp Analysis for Nonlinear Optimization

Invited Session

Optimization in Data Science

Chair: Yao Ji, Georgia Tech, Atlanta, United States

1 - Further Analysis on the Accelerated Gradient Method

Yuyuan Ouyang, Clemson University, Clemson, SC, United States, Yan Wu, Lu Liu, Yipeng Zhang

A substantial body of research has examined the fundamental ideas underlying the optimal complexity of the accelerated gradient method. In the original formulation, the method generates two distinct sequences of iterates: gradient evaluations are performed only on one sequence, while the approximate solution is only selected from the other. Subsequent developments reveal that in unconstrained problems the sequence for gradient evaluations could itself serve as approximate solutions, while still maintaining the optimal order of iteration complexity. However, such analysis does not extend to projected-based accelerated gradient descent method for problems with projection-friendly feasible sets, leaving open the research question of whether the gradient evaluation sequence could also achieve the same optimal order of iteration complexity. Motivated by computer-aided algorithm analysis, we provide a positive result that answers this open problem affirmatively.

2 - On Tackling High-Dimensional Nonconvex Stochastic Optimization via Stochastic First-Order Methods with Non-smooth Proximal Terms and Variance Reduction

Hongcheng Liu, University of Florida, Gainesville, FL, United States

In nonconvex stochastic programming, the sample complexity of stochastic first-order methods may depend linearly on the problem dimension, which is undesirable for large-scale problems. To alleviate this linear dependence, we adopt non-Euclidean settings and propose two choices of non-smooth proximal terms when taking the stochastic gradient steps. This approach leads to stronger convergence metric, incremental computational overhead, and potentially dimension-insensitive sample complexity. We also consider further acceleration through variance reduction which achieves near optimal sample complexity and, to our best knowledge, is the first such result in the $\ell 1/\ell \infty$ setting. Since the use of non-smooth proximal terms is unconventional, the convergence analysis deviates much from algorithms in Euclidean settings or employing Bregman divergence, providing tools for analyzing other non-Euclidean choices of distance functions. Efficient resolution of the subproblems in various scenarios is also discussed and simulated. We illustrate the dimension-insensitive property of the proposed methods via preliminary numerical experiments.

3 - Accelerated Prox-level Method for Unknown Piecewise Smooth Problems

Zhenwei Lin, Purdue University, West Lafayette, IN, United States

In this work, we address the challenge of designing an algorithm that achieves the optimal convergence rate for optimizing unknown piecewise smooth (PWS) functions. PWS optimization refers to a special case of nonsmooth optimization in which the domain is partitioned into multiple regions, and the objective function is smooth within each region. The "unknown" setting means we have no prior knowledge of the locations or structure of these smooth regions; we treat the problem entirely as a black box.

Our approach builds upon the bundle level method. We rigorously show that when the number of cutting planes (cuts) exceeds the number of smooth pieces in the PWS function, the algorithm attains a convergence rate of $k\sqrt{L/\mu^*}\log(1/\varepsilon)$, where k is the number of pieces, L is the Lipschitz constant, μ^* is the quadratic growth (QG) modulus, and ε is the target accuracy. Furthermore, we establish that this rate matches the corresponding lower complexity bound, thereby proving the optimality of our method. Notably, our algorithm is nearly parameter-free: apart from setting the number of cuts greater than the number of pieces, no additional problem-specific information is required.

4 - Nonlinear Tomographic Reconstruction via Nonsmooth Optimization

Vasilis Charisopoulos, University of Washington, Seattle, WA, United States

Computed tomography (CT) is an imaging modality with widespread use in biomedical imaging, non-destructive materials testing, archaeology, etc. Standard approaches for inverting CT measurements, which rely on a numerically poorly conditioned preprocessing step, often struggle to reconstruct signals with high dynamic range (e.g., X-ray imaging of tissue with embedded metal). In this talk, I will present an iterative method based on nonsmooth optimization that sidesteps the aforementioned preprocessing step to obtain faster and more accurate reconstructions, and enjoys a linear rate of convergence under certain statistical assumptions.

Joint work with R. Willett.

5 - On the Global Optimality of Algorithms for Solving Nonconvex Optimization

Ying Cui, University of California, Berkeley, Berkeley, CA, United States

Convexification is a classical technique in global optimization, but standard convex relaxations are typically one-shot: they provide a single convex surrogate that cannot be systematically tightened. As a result, convexification plays a limited role in nonlinear programming, where progressively improving approximations are essential in the latter area. In this talk, we introduce a parametric relaxation framework for a broad class of nonconvex optimization problems. The relaxation is governed by a parameter \gamma that explicitly trades off between the convexity of the surrogate model and its tightness to the original objective. A key feature of this framework is its hierarchical approximation property: as \gamma increases, the relaxed problem first touches the set of global minimizers, then all local minimizers, and eventually all stationary points of the original problem. We show that this metric provides a principled way to assess the tightness of stationary points and to guide algorithmic strategies for approaching globally optimal solutions. Algorithms and convergence rates to sequentially solve the \gamma-relaxation will also be discussed.

FD02

Georgian Room

Optimization in the Era of Large AI Models: from Pre-training to Post-training.

Invited Session

Optimization in Data Science

Chair: Wenlong Mou, University of Toronto, Toronto, ON, Canada

1 - Stochastic Approaches to Conditional Diffusion Guidance and Applications

Wenpin Tang, Columbia University, New York, NY, United States

Recently, there has been growing interest in guiding, or fine tuning pretrained diffusionmodels for specific purposes, e.g., aesthetic quality of images, functional property of proteins, and downstream tasks in finance and operations management. In this talk, I will present a novel approach to conditional diffusion guidance in the context of classifier guidance. The approach is probability-theoretic, relying on various techniques such as martingale, quadratic variations, etc. I will also discuss some applications including the generation of synthetic queueing and financial data.

2 - Preconditioning Benefits of Spectral Orthogonalization in Muon

Yuxin Chen, University of Pennsylvania, Philadelphia, PA, United States

The Muon optimizer, a matrix-structured algorithm that leverages spectral orthogonalization of gradients, is a milestone in the pretraining of large language models. However, the underlying mechanisms of Muon—particularly the role of gradient orthogonalization—remain poorly understood, with very few works providing end-to-end analyses that rigorously explain its advantages in concrete applications. We take a step by studying the effectiveness of a simplified variant of Muon through two case studies: matrix factorization, and in-context training of linear transformers. For both problems, we prove that Muon converges linearly with iteration complexities independent of the relevant condition number, provably outperforming gradient descent and Adam. Our analysis reveals that the Muon dynamics decouple into a collection of independent scalar sequences in the spectral domain, each exhibiting similar convergence behavior. Our theory formalizes the preconditioning effect induced by spectral orthogonalization, offering insight into Muon's effectiveness in these matrix optimization problems and potentially beyond.

3 - Stepwise Guided Policy Optimization: Coloring your Incorrect Reasoning in GRPO

Xiaopeng Li, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, People's Republic of, Peter Chen, Ziniu Li, Xi Chen, Tianyi Lin

Reinforcement learning (RL) has proven effective in strengthening the reasoning capabilities of large language models (LLMs). A widely adopted method, Group Relative Policy Optimization (GRPO), has shown strong empirical results in training DeepSeek-R1. However, GRPO fails to update the policy when all responses within a group are incorrect (i.e., *all-negative-sample* groups). This limitation underscores a key gap between artificial and human intelligence: unlike humans, who can learn from mistakes, GRPO discards these signals. Our first contribution is to introduce a simple framework that mitigates the all-negative-sample issue by incorporating response diversity within groups using a *step-wise* judge model, which can be either directly trained or adapted from existing LLMs. We prove that this diversification can accelerate GRPO's learning dynamics in a simplified setting. We also empirically validate the proposed stepwise guided policy optimization (SGPO) method, demonstrating consistent gains across model sizes (7B, 14B, 32B) in offline and online training on 9 benchmarks, including base and distilled variants. Our results highlight two advantages: (i) SGPO surpasses GRPO, especially in the early and mid-training stages where all-negative-sample groups are prevalent; and (ii) SGPO does not require judge models to generate correct answers, differentiating it from knowledge distillation methods.

4 - Learning Quadratic Neural Networks in High-dimensions

Murat Erdogdu, University of Toronto, Toronto, ON, Canada

We study the optimization and sample complexity of gradient based training of a two-layer student neural network with quadratic activation function in the high-dimensional regime, where the input is Gaussian and the response is generated from a two-layer teacher network with quadratic activation, and the power law decay on the second layer coefficients. We provide a sharp analysis of the SGD dynamics in the feature learning regime, and derive scaling laws for the prediction risk that highlight the power-law dependences on the optimization time, sample size, and model width.

FD03

Plaza I

Advanced Methods for Large-scale Nonlinear and Stochastic Optimization III

Invited Session

Nonlinear Optimization

Chair: Qi Wang, University of Michigan, Ann Arbor, MI, United States

1 - Projected Stochastic Diagonal-Scaling Methods for Nonlinear Equality-Constrained Optimization

Yunlang Zhu, Lehigh University, Bethlehem, PA, United States, Qi Wang, Christian Piermarini, Frank E. Curtis

It is general practice to apply a regularization-based approach in supervised machine learning to solve continuous optimization problems with equality constraints. However, this approach requires extra effort to design the loss function, and there is no guarantee the constraints will ultimately be satisfied in practice. Another approach is to impose hard constraints directly. We propose two algorithms for solving continuous optimization problems with equality constraints. Each is an extension of a stochastic diagonal-scaling method from the unconstrained setting to the setting of a stochastic Newton-SQP-type algorithm for solving equality-constrained problems. One is an extension of the heavy-ball method and the other is an extension of the Adam optimization method. A critical feature of each method is that the momentum terms are implemented with projected gradient estimates, rather than the gradient estimates themselves. We provide convergence guarantees comparable to their unconstrained counterparts for both methods. Our experiments on solving informed supervised machine learning problems show benefits of employing hard constraints rather than the regularization-based approach.

2 - ProxSTORM: A Stochastic Trust Region Algorithm for Nonsmooth Optimization

Aurya Javeed, Sandia National Laboratories, Albuquerque, NM, United States, Robert Baraldi, Drew Kouri, Katya Scheinberg

This talk is about minimizing a smooth term plus a convex nonsmooth term. We present a stochastic proximal Newton trust region algorithm that assumes models and estimates of the objective are sufficiently accurate, sufficiently often. Like STORM (work on stochastic optimization with random models), we use facts about martingales to prove our algorithm is globally convergent with probability one.

3 - A Stochastic LineseArch Method (SLAM) for Minimizing Expectation Residuals

Qi Wang, University of Michigan, Ann Arbor, MI, United States, Yue Xie, Uday Shanbhag

Most existing rate and complexity guarantees for stochastic gradient methods in L-smooth settings impose restrictive requirements on steplength sequences, mandating that such sequences be non-adaptive, non-increasing, and abide by an upper bound given by 1/L. This severely limits the applicability and adversely affects the performance of such schemes for three key reasons: (i) requires knowing L; (ii) precludes larger steps when L is large; and (iii) employs non-increasing steplength sequences. Motivated by these shortcomings, we present an Armijo-enabled stochastic linesearch framework with standard stochastic zeroth- and first-order oracles. The resulting steplength sequence is non-monotonic. We provide the iteration and sample complexity of the proposed method and the sample complexity matches the standard SGD method. Preliminary numerical experiments are seen promising as the scheme is observed to compete well with de facto choices such as SGD and Adam across a range of problem types and instances.

4 - Adaptive Sampling Methods for Expectation-Constrained Stochastic Optimization

Yash Kumar, The University of Texas at Austin, Austin, TX, United States, Raghu Bollapragada

We study constrained stochastic optimization problems in which both the objective and the constraints involve expectations that are expensive to evaluate exactly. Traditional Augmented Lagrangian methods can be inefficient in this setting due to the high cost of sampling and the noise introduced by stochastic evaluations. In this talk, we present an adaptive sampling algorithm that dynamically adjusts the sampling effort across iterations of an Augmented Lagrangian framework. We establish optimal theoretical convergence and complexity guarantees under mild assumptions. Further, we demonstrate empirically that adaptive sampling improves both robustness and efficiency compared to fixed-sample strategies, using benchmark stochastic constrained optimization problems.

FD04

Plaza II

Recent Advances in First-Order Methods and ADMM

Invited Session

Nonlinear Optimization

Chair: Leandro F. Maia, Oregon State University, Corvallis, OR, United States

1 - A Parallel Proximal ADMM Framework for Nonconvex Composite Problems

Leandro F. Maia, Oregon State University, Corvallis, OR, United States, Renato D C Monteiro

In this talk, I will present a *parallel adaptive proximal ADMM* algorithm for solving a linearly constrained, non-smooth, and nonconvex composite optimization problems. The proposed method is adaptive to all problem parameters, and allows each of its block proximal subproblems to be inexactly solved. A novel adaptive test is proposed to decide whether to perform a full Lagrange multiplier and/or penalty parameter update(s). Without any rank assumptions on the constraint matrices, it is shown that our algorithm obtains an approximate first-order stationary point of problem in a number of iterations that matches the state-of-the-art complexity for the class of proximal ADMM type methods. If time allows, I will present proof-of-concept numerical experiments with the benefits of the proposed framework.

2 - Projection-free Algorithms for Minimax Problems

Soroosh Shafiee, Cornell University, Ithaca, NY, United States

While projection-based and projection-free methods are well-studied in standard convex minimization, their counterparts for saddle-point optimization remain underexplored. This work addresses this gap by studying constrained smooth saddle-point problems. We propose and analyze single-loop algorithms that employ a linear minimization oracle (LMO) over one or both feasible sets, providing projection-free alternatives to classical approaches.

3 - Solving Variational Inequality under Generalized Lipschitz Conditions

Digvijay Boob, Southern Methodist University, Dallas, TX, United States, Qi Deng

We present efficient methods for solving monotone variational inequality problems when the operator is not uniformly Lipschitz continuous, and the feasible set may be unbounded. In recent years, optimization problems under generalized Lipschitz conditions have gained traction. However, there is limited literature on generalized conditions for VI problems. We introduce a new type of generalization of Lipschitz conditions for the VI problem that includes convex function-constrained optimization and variational inequality problems, average reward MDPs, and general robust optimization with semi-infinite (or robust) function constraints. In our problem setup, a uniform Lipschitz smoothness constant may not hold, especially since we allow the feasible set to be unbounded. Under the assumption that a VI solution exists and after carefully designing the algorithm, we show that our iterates remain bounded and establish state-of-the-art convergence rate guarantees for various cases, including monotone or strongly monotone operators with possibly generalized nonsmoothness. When the operator is stochastic, we show expected convergence guarantees. We also establish that the iterates are bounded in expectation. Our convergence criterion uses a Nesterov-type sufficiently large radius ball that contains the VI solution. However, unlike Nesterov, we do not need to know this radius to implement our methods for the abovementioned cases.

4 - HALLaR: A New Solver with GPU Acceleration for Large-Scale Semidefinite Programming

Arnesh Sujanani, University of Waterloo, Waterloo, ON, Canada, Renato D C Monteiro, Diego Cifuentes

This talk presents a new first-order method for solving large-scale semidefinite programs (SDPs) with bounded domain. It is an inexact augmented Lagrangian (AL) method where the AL subproblems are solved by a hybrid low-rank method. In contrast to the classical low-rank method, the new method finds a near-optimal solution (with provable complexity bounds) of SDP instances satisfying strong duality. Computational results comparing the new method to state-of-the-art solvers on several large SDP instances show that the former finds higher accurate solutions in substantially less CPU time than the latter ones. For example, in less than 20 minutes, our method can solve (on a personal laptop) a maximum stable set SDP with 1 million vertices and 10 million edges within 1e-5 relative accuracy. Our computational results show that our HALLaR achieves 100 to 150 times speedup over its CPU version using the power of GPU programming.

FD05

Plaza III

PDE-Constrained Optimization Under Uncertainty: Models & Algorithms

Invited Session

Optimization under Uncertainty

Chair: Anton Malandii, Brown University, Providence, RI, United States

1 - Gaussian Mixture Taylor Expansion for Risk-averse Optimization under Uncertainty

Wenbo Hao, Georgia Institute of Technology, Atlanta, GA, United States, Peng Chen

I will present a scalable computational framework for solving risk-averse optimization problems governed by large-scale partial differential equations (PDEs) with high-dimensional random parameters. Such optimization under uncertainty (OUU) problems can be computationally prohibitive using classical methods, since evaluating risk measures of quantities of interest (QoIs) associated with the PDE solution often requires a large number of PDE solves. This challenge is exacerbated by the nonlinear dependence of the QoI on uncertain parameters and the high dimensionality of the parameter space, which make direct sampling-based uncertainty quantification prohibitively expensive.
To address this challenge, we propose a Gaussian Mixture Taylor Expansion framework for efficient evaluation of QoI statistics in PDE-constrained OUU problems. The method combines Gaussian mixture approximations of the parameter distribution with local Taylor expansions of the parameter-to-QoI map. While Taylor expansions around parameter means significantly reduce the number of required PDE solves, they are accurate only in a local neighborhood. By decomposing the global parameter distribution into a set of locally concentrated Gaussian components and performing componentwise Taylor expansions, the proposed framework achieves accurate global approximations while maintaining low

computational cost.
We demonstrate the effectiveness of the proposed approach through numerical experiments involving large-scale PDE-constrained OUU problems. The results show that the Gaussian Mixture Taylor Expansion framework enables efficient and scalable estimation of risk measures while requiring significantly fewer PDE solves than conventional sampling-based approaches

2 - A Projected Mirror Descent Method for Density-based Multi-material Topology Optimization

Dohyun Kim, Brown University, Providence, RI, United States

In this talk, we introduce a novel, first-order method for solving density-based multi-material topology optimization problems. A fundamental challenge in multi-material design is ensuring that the material indicator field satisfies complex local constraints—specifically, that material fractions are non-negative and sum up to one. For arbitrary material sets, these constraints define a high-dimensional convex polytope, making traditional projection-based algorithms computationally cumbersome and ill-suited for general interpolation domains.Our proposed method is a major extension of the single-material SiMPL algorithm. The key innovation is the generalization of the Bregman divergence -- a concept of generalized distance -- to implicitly enforce the constraints of any arbitrary polytope. This is achieved by deriving a new mirror map that establishes a bijection between the physical material distribution and an unconstrained latent variable. This map allows the optimization process to be performed entirely in the unconstrained latent space, leading to a simple, explicit update rule based only on the objective function's first-order derivative. Because we discretize this unconstrained latent field, and the mirror map guarantees feasibility, the method robustly produces a sequence of pointwise-feasible iterates regardless of the finite element order used, guaranteeing that all material mixtures respect the design space throughout the optimization. We will demonstrate the method's superior performance and robustness across several benchmark problems, showing rapid convergence and high-contrast, manufacturable designs.

3 - Bregman Proximal Point in Function Spaces

Noe Reyes Rivas, Brown University, Providence, RI, United States, Brendan Keith

In this work, we investigate the convergence of the Bregman proximal point method for convex optimization problems in function spaces. In particular, we establish sufficient conditions for the convergence of the method when applied to proper, convex, and lower-semicontinuous functionals defined on Lp and W1,p, highlighting the differences between the two settings. Our analysis leverages properties of Legendre functions and Bregman distances to derive convergence rates and stability results. We also provide illustrative examples demonstrating the applicability of our theoretical findings to practical optimization problems, including the minimization of the Dirichlet energy under general pointwise inequality constraints. The key contribution is a novel cut-off argument, which allows us to handle non-decomposable spaces.

4 - Adaptive Exponential Smoothing and Importance Sampling for CVaR Optimization

Boyan Lazarov, Lawrence Livermore National Laboratory, Livermore, CA, United States, Will Asness , Brendan Keith, Anton Malandii, Stanislav Uryasev

We present a stochastic algorithm for Conditional Value-at-Risk (CVaR) optimization that tightly couples adaptive importance sampling with a Bregman proximal-point scheme. Starting from the dual representation of CVaR as a supremum of expectations over a risk envelope, the proposed algorithm alternates between (i) solving a stochastic primal subproblem and (ii) updating dual probabilities via a proximal step. The evolving dual distribution concentrates sampling mass in the tail that determines CVaR, dramatically improving efficiency at high confidence levels where naïve sampling is wasteful. A key design choice is the introduction of a Legendre-type function based on a generalized Fermi–Dirac entropy, which induces a Bregman divergence respecting the box constraints of the dual feasible set. This yields closed-form, logistic-style updates that keep iterates in the interior and provide a built-in, data-driven importance sampler whose likelihood ratio converges to the risk identifier of the optimal solution. Consequently, the primal stages can be solved with few informative samples while the dual stages steer sampling toward the active tail. We prove convergence in function values for general convex objectives and almost sure convergence of iterates under mild assumptions, and we derive rates for smooth and strongly convex subproblems. Numerical experiments demonstrate strong performance across representative tasks in financial portfolio selection, support vector regression, and risk-averse large-scale topology optimization, where the algorithm outperforms standard stochastic approximation baselines, particularly as the CVaR level approaches one. Overall, the work provides a practical approach to scalable CVaR optimization via entropy-regularized proximal updates and automatically tuned tail sampling.

FD06

Director's and Lounge

Recent Advances in Global Optimization

Invited Session

Global Optimization

Chair: Anna Deza, Georgia Tech, Atlanta, GA, United States

1 - Convexification for the Forbidden Vertices Problem

Jisun Lee, Georgia Tech, Atlanta, GA, United States, Santanu Dey, Oktay Gunluk

We revisit the forbidden vertices problem in the n-dimensional hypercube. Given a collection of n-dimensional binary points X = {x_1,...,x_k}, we study the convex hull of {0,1}^n \ X. Extended formulations of size O(nk) were proposed by Angulo et al. (2015), and Cornuéjols and Lee (2018) introduced a representation in the original variable space for X, where the skeleton G(X) has treewidth at most 2.
In this work, we address the problem in both directions. First, we employ a decomposition method that yields extended formulations of the same asymptotic size O(nk), but which can be computationally more efficient in practice. Second, we analyze when the convex hull can be characterized directly in the original space, focusing on instances where the skeleton G(X) has treewidth 3. For these cases, we provide explicit convex-hull descriptions and identify the combinatorial features of G(X) that govern the representation size.

2 - Lagrangian Dual for Integer Optimization with Zero Duality Gap that Admits Decomposition

Santanu Dey, ISyE Georgia Tech, Atlanta, GA, United States, Jingye Xu, Diego Cifuentes

For mixed integer programs (MIPs) with block structures and coupling constraints, on dualizing the coupling constraints the resulting Lagrangian relaxation becomes decomposable into blocks which allows for the use of parallel computing. However, the resulting Lagrangian dual can have non-zero duality gap due to the inherent non-convexity of MIPs. In this paper, we propose two reformulations of such MIPs by adding redundant constraints, such that the Lagrangian dual obtained by dualizing the coupling constraints and the redundant constraints have zero duality gap while still remaining decomposable. One of these reformulations is similar, although not the same as the RLT hierarchy. In this case, we present multiplicative bounds on the quality of the dual bound at each level of the hierarchy for packing and covering MIPs. We show our results are applicable to general sparse MIPs, where decomposability is revealed via the tree-decomposition of the intersection graph of the constraint matrix. In preliminary experiments, we observe that the proposed Lagrangian duals give better dual bounds than classical Lagrangian dual and Gurobi in equal time, where Gurobi is not exploiting decomposability.

3 - A Parametric Approach for Solving Convex Quadratic Optimization with Indicators over Trees

Aaresh Bhathena, University of Michigan, Ann Arbor, MI, United States, Salar Fattahi, Andres Gomez, Simge Kucukyavuz

This paper investigates convex quadratic optimization problems involving n indicator variables, each associated with a continuous variable, particularly focusing on scenarios where the matrix Q defining the quadratic term is positive definite and its sparsity pattern corresponds to the adjacency matrix of a tree graph. We introduce a graph-based dynamic programming algorithm that solves this problem in time and memory complexity of $O(n^2)$. Central to our algorithm is a precise parametric characterization of the cost function across various nodes of the graph corresponding to distinct variables. Our computational experiments conducted on both synthetic and real-world datasets demonstrate the superior performance of our proposed algorithm compared to existing algorithms and state-of-the-art mixed-integer optimization solvers. An important application of our algorithm is in the real-time inference of Gaussian hidden Markov models from data affected by outlier noise. Using a real on-body accelerometer dataset, we solve instances of this problem with over 30,000 variables in under a minute, and its online variant within milliseconds on a standard computer.

4 - Closing the Gap: Efficient Algorithms for Discrete Wasserstein Barycenters

Weijun Xie, Georgia Institute of Technology, Atlanta, GA, United States, Jiaqi Wang

The Wasserstein barycenter problem seeks a probability measure that minimizes the weighted average of the Wasserstein distances to a given collection of probability measures. We study the discrete setting, where each measure has finite support — a regime that frequently arises in machine learning and operations research. The discrete Wasserstein barycenter problem is known to be NP-hard, which motivates us to study approximation algorithms with provable guarantees. The best-known algorithm to date achieves an approximation ratio of two. We close this gap by developing a polynomial-time approximation scheme (PTAS) for the discrete Wasserstein barycenter problem that generalizes and improves upon the 2-approximation method. In addition, for the special case of equally weighted measures, we obtain a strictly tighter approximation guarantee. Numerical experiments show that the proposed algorithms are computationally efficient and produce near-optimal barycenter solutions.

FD07

Congress Room

Recent Advances in Non-uniform Network Design Models

Invited Session

Discrete Optimization

Chair: Ishan Bansal, Amazon, Lynnwood, WA, United States

1 - Non-uniform Vertex Deletion to Bounded Network Density

Shubhang Kulkarni, Amazon, Redmond, WA, United States

We study a vertex deletion problem motivated by network robustness: remove a minimum-cost set of vertices to reduce network density below a given threshold. This measures the sensitivity of the densest subgraph to non-uniform vertex deletions. We show this problem unifies several classical problems in combinatorial optimization and present a simple bicriteria approximation algorithm.

2 - Almost Tight Additive Guarantees for k-Edge Connectivity

Chaitanya Swamy, University of Waterloo, Waterloo, ON, Canada, Nikhil Kumar

We consider the k-edge connected spanning subgraph (kECSS) problem, where we are given an undirected graph with nonnegative edge costs, and we seek a minimum-cost subgraph that is k-edge connected, i.e., there are k edge-disjoint paths between every pair of nodes.

For even k, we present a polytime algorithm that computes a (k-2)-edge connected subgraph of cost at most the optimal value of the natural LP-relaxation for kECSS; for odd k, this leads to a (k-3)-edge connected subgraph of cost at most the LP optimum. Since kECSS is APX-hard for all k≥2, the tightest connectivity guarantee one can hope for without exceeding the optimal value is (k-1)-edge connectivity, so our results are nearly tight. They also significantly improve upon the recent work of Hershkowitz et al., both in terms of solution quality and the simplicity of algorithm and its analysis. Our result is obtained using iterative rounding, with additional insights involving uncrossing tight sets for kECSS. A slight variant of our approach yields (k-1)-edge connectivity incurring cost at most 1.5 times the LP optimum; with unit edge costs, the cost guarantee improves to (1+4/3k) times the LP optimum, which improves upon the state-of-the-art for unit edge costs, but with a unit loss in edge connectivity. The kECSS-result also yields results for the k-edge connected spanning multigraph (kECSM) problem, where multiple copies of an edge can be selected. Our techniques extend to the degree-bounded versions of kECSS and kECSM.

Joint work with Nikhil Kumar.

3 - Exponentially Improved Approximation Guarantees for Capacitated Network Design

Ishan Bansal, Amazon, Lynnwood, WA, United States

We study the capacitated-k-edge-connected problem where the goal is to find a cheapest subset of edges in a graph with edge costs and capacities such that the minimum cut in the purchased graph is at least k. This is a classical NP-Hard problem introduced in the 1990s to study telecommunication networks. With arbitrary capacities, O(k) was the best known approximation factor prior to our work, while no super constant hardness of approximation is known. We partially bridge this gap by exponentially improving the factor to obtain an O(log k)-approximation algorithm. Our techniques rely on a combination of various tools developed over the years including knapsack cover constraints, primal-dual methods beyond uncrossable families and iterative rounding. Our methodology also provides the first known constant factor approximation for the recently introduced (1,q) -flexible graph connectivity problem.

4 - Approximation Algorithms for Steiner Connectivity Augmentation

Daniel Hathcock, Carnegie Mellon University, Pittsburgh, PA, United States, Michael Zlatin

We consider connectivity augmentation problems in the Steiner setting.
In the Steiner Augmentation of a Graph problem (k-SAG), we are given a k-edge-connected graph H which we seek to augment by including links of minimum cost so that the edge-connectivity between nodes of H increases by 1. Unlike the standard Connectivity Augmentation Problem, links to Steiner nodes outside H are available for the augmentation. If H is not assumed to be globally k-edge-connected but rather Steiner k-edge-connected on some set of terminals R, we obtain the more general Steiner Connectivity Augmentation Problem (k-SCAP).

We give a (1 + ln(2) +\varepsilon)-approximation for the Steiner Ring Augmentation Problem (SRAP). This yields a polynomial time algorithm with approximation ratio (1 + ln(2) + \varepsilon) for 2-SCAP. We obtain an improved approximation guarantee for SRAP when the ring consists of only terminals, yielding a (1.5+ \varepsilon)-approximation for k-SAG for any k. This talk is based on joint work with Michael Zlatin.

FD08

Committee Room

Optimization Advances for Large-Scale Power System Planning

Invited Session

Emerging Applications of Optimization

Chair: Thomas Lee, MIT, Cambridge, MA, United States

Co-Chair: Xu Sun, Massachusetts Institute of Technology, Cambridge, MA, United States

1 - CANOPI: Contingency-Aware Nodal Optimal Power Investments with High Temporal Resolution

Thomas Lee, MIT, Cambridge, MA, United States, Andy Sun

We present CANOPI, a novel algorithmic framework, for solving the Contingency-Aware Nodal Power Investments problem, a large-scale nonlinear optimization problem that jointly optimizes generation, storage, and transmission expansion. The underlying problem is nonlinear due to the impact of transmission upgrades on impedances, and the problem's large scale arises from the confluence of spatial and temporal resolutions. We propose algorithmic approaches to address these computational challenges. We pose a linear approximation of the overall nonlinear model, and develop a fixed-point algorithm to adjust for the nonlinear impedance feedback effect. We solve the large-scale linear expansion model with a specialized level-bundle method leveraging a novel interleaved approach to contingency constraint generation. We introduce a minimal cycle basis algorithm that improves the numerical sparsity of cycle-based DC power flow formulations, accelerating solve times for the operational subproblems. CANOPI is demonstrated on a 1493-bus Western Interconnection test system built from realistic-geography network data, with hourly operations spanning 52 week-long scenarios and a total possible set of 20 billion individual transmission contingency constraints. Numerical results quantify the reliability and economic benefits of fully incorporating transmission contingencies in integrated planning models and highlight the computational advantages of the proposed methods.

2 - Optimization Approaches for interregional Transmission Planning

Patrick Brown, NREL, Golden, CO, United States

A variety of challenges face the U.S. electricity system: backlogged generator interconnection queues, a rapidly changing resource mix, increasingly impactful extreme weather events, and the prospect of increasing electricity demand after decades of little growth. Transmission expansion has the potential to address many of these issues, but traditional approaches to transmission planning—which have historically been regional and reactive in nature—do not capture the many benefits that transmission can provide. Here, we describe an integrated optimization approach to facilitate wide-area, multi-value transmission planning. The core of the approach uses two coupled models: a capacity expansion model (CEM) that cooptimizes the capacity and operations of generation, storage, and transmission at zonal resolution across the U.S., and a resource adequacy model that assesses the reliability of the resulting systems at high time resolution with Monte Carlo sampling of individual generator outages, iteratively identifying “stress periods” to guide capacity investments in the CEM. Three additional optimization approaches feed into the CEM: The selection of representative time periods is optimized to minimize regional errors in electricity demand and solar/wind capacity factors; zonal interface transfer limits are determined via a nodal power flow model that that optimizes interregional transfers; and a “transmission upgrade supply curve” is determined for each zonal interface, wherein congested transmission branches are sequentially upgraded (via reconductoring, rebuilds, voltage upgrades, and new parallel circuits/transformers) to minimize the cost per megawatt of interregional transfer capability. This integrated planning approach identifies generation/transmission portfolios that minimize system costs while meeting reliability targets nationwide.

3 - Large-load Siting in Modal Power System Capacity Expansion Planning

Tomas Valencia Zuluaga, Lawrence Livermore National Laboratory, San Francisco, CA, United States, Simon Pang, Jean-Paul Watson

Traditional power grid expansion planning studies assume that load growth is an exogenous parameter of the optimization model. The proliferation of datacenters supporting the AI revolution, as well as industrial facilities built for direct air capture of CO2 are challenging this view. These loads can be sited with some flexibility and are large enough to have an impact on the electricity network, even at the transmission level. In this work, we incorporate large-scale loads with flexible siting as part of the decision variables in a nodal capacity expansion planning model. The operational flexibility of large loads is modeled through tranches with differentiated reliability and incorporated into the model as expectation constraints. We develop an augmented Progressive Hedging algorithm that uncouples expectation constraints to solve this problem exploiting parallel computing. We show through some numerical testcases on realistic systems based on California and South Carolina the impact of proactive load siting on the need for transmission investments with different power flow formulations to highlight the importance of incorporating them into software planning tools.

4 - Modeling Adversarial Wildfires for Power Grid Outages

Matthew Brun, Operations Research Center, MIT, Cambridge, MA, United States, Xu Sun, Jean-Paul Watson

Electric power infrastructure faces increasing risks of damage and disruption due to wildfires. The planning and operation of electric grids in wildfire-prone regions must consider the potential impact of unpredictable fires. However, traditional wildfire models do not easily answer how and when wildfires may disable some set of power grid elements. To address these questions, we propose an optimization model to characterize an adversarial wildfire that targets specific infrastructure while respecting physical fire spread dynamics. Using the Rothermel fire spread model, we design a feasible fire spread region that varies with wind velocity. We develop principled convex relaxations of the spread region, introducing new relaxations of the inner product over Euclidean balls. We formulate and solve a mixed-integer conic program to characterize the worst-case time to outage and load shed of a realistic wildfire, and present test cases derived from the recent Park, Eaton, and Palisades fires in California.

FD09

Cabinet Room

Methods Applications for Linear and Conic Programming

Contributed Session

Linear and Conic Optimization

Chair: Jackson Forner, Southern Methodist University

1 - Structural Low-Rank Solutions in the Block-LMI Semidefinite Programming

Hongxi Bai, University of Arizona, Tucson, AZ, United States, Jianqiang Cheng

The Barvinok–Pataki bound provides a general upper bound on the rank of optimal solutions to semidefinite programming (SDP) problems, depending solely on the number of affine constraints. However, this bound is often not tight in practice. As a result, we develop structural rank bounds for a broad class of SDPs with block linear matrix inequality (LMI) constraints. By exploiting the underlying block structure—rather than relying only on the number of affine constraints—our approach yields significantly tighter guarantees. Numerical experiments demonstrate that these refined bounds lead to substantially faster solution times while maintaining comparable objective values in low-rank optimization algorithms for SDP problems.

2 - Decision-Aware Predictions for Right-Hand Side Parameters in Linear Programs

Jackson Forner, Southern Methodist University, Dallas, TX, United States, Miju Ahn, Harsha Gangammanavar

We present an integrated learning and optimization problem in which a prediction model estimates the right-hand-side parameters of a linear program (LP) using a contextual vector. Considering that such a prediction alters the feasible region of the LP, we aim to estimate the constraint set to contain the optimal solution of the underlying LP, given by the true right-hand side parameters. We propose formulations for training a prediction model by minimizing the decision error while accounting for feasibility, measured by a collection of historical primal and dual solutions. Our analysis identifies conditions under which a resulting predicted feasible region contains the true solution, and whether the latter solution achieves optimality for the predicted problem. To solve the alternative training problems, we employ existing LP and nonconvex programming solution methods. We conduct numerical experiments on a synthetic LP and a network optimization problem. Our results indicate that the proposed methods effectively implement the desired feasibility, compared to standard regression models.

3 - Efficient Multi-Column LU Factorization Updates for Exact Linear Programming Solvers

Mason Gao, Tufts University, Medford, MA, United States, Joshua Grassel, Adolfo Escobedo

This paper focuses on core subroutines in linear programming (LP) solvers that combine numerical computations and exact arithmetic to guarantee their outputs' correctness. These solvers use floating-point arithmetic to perform the steps of the simplex algorithm, culminating in a basic solution and terminating status (optimality, infeasibility, etc.). They then use exact arithmetic to solve systems of linear equations --- typically via LU factorization algorithms --- to validate whether the purposed terminating status is correct or to indicate if further adjustments and floating-point arithmetic iterations are needed.

To improve the performance of these solvers on highly ill-conditioned problems requiring multiple validation steps, this work develops a symbolic/numerical LU factorization update algorithm for reflecting the replacement of multiple columns in the basis. The process entails performing a valid sequence of single-column replacement updates. We demonstrate that such sequences can be determined efficiently at runtime. We incorporate the methodology into the exact solver QSopt_ex and apply it on well-known benchmark datasets and additional numerically challenging problems generated herein using perturbations. We utilize the results to propose and test further enhancements for expediting the exact solution of numerically challenging LPs.

4 - Stress Testing the Numerical Stability of LP and MIP Solvers

Joshua Grassel, North Carolina State University, Raleigh, NC, United States, Adolfo Escobedo

Modern solvers for linear programming (LP) and mixed-integer programming (MIP) are indispensable tools in optimization, yet they can yield incorrect results due to roundoff errors from floating-point arithmetic. These errors, such as reporting suboptimal solutions as optimal or misidentifying problems as infeasible, cast doubt on solver reliability. Their impact is particularly acute in domains requiring high numerical precision, such as biochemical modeling, astrophysics simulations, compiler optimization, and mathematical proofs.

To mitigate these issues, various strategies have been developed, including presolve scaling, algorithm tuning, and exact arithmetic subroutines. However, current evaluation practices rely heavily on standard problem libraries, which may not expose numerical weaknesses in a consistent or generalizable way.

This research proposes a novel framework for generating synthetic LP instances that systematically stress test solvers' numerical robustness. Inspired by worst-case complexity constructions, the framework allows control over key parameters, such as variable count and coefficient precision, enabling exploration of solver behavior under varying numerical stress. Each generated problem is guaranteed to have an exactly representable input data in floating-point, ensuring that any observed errors originate from solver internals rather than input encoding.

This approach provides a systematic method for testing solver reliability and benchmarking numerical stability. The presentation showcases results from applying this framework to a general purpose commercial solver, and two open-source exact solvers, highlighting their limitations and computational differences when facing numerically challenging LP instances.

FD10

Caucus Room

Discrete and Stochastic Optimization for Networked Systems

Invited Session

Computational Optimization

Chair: Cheng Guo, Clemson University, Clemson, United States

1 - Optimization of Cascaded Hydroelectric Systems under Decision Dependent Uncertainty

Eliza Cohn, Columbia University, New York, NY, United States, Ning Qi, Upmanu Lall, Bolun Xu

Quantifying the impact of streamflow forecast uncertainty on short-term power generation in cascaded hydropower systems is critical for designing optimal operation strategies. Standard streamflow models determine the exogenous uncertainty, i.e. error occurring from environmental factors that are inputs into system, and is a form of decision independent uncertainty (DIU). In this work, we propose a formulation based on endogenous uncertainty, which is achieved by conditioning the uncertainty in streamflow forecasts on release decisions from upstream units. This representation is a form of decision dependent uncertainty (DDU) and is incorporated when constructing the inflow forecasts into each hydropower unit. We formulate a system-wide optimization problem and use joint chance-constraints to constrain the random variables introduced from the inflow. We propose two different solution algorithms to demonstrate how our enhanced uncertainty representation increases system reliability and generates more energy. These results are verified under various network configurations and streamflow conditions. This work paves the way for the next generation of dynamic cascaded hydropower management.

2 - Value of Covariate Information in Contextual Stochastic Optimization with Application to Networked Systems

Pouria Shahmiri, The Ohio State University, Columbus, OH, United States, Guzin Bayraksan

Contextual Stochastic Optimization (CSO) integrates machine learning and predictive algorithms with optimization under uncertainty. CSO problems depend on multiple contextual factors, yet these factors do not contribute equally to the quality of the final decision. Having insights into the value of each context can help prioritize which contexts are worth modeling and avoid investing effort in contexts that are unlikely to produce high quality outcomes. In this study, we analyze the Value of Contextual (covariate) Information (VCI) in CSO problems, where decisions can depend on observed features, such as forecasts or system states. We derive structural bounds on (i) VCI for a given context and on (ii) the expected VCI across the full range of contexts for a class of CSO problems. Our bounds on VCI depend on both probabilistic and problem-specific properties, and we apply them to networked systems such as the unit commitment problem and capacity expansion problem. We numerically evaluate our results, illustrating how different types of contextual information contribute to improved decision quality and system performance.

3 - Multi-Stage Robust Optimization with Conformal Prediction for Long-Term Grid Resilience Planning

Wei Gu, Carnegie Mellon University, Pittsburgh, PA, United States

Frequent extreme weather events and cyber threats pose increasing needs for grid resilience planning. Most existing approaches to enhance grid resilience employ short-term planning models and simplified uncertainty quantifications. In this research, we develop a novel multi-stage robust optimization model that integrates proactive actions, adversarial disruptions, and reactive responses over time. The adversarial uncertainty sets are constructed using conformal prediction with theoretical guarantees. Reformulation and approximation techniques are proposed in order to solve the model efficiently. The effectiveness of our methods are validated in numerical experiments, which demonstrates the advantage of our methods in prompting energy security and sustainability.

4 - Pricing Discrete and Nonlinear Markets With Semidefinite Relaxations

Cheng Guo, Clemson University, Clemson, SC, United States, Lauren Henderson, Ryan Cory-Wright, Boshi Yang

Nonconvexities in markets with discrete decisions and nonlinear constraints make efficient pricing challenging, often necessitating subsidies. A prime example is the unit commitment (UC) problem in electricity markets, where costly subsidies are commonly required. We propose a new pricing scheme for nonconvex markets with both discreteness and nonlinearity, by convexifying nonconvex structures through a semidefinite programming (SDP) relaxation and deriving prices from the relaxation's dual variables. When the choice set is bounded, we establish strong duality for the SDP, which allows us to extend the envelope theorem to the value function of the relaxation. This extension yields a marginal price signal for demand, which we use as our pricing mechanism. We demonstrate that under certain conditions - for instance, when the relaxation's right hand sides are linear in demand - the resulting lost opportunity cost is bounded by the relaxation's optimality gap. This result highlights the importance of achieving tight relaxations. The proposed framework applies to nonconvex electricity market problems, including for both direct current and alternating current UC. Our numerical experiments indicate that the SDP relaxations are often tight, reinforcing the effectiveness of the proposed pricing scheme. Across a suite of IEEE benchmark instances, the lost opportunity cost under our pricing scheme is, on average, 46% lower than that of the commonly used fixed-binary pricing scheme.

5 - Multi-echelon Batch Inventory Networks

Levi DeValve, Chicago Booth, Chicago, IL, United States, Maximiliano Stock

We study the Capacitated Multi-echelon Batch Inventory problem, which models batch inventory ordering in a two-echelon fulfillment network with capacity constraints at the lower echelon, as often seen in e-retailing. A centralized upper echelon places batch orders from a supplier and allocates units to the lower echelon to minimize total costs, including ordering, allocation, and expected fulfillment costs.

The hard capacity constraints make the problem structurally similar to capacitated facility location, which is well known to be challenging to approximate using linear relaxations. We address these challenges by exploiting key features of our setting: the allocation and fulfillment costs are metric, and the upper echelon is uncapacitated.

Leveraging these features, we develop a new clustering technique that uses information from the dual program to group nearby distribution centers in the lower echelon with enough inventory. This clustering enables us to round the LP solution while satisfying the capacity constraints and leads to a constant-factor approximation algorithm for the offline two-stage version of the problem.

We further extend our approach to an online setting in which demand arrives sequentially over a finite time horizon. To address this, we develop a framework that integrates our approximation algorithm with tools from the online decision-making literature to decide on the ordering, allocation, and online fulfillment. Our framework achieves an additive regret bound that is constant relative to the time horizon, making the theoretical guarantee of the online solution converge linearly to that of the offline solution as the horizon grows.

FD11

Charter Room

Optimization and Learning in Complex Networks

Invited Session

Network Optimization

Chair: Qi Luo, University of Iowa, Iowa City, IA, United States

Co-Chair: Yan Wu, Clemson University, Clemson, SC, United States

Co-Chair: Yuyuan Ouyang, Clemson University, Clemson, SC, United States

1 - Contextual Distributionally Robust Optimization under Multi-Modal Decision-Dependent Uncertainty for Facility Location Problems

Huangrong Sun, The Ohio State University, Columbus, OH, United States, Beste Basciftci, Xian Yu

We study a two-stage facility location problem under uncertain customer demand, where the demand depends on exogenous factors and endogenous facility location decisions and may exhibit distinct modes in different regions. To calibrate this decision-dependent and multi-modal uncertainty based on historical data, we first formulate a mathematical optimization problem to jointly optimize the number of modes, the assignment of data points to modes, and the regression function in each mode. Given the complexity of this optimization model, we apply cluster-wise regression algorithms to solve it more efficiently. Leveraging the estimated mode probabilities, regression functions, and the empirical residuals obtained from the regression step, we formulate a two-stage empirical residuals-based distributionally robust optimization (ER-DRO) model with both moment-based and Wasserstein distance-based ambiguity sets and derive their tractable reformulations. For the Wasserstein case, we further propose a separation-based algorithm to enhance computational efficiency. We compare the performance of our cluster-wise regression algorithms with various regression benchmarks through extensive numerical experiments using synthetic datasets and real-world data for electric vehicle charging network from New York State. We also illustrate the benefit of considering decision-dependency and multi-modality by comparing our DRO model with decision-independent or single-modal counterparts.

2 - Anticipating Infeasibility: Decision Diagram-Guided Reinforcement Learning for Routing Problems

Macarena Navarro, Carnegie Mellon University, Pittsburgh, PA, United States, Karan Singh, Willem-Jan van Hoeve, Segev Wasserkrug

Recent advances in neural combinatorial optimization have shown that deep reinforcement learning can solve combinatorial optimization problems by training policies to generate near-optimal solutions. However, existing approaches struggle with constraints that require global reasoning on the entire decision space, such as time window constraints augmenting the traveling salesperson problem. Most existing approaches only handle simpler constraints, such as permutation structures, where one can ensure feasibility incrementally while generating the solution.

We address this limitation by integrating relaxed decision diagrams into the reinforcement learning framework. Our approach equips the learner with structured lookahead capabilities that enable it to anticipate and avoid decisions leading to downstream infeasibility during solution generation. We embed these decision diagrams directly into attention-based neural architectures and evaluate their effectiveness on the traveling salesperson with time windows problem and precedence constraints.

3 - Electrification of Heavy-Duty Vehicles: Challenges and New Algorithmic Insights

Haripriya Pulyassary, Cornell University, Ithaca, NY, United States, David Shmoys, Manxi Wu

While the adoption of electric vehicles (EV) offers substantial environmental benefits, it also introduces new challenges in vehicle routing due to range anxiety, long charging times, and sparse charging infrastructure. In this work, we approach EV routing through the lens of network flow theory. We begin with the analogs of three classical network flow problems: the problem of optimally routing a single EV, the maximum EV flow problem, and the minimum-cost EV flow problem. While these problems are NP-hard in the general case, we identify conditions under which they can be solved in polynomial time, and give exact, polynomial-time algorithms for this setting.

We then consider fleet electrification in long-distance logistics under various recharging modes. To model these, we introduce a general flow-based formulation defined
over a charge-and-time augmented network. Due to the density of this network, directly solving the full integer program is impractical. Instead, we propose a variant of
dynamic discretization discovery that significantly improves computational time. We also apply these methods in an empirical study of an instance constructed from real-
world data.

4 - Wasserstein Motifs: Non-deterministic Alignment of Ecological Networks

Yifan Xu, Rice University, Houston, TX, United States, Carlos Taveras, Lydia Beaudrot, Cesar A. Uribe

We study the problem of ecological network (food webs) alignment, where we seek to identify structural equivalences of species and uncover backbones of interactions representing shared functional substructures. These fundamental properties reveal the functional relationships that sustain ecosystems, enabling more accurate predictions of biodiversity responses to environmental change. Existing methods are computationally expensive, not scalable, and hard to interpret ecologically. We provide a first rigorous formalization of food web alignment based on network motifs, and show existing methods popularized in the ecological community are equivalent to minimizing a Fused Gromov-Wasserstein-like cost functional, termed Wasserstein Motifs. Moreover, we propose an interpretable and provably correct algorithm that efficiently computes non-deterministic alignments between food webs by leveraging their representation as feature measure networks. As a byproduct, we introduce a novel approach for the identification of non-deterministic backbones of interactions. Experiments on a continental-scale dataset of 129 Sub-Saharan African mammal food webs demonstrate significant gains in accuracy, scalability, and interpretability over state-of-the-art methods. Our results establish a principled bridge between ecological network science and optimal transport, opening avenues for the analysis of complex structured data.

March 21, 2026, 8:00 AM - 9:30 AM

SatA01

Grand Ballroom

Adaptive Methods for Nonlinear and Stochastic Optimization I

Invited Session

Optimization in Data Science

Chair: Tianjiao Li, Massachusetts Institute of Technology, Cambridge, MA, United States

1 - High-order Accumulative Regularization Methods for Gradient Minimization in Convex Programming

Yao Ji, Georgia Institute of Technology, Atlanta, GA, United States, Guanghui Lan

This paper develops a unified high-order accumulative regularization (AR) framework for convex and uniformly convex gradient norm minimization. Existing high-order methods often exhibit a gap: the function-value residual decreases fast, while the gradient norm converges much slower. To close this gap, we introduce AR that systematically transforms the fast function-value residual convergence rate into a fast (matching) gradient norm convergence rate.

Specifically, for composite convex problems, to compute an approximate solution such that the norm of its (sub)gradient does not exceed $\varepsilon,$ the proposed AR methods match the best corresponding convergence rate for the function-value residual. We further extend the framework to uniformly convex settings, establishing linear, superlinear, and sublinear convergence of the gradient norm under different lower curvature conditions. Moreover, we design parameter-free algorithms that require no input of problem parameters, e.g., the Lipschitz constant of the p-th-order gradient, the initial optimality gap and the uniform convexity parameter, and allow an inexact solution for each high-order step. To the best of our knowledge, no parameter-free methods can attain such a fast gradient norm convergence rate which matches that of the function-value residual in the convex case, and no such parameter-free methods for uniformly convex problems exist.These results substantially generalize existing parameter-free and inexact high-order methods and recover first-order algorithms as special cases, providing a unified approach for fast gradient minimization across a broad range of smoothness and curvature regimes.

2 - Mathjax-A Practical Adaptive Subgame Perfect Gradient Method

Benjamin Grimmer, Johns Hopkins University, Baltimore, MD, United States

We present a performant gradient method for smooth convex optimization, drawing inspiration from several recent advances in the field. Our algorithm, the Adaptive Subgame Perfect Gradient Method (ASPGM) is based on the notion of subgame perfection, attaining a dynamic strengthening of minimax optimality. At each iteration, ASPGM makes a momentum-type update, optimized dynamically based on a (limited) memory/ bundle of past first-order information. ASPGM is linesearch-free, parameter-free, and adaptive due to its use of recently developed auto-conditioning, restarting, and preconditioning ideas. We show that ASPGM is competitive with state-of-the-art L-BFGS methods on a wide range of smooth convex problems. Unlike quasi-Newton methods, however, our core algorithm underlying ASPGM has strong, subgame perfect, non-asymptotic guarantees, providing certificates of solution quality, resulting in simple stopping criteria and restarting conditions.

3 - Inexactly Smooth Performance Estimation and a Universal Optimized Gradient Method

Aaron Zoll, Johns Hopkins University, Baltimore, MD, United States, Benjamin Grimmer

Universal first-order methods are methods with performance guarantees across a wide range of distinct problem families. In this paper, we consider a general class of ``inexactly smooth'' convex functions which contains as special cases families like $L$-smooth functions, $M$-Lipschitz nonsmooth functions, H\"older smooth functions, and any combination thereof. We show that such ``inexactly smooth'' functions can be addressed within the Performance Estimation Problem (PEP) framework by providing interpolation theorems that are necessary and sufficient up to modest universal constants. These PEPs enable the design of new optimized gradient methods. In particular, for convex functions with bounded difference in subgradients, we provide the first (exactly) minimax optimal method. For H\"older smooth problems, we provide a method with the best known convergence guarantee by constants. Our approach is amenable to backtracking, enabling the design of a parameter-free Universal Optimized Gradient Method for any ``inexactly smooth'' convex minimization problem.

4 - A New Termination Certificate for First-Order Methods

Jimmy Zhang, Purdue University, West Lafayette, IN, United States

First-order methods excel at delivering the medium-to-low accuracy solutions required for large-scale optimization. However, it is precisely in this regime that standard termination certificates, like the gradient norm or KKT conditions, prove unreliable. Their failure to measure true algorithmic progress leads to premature termination or wasted computation, and critically, prevents their use in designing adaptive algorithms.

We introduce a novel, verifiable termination certificate that resolves this shortcoming. Our certificate enables the use of the notoriously difficult-to-estimate quadratic growth (or weak convexity) parameter to develop parameter-free algorithms with nearly optimal complexity bounds. We demonstrate its power across a wide range of problems, including non-smooth, weakly smooth, piecewise smooth, and constrained optimization.

SatA02

Georgian Room

Interplay of Optimization, Learning, and Data Science

Invited Session

Optimization in Data Science

Chair: Liwei Jiang, Purdue University, West Lafayette, IN, United States

1 - ASGO: Adaptive Structured Gradient Optimization

Kang An, Rice University, Houston, TX, United States, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, Tong Zhang

Training deep neural networks is a structured optimization problem, because the parameters are naturally represented by matrices and tensors rather than by vectors. Under this structural representation, it has been widely observed that gradients are low-rank and Hessians are approximately block diagonal. These structured properties are crucial for designing efficient optimization algorithms, but are not utilized by many current popular optimizers like Adam. In this paper, we present a novel optimization algorithm ASGO that capitalizes on these properties by employing a preconditioner that is adaptively updated using structured gradients. By a fine-grained theoretical analysis, ASGO is proven to achieve superior convergence rates compared to existing structured gradient methods. Based on this convergence theory, we further demonstrate that ASGO can benefit from low-rank gradients and block diagonal Hessians. We also discuss practical modifications of ASGO and empirically verify ASGO's effectiveness on language model tasks.

2 - Accurate, Provable, and Fast Polychromatic Tomographic Reconstruction: A Variational Inequality Approach

Mengqi Lou, Georgia Institute of Technology, Atlanta, GA, United States, Kabir Verchand, Sara Fridovich-Keil, Ashwin Pananjady

We consider the problem of signal reconstruction for computed tomography (CT) under a nonlinear forward model that accounts for exponential signal attenuation, a polychromatic X-ray source, general measurement noise (e.g. Poisson shot noise), and observations acquired over multiple wavelength windows. We develop a simple iterative algorithm for single-material reconstruction, which we call EXACT (EXtragradient Algorithm for Computed Tomography), based on formulating our estimate as the fixed point of a monotone variational inequality. We prove guarantees on the statistical and computational performance of EXACT under realistic assumptions on the measurement process. We also consider a recently introduced variant of this model with Gaussian measurements, and present sample and iteration complexity bounds for EXACT that improve upon those of existing algorithms. We apply our EXACT algorithm to a CT phantom image recovery task and show that it often requires fewer X-ray projection exposures, lower source intensity, and less computation time to achieve similar reconstruction quality to existing methods.

3 - Parameter-Free Strongly Convex Optimization

Ke Tang, Purdue University, West Lafayette , IN, United States, Liwei Jiang, Jimmy Zhang

Parameter-free optimization methods have been extensively studied in convex optimization, and optimal algorithms are known in the smooth, weakly smooth, and nonsmooth settings. However, none of the existing approaches achieve the optimal rate for strongly convex problems without prior knowledge of the strong convexity modulus. In this talk, we introduce a variant of the bundle-level method that is parameter-free, anytime (i.e., it does not require specifying the target accuracy or the total number of iterations), and universally optimal for smooth, weakly smooth, and nonsmooth problems for almost all accuracy levels. The only nearly optimal case is in the nonsmooth regime over a certain interval of accuracies, where our

complexity has an additional log log factor.

4 - Effective Uncertainty of General Priors for Memorization in Overparameterized Linear Models

Chen Cheng, University of Chicago, Chicago, IL, United States

We examine the necessity of interpolation in overparameterized linear models---also known as memorization---when achieving optimal predictive risk in machine learning problems requires (nearly) interpolating the training data. We consider a Bayesian setup with the ground truth linear predictor θ in R^d following some prior P_θ. In the first work, we demonstrate memorization in a simple isotropic prior in the low noise regime. In the second work, we consider general priors, including Gaussian mixtures, and show universality of such phenomenon, and identify a characterization of effective uncertainty vs. noise level for memorization to occur. In both cases, our conclusion implies when the effective uncertainty of the prior is much larger than the observational uncertainty (noise level), optimal performance requires fitting training data to substantially higher accuracy than the noise level of the problem.

SatA03

Plaza I

Systematic Analysis of Optimization Algorithms Leveraging Computer-Assisted Proofs

Invited Session

Nonlinear Optimization

Chair: Jaewook J. Suh, Rice University, Houston, TX, United States

1 - Optimized Methods for Composite Optimization: A Reduction Perspective

Jinho Bok, University of Pennsylvania, Philadelphia, PA, United States, Jason Altschuler

Recent advances in convex optimization have leveraged computer-assisted proofs to develop optimized first-order methods that improve over classical algorithms. However, each optimized method is specially tailored for a particular problem setting, and it is a well-documented challenge to extend optimized methods to other settings due to their highly bespoke design and analysis. We provide a general framework that derives optimized methods for composite optimization directly from those for unconstrained smooth optimization. The derived methods naturally extend the original methods, generalizing how proximal gradient descent extends gradient descent. The key to our result is certain algebraic identities that provide a unified and straightforward way of extending convergence analyses from unconstrained to composite settings. As concrete examples, we apply our framework to establish (1) the phenomenon of stepsize acceleration for proximal gradient descent; (2) a convergence rate for the proximal optimized gradient method which is faster than FISTA; (3) a new method that improves the state-of-the-art rate for minimizing gradient norm in the composite setting.

2 - H-invariance Theory: A Complete Characterization of Minimax Optimal Fixed-point Algorithms

TaeHo Yoon, Johns Hopkins University, Baltimore, MD, United States, Ernest Ryu, Benjamin Grimmer

For nonexpansive fixed-point problems, Halpern's method with optimal parameters, its so-called H-dual algorithm, and in fact, an infinite family of algorithms containing them, all exhibit the exactly minimax optimal convergence rates. In this work, we provide a characterization of the complete, exhaustive family of distinct algorithms using predetermined step-sizes, represented as lower triangular H-matrices, which attain the same optimal convergence rate. The characterization is based on polynomials in the entries of the H-matrix that we call H-invariants, whose values stay constant over all optimal H-matrices, together with H-certificates, of which nonnegativity precisely specifies the region of optimality within the common level set of H-invariants. The H-invariance theory we present offers a novel view of optimal acceleration in first-order optimization as a mathematical study of carefully selected invariants, certificates, and structures induced by them.

3 - Data-driven Analysis of First-Order Methods via Distributionally Robust Optimization

Jisun Park, Princeton University, Princeton, NJ, United States, Vinit Ranjan, Bartolomeo Stellato

We consider the problem of analyzing the probabilistic performance first-order methods, when solving a distribution of parametric convex optimization problems. By combining performance estimation (PEP) and Wasserstein distributionally robust optimization (DRO), we formulate the analysis as a tractable semidefinite program. Our approach unifies worst-case and average-case analyses by incorporating data-driven information from the observed convergence of first-order methods on a limited number of problem instances. This yields probabilistic, data-driven performance guarantees in terms of the expectation or conditional value-at-risk of the selected performance metric. Experiments on smooth convex minimization, logistics regression, and Lasso show that our method significantly reduces the conservatism of classical worst-case bounds and narrows the gap between theoretical and empirical performance.

SatA04

Plaza II

Recent Topics in Stochastic Optimization

Invited Session

Optimization under Uncertainty

Chair: Yifan Hu, Rutgers, Piscataway, NJ, United States

Co-Chair: Bingcong Li, ETH Zurich, Zurich, N/A, Switzerland

1 - Zeroth-Order Optimization Finds Flat Minima

Liang Zhang, ETH Zurich, Zurich, Switzerland, Bingcong Li, Kiran Thekumparampil, Sewoong Oh, Michael Muehlebach, Niao He

Zeroth-order methods are extensively used in machine learning applications where gradients are infeasible or expensive to compute, such as black-box attacks, reinforcement learning, and language model fine-tuning. Existing optimization theory focuses on convergence to an arbitrary stationary point, but less is known on the implicit regularization that provides a fine-grained characterization on which particular solutions are finally reached. We show that zeroth-order optimization with the standard two-point estimator favors solutions with small trace of Hessian, which is widely used in previous work to distinguish between sharp and flat minima. We further provide convergence rates of zeroth-order optimization to approximate flat minima for convex and sufficiently smooth functions, where flat minima are defined as the minimizers that achieve the smallest trace of Hessian among all optimal solutions. Experiments on binary classification tasks with convex losses and language model fine-tuning support our theoretical findings.

2 - When Scores Learn Geometry: Rate Separations under the Manifold Hypothesis

Xiang Li, ETH Zurich, Zurich, Switzerland

Score-based methods, such as diffusion models and Bayesian inverse problems, are often interpreted as learning the data distribution in the low-noise limit ($\sigma \to 0$). In this work, we propose an alternative perspective: their success arises from implicitly learning the data manifold rather than the full distribution. Our claim is based on a novel analysis of scores in the small-σ regime that reveals a sharp separation of scales: information about the data manifold is $\Theta(\sigma^{-2})$ stronger than information about the distribution. We argue that this insight suggests a paradigm shift from the less practical goal of distributional learning to the more attainable task of geometric learning, which provably tolerates $O(\sigma^{-2})$ larger errors in score approximation. We illustrate this perspective through three consequences: i) in diffusion models, concentration on data support can be achieved with a score error of $o(\sigma^{-2})$, whereas recovering the specific data distribution requires a much stricter $o(1)$ error; ii) more surprisingly, learning the uniform

distribution on the manifold-an especially structured and useful object-is also O(σ^{−2}) easier; and iii) in Bayesian inverse problems, the maximum entropy prior is O(σ^{−2}) more robust to score errors than generic priors. Finally, we validate our theoretical findings with preliminary experiments on large-scale models, including Stable Diffusion.

3 - Ensemble Methods for Multi-sourced Stochastic Optimization

Yanru Guo, University of Michigan, Ann Arbor, MI, United States, Ruiwei Jiang, Siqian Shen

Many decision problems require acting after observing covariates, while predictions arrive from multiple models of varying reliability. We propose a trust-weighted ensemble that lifts stacking to the decision layer via an empirical residual-based sample average approximation (ER-SAA) method. Given T bases and a simplex weight w, each base contributes a residual pool from training data. At a new context x we form scenarios by pairing current prediction of each base and residuals from its own pool and solve a trust-weighted ER-SAA by allocating probability mass across bases according to wWe operate online in a “decide – observe – learn” protocol that updates w by either a prediction-driven or a decision-aware strategy. The framework adapts trust to source reliability and converts multi-source reference information into decisions that respect both central tendency and structured uncertainty.

4 - Accelerated Dual Methods for Stochastic Distributed Optimization with Local Updates

Junchi Yang, Chinese University of Hong Kong, Shenzhen, Shenzhen, China, People's Republic of

This talk presents a method for stochastic, strongly convex distributed optimization that addresses the challenges of training across multiple agents with heterogeneous data distributions. The approach combines accelerated gradient ascent on dual variables with multi-step stochastic gradient descent (SGD) on primal variables in a Lagrangian framework. The design naturally supports extensive local computation, as the inner SGD loops require no inter-agent communication, enabling flexible trade-offs between computation and communication. The proposed algorithm is proven to converge for any number of local updates and achieves optimal communication complexity when local computation is sufficient. The analysis builds on an inexact accelerated gradient framework, interpreting partial dual gradients as inexact gradients of the dual function. Within this framework, there is a trade-off between gradient exactness and convergence rate. In distributed optimization, this theoretical trade-off translates directly into the trade-off between local computation and communication complexity.

SatA05

Plaza III

Adaptive and Robust Methods in Stochastic and Multistage Optimization

Contributed Session

Optimization under Uncertainty

Chair: Ahmadreza Marandi, Eindhoven University of Technology, Eindhoven, Netherlands

1 - K-Adaptability in Stochastic Optimization with Recourse

Jim Luedtke, University of Wisconsin-Madison, Madison, WI, United States, Arnav Aakash, Carla Michini

Many optimization problems solved repeatedly in practice—such as in power systems or logistics—share a fixed structure but with different parameter values. Re-optimizing each instance from scratch can be computationally prohibitive, while adopting a single solution for all realizations is overly conservative. The K-adaptability paradigm offers a compromise by precomputing a small set of candidate solutions and selecting the best once uncertainty is revealed. When K is small, this approach enables real-time decision making and provides human decision-makers with a manageable set of implementable solutions. However, it is often impossible to identify a small number of representative solutions that remain feasible across all parameter realizations. We introduce K-adaptability with recourse, in which the goal is to precompute K candidate assignments for the discrete variables while allowing the continuous recourse decisions to be optimized online after uncertainty is observed. This structure preserves online computational tractability while substantially improving adaptability. We present formulations of this model and two additional solution approaches: a scenario-based heuristic inspired by scenario decomposition, and an exact column and constraint generation method. Preliminary results show that the heuristic consistently identifies solutions close to the optimum and demonstrate the computational efficiency of the column and constraint generation approach.

2 - K-revision Approach in Multistage Stochastic Programming

Alexander Estes, University of Maryland, College Park, MD, United States, Chengwenjian Wang, Jean-Philippe Richard

A standard assumption in multistage stochastic programming is that decisions are made after observing the uncertainty from the prior stage. The resulting solutions can be difficult to implement in practice, as they leave practitioners ill-prepared for future stages. To provide better foresight to practitioners, we introduce the K-revision approach. This new framework requires plans to be specified in advance. To maintain flexibility, we allow plans to be revised a maximum of K of times as new information becomes available. We analyze the complexity of K-revision problems, showing NP-hardness even in a simple setting. We examine, both theoretically and computationally, the impact of the K-revision approach on the objective compared with classical multistage stochastic programming models and the recently introduced partial adaptive approach. We develop two MIP formulations, one directly from our definition and the other based on a combinatorial characterization. We analyze the tightness of these formulations and propose several methods to strengthen them using polyhedral approaches. Computational experiments on synthetic problems and practical applications demonstrate that our approach is both computationally tractable and effective in reaching near-optimal performance while increasing the predictability of the solutions produced.

3 - Flexible Resource Allocation using k-Adaptable Policies

Reem Khir, Purdue University, West Lafayette, IN, United States, Olivia Wang

Designing resource allocation policies that perform well across a wide range of demand conditions remains a central challenge in modern logistics operations. Fully dynamic approaches can, in principle, tailor decisions to each demand outcome but are often too complex to implement at scale. Static allocations are easy to deploy yet may perform poorly when demand deviates from expectations. We study k-adaptable policies as an intermediate solution: the system pre-computes k allocation designs offline and then selects which one to operate after demand is realized. We characterize when an individual design is effective and how multiple k-designs can be generated to complement one another, yielding a tractable optimization procedure that performs well across different instance sizes and compositions. Computational experiments in parcel-sorting and zone-picking settings show that small-k designs substantially outperform static allocations while capturing most benefits of fully dynamic policies, with the magnitude of these gains varying significantly with system structure, scale, and load conditions.

4 - Quadratic Optimization Through the Lens of Adjustable Robust Optimization

Ahmadreza Marandi, Eindhoven University of Technology, Eindhoven, Netherlands, Abbas Khademi

Quadratic optimization (QO) has been studied extensively in the literature due to its applicability in many practical problems. While practical, it is known that QO problems are generally NP-hard. So, researchers developed many approximation methods to find good solutions. In this paper, we analyze QO problems using robust optimization techniques. To this end, we first show that any QO problem can be reformulated as a disjoint bi-convex QO problem. Then, we provide an equivalent adjustable robust optimization (ARO) reformulation and leverage the methods available in the literature on ARO to approximate this reformulation. More specifically, we show that using a so-called decision rule technique to approximate the ARO reformulation is equivalent to using a reformulation-linearization technique on the original QO problem. Additionally, we design an algorithm that can find a close-to-optimal solution based on our new reformulations. Our numerical results demonstrate the efficiency of our algorithm to find near-optimal solutions, particularly for large-sized instances, compared with off-the-shelf solvers and state-of-art approaches.

5 - Circumventing Compactness in Conditional-Moment Optimal Transport Duality: A Perturbation Approach

Jake Roth, University of Minnesota, Minneapolis, MN, United States, Louis Chen, Johannes Royset

Recent work by Blanchet et al (2025) introduced the conditional-moment optimal transport (CM OT) problem, which generalizes several popular Distributionally Robust Optimization (DRO) frameworks including: Phi-divergence DRO, Sinkhorn-DRO, and Wasserstein DRO. A dual problem was derived, and it was conjectured that certain compactness assumptions were crucial for duality to hold. Indeed, if CM OT, which already lacks a feasible interior, is further complicated by the removal of compactness, then various popular tools (e.g., Sion's minimax, Slater's condition) cannot be leveraged. However, we show that the desired duality can in fact hold without compactness, under some (mild) alternative assumptions. These assumptions and our techniques are inspired by the classical perturbation approach to duality. We find that the perturbation approach is not only natural for devising and establishing duality for DRO problems but also useful for providing arguably simpler and shorter proofs.

SatA06

Director's and Lounge

Advances in Deterministic Methods I

Invited Session

Global Optimization

Chair: Matthew Stuber, University of Connecticut, Mansfield, CT, United States

Co-Chair: Joseph Scott, Georgia Institute of Technology, Atlanta, United States

1 - Automated Discovery of Optimization Algorithms

Nikolaos Sahinidis, Georgia Tech, Atlanta, GA, United States, Ilias Mitrai

By modeling iterative algorithms as expression trees, we propose an MINLP model for automated discovery of optimization algorithms. The formulation seeks functions that satisfy a given convergence criterion while a desired objective, such as the complexity of the function, is optimized. We solve the MINLP with BARON and present results for the design of first-order algorithms and systems of equations.

2 - New Discretization Methods for the Global Optimization of Nonconvex Semi-Infinite Programs

Rohit Kannan, Virginia Tech, Blacksburg, VA, United States

We introduce two new discretization methods for the global optimization of nonconvex semi-infinite programs (SIPs). One method includes an additional local optimization step at each iteration to guide the discretization process. We establish convergence results showing that the bounds produced by both methods converge to the optimal objective of the SIP. Computational experiments on standard nonconvex SIP benchmarks demonstrate that the proposed methods substantially reduce the number of iterations required for convergence compared to existing discretization approaches.

3 - MILP Models for Capturing Nuanced, Multi-Period State Transitions in the Context of the Unit Commitment Problem

Nathan Barrett, Princeton University, Princeton, NJ, United States, Christos Maravelias

This work introduces several novel mixed-integer linear programming (MILP) unit commitment models for thermal generators (e.g. bioenergy, natural gas, coal, etc.) featuring the ability to model adaptable power output trajectories during multi-period startups and shutdowns. These models capture various degrees of fidelity to fine-resolution dynamic models including variable power output and variable transition duration along with accurate representations of transition costs. We demonstrate that these models yield notably improved objective function values over existing state-of-the-art models while maintaining adequate computational performance. While the body of this work is performed within the context of electrical power grid operation, we demonstrate that these models are easily generalized to other sectors that involve nuanced state transitions such as chemical production scheduling and optimal control.

4 - Improvements to Decomposition Methods for the Global Solution of Stochastic Programs

Joseph Scott, Georgia Institute of Technology, Atlanta, GA, United States, Jingzhi Yang

This talk presents an accelerated scheme for obtaining high-quality Lagrangean multipliers for the LG method, one of the few existing decomposition algorithms for solving two-stage stochastic mixed-integer nonlinear programs (SMINLPs) to guaranteed global optimality.
The LG method decomposes the original problem into scenario subproblems, and its lower bound is computed by solving a Lagrangean dual problem in which multipliers are updated using a standard subgradient method. However, existing decomposition algorithms struggle to compete with global solvers such as Gurobi or BARON. Prior theoretical analysis indicates that achieving competitive performance requires satisfying a key convergence property, which in turn depends on obtaining optimal multipliers for the LG method. Unfortunately, a recent numerical study from our group shows that obtaining such optimal multipliers is generally impractical for realistic SMINLP problems.
To address this challenge, we propose a subgradient-based update scheme inspired by the Augmented Lagrangean method. Our proposed approach accelerates convergence toward good-quality multipliers early in the iteration process, even potentially achieving the true optimal multiplier when the search space is small. In this talk, we will present the details of our accelerated scheme and provide numerical evidence demonstrating its performance relative to the standard subgradient approach used by the LG method.

SatA07

Congress Room

Structural and Complexity Perspectives on Mixed-Integer Programming Algorithms

Invited Session

Discrete Optimization

Chair: Hongyu Cheng, Johns Hopkins University, Baltimore, United States

1 - How Hard is Learning to Cut? Trade-Offs and Sample Complexity

Sammy Khalife, Amazon, New York City, NY, United States

In the recent years, branch-and-cut algorithms have been the target of data-driven approaches designed to enhance the decision making in different phases of the algorithm such as branching, or the choice of cutting planes. In particular, for cutting plane selection two score functions have been proposed in the literature to evaluate the quality of a cut: branch-and-cut tree size and gap closed. In this talk, I will present new sample complexity lower bounds, valid for both scores. For a wide family of classes that maps an instance to a cut, learning over an unknown distribution of the instances to minimize those scores requires at least (up to multiplicative constants) as many samples as learning from the same class function any generic target function. Those results also extend to the case of learning from a restricted set of cuts, namely those from the

Simplex tableau, and constitute the first lower bounds for the learning-to-cut framework. In the case of neural networks, those upper bounds are nearly tight. I will illustrate those results experimentally on various integer programming models to show that the gap closed score is an effective proxy to minimize the branch-and-cut tree size.

2 - Theoretical Foundations and Challenges in Machine Learning for Branch-and-Cut

Hongyu Cheng, Johns Hopkins University, Baltimore, MD, United States, Amitabh Basu

This talk explores the theoretical foundations and challenges of using machine learning (ML) to enhance the Branch-and-Cut (B&C) framework for solving mixed-integer programs. While ML, particularly neural networks, has shown empirical promise in learning policies for key decisions like branching and cut selection, their theoretical underpinnings remain underdeveloped. We first address generalization by establishing rigorous sample complexity bounds. We introduce a framework based on the piecewise structure of the performance metric (e.g., tree size) with respect to policy parameters. This framework unifies the analysis of traditional linear heuristics and modern neural networks (e.g., ReLU MLPs). Second, we investigate the approximation error and sensitivity of learned policies. We demonstrate a critical challenge: high imitation accuracy does not guarantee good performance. We prove that small local deviations from an expert policy can lead to exponential differences in the global tree size for both branching and cutting plane selection. Finally, we examine the approximation ability of common scoring functions. We show the limitations of standard linear combinations used in solvers for cut selection and propose theoretically grounded non-linear alternatives.

3 - Relative Complexity of Cutting Planes and Branch-and-Cut in Convex 0/1 Integer Programming

Ziyi Su, Georgia Institute of Technology, Atlanta, GA, United States

We study the effectiveness of cutting-plane (CP) methods relative to branch-and-cut (BC) techniques for convex 0/1 integer programming under variable disjunctions. A recent work of Basu et al. shows that for every valid inequality an optimal CP proof is at least as efficient as an BC proof, but possibly only up to an arbitrarily small epsilon slack in the bound. This raises a further question: can one obtain zero slack, at least in some structured settings, while keeping the proof size under control?

We give a positive answer in three dimensions. For convex 0/1 sets C contained in [0,1]^3 with variable disjunctions, we prove that every valid inequality that admits a BC proof of size N also admits a pure CP proof of size at most N, with zero slack. Thus, in 3D, branching does not provide additional proof power beyond what can be achieved with cutting planes alone.

The proof proceeds by eliminating maximal-depth branching nodes via structural induction, together with a geometric analysis of five types of interactions between the cube and the cutting planes. I will describe this construction, place the result within the Basu-et-al. landscape, and discuss obstacles and possible approaches to extending such zero-slack CP-BC efficiency guarantees beyond three dimensions.

4 - Solving Separable Optimization Problems with Structured Sparsity Under Indicator Conditions

Tongtong Chen, Columbia University, New York, NY, United States

We study convex optimization problems where disjoint blocks of variables are controlled by binary indicator variables, which themselves are subject to additional conditions such as cardinality constraints. Several classes of important examples can be formulated in such a way that both the objective and the constraints are block-separable convex functions. We introduce a new metric for sparsity, namely, the treewidth of the constraint-block intersection graph. We provide complexity results and a family of polynomial-time approximation algorithms for those types of problems.

SatA08

Committee Room

Optimization for Societal Systems

Invited Session

Emerging Applications of Optimization

Chair: Elijah Pivo

1 - Advancing the Design of U.S. Organ Transplant Policy with Optimization

Elijah Pivo, Massachusetts Institute of Technology, Cambridge, MA, United States, Thomas Athey, Maura Hegarty, Nikolaos Trichakis, Dimitris Bertsimas

The U.S. organ transplant system has long contended with the challenge of fairly and effectively allocating transplantable organs. A national organ allocation policy defines how patients are prioritized for each donated organ. However, understanding the implications of an allocation policy, much less designing one capable of improving the system, is a formidable task. The waitlist consists of over 100,000 individuals spread across the country with compatible donor pools that vary drastically, each transplant involves complex and time-critical logistics, and policy changes may influence how patients and transplant centers judge transplant offers. This talk will focus on new analytical tools our group has been developing in partnership with UNOS and the OPTN, the main organizations that manage the national transplant system, in order to better support policymakers. These tools include clinical, logistical and behavioral models, a high-speed transplant system simulator, and multi-objective policy optimization. Together, these methods represent a major advance in the computational methods available for the design of organ allocation policy and an opportunity to improve the system for hundreds of thousands of current and future patients.

2 - Mission vs. Profit: The Interplay Between Social Service Agencies and Private Service Providers

Gulten Busra Karkili, University of Massachusetts Amherst, Amherst, MA, United States, Senay Solak

Expanding access to essential services such as childcare and housing remains a central challenge for social service agencies, particularly when delivery depends on private providers. These providers make participation decisions based on financial incentives, market competition, and expected demand, which can limit access for low-income individuals and families even when subsidies are available. In this paper, we study how a mission-driven agency can strategically invest its limited budget to influence provider behavior and increase service availability for voucher subsidy recipients. We develop an optimization framework that accounts for provider heterogeneity, demand uncertainty, and competitive dynamics. Our results show that some providers may participate without direct investment when voucher demand improves capacity utilization, while others require targeted funding. We characterize the minimum investment needed to induce participation and identify conditions where additional spending no longer increases access. Moreover, we show that funding can have non-monotonic effects, as quality upgrades may draw in more private-pay customers and reduce subsidized availability. A numerical study using Massachusetts childcare data demonstrates that our approach can increase voucher-based participation by over 10 percent without raising the overall budget. Our findings offer practical guidance for agencies seeking to expand access through targeted investments in private providers.

3 - Achieving Rawlsian Justice in Food Rescue

Gerdus Benade, Boston University, Boston, MA, United States, Aydin Alptekinoglu

We study a problem faced by a national food rescue platform that matches each donation to the first recipient who claims it. Recipients have very different response rates, leading to a few highly responsive recipients claiming the bulk of the donations. We ask whether priority lists, which control when the donation is announced to each recipient, are a remedy for inequitable outcomes. We give efficient algorithms to find the n-stage and binary priority lists that optimize a class of Rawlsian objective functions focusing on the worst-off recipients. The simple idea is to give higher priority to recipients who have received less in the past and to those who were slower in responding to notifications. This can be codified into an index by which to rank order eligible recipients. Computational experiments calibrated by historical data confirm that even binary priority lists lead to significantly more fair allocations than the existing first-come-first-serve allocation system.

SatA09

Cabinet Room

Advances on Convex and Conic Optimization I

Invited Session

Linear and Conic Optimization

Chair: Lijun Ding, University of California San Diego, La Jolla, CA, United States

Co-Chair: Yang Zheng, University of California San Diego, San Diego, CA, United States

Co-Chair: Feng-Yi Liao, University of California, San Diego, La Jolla, CA, United States

1 - Subgame Perfect Algorithms in Convex Optimization

Alex Wang, Purdue University, West Lafayette, IN, United States, Benjamin Grimmer, Kevin Shu

The study of convex optimization has historically been concerned with worst-case a priori convergence rates. The development of the Optimized Gradient Method (OGM), due to Drori and Teboulle, Kim and Fessler, marked a major milestone in this study, as OGM achieves the optimal worst-case convergence rate among all first-order methods in the smooth unconstrained setting. However, this notion of worst-case optimality is relatively coarse and allows OGM to perform at its worst-case rate even when better guarantees become dynamically available. This talk introduces a stronger notion of optimality for first-order methods that requires a method to give optimal dynamic guarantees that exploit "non-adversarialness" in the observed first-order information. We then give an algorithm which achieves this stronger optimality notion, the Subgame Perfect Gradient Method (SPGM), and illustrate its numerical performance.

This talk is based on joint work with Benjamin Grimmer (JHU) and Kevin Shu (CalTech).

2 - Recent Advance in Relative-type Inexact Proximal ALMs

Ling Liang, The University of Tennessee, Knoxville, Knoxville, TN, United States, Lei Yang, Jiayi Zhu, Kim-Chuan Toh

Inexact proximal augmented Lagrangian methods (pALMs) are particularly appealing for tackling convex constrained optimization problems because of their elegant convergence properties and strong practical performance. To solve the associated pALM subproblems, efficient methods such as Newton-type methods are essential. Consequently, the effectiveness of the inexact pALM hinges on the error criteria used to control the inexactness when solving these subproblems. However, existing inexact pALMs either rely on absolute-type error criteria (which may complicate implementation by necessitating the pre-specification of an infinite sequence of error tolerance parameters) or require an additional correction step to guarantee convergence when using relative-type error criteria (which can potentially degrade the practical performance of pALM). To address these deficiencies, this paper proposes ripALM, a relative-type inexact pALM, for linearly constrained convex optimization. This method can simplify practical implementation while preserving the appealing convergence properties of the classical absolute-type inexact pALM. To the best of our knowledge, ripALM is the first relative-type inexact version of the vanilla pALM with provable convergence guarantees. Numerical experiments demonstrate the competitive efficiency of the proposed method compared to existing state-of-the-art methods. As our model encompasses a wide range of application problems, the proposed ripALM offers broad applicability and has the potential to serve as a basic optimization toolbox.

3 - On Squared-Variable Formulations for Nonlinear Semidefinite Programming

Lijun Ding, University of California San Diego, La Jolla, CA, United States, Stephen Wright

We study squared-variable formulations for nonlinear semidefinite programming. We show an equivalence result of second-order stationary points of the nonsymmetric-squared-variable formulations and the nonlinear semidefinite programs. We also show that such an equivalence fails for the local minimizers and second-order stationary points of the symmetric-squared-variable formulations and the nonlinear semidefinite programs, correcting a false understanding in the literature and providing sufficient conditions for such a correspondence to hold.

4 - A Bundle-based Augmented Lagrangian Framework: Algorithm, Convergence, and Primal-dual Principles

Feng-Yi Liao, University of California, San Diego, La Jolla, CA, United States, Yang Zheng

We propose a new bundle-based augmented Lagrangian framework for solving constrained convex problems. Unlike the classical (inexact) augmented Lagrangian method (ALM) that has a nested double-loop structure, our framework features a single-loop process. Motivated by the proximal bundle method (PBM), we use a bundle of past iterates to approximate the subproblem in ALM to get a computationally efficient update at each iteration. We establish sub-linear convergences for primal feasibility, primal cost values, and dual iterates under mild assumptions. With further regularity conditions, such as quadratic growth, our algorithm enjoys linearconvergences. Importantly, this linear convergence can happen for a class of conic optimization problems, including semidefinite programs. Our proof techniques leverage deep connections with inexact ALM and primal-dual principles with PBM.

SatA10

Caucus Room

Recent Advances in GPU-Accelerated Optimization Algorithms

Invited Session

Computational Optimization

Chair: Inho Sin, Arizona State University, Tempe, AZ, United States

Co-Chair: Minseok Ryu, Arizona State University, Tempe, AZ, United States

Co-Chair: Inho Sin, Arizona State University, Tempe, AZ, United States

1 - GPU Implementation of Second-Order Linear and Nonlinear Programming Solvers

Sungho Shin, Massachusetts Institute of Technology, Cambridge, MA, United States, Alexis Montoison, François Pacaud, Mihai Anitescu

In recent years, GPU-accelerated optimization solvers based on second-order methods (e.g., interior-point methods) have gained momentum with the advent of mature and efficient GPU-accelerated direct sparse linear solvers, such as cuDSS. This paper provides an overview of the state of the art in GPU-based second-order solvers, focusing on pivoting-free interior-point methods for large and sparse linear and nonlinear programs. We begin by highlighting the capabilities and limitations of the currently available GPU-accelerated sparse linear solvers. Next, we discuss different formulations of the Karush-Kuhn-Tucker systems for second-order methods and evaluate their suitability for pivoting-free GPU implementations. We also discuss strategies for computing sparse Jacobians and Hessians on GPUs for nonlinear programming. Finally, we present numerical experiments demonstrating the scalability of GPU-based optimization solvers. We observe speedups often exceeding 10x compared to comparable CPU implementations on large-scale instances when solved up to medium precision. Additionally, we examine the current limitations of existing approaches.

2 - cuPDLPx: A Further Enhanced GPU-Based First-Order Solver for Linear Programming

Zedong Peng, MIT, Cambridge, MA, United States, Haihao Lu, Jinwen Yang

We introduce cuPDLPx , a further enhanced GPU-based first-order solver for linear programming. Building on the recently developed restarted Halpern PDHG for LP, cuPDLPx incorporates a number of new techniques, including a new restart criterion and a PID-controlled primal weight update. These improvements are carefully tailored for GPU architectures and deliver substantial computational gains. Across benchmark datasets, cuPDLPx achieves 2.5×–5× speedups on MIPLIB LP relaxations and 3×–6.8× on Mittelmann's benchmark set, with particularly strong improvements in high-accuracy and presolve-enabled settings. The solver is publicly available at https://github.com/MIT-Lu-Lab/cuPDLPx.

3 - Solving Large Multicommodity Network Flow Problems on GPUs

Fangzhao Zhang, Stanford University, Stanford, CA, United States, Stephen Boyd

We consider the all-pairs multicommodity network flow problem on a network with capacitated edges. The usual treatment keeps track of a separate flow for each source destination pair on each edge; we rely on a more efficient formulation in which flows with the same destination are aggregated, reducing the number of variables by a factor equal to the size of the network. Problems with hundreds of nodes, with a total number of variables on the order of a million, can be solved using standard generic interior point methods on CPUs; we focus on GPU-compatible algorithms that can solve such problems much faster, and in addition scale to much larger problems, with up to a billion variables. Our method relies on the primal-dual hybrid gradient algorithm, and exploits several specific features of the problem for efficient GPU computation. Numerical experiments show that our primal-dual multicommodity network flow method accelerates state of the art generic commercial solvers by 100× to 1000×, and scales to problems that are much larger.

4 - Disjunctive Benders Decomposition with a GPU-accelerated Oracle

Inho Sin, Arizona State University, Tempe, AZ, United States, Minseok Ryu, Geunyeong Byeon

We present a *Disjunctive Benders Decomposition* (DBD) framework augmented with GPU-accelerated cut-generating oracles. Unlike standard Benders cuts—which are typically tight only with respect to the LP relaxation of the Benders reformulation—DBD produces *split cuts* that are stronger and cut deeper relative to the convex hull. Leveraging recent advances in GPU-based large-scale convex optimization, we accelerate the cut-generation process within the DBD framework. The resulting GPU-integrated DBD is designed to enable efficient generation of split cuts while addressing the information loss inherent to decomposition and preserving its modularity benefits. Numerical experiments on large-scale capacitated facility location problem instances compare CPU and GPU implementations, within both BD and DBD frameworks.

SatA11

Charter Room

Topics in Optimization on Networks

Invited Session

Network Optimization

Chair: Alexander Vinel, Auburn University, Auburn, AL, United States

1 - Workload-based Constraints in Mixed-Integer Linear Programming Approaches for Resource Allocation in Project Networks

Norbert Trautmann, University of Bern, Bern, Switzerland, Nicklas Klein

In an activity-on-node project network, the nodes represent the activities of a project and the edges represent the prescribed precedence relations between these activities. Executing an activity requires a prescribed amount of time and specific resources of different types. At any given time, the total consumption of each resource type cannot exceed its capacity. The resource-constrained project scheduling problem (RCPSP) involves allocating resources to project activities over time in order to minimize time-to-market or project duration. Various formulations of RCPSP as a mixed-integer linear programming model have been discussed in the literature. We propose enhancing existing models by adding constraints that evaluate the workload to be processed before each activity begins. Our computational analysis of a standard test set from the literature confirms that these constraints considerably improve the relaxation-based lower bound.

2 - Trail-based Moving Target Search on a Network

Alexander Vinel, Auburn University, Auburn, AL, United States, Joseph Kennedy

This study addresses what we refer to as the Path-Constrained Trail-Based Moving Target Problem (PCTBMT), a novel variant of a search theory problem where the objective is to detect a moving target that leaves a trail of material, such as scent, by a search team equipped with a canine-like sensor. The problem accounts for the trail's decay over time, and the target's movement governed by a stochastic process over a pre-defined area. We adapt branch-and-bound methods developed in the existing literature for similar search problems. We also systematize the design of the existing longest-path-based bounds with a common framework and formulate a new bound. We also introduce and demonstrate the versatility of an Ant Colony Optimization (ACO) algorithm for heuristically solving the PCTBMT problem. To show the value of the ACO algorithm, we present several computational experiments and examine the performance of the methods considered. We provide results for simplistic search spaces, representing idealized areas, and for a larger-scale case instance inspired by a more realistic setting.

3 - Continuous Cubic Formulations for the Maximum 2-Plex Problem

Ashwini Ravindran, Texas A&M University, College Station, TX, United States, Sergiy Butenko

We study alternative continuous cubic formulations for a class of fundamental cluster-detection problems in networks, including the maximum clique, s-defective clique, s-plex, and the independent union of cliques problems. While our broader work pertains to each of these problems, this talk will focus in particular on the maximum 2-plex problem. These combinatorial problems, central to graph theory and network analysis, have wide-ranging applications in social network analysis, bioinformatics, and communication systems. The considered formulations capture the structural properties of these discrete problems while enabling their treatment within a continuous optimization framework. Each problem is modeled as the maximization of a degree-three multi-linear polynomial objective over a polyhedral feasible set defined by either the standard simplex or unit hypercube.

These continuous models provide new perspectives for theoretical analysis and algorithmic development. In particular, we derive analytical bounds that offer insight into the quality of feasible solutions and can be used to guide solution refinement. Additionally, we leverage the structure of the cubic formulations to design local and global solution methods based on existing nonlinear optimization solvers.

To assess the practical utility of our approach, we conduct a comprehensive computational study on standard benchmark instances. We compare the performance of multiple solvers in terms of solution quality, computational efficiency, and scalability. Our numerical results demonstrate the effectiveness of the proposed formulations in yielding high-quality solutions and suggest that continuous polynomial models can serve as a powerful alternative to traditional discrete or integer programming techniques for cluster detection in complex networks.

4 - Multi-stage Stochastic Planning of Mobile Additive Manufacturing Factories with Drone-enhanced Delivery

Daniela Granados Rivera, Auburn University, Auburn, AL, United States, Daniel Silva, Alice Smith

Mobile additive manufacturing (MAM) factories are beginning to be widely explored as a decentralized system to supply spare parts in a flexible and timely manner. MAM factories can be located close to customers, where parts may be delivered by drone according to the urgency of the demand. However, relocation of MAM factories is generally confined to pre-determined sites. Due to weather, insecurity politically and militarily, and disasters, site locations can change their availability over time. These variations are considered a disruption to the supply that, to the best of our knowledge, has not been explored in the literature. Consequences can be high costs to unfulfilled demand, loss of a factory, or even loss of life.. Considering this uncertainty and the mobile nature of the MAM factories, we propose a multi-stage stochastic programming model where the availability of the locations is defined with some probability of occurrence. These are defined using a scenario tree considering states and transition probabilities. Preliminary results reveal the complexity of the model based on the tree expansion and the transition probabilities.

March 21, 2026, 10:00 AM - 10:45 AM

Plenary 3

Grand Ballroom

Paths to Minima Selection

Plenary

Plenary

Chair: Nikolaos Sahinidis, Georgia Tech, Atlanta, GA, United States

1 - Paths to Minima Selection

Niao He, ETH Zürich, Zurich, Switzerland

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.

March 21, 2026, 10:45 AM - 11:30 AM

Plenary 4

Grand Ballroom

Data-Driven Approaches to Enhance Branch-and-Cut: Guarantees and Pitfalls

Plenary

Plenary

Chair: Oktay Gunluk, Georgia Tech, Atlanta, GA, United States

1 - Data-Driven Approaches to Enhance Branch-and-Cut: Guarantees and Pitfalls

Amitabh Basu, Johns Hopkins University, Baltimore, MD, United States

Branch-and-cut is arguably the most effective algorithm to solve large scale mixed-integer optimization problems in practice. In spite of tremendous progress in both the mathematical underpinnings and the computational insights required to deploy branch-and-cut in practice, several aspects of this important algorithm remain poorly understood. In particular, several crucial decisions (e.g., cutting plane or branching variable selection) that need to be made during the execution of branch-and-cut lack a satisfactory foundation, and clear recipes are difficult to formulate given the wide diversity of instances that these algorithms need to tackle. This has led several researchers to explore data driven, machine learning tools to help guide these decisions. The philosophy is simple: the decisions that work well for a given sample of instances should have good performance on a larger class of instances, assuming the sample is "representative" of the larger class. In this talk, we explore the foundational principles that formalize this idea, focusing on two issues: 1) The size of the sample required for high quality performance on unseen instances, and 2) Pitfalls to avoid when selecting a particular machine learning paradigm for branch-and-cut. Both these lines of investigations uncover insights into mixed-integer optimization which we believe are useful beyond the principled use of machine learning for branch-and-cut.

March 21, 2026, 12:45 PM - 2:15 PM

SatB01

Grand Ballroom

Applications of Stochastic Programming in Machine Learning

Invited Session

Optimization in Data Science

Chair: Anton Malandii, Brown University, Providence, RI, United States

1 - Understanding Measure Consistency Regularization via Neural Net Distance

Yinsong Wang, Georgia Institute of Technology, Atlanta, GA, United States, Shahin Shahrampour

The problem of corrupted data, missing feature or missing modality is constantly plaguing modern machine learning landscape. To address this issue, a class of regularization methods that enforce the consistency between imputed data and fully observed data has emerged as a promising technique to enhance model generalization, particularly when dealing with partially observed data. We refer to this class of methods, \textbf{Measure Consistency Regularization (MCR)}. Despite its empirical success in various applications, such as image inpainting, data imputation and general semi-supervised learning, a fundamental understanding of its theoretical underpinnings remains limited. This paper aims to bridge this gap by providing theoretical insights into why, when, and how MCR, when combined with neural net distance, improves imputation quality in the presence of partially observed data.

Our theoretical analysis pinpoints the term responsible for the MCR generalization advantage, and we extend the theoretical analysis into imperfect training regime to demonstrate that the advantage of MCR is not always guaranteed. Following our theoretical findings, we propose a novel training protocol that monitors duality gap to inform the earliest stopping point maintaining generalization advantage. We provide detailed empirical evidence to support our theoretical claims and to show the effectiveness and accuracy of our proposed stopping condition. We further provide a set of real-world data simulations to show the versatility of MCR under different model architectures designed for different data sources.

2 - Approximating Rockafellians Mitigate Distributional Perturbations: Discontinuous Integrands and Chance-Constrained Applications

Johannes Royset, University of Southern California, Los Angeles, CA, United States, Lai Tian

In this presentation, we show how approximating Rockafellians serve as a principled and effective alternative for improving the stability of stochastic programs under distributional changes. Unlike previous efforts that focus on special distributions and continuous integrands, our results accommodate general probability distributions and discontinuous integrands. Thus, our results apply to chance-constrained programs, for which we obtain improved qualitative and quantitative stability results under weaker assumptions pertaining to metric subregularity and upper outer-Minkowski content.

3 - Failure of Uniform Laws of Large Numbers for Subdifferentials and Beyond

Lai Tian, University of Southern California, Los Angeles, CA, United States, Johannes Royset

We provide counterexamples showing that uniform laws of large numbers do not hold for subdifferentials under natural assumptions. Our results apply to random Lipschitz functions and random convex functions with finite number of smooth pieces. Consequently, they resolve in the negative the questions posed by Shapiro and Xu [J. Math. Anal. Appl., 325(2), 2007] and highlight the obstacles nonsmoothness poses to uniform results.

4 - Mixed-Integer Programming for Changepoint Detection

Apoorva Narula, Georgia Institute of Technology, H. Milton Stewart School of Industrial and Systems Engineering, Atlanta, GA, United States, Yao Xie, Santanu S. Dey

We study continuous piecewise linear (PWL) fitting as a fundamental framework for offline changepoint detection and develop strengthened mixed-integer programming formulations based on an extended representation of the standard segment-assignment variables. A key theoretical result shows that the linear programming relaxation admits an integral projection onto these variables, yielding tighter relaxations than existing models. Experiments across a diverse set of datasets demonstrate markedly faster solves for L1 fitting and competitive performance for L2 fitting. We further extend the framework to multidimensional PWL models with shared breakpoints and to sparse changepoint detection that identifies which coordinates change, highlighting that our extended formulation can be leveraged to enhance both statistical interpretability and computational performance.

SatB02

Georgian Room

Advances in Nonlinear Optimization Methods: Convex Optimization and Beyond

Invited Session

Optimization in Data Science

Chair: Yuyuan Ouyang, Clemson University, Clemson, SC, United States

Co-Chair: Yan Wu, Clemson University, Clemson, SC, United States

Co-Chair: Qi Luo, University of Iowa, Iowa City, IA, United States

1 - An Invitation to Higher-Order Riemannian Optimization

David Gutman, Texas A&M University, College Station, TX, United States, George Lobo

In this talk, we propose a class of implementable, optimal-rate methods for nonconvex optimization over Riemannian manifolds when derivative oracles up to order p are available. Our analysis proves that, under this regime, Riemannian optimization is no harder than Euclidean optimization: the Riemannian convergence rates match the optimal Euclidean rates. Furthermore, we comprehensively characterize how the regularity of the retraction and objective function jointly impacts that of the pullback functions. The engine driving this characterization is a novel Faà di Bruno–type formula for pullback derivatives, constructed using the rarely applied higher-order covariant differential calculus. Moreover, we show that the entire class of efficient retractions of Gawlik *et al*. can be applied within these methods.

Our theoretical framework yields implementable third-order methods when equipped with suitable subproblem solvers. The mechanism enabling implementation is a symmetric Krylov tensor subspace method. Building on the Savas–Eldén approach, it can hybridize with recent techniques for minimizing quartically regularized third-order polynomials.

2 - Metric Entropy-Free Sample Complexity Bounds for Sample Average Approximation in Convex Stochastic Programming

Jindong Tong, University of Florida, Gainesville, FL, United States, Hongcheng Liu

This work studies sample average approximation (SAA) in solving convex or strongly convex stochastic programming (SP) problems. In estimating SAA's sample efficiency, existing bounds often entail metric entropy terms, which often grow polynomially with problem dimensionality. Although metric entropy-free complexity results are known, they based on uniform Lipschitz conditions, which can be overly critical in many settings. Our work presents metric entropy-free sample complexity results for the SAA without uniform Lipschitz condition. The new results often lead to an $O(d)$-improvement in the complexity rate than the state-of-the-art, which also shows the matching sample complexity of SAA and SMD (stochastic mirror descent). Our numerical experiment results on SAA for solving a simulated SP problem align with our theoretical findings.

3 - Multi-cut Stochastic Approximation Methods for Solving Stochastic Convex Composite Optimization

Honghao Zhang, Georgia Institute of Technology , Atlanta, GA, United States, Renato D C Monteiro, Jiaming Liang

The development of a multi-cut stochastic approximation (SA) method for solving stochastic convex composite optimization (SCCO) problems has remained an open challenge. The difficulty arises from the fact that the stochastic multi-cut model, constructed as the pointwise maximum of individual stochastic linearizations, provides a biased estimate of the objective function, with the error being uncontrollable. This paper introduces multi-cut SA methods for solving SCCO problems, achieving near-optimal convergence rates. The cutting-plane models used in these methods are the pointwise maxima of appropriately chosen one-cut models. To the best of our knowledge, these are the first multi-cut SA methods specifically designed for SCCO problems. Finally, computational experiments demonstrate that these methods generally outperform both the robust stochastic approximation method and the stochastic dual averaging method across all instances tested.

4 - On the Analysis of Misspecified Variational Inequalities with Nonlinear Constraints

Novel Kumar Dey, University of Arizona, Tucson, AZ, United States, Mohammad Mahdi Ahmadi, Erfan Yazdandoost Hamedani, Afrooz Jalilzadeh

In this paper, we study a class of misspecified variational inequalities (VIs) where both the monotone operator and nonlinear convex constraints depend on an unknown parameter learned via a secondary VI. Existing data-driven VI methods typically follow a decoupled learn-then-optimize scheme, causing the approximation error from the learning to propagate the main decision-making problem and hinder convergence. We instead consider a simultaneous approach that jointly solves the main and secondary VIs. To efficiently handle nonlinear constraints with parameter misspecification, we propose a single-loop inexact Augmented Lagrangian method that simultaneously updates the primal decision variables, dual multipliers, and the misspecified parameter. The method combines a forward-reflected-backward step with an Augmented Lagrangian penalty, and explicitly handles misspecification on both the operator and constraint functions. Moreover, we introduce a relaxed performance metric based on the Minty VI gap combined with an aggregated infeasibility metric. By proving boundedness of the dual iterates, we establish O(1/K) ergodic convergence rates for these metrics. Numerical Experiments are provided to showcase the superior performance of our algorithm compared to state-of-the-art methods.

SatB03

Plaza I

Recent Advances in Derivative-Free Optimization I

Invited Session

Nonlinear Optimization

Chair: Zikai Xiong, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Abraar Chaudhry, Georgia Tech, Atlanta, GA, United States

Co-Chair: Katya Scheinberg, Georgia Tech, Atlanta, GA, United States

1 - Derivative-free Mirror Markov Chain Monte Carlo

Raghu Pasupathy, Purdue University, West Lafayette, IN, United States, Antik Chakraborty, Ruiting Tong

Constrained sampling --- the problem of sampling from a density corresponding to a potential function that is supported on a constrained subset of a normed space --- arises prominently in Bayesian inference, uncertainty quantification, statistical estimation, and modern machine learning. The derivative-free version of this problem stipulates that we have access only to a zero-order oracle of the potential function. A leading recent approach, *Mirror MCMC*, operates by (i) mapping points from the constrained (primal) space to an unconstrained dual space via the gradient of a mirror map, (ii) applying a standard unconstrained MCMC update in the dual space, and finally (iii) mapping back to the primal space via the inverse mirror map. In this talk, we develop the derivative-free version of Mirror MCMC. After showing that the Mirror MCMC iterates admit an equivalent *primal* representation where each update corresponds to the solution of a stochastic mirror-optimization subproblem, we construct an implementable inexact sequence using the zero-order oracle. We then establish sufficient conditions under which the resulting derivative-free sampler achieves *condition-number–free* mixing rates, and characterize how much inexactness can be tolerated in the subproblem solves without degrading these rates. For furthering intuition, we will illustrate with examples.

2 - ComPO: Preference Alignment via Comparison Oracles

Tianyi Lin, Columbia University, New York, NY, United States

Direct alignment methods are increasingly used for aligning large language models (LLMs) with human preferences. However, these methods suffer from the likelihood displacement, which can be driven by noisy preference pairs that induce similar likelihood for preferred and dis-preferred responses.To address this issue, we consider doing derivative-free optimization based on comparison oracles. First, we propose a new preference alignment method via comparison oracles and provide convergence guarantees for its basic mechanism. Second, we improve our method using some heuristics and conduct the experiments to demonstrate the flexibility and compatibility of practical mechanisms in improving the performance of LLMs using noisy preference pairs. Evaluations are conducted across multiple base and instruction-tuned models with different benchmarks. Experimental results show the effectiveness of our method as an alternative to addressing the limitations of existing methods. A highlight of our work is that we evidence the importance of designing specialized methods for preference pairs with distinct likelihood margin.

3 - Function-free Optimization via Comparison Oracles

Zikai Xiong, Georgia Institute of Technology, Atlanta, GA, United States, Katya Scheinberg

In this work, we study optimization specified only through a comparison oracle: given two points, it returns which one is preferred. We call it function-free optimization because we do not assume access to, nor the existence of, a canonical objective function; instead, our goal is to find points that are not preferred over by any other point, which we denote as the optimal solutions. This model arises in preference- and ranking-based settings, where the objective values and derivatives are unavailable, meaningless, or non-identifiable. We develop a systematic, analytical, and algorithmic framework based on the geometry of the preference level set, as it remains well-defined from comparisons alone. We introduce a new optimality measure, called the “level set optimality gap,” which denotes the distance between the level set and the optimal solutions. We then propose a method for estimating normal directions to level sets with complexity $O(d \cdot \log(1/

\epsilon))$, and show that the method is nearly optimal. Leveraging these normal direction estimates, we show that a normal direction descent method reaches level-set optimality gap at most ϵ in $O(d/\epsilon^2)$ iterations when the level sets are smooth and convex. The underlying representative function of our algorithm may be nonsmooth, nonconvex, or even discontinuous. Since prior knowledge in practical applications of function-free optimization is usually very limited, we also give parameter-free schemes for normal direction estimation and solving the optimization problem with nearly the same order of complexity.

4 - Reducing Sample Complexity in Stochastic Derivative-Free Optimization via Tail Bounds and Hypothesis Testing.

Luis Nunes Vicente, Lehigh University, Bethlehem, PA, United States, Anjie Ding, Francesco Rinaldi, Damiano Zeffiro

We introduce and analyze new probabilistic strategies for enforcing sufficient decrease conditions in stochastic derivative-free optimization, with the goal of reducing sample complexity and simplifying convergence analysis. First, we develop a new tail bound condition imposed on the estimated reduction in function value, which permits flexible selection of the power used in the sufficient decrease test, q in (1,2]. This approach allows us to reduce the number of samples per iteration from the standard O(delta^{−4}) to O(delta^{-2q}), assuming that the noise moment of order q/(q-1) is bounded. Second, we formulate the sufficient decrease condition as a sequential hypothesis testing problem, in which the algorithm adaptively collects samples until the evidence suffices to accept or reject a candidate step. This test provides statistical guarantees on decision errors and can further reduce the required sample size, particularly in the Gaussian noise setting, where it can approach O(delta^{−2-r}) when the decrease is of the order of delta^r. We incorporate both techniques into stochastic direct-search and trust-region methods for potentially non-smooth, noisy objective functions, and establish their global convergence rates and properties.

SatB04

Plaza II

Stochastic Optimization Theory and Application

Invited Session

Optimization under Uncertainty

Chair: Zhaolin Hu, Tongji University, Shanghai, N/A

1 - Decision-Focused Learning under Stochastic Bilevel Optimization

Siqi Zhang, Nanjing University, Nanjing, China, People's Republic of

Decision-focused learning (DFL) seeks to train predictive models explicitly for the downstream optimization problems, and it has attracted increasing attention in recent years. In this talk, we study the DFL problem through the lens of stochastic bilevel optimization, motivated by a generalized assignment problem. We will discuss several approaches to solving the problem, present empirical results that highlight practical challenges arising from uncertainty and model misspecification. Our results suggest a broader research direction at the intersection of DFL, stochastic bilevel optimization and related topics.

2 - Power Distribution Systems under Wildfire Risks: Chance-Constrained Model with Decision-Dependent Probabilities

Sainan Zhang, George Washington University, Washington, DC, United States, Miguel Lejeune, Payman Dehghanian

Wildfires pose a significant threat to the reliability and security of power distribution systems where wildfire-induced power line failures exhibit spatial correlations and are influenced by operational decisions. To address this challenge, we introduce a new joint chance-constrained stochastic programming model that accounts for both spatial dependency and decision-dependent uncertainty. Power line failures are affected by wildfires and their propagation (exogenous uncertainty) and by power flow decisions (decision-dependent uncertainty). The model simultaneously reconfigures the network topology in the face of wildfire ignition risks and minimizes the total costs associated with power generation, switching actions, and penalties for power imbalance while ensuring the operational reliability of the network. We incorporate an Archimedean copula to capture spatial dependency across power line failures, and introduce a decision-dependent distortion function to model how power flows impact line failure probabilities. We develop a scenario generation method providing scenarios representative of how wildfires affect power line failures and derive a deterministic and equivalent mixed-integer nonlinear programming reformulation of the chance-constrained model. Additionally, we approximate the truncated generator of the parameterized copula using a piecewise linear function in order to obtain a tractable mixed-integer linear programming formulation. The numerical tests showcase the applicability and effectiveness of the proposed reformulation and algorithmic approach, which outperforms alternative modeling approaches by (i) reducing the out-of sample loss of loads, (ii) increasing the reliability of the network, (iii) lowering the loading percentage of power lines in the most vulnerable areas, and (iv) decreasing the probability of electrically-induced wildfires.

3 - Schroudinger Bridge as Robustified Optimal Transport Flows

Jinxin Wang, Booth School of Business, the University of Chicago, Chicago, IL, United States, Ya-Ping Hsieh, Bahar Taskesen

Generative modeling has seen remarkable progress through two main paradigms: deterministic flow-based methods and diffusion models. While flow-based approaches offer efficient training and fast sampling, enabling scaling to extremely large architectures, diffusion models consistently achieve superior sample quality under comparable parameter budgets. In this work, we investigate the theoretical source of this discrepancy and argue that the key lies in the sampling trajectories rather than marginal distributions alone. Our main result shows that the sampling trajectories of diffusion models can be understood as first identifying the most efficient flows under the Wasserstein loss and then implicitly robustifying these OT-based flows via solving a distributionally robust variant of the Kantorovich dual problem. This perspective naturally explains their improved generalization and stability under noisy or imperfect training, and suggests broader implications for robust generative modeling.

4 - Robust Simulation with Distribution Parameter Uncertainty

Zhaolin Hu, Tongji University, Shanghai, China, People's Republic of, Lewei Shi, Ying Zhong

We study robust simulation for the complex system where the aim is to assess the worst-case (maximal) expectation of the system output when there is uncertainty on the parameters of the underlying distribution. We derive bounds for the worst-case expectation and also build the statistical confidence for the bounds. To derive the upper bound, we approximate the parameter uncertainty with distributional uncertainty involving statistical divergence and compute the worst-case expectation under distributional uncertainty. To obtain the lower bound, we partition the parameter space and implement the ranking and selection procedures. We conduct numerical experiments to show the effectiveness of our approach. We also conduct robust simulation for the Dynamic Integrated model of Climate and the Economy (DICE), a famous integrated assessment model in climate change, to show the practical application of our approach.

SatB05

Plaza III

Stochastic and Robust Approaches with Contextual Optimization and Decision-Dependent Uncertainties

Invited Session

Optimization under Uncertainty

Chair: Beste Basciftci, Tippie College of Business, University of Iowa, Iowa City, IA, United States

1 - Residuals-Based Contextual Distributionally Robust Optimization with Decision-Dependent Uncertainty

Guzin Bayraksan, The Ohio State University, Columbus, OH, United States, Zoey Zhu, Xian Yu

We present a methodology to handle decision dependency via a residuals-based distributionally robust optimization model, where the underlying uncertainty depends on both covariate information and the decisions. We adopt regression models to learn the latent decision dependency and construct a nominal distribution (thereby ambiguity sets) around the learned model using empirical residuals from the regressions. Ambiguity sets can be formed via the Wasserstein distance, a sample robust approach, or with the same support as the nominal empirical distribution (e.g., phi-divergences), where both the nominal distribution and the radii of the ambiguity sets could be decision- and covariate-dependent. We establish conditions under which desired statistical properties, such as asymptotic optimality, rates of convergence, and finite sample guarantees, are satisfied. We present decomposition-based solution methods to efficiently solve the resulting (typically nonconvex) models under parametric and nonparametric regression models and devise data driven cross-validation schemes to size the ambiguity sets. Through numerical experiments, we illustrate the effectiveness of our approach and the benefits of integrating decision dependency into a residuals-based contextual DRO framework.

2 - Distributionally Robust Hydrogen Network Expansion Planning under Decision-dependent Uncertainty

Beste Basciftci, Tippie College of Business, University of Iowa, Iowa City, IA, United States

Transition to a green hydrogen economy faces a critical challenge known as the `chicken-and-egg dilemma', wherein establishing a hydrogen supply network relies on demand, while demand only grows with reliable supply. In addition, as the hydrogen market is in the early stage, predicting demand distributions is challenging due to lack of data availability. This talk addresses these complex issues through a risk-averse framework with the introduction of a distributionally robust hydrogen network expansion planning problem under decision-dependent demand ambiguity. The problem optimizes location and production capacity decisions of the suppliers considering the moments of the stochastic hydrogen demand as a function of these investment decisions. To obtain tractable representations of this problem, we derive a generic reformulation under polyhedral support sets and further derive tailored reformulations for box-type and discrete hydrogen demand support sets under various forms of decision dependencies. To efficiently solve the reformulations, we develop a tailored algorithm based on the column-and-constraint generation approach, and enhance the computational performance through solving the master problems to a relative optimality gap, decomposing the subproblems, and integrating pre-generated columns and constraints. Comparative experiments demonstrate the superiority of our algorithm over the classical column-and-constraint generation algorithm and cutting-plane approaches with significant speedups. To validate the effectiveness of our approach in decision-making, we investigate a real case study leveraging data from the Hydrogen Energy Applications in Valley Environments for Northern Netherlands (HEAVENN) project, providing critical insights for policymakers based on the degree of decision dependency.

3 - Tractable Mixed-Integer Programming for Contextual Stochastic Optimization with Decision-Dependent Uncertainty via Nonparametric Regression

Xian Yu, The Ohio State University, Columbus, OH, United States, Huangrong Sun

We study a general data-driven contextual stochastic optimization problem where the uncertainty depends on both exogenous contextual information and endogenous decisions to be optimized. To learn this complex relationship, we employ nonparametric regression models (e.g., k-nearest neighbors, classification and regression trees, and neural networks) and embed them into the downstream optimization model using the empirical residuals-based decision-dependent sample average approximation (ER-DD-SAA) framework. Since these nonparametric regression models often lack a simple functional form, we derive exact mixed-integer linear and nonlinear programming formulations for the resulting optimization problems. We establish the consistency and asymptotic optimality of the ER-DD-SAA framework with these nonparametric regression models under mild assumptions. To validate our proposed framework, we consider a newsvendor problem and conduct experiments on a synthetic dataset using the proposed formulations. We compare the results with those obtained by using linear regression models. The results demonstrate the superior performance of our approach when incorporating nonparametric regression models.

4 - Deep Learning-Driven Contextual Stochastic Optimization for Real-Time Order Fulfillment

Tinghan Ye, Georgia Institute of Technology, Atlanta, GA, United States, Shuaicheng Tong, Changkun Guan, Beste Basciftci, Pascal Van Hentenryck

Order fulfillment optimization is a fundamental challenge in large-scale e-commerce, requiring real-time decisions for every incoming order. For enterprises with extensive fulfillment networks, selecting the optimal fulfillment plan demands balancing operational costs with strict service-level guarantees under uncertainty. To model this problem, this work introduces a two-stage contextual stochastic optimization framework explicitly capturing two sources of uncertainty, delivery timeliness and future inventory consumption. To enable real-time deployment in peak hours, where traditional solvers are computationally prohibitive, an optimization proxy is developed, training deep neural networks to approximate solutions of the underlying stochastic program with high fidelity. Computational experiments on a large-scale JD.com transactional dataset demonstrate that the proposed approach achieves orders-of-magnitude speedups compared to a state-of-the-art commercial solver while preserving similar solution quality. The results establish a scalable paradigm for real-time stochastic optimization in e-commerce logistics, bridging rigorous optimization with deep learning to deliver industrial-scale efficiency.

SatB06

Director’s and Lounge

Nonlinear Optimization and Applications

Invited Session

Global Optimization

Chair: Bissan Ghaddar, Ivey Business School, London, ON, Canada

1 - MIP Restrictions for nonconvex MINLPs

Hassan Hijazi, Gurobi Optimization, Los Alamos, NM, United States

MIP restrictions are a promising tool for finding good quality feasible solutions to hard non-convex MINLPs. The feasible region of the restriction is guaranteed to be a subset of the original problem's feasible region. This property guarantees feasibility of the solutions returned during the Branch & Bound exploration, which is crucial for applications with tight time limits.
In this talk, we will define MIP restrictions and discuss ways of automatically generating them for generic MINLPs. We will also look at specific use cases such as models drawn from the ARPAe Grid Optimization competition.

2 - PolyOpt.jl: A Polynomial Optimization Tool

Youssouf Emine, Ivey Business School, London, ON, Canada, Bissan Ghaddar

Hierarchies of semidefinite programming (SDP) relaxations based on sums of squares (SOS) polynomials provide asymptotically exact approximations for general polynomial optimization problems (POPs). Yet, their applicability to large-scale instances remains limited by the high computational cost of solving SDPs. Recent work has therefore focused on second-order cone programming (SOCP) hierarchies obtained by restricting the SOS condition, offering a more scalable alternative. In this work, we investigate new ways to leverage these SOCP restrictions and show that they can serve as practical alternatives to SDP relaxations while achieving comparable performance. We further present results on a collection of benchmark instances arising from energy applications to illustrate the effectiveness of these methods in practice. These developments will be gathered into a unified tool, **PolyOpt.jl**, a Julia package that integrates the various relaxation hierarchies and enhaces them with sparsity-exploiting techniques to enable scalable optimization of large, structured POPs.

3 - Storage Participation in Electricity Markets: Arbitrage and Ancillary Services

Dirk Lauinger, MIT, Cambridge, MA, United States, Luc Cote, Xu Sun

Electricity storage is used for intertemporal price arbitrage and for ancillary services that balance unforeseen supply and demand fluctuations via frequency regulation. We present an optimization model that computes bids for both arbitrage and frequency regulation and ensures that storage operators can honor their market commitments at all times for all fluctuation signals in an uncertainty set inspired by market rules. This requirement, initially expressed by an infinite number of nonconvex functional constraints, is shown to be equivalent to a finite number of deterministic constraints. The resulting formulation is a mixed-integer bilinear program that admits mixed-integer linear relaxations and restrictions. Empirical tests on European electricity markets show a negligible optimality gap between the relaxation and the restriction. The model can account for intraday trading and, with a solution time of under 5 seconds, may serve as a building block for more complex trading strategies. Such strategies become necessary as battery capacity exceeds the demand for ancillary services. In a backtest from 1 July 2020 through 30 June 2024 joint market participation more than doubles profits and almost halves energy storage output compared to arbitrage alone.

4 - Hierarchies of QCQP Relaxations

Dahye Han, Georgia Institute of Technology, Atlanta, GA, United States, Santanu Dey, Jingye Xu

The tightness of convex relaxations directly affects the quality of dual bounds for QCQPs. Several relaxation methods are available, including semidefinite programming (SDP), the reformulation-linearization technique (RLT), and classical Lagrangian relaxation. Another approach is Lagrangian relaxation based on variable splitting, where copies of variables are introduced for each quadratic constraint. In this work, we establish a theoretical hierarchy among these relaxations for QCQPs with bounded variables and study the computational efficiency. Specifically, we show that Lagrangian relaxation with variable splitting gives tighter bounds than traditional Lagrangian relaxation, which in turn is also tighter than the combination of SDP and RLT. However, tighter convex relaxations do not always mean better computational efficiency. We examine the trade-off between the quality of bounds and the cost of computation to offer practical advice.

SatB07

Congress Room

Algorithmic Advances in Discrete and Data-Driven Optimization

Invited Session

Discrete Optimization

Chair: Lin An, Carnegie Mellon University, Pittsburgh, PA, United States

1 - Minimum Cut Representability of Stable Matching Problems

Chengyue He, Columbia University, New York, NY, United States, Yuri Faenza, Ayoub Foussoul

We introduce and study *Minimum Cut Representability*, a framework to solve optimization and feasibility problems over stable matchings by representing them as minimum s-t cut problems on digraphs over rotations. We provide necessary and sufficient conditions on objective functions and feasibility sets for problems to be minimum cut representable. In particular, we define the concepts of first and second order differentials of a function over stable matchings and show that a problem is minimum cut representable if and only if, roughly speaking, the objective function can be expressed solely using these differentials, and the feasibility set is a sublattice of the stable matching lattice. To demonstrate the practical relevance of our framework, we study a range of real-world applications, including problems involving school choice with siblings and a two-stage stochastic stable matching problem. We show how our framework can be used to help solve these problems.

2 - Near-Optimal Real-Time Recommendation with Simple Transformers

Lin An, Carnegie Mellon University, Pittsburgh, PA, United States, Andrew Li, Vaisnavi Nemala, Gabriel Visotsky

Real-time personalization has advanced significantly in recent years, with platforms utilizing machine learning models to predict user preferences based on rich behavioral data on each individual user. Traditional approaches usually rely on embedding-based machine learning models to capture user preferences, and then reduce the final optimization task to nearest-neighbors, which can be performed extremely fast. However, these models struggle to capture complex user behaviors, which are essential for making accurate recommendations. Transformer-based models, on the other hand, are known for their practical ability to model sequential behaviors, and hence have been intensively used in personalization recently to overcome these limitations. However, optimizing recommendations under transformer-based models is challenging due to their complicated architectures. In our paper, we address this challenge by considering a specific class of transformers, showing its ability to represent complex user preferences, and developing efficient algorithms for real-time personalization.
We focus on a particular set of transformers, called simple transformers, which contain a single self-attention layer. We show that simple transformers are capable of capturing complex user preferences. We then develop an algorithm that enables fast optimization of recommendation tasks based on simple transformers. Our algorithm achieves near-optimal performance in sub-linear time. Finally, we demonstrate the effectiveness of our approach through an empirical study on datasets from Spotify and Trivago. Our experiment results show that (1) simple transformers can model/predict user preferences substantially more accurately than non-transformer models and nearly as accurately as more complex transformers, and (2) our algorithm completes simple-transformer-based recommendation tasks quickly and effectively.

3 - Transformer-Based Next-Step Prediction for Queue Length Distribution

Jieqi Di, Georgia Institute of Technology, Atlanta, GA, United States, Jiecheng Lu, Runhua Wu, Yuwei Zhou

Traditional approaches to model queueing dynamics from real-world setting rely on model selection, parameter estimation, and simulation. These methods struggle with complexities like time-varying arrivals, customer abandonment, and routing interactions, while parameter estimation errors and time-consuming simulations further limit their predictive accuracy and increase their costs.
We propose a data-driven alternative that learns queueing dynamics directly from historical data using decoder-only Transformers. Our approach uses the autoregressive nature of queueing systems, where future states depend causally on past sequences through recursions. We treat queue evolution as a sequence prediction problem: inputting queue length distributions and predicting next-step distributions. Experiments on synthetic data demonstrate that our Transformer-based model successfully reproduces queueing dynamics across different queueing models and outperforms traditional sequence models like Recurrent Neural Networks (RNNs). This model-free approach eliminates the need for explicit model specification and parameter estimation while maintaining accuracy in capturing complex system behaviors. Future work will validate this methodology on real-world ride-hailing datasets.

4 - Weighted Chairman Assignment and Flow-Time Scheduling

Siyue Liu, Carnegie Mellon University, Pittsburgh, PA, United States, Victor Reis

Given positive integers m, n, a fractional assignment $x \in [0,1]^{m \times n}$ and weights $d \in \mathbb{R}^n_{>0}$, we show that there exists an assignment $y \in \{0,1\}^{m \times n}$ so that for every $i \in [m]$ and $t \in [n]$,

$$\left| \sum_{j \in [t]} d_j (x_{ij} - y_{ij}) \right| < \max_{j \in [n]} d_j \,.$$

This generalizes a result of Tijdeman (1973) on the unweighted version, known as the chairman assignment problem. This also confirms a special case of the single-source unsplittable flow conjecture with arc-wise lower and upper bounds due to Morell and Skutella (IPCO 2020). As an application, we consider a scheduling problem where jobs have release times and machines have closing times, and a job can only be scheduled on a machine if it is released before the machine closes. We give a 33-approximation algorithm for maximum flow-time minimization.

SatB08

Committee Room

Interplay of Optimization and Statistics

Invited Session

Emerging Applications of Optimization

Chair: Yifan Hu, Rutgers, Piscataway, NJ, United States

Co-Chair: Xiang Li, N/A

1 - Online Covariance Estimation in Nonsmooth Stochastic Approximation

Abhishek Roy, Texas A&M University, College Station, TX, United States

We consider applying stochastic approximation (SA) methods to solve nonsmooth variational inclusion problems.
Existing studies have shown that the averaged iterates of SA methods exhibit asymptotic normality, with an optimal limiting covariance matrix in the local minimax sense of H\'{a}jek and Le Cam. However, no methods have been proposed to estimate this covariance matrix in a nonsmooth and potentially non-monotone (nonconvex) setting.
In this paper, we study an~online batch-means covariance matrix estimator introduced in \cite{zhu2023online}. The estimator groups the SA iterates appropriately and computes the sample covariance among batches as an estimate of the limiting covariance. Its construction does not require prior knowledge of the total sample size, and updates can be performed recursively as new data arrives.
We establish that, as long as the batch size sequence is properly specified (depending on the stepsize sequence), the estimator achieves a convergence rate of order $O(\sqrt{d}n^{-1/8+\varepsilon})$ for any $\varepsilon>0$, where d and n denote the problem dimensionality and the number of iterations (or samples) used. Although the problem is nonsmooth and potentially non-monotone (nonconvex), our convergence rate matches the best-known rate for covariance estimation methods using only first-order information in smooth and strongly-convex settings. The consistency of this covariance estimator enables asymptotically valid statistical inference, including constructing confidence intervals and performing hypothesis testing.

2 - Online Statistical Inference of Newton Methods with Nesterov's Accelerated Sketching

Xinchen Du, Georgia Institute of Technology, Atlanta, GA, United States

The widespread availability of streaming data has made online algorithms fundamental for parameter estimation, with second-order methods offering notable advantages in numerical stability. In this work, we investigate an online sketched Newton method that employs a sketching technique—the accelerated sketch-and-project algorithm with Nesterov's momentum—to approximate the Newton direction at each iteration, thereby mitigating the computational bottleneck inherent in classical second-order methods. Compared with the standard sketch-and-project algorithm, the accelerated randomized sketching framework has been shown to achieve both superior convergence rates and lower computational costs. Building on this design, we first establish the asymptotic normality of the averaged iterates and characterize the corresponding limiting covariance matrix. We demonstrate that this limiting covariance reduces to that of the standard sketch-and-project Newton method under a particular choice of the acceleration parameters. Furthermore, rather than directly estimating the limiting covariance matrix, we study an online inference procedure called random scaling. Specifically, we construct a pivotal test statistic by appropriately rescaling the averaged iterates so that its limiting distribution is parameter-free, enabling asymptotically valid online statistical inference.

3 - Estimate–Then–Optimize for Contextual Chance Constraints: A Unified Framework with Finite-Sample Guarantees

Mohamed el amine Lakhnichi, Columbia University, NYC, NY, United States, Henry Lam

We present an Estimate–Then–Optimize (ETO) framework for contextual chance-constrained decision problems. The method first fits a conditional simulator from data and then solves the downstream chance-constrained program using samples from that simulator. Our main theoretical contribution is a finite-sample guarantee that quantitatively links estimator quality to decision reliability: the true conditional violation of the ETO solution is controlled by the estimator's approximation error and its finite-sample uncertainty. From this result we derive explicit sample-size prescriptions for common downstream procedures (scenario optimization, SAA, and discard variants), showing how estimation error reduces the available risk budget. We further provide finite-sample goodness-of-fit bounds for parametric MLEs, kernel conditional estimators, and modern conditional generative models, thereby clarifying the bias–feasibility trade-off and delivering concrete guidance on data and sampling requirements. The framework yields computational gains in convex settings and is directly applicable to power systems, routing, and public-health planning.

4 - On the Benefits of Weight Normalization for Over-parametrized Matrix Sensing

Bingcong Li, ETH Zurich, Zurich, Switzerland, Yudong Wei, Liang Zhang, Niao He

While normalization techniques are widely used in deep learning, their theoretical understanding remains relatively limited. In this work, we establish the benefits of (generalized) weight normalization (WN) applied to the overparameterized matrix sensing problem. We prove that WN with Riemannian optimization achieves linear convergence, yielding an exponential speedup over standard methods that do not use WN. Our analysis further demonstrates that both iteration and sample complexity improve polynomially as the level of overparameterization increases. To the best of our knowledge, this work provides the first characterization of how WN leverages overparameterization for faster convergence in matrix sensing.

5 - Distributionally Robust Optimization with Incomplete Covariate Information

Qingyuan Xu, University of Michigan, Ann Arbor, MI, United States, Ruiwei Jiang

Modern decision-making systems increasingly rely on contextual features (covariates) to improve optimization under uncertainty. However, in many datasets covariate information is incomplete. This talk proposes a method based on distributionally robust optimization to leverage those incomplete covariates for robust decision. We discuss the computational tractability and statistical guarantees for the proposed approach and demonstrate its performance in comparison with several impute-then-optimize alternatives.

SatB09

Cabinet Room

Advances on Convex and Conic Optimization II

Invited Session

Linear and Conic Optimization

Chair: Lijun Ding, University of California San Diego, La Jolla, CA, United States

Co-Chair: Feng-Yi Liao, University of California, San Diego, La Jolla, CA, United States

Co-Chair: Yang Zheng, University of California San Diego, San Diego, CA, United States

1 - On the Complexity of Second-Order Methods for Semidefinite Least-Squares Problems with Chordal Sparsity

Hong-Ming Chiu, University of Illinois Urbana-Champaign, Urbana, IL, United States, Renato D C Monteiro, Richard Zhang

High-accuracy solutions to sparse semidefinite least-squares problems (SDLS) $\min_{X\succeq 0}\left<C,X\right> + \frac{1}{2}\|\mathcal{A}(X)-b\|_2^2$ are critical in many real-world applications. First-order methods offer linear per-iteration cost but suffer from sublinear convergence, requiring an excessive number of iterations to reach high accuracy. In contrast, second-order methods achieve linear to quadratic convergence but are typically limited to medium-scale problems due to the cubic cost of Hessian inversion. In this paper, we establish the first sufficient conditions under which second-order methods achieve linear per-iteration cost for solving both SDLS and its Burer--Monteiro reformulation obtained by substituting $X = UU^T$. Specifically, we prove that when the extended sparsity pattern is chordally sparse, interior-point methods achieve linear per-iteration complexity for solving SDLS. Moreover, when the solution is low-rank, we show that trust-region and adaptive cubic regularization methods achieve a further reduced per-iteration cost for solving the Burer--Monteiro reformulated SDLS. The chordal sparsity and low-rank structure commonly arise in many practical SDLS problems, including matrix completion, sensor network localization, and power system state estimation.

2 - Disjunctive Sum of Squares

Amir Ali Ahmadi, Princeton University, Princeton, NJ, United States, Sanjeeb Dash, Yixuan Hua, Bartolomeo Stellato

We introduce the concept of disjunctive sum of squares for certifying nonnegativity of polynomials. Unlike the popular sum of squares approach where nonnegativity is certified by a single algebraic identity, the disjunctive sum of squares approach certifies nonnegativity with multiple algebraic identities. Our main result is a disjunctive Positivstellensatz proving that we can keep the degree of each algebraic identity as low as the degree of the polynomial whose nonnegativity is in question. Based on this result, we construct a semidefinite programming based converging hierarchy of lower bounds for the problem of minimizing a polynomial over a compact basic semialgebraic set, where the size of the largest semidefinite constraint is fixed throughout the hierarchy. We further prove a second disjunctive Positivstellensatz which leads to an optimization-free hierarchy for polynomial optimization. We specialize this result to the problem of proving copositivity of matrices. Finally, we describe how the disjunctive sum of squares approach can be combined with a branch-and-bound algorithm and we present numerical experiments on polynomial, copositive, and combinatorial optimization problems.

3 - Lagrangian Dual Sections: A Topological View of Hidden Convexity

Kevin Shu, California Institute of Technology, Pasadena, CA, United States

Convex relaxations are of central interest in optimization, and it is typically challenging to determine whether a given convex relaxation will be tight for a given problem. We introduce a topological framework for analyzing situations in which a constrained optimization problem over a nonconvex set (such as a manifold) has a tight convex relaxation. In particular, we give a criterion for the existence of such a tight convex relaxation in terms of the existence of a continuous function of Lagrange multipliers for the constrained problem maximizing the corresponding Lagrangian. We call such a function a Lagrangian dual section, in reference to the topological notion of a section of a bundle.

As a corollary of this result, we will give new criteria for the exactness of SDP relaxations for Stiefel manifold optimization and inverse eigenvalue problems in terms of linear subspaces of matrices satisfying spectral properties such as being nonsingular. We will also illustrate a homotopy continuation style algorithm with global optimality guarantees with applications to the unbalanced procrustes problem.

Joint work with Venkat Chandrasekaran, Jose Israel Rodriguez, Tim Duff.

4 - Online Wasserstein Distributionally Robust Optimization: Frameworks, Performance Guarantees, and Efficient Solvers

Guixian Chen, University of Michigan, Ann Arbor, MI, United States, Soroosh Shafiee, Salar Fattahi

We consider the problem of Online Wasserstein Distributionally Robust Optimization (WDRO), where a decision-maker must learn sequentially from data arriving in an online fashion. We model this problem as an online adversarial game, departing from standard stochastic or benign settings to explicitly account for an active, intelligent adversary. This approach requires algorithms that can simultaneously adapt to new information while protecting against worst-case distributional shifts within a Wasserstein ball. In this paper, we develop two novel online distributionally robust optimization frameworks that achieve this goal. We provide rigorous theoretical guarantees on their performance, bounding the regret against the adaptive adversary. A significant practical barrier to deploying WDRO in online settings is the computational expense of repeatedly solving the inner robust optimization problem. To address this, we propose a novel and highly efficient computational approach for a critical class of WDRO problems where the loss function is piece-wise concave. In particular, we design specialized algorithms for the common piece-wise linear and piece-wise quadratic cases. Our method demonstrates substantial computational speedups compared to state-of-the-art solvers such as Gurobi. Together, our contributions provide a theoretically sound and computationally practical solution for robust sequential decision-making in adversarial environments.

SatB10

Caucus Room

Advances in Decomposition Methods and Applications

Invited Session

Computational Optimization

Chair: Sean Lo, Massachusetts Institute of Technology, Cambridge, MA, United States

1 - Subpath-Based Column Generation for the Electric Vehicle Routing Problem with Time Windows

Sean Lo, Massachusetts Institute of Technology, Cambridge, MA, United States, Alexandre Jacquillat

We consider the Electric Vehicle Routing Problem, where customers are served by an electrified fleet that recharges at charging stations. The set-partitioning formulation for the EVRP has infinitely many path-based variables due to continuous charging decisions. We develop a column generation algorithm, where the pricing problem is decomposed into two phases: (i) generating subpaths between two consecutive charging stations, and (ii) combining subpaths into paths. We formalize subpath-based domination properties to establish the finite convergence and exactness of the column generation algorithm. We design domination criteria and prove that the resulting two-phase column generation algorithm can solve several EVRP variants, including time windows, nonlinear charging functions, heterogenous charging costs, and relaxation tightening strategies (e.g., *ng*-routes, limited-memory subset-row inequalities). Real-world experiments show that the two-phase algorithm outperforms path-based label-setting benchmarks, and can scale to large problem instances.

2 - The Surprising Performance of Random Partial Benders Decomposition

Yupeng Wu, London Business School, London, United Kingdom, Jean Pauphilet

Benders decomposition is a technique to solve large-scale mixed-integer optimization problems by decomposing them into a pure-integer master problem and a continuous separation subproblem. To accelerate convergence, we propose Random Partial Benders Decomposition (RPBD), a decomposition method that randomly retains a subset of the continuous variables within the master problem. Unlike existing problem-specific approaches, RPBD is simple to implement and universally applicable. We present both computational and theoretical evidence to support its effectiveness. For example, in extensive numerical experiments on network design and facility location problems, we find that (i) RPBD accelerates the convergence of Benders decomposition for problems with and without relatively complete recourse; (ii) RPBD usually halves the optimality gap at termination compared with a standard Benders approach; (iii) a random retention strategy is just as effective as problem-specific approaches proposed in the literature.

3 - Optimization of Sample-Path Clustering for Context-Sensitive Markov Decision Processes

Allison Grimsted, University of Michigan, Ann Arbor, MI, United States, Agni Orfanoudaki, Brian Denton

We address the problem of estimating transition probability matrices for heterogeneous longitudinal data representing sample-paths for stochastic systems. We propose a new clustering approach to partition sample-paths into a parsimonious set of clusters, each associated with its own transition probability matrix, capturing contextual variation in the system state transition probabilities. Our clustering method utilizes a set partitioning formulation with column generation to maximize the likelihood that the sample-paths were generated by their assigned class transition probability matrices. To address the computational challenges of solving a mixed integer nonlinear and nonconvex pricing problem exactly, we describe approximation algorithms that generate effective warm-start columns for column generation. The resulting clusters are used to estimate context-specific transition probability matrices for Markov decision processes (MDPs). We benchmark our approach against a baseline maximum likelihood estimation method that pools all sample paths to estimate a single "one-size-fits-all" transition matrix. We validate our proposed method in a real-world case study on maintenance and repair decisions for highway bridges, leveraging data from the U.S. Department of Transportation.

4 - Let's Have a Conversation: Designing and Evaluating AI Agents for Interactive Optimization

Joshua Drossman, Massachusetts Institute of Technology , Cambridge, MA, United States, Alexandre Jacquillat, Sebastien Martin

Large language models (LLMs) offer new opportunities for bringing optimization techniques into human-centered decision processes. Whereas existing approaches largely leverage LLMs to formulate an optimization problem, real-world decision-making often relies on dynamic interactions between a user and an optimization model. In response, we introduce a conversational chatbot to deploy optimization capabilities in non-technical areas. We develop LLM-based optimization agents equipped with structured toolkits that enable safe, interpretable, and iterative interaction with mixed-integer optimization models. We propose a conversational assessment approach built on LLM-based decision agents, which replicate real-world stakeholders. Unlike traditional one-shot evaluations, this framework measures an optimization agent's ability to elicit preferences, navigate trade-offs, and guide decision agents to high-quality solutions over a conversation. Using a school-scheduling problem, we construct a dataset of 436 decision agents and compare four optimization-agent designs. Our results demonstrate that conversations are critical in assessing the real-world performance of an optimization agent. We also show that structured, tool-based agents substantially outperform free-form code-based baselines in solution quality, reliability, and interaction efficiency. Ultimately, this framework contributes a new LLM-powered decision support system at the human-AI-optimization interface.

SatB11

Charter Room

Adaptive Learning and Optimization in Multi-Agent Networks

Invited Session

Network Optimization

Chair: Shahin Shahrampour, Northeastern University, Boston, MA, United States

1 - Decentralized Online Riemannian Optimization Beyond Hadamard Manifolds

Muhammet Emre Sahinoglu, Northeastern University, Boston, MA, United States, Shahin Shahrampour

We study decentralized online Riemannian optimization over manifolds with possibly positive curvature, going beyond the Hadamard manifold setting. Decentralized optimization techniques rely on a consensus step that is well understood in Euclidean spaces because of their linearity. However, in positively curved Riemannian spaces, a main technical challenge is that geodesic distances may not induce a globally convex structure. In this work, we first analyze a curvature-aware Riemannian consensus step that enables a linear convergence beyond Hadamard manifolds. Building on this step, we establish a $O(\sqrt{T})$ regret bound for the decentralized online Riemannian gradient descent algorithm. Then, we investigate the two-point bandit feedback setup, where we employ computationally efficient gradient estimators using smoothing techniques, and we demonstrate the same $O(\sqrt{T})$ regret bound through the subconvexity analysis of smoothed objectives.

2 - Riemannian Dueling Optimization

Yuxuan Ren, Rice University, Houston, TX, United States

We study optimization on Riemannian manifolds when the learner has access to only \emph{dueling feedback}, i.e., binary response to the given query ``Do you prefer A over B?". For several target applications where dueling feedback arises naturally, e.g., recommendation systems, and robotics, the decision space is often constrained and/or inherently a low-dimensional Riemannian manifold. However, prior work on optimization with dueling feedback focuses mainly on unconstrained problems in Euclidean settings. In this work, we bridge this gap by proposing Riemannian optimization algorithms with dueling feedback for constrained problems with convergence rate that adapts to the low-dimensional structure of the problem. In particular, we propose a (projected)-Riemannian Dueling Normalized Gradient Descent (RDNGD) method that attains non-asymptotic iteration complexity of $\mathcal{O}(\frac{d}{\epsilon^2})$ and $\mathcal{O}(\frac{d}{\epsilon})$ for nonconvex and geodesically convex objectives respectively matching the best Euclidean counterparts; our analysis covers both constant and diminishing step sizes. In the geodesically strongly convex case, our algorithm attains the iteration complexity $\mathcal{O}(d\log(\frac{1}{\epsilon}))$. To handle costly or intractable projections, we further develop a projection-free Riemannian Dueling Frank–Wolfe (RDFW) method and prove an iteration complexity of $\mathcal{O}(\frac{d}{\epsilon})$ and oracle complexity of $\mathcal{O}(\frac{d}{\epsilon^2})$ for geodescially convex function. Interestingly, our analysis and the theoretical guarantee extend trivially to the Euclidean setting. We illustrate the effectiveness of the proposed algorithms through extensive numerical experiments on benchmark problems.

3 - Decentralized Adaptive Optimization Algorithms

Gesualdo Scutari, Purdue University, West Lafayette, IN, United States, Xiaokai Chen

We consider decentralized optimization where multiple agents minimize the average of their (strongly) convex, smooth losses over a communication graph. Convergence of the existing decentralized methods generally hinges on an apriori, proper selection of the stepsize. Choosing this value is notoriously delicate: (i) it demands global knowledge from all the agents of the graph's connectivity and every local smoothness/ strong-convexity constants--information they rarely have; (ii) even with perfect information, the worst-case tuning forces an overly small stepsize, slowing convergence in practice; and (iii) large-scale trial-and-error tuning is prohibitive. This work introduces a decentralized algorithm that is fully adaptive in the choice of the agents' stepsizes, without any global information and using only neighbor-to-neighbor communications--agents need not even know whether the problem is strongly convex. The algorithm retains strong guarantees:

it converges at linear rate when the losses are strongly convex and at sublinear rate otherwise, matching the best-known rates of parameter-dependent methods. Time permitting, extensions to composite optimization and stochastic optimization will be discussed.

4 - RESIST: Resilient Decentralized Learning Using Consensus Gradient Descent

Cheng Fang, Rutgers University, Piscataway, NJ, United States, Rishabh Dixit, Waheed Bajwa, Mert Gurbuzbalaban

Empirical risk minimization (ERM) is a cornerstone of modern machine learning (ML), supported by advances in optimization theory that ensure efficient solutions with provable algorithmic convergence rates and statistical learning rates. Privacy, memory, computational, and communications constraints increasingly necessitate data collection, processing, and storage across network-connected devices. In many applications, these networks operate in decentralized settings where a central server cannot be assumed, requiring decentralized ML algorithms that are both efficient and resilient. Decentralized learning, however, faces significant challenges, including an increased attack surface for adversarial interference during decentralized learning processes. This paper focuses on the man-in-the-middle (MITM) attack, wherein adversaries exploit communication vulnerabilities between devices, potentially causing models to deviate significantly from their intended ERM solutions. To address this challenge, we propose RESIST (Resilient dEcentralized learning using conSensus gradIent deScenT), an optimization algorithm designed resilient to compromised communication links. Unlike existing adversarially robust decentralized learning methods, which often (i) guarantee convergence only to a neighborhood of the solution, (ii) lack guarantees of linear convergence for strongly convex problems, or (iii) fail to ensure statistical consistency as sample sizes grow, RESIST overcomes all three limitations. It achieves algorithmic and statistical convergence for strongly convex, Polyak–Łojasiewicz, and nonconvex ERM problems by employing a multistep consensus gradient descent framework and robust statistics-based screening methods to mitigate the impact of MITM attacks. Experimental results demonstrate the robustness and scalability of RESIST across diverse attack strategies, screening methods, and loss functions, confirming its suitability for real-world decentralized optimization and learning in adversarial environments.

March 21, 2026, 2:45 PM - 4:15 PM

SatC01

Grand Ballroom

Optimization, Statistics, and Learning Dynamics

Invited Session

Optimization in Data Science

Chair: Yao Xie, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Xiuyuan Cheng, Duke University, Durham, NC, United States

1 - Robust Learning in Online Markets: Algorithm and Optimal Regret Guarantee

Zhengli Wang, The University of Hong Kong, Hong Kong, Hong Kong

We study the robust learning in online markets under the contamination model. Here, a platform sequentially allocates advertiser demand to creators, collects rewards, and incentivizes them through revenue sharing. We develop an algorithm based on adaptive elimination, for which we can establish a tight regret bound. The algorithm is simple, interpretable, and provides clear insight into the optimal explore-exploit trade-off under contamination.

2 - Wasserstein-Cramér-Rao Theory of Unbiased Estimation

Nicolas Garcia Trillos, University of Wisconsin Madison, Madison, WI, United States, Adam Quinn

Jaffe, Bodhisattva Sen

The quantity of interest in the classical Cramér-Rao theory of unbiased estimation (e.g., the Cramér- Rao lower bound, exact efficiency in exponential families, and asymptotic efficiency of maximum likelihood estimation) is the variance, which represents the instability of an estimator when its value is compared to the value for an independently sampled data set from the same distribution. In this talk, we will be interested in a quantity that represents the instability of an estimator when its value is compared to the value for an infinitesimal additive perturbation of the original data set; we refer to this as the "sensitivity" of an estimator. The resulting theory of sensitivity is based on the Wasserstein geometry in the same way that the classical theory of variance is based on the Fisher-Rao (equivalently, Hellinger) geometry. I'll present a collection of results which are analogous to the classical case: a Wasserstein-Cramér-Rao lower bound for the sensitivity of any unbiased estimator, a characterization of models in which there exist unbiased estimators achieving the lower bound exactly, and a guarantee that Wasserstein projection estimators achieve the lower bound asymptotically. I'll discuss some simple statistical examples to illustrate the theory, sometimes revealing new optimality properties for existing estimators and other times revealing entirely new ones. I'll also discuss some of the many open questions that this work (and in fact the whole perspective this work is based on) motivates.

3 - Stochastic Online Fisher Markets: Static Pricing Limits and Adaptive Enhancements

Devansh Jalota, Columbia University, New York, NY, United States, Yinyu Ye

Fisher markets are one of the most fundamental models for resource allocation. However, the problem of computing equilibrium prices in Fisher markets typically relies on complete knowledge of users' budgets and utility functions and requires transactions to happen in a static market where all users are present simultaneously. Motivated by these practical considerations, we study an online variant of Fisher markets, wherein users with privately known utility and budget parameters, drawn independently and identically (i.i.d.) from a distribution, arrive sequentially. In this setting, we first study the limitations of static pricing algorithms, which set uniform prices for all users, along two performance metrics: (i) regret, that is, the optimality gap in the objective of the Eisenberg-Gale program between an online algorithm and an oracle with complete information, and (ii) capacity violations, that is, the overconsumption of goods relative to their capacities. Given the limitations of static pricing, we design adaptive posted-pricing algorithms, one with knowledge of the distribution of users' budget and utility parameters and another that adjusts prices solely based on past observations of user consumption, that is, revealed preference feedback, with improved performance guarantees. Finally, we present numerical experiments to compare our revealed preference algorithm's performance to several benchmarks.

4 - Beyond Maximum Likelihood: Variational Inequality Estimation for Generalized Linear Models

Jonghyeok Lee, Georgia Institute of Technology, Atlanta, GA, United States, Linglingzhi Zhu, Yao Xie

Generalized linear models (GLMs) are fundamental tools for statistical modeling, with maximum likelihood estimation (MLE) serving as the classical method for parameter inference. While MLE performs well in canonical GLMs, it can become computationally inefficient near the true parameter value. In more general settings with non-canonical or fully general link functions, the resulting optimization landscape is often non-convex, non-smooth, and numerically unstable. To address these challenges, we investigate an alternative estimator based on solving the variational inequality (VI) formulation of the GLM likelihood equations, originally proposed by Juditsky and Nemirovski (2019) as an alternative for solving nonlinear least-squares problems. Unlike their focus on algorithmic convergence in monotone settings, we analyze the VI approach from a statistical perspective, comparing it systematically with the MLE. We also extend the theory of VI estimators to a broader class of link functions, including non-monotone cases satisfying a strong Minty condition, and show that it admits weaker smoothness requirements than MLE, enabling faster, more stable, and less locally trapped optimization. Theoretically, we establish both non-asymptotic estimation error bounds and asymptotic normality for the VI estimator, and further provide convergence guarantees for fixed-point and stochastic approximation algorithms. Numerical experiments show that the VI framework preserves the statistical efficiency of MLE while substantially extending its applicability to more challenging GLM settings.

5 - Hamiltonian Descent Algorithms for Optimization: Accelerated Rates via Randomized Integration Time

Andre Wibisono, Yale, New Haven, CT, United States

We study the Hamiltonian flow for optimization (HF-opt), which simulates the Hamiltonian dynamics for some integration time and resets the velocity to 0 to decrease the objective function; this is the optimization analogue of the Hamiltonian Monte Carlo algorithm for sampling. We show that for short integration time, HF-opt has the same convergence rates as gradient descent for strongly and weakly convex objective functions. We also show that by randomizing the integration time in HF-opt, the resulting randomized Hamiltonian flow (RHF) achieves accelerated convergence rates in continuous time, similar to the continuous-time rates for the accelerated gradient flow. We discretize RHF as the randomized Hamiltonian gradient descent (RHGD) algorithm, and prove that RHGD achieves the same accelerated convergence rates as Nesterov's accelerated gradient descent for smooth strongly and weakly convex objective functions.

6 - Convex Optimization Algorithms for Fair Classification with POE Parity

Joonhyuk Suh, Stony Brook University, East setauket, NY, United States, Stanislav Uryasev

We study fair binary classification where decision is made when score exceeds some threshold. We considered the Probability-of-exceedance (POE) parity, i.e., the probability that the score exceeds a threshold is equal for sensitive groups. POE parity is a difficult nonconvex optimization constraint. We show that a simple convex program combining a standard convex loss (e.g., logistic regression) with convex constraints can achieve the exact POE parity. The modeling involved group-specific intercepts ensuring a monotone response of each group tail probability to its own intercept. The equality POE constraints are replaced with three types of convex constraints on: (1) buffered probability of exceedance (bPOE), (2) conditional value-at-risk (CVaR), (3) expectations. The optimization problems are quickly solved with convex solvers. We demonstrated the approach with with four benchmark datasets: Adult Income, COMPAS, Default Credit, and Law School. The approach:
(1) attains essentially zero empirical POE gap (perfect fairness at the target threshold),
(2) achieves lower mean-squared error (compared to other published approaches),
(3) uses simple constraints, and
(4) is much faster than nonconvex POE formulations.

SatC02

Georgian Room

Experimental Design as Decision-Making under Uncertainty

Invited Session

Optimization in Data Science

Chair: Jinglong Zhao, Boston University, Boston, MA, United States

1 - Efficient Active Learning Strategies for Computer Experiments

Difan Song, Georgia Institute of Technology, Atlanta, GA, United States, Roshan Joseph

Active learning aims at allocating resources in an intelligent manner based on the already observed data to satisfy certain objectives such as surrogate modeling or optimizing a computationally expensive function. When a Gaussian process model is used as a surrogate, space-filling designs are commonly employed to initialize active learning. Here we propose screening designs as initial designs and a new correlation function for the Gaussian process. Specifically, we propose using the maximum one-factor-at-a-time (MOFAT) design as the initial design and a multiplicative inverse multiquadric (MIM) kernel for the correlation function. The ideas behind them are known in other fields, such as sensitivity analysis or kernel theory, but they never seem to have been used for active learning in computer experiments. We also propose an integrated MOFAT-MIM strategy that automatically incorporates screening in the model estimation step. We show that these strategies provide substantial improvement to the state-of-the-art methods for both emulation and optimization objectives. We support our findings through theory and simulations, and a real experiment on the vapor-phase infiltration process.

2 - Randomization Inference for Always-Reporter Average Treatment Effect

Haoge Chang, Columbia University, Cambridge, MA, United States, Arthur Yu

This article studies randomization inference for treatment effects in randomized controlled trials with attrition, where outcomes are observed for only a subset of units. We assume monotonicity in reporting behavior and focus on the average treatment effect for always-reporters (AR-ATE), defined as units whose outcomes are observed under both treatment and control.
Because always-reporter status is only partially revealed by observed assignment and response patterns, we propose a worst-case randomization test that maximizes the randomization p-value over all always-reporter configurations consistent with the data, with an optional pretest to prune implausible configurations. Using studentized Hajek- and chi-square–type statistics, we show the resulting procedure is finite-sample valid for the sharp null and asymptotically valid for the weak null. We also discuss computational implementations for discrete outcomes and integer-programming-based bounds for continuous outcomes.

3 - Automated Experimental Design with Optimization from Historical Data Simulations

Ruoxuan Xiong, Emory University, Atlanta, GA, United States

We study the problem of designing randomized experiments to estimate the causal effects of new interventions. The design problem involves selecting a treatment assignment mechanism, including assignment probabilities, in order to optimize a user-defined objective (e.g., maximizing the precision of a causal-effect estimator). A key challenge is that the objective is typically a black-box function of the design, depending on unknown outcome-generating processes and causal effects. To tackle this challenge, we propose a novel automated experimental design (AutoED) approach. AutoED evaluates the objective value of a candidate design by simulating experiments based on stationary historical control data and prior distributions of causal effects. Then a gradient-free method is used to iteratively optimize the design. We rigorously analyze the convergence behavior of AutoED and show how it depends on temporal and cross-unit correlations in historical data as well as on implementation specifics. In synthetic experiments on three application domains—digital platforms, health, and energy—AutoED can reduce the estimation error of treatment effects by up to 25% compared to the state-of-the-art benchmark designs. Our results reveal new design insights to improve efficiency. In particular, for multi-unit crossover experiments with multiple interventions, we find that efficiency is improved by sequentially rolling out (a) the same intervention across units and (b) different interventions on the same unit.

SatC03

Plaza I

Recent Advances in Derivative-Free Optimization II

Invited Session

Nonlinear Optimization

Chair: Zikai Xiong, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Abraar Chaudhry, Georgia Tech, Atlanta, GA, United States

Co-Chair: Katya Scheinberg, Georgia Tech, Atlanta, GA, United States

1 - Managing Multiple Models in Derivative-Free Optimization with Contextual Bandits

Matt Menickelly, Argonne National Laboratory, Lemont, IL, United States

In computational science workflows, it is often the case that 1) simulations do not permit readily the algorithmic computation of derivatives, 2) objective functions for optimization involve multiple simulation outputs, and 3) those simulation outputs can be performed (at least partially) in parallel. While the Argonne-maintained derivative-free optimization (DFO) solvers in the Interpolation-Based Composite Derivative-Free Optimization (IBCDFO) repository are specifically designed to address problems exhibiting the first two of these characteristics, they do not support the third at the time of writing this abstract.

In this work, we reexamine our past work on a class of randomized algorithms, stochastic average model (SAM) methods. SAM methods are conceptually similar to stochastic average gradient methods, and effectively require that only randomized subsets of simulation outputs be locally modeled in each iteration of a model-based DFO method. From the perspective of variance reduction, given a batch size, the optimal sampling distribution for a batch within a SAM method permits a closed-form, but the closed-form distribution contains unknowable parameters.

However, clues concerning the values of the unknowable parameters are often available either through domain expertise, alternative lower-fidelity simulations, or simply regression models. Motivated by this, we consider the problem of determining sampling distributions for SAM methods as a contextual bandit problem and we apply the Exponential weights algorithm for Exploration and Exploitation with Experts (Exp4). Results are demonstrated on both synthetic problems and a problem of density functional theory calibration.

2 - Stochastic Trust Region Optimization with Quadratic Regularization

Sara Shashaani, North Carolina State University, Raleigh, NC, United States, Yunsoo Ha, Quoc Tran-Dinh

Trust region methods have shown much stability for noisy non-convex problems. A subclass of these methods, called ASTRO, uses adaptive sampling to build quadratic local approximations and move to new iterates. ASTRO attains competitive worst-case complexity results via exploiting dependence of noisy observations and underlying problem's sample-path structure. When the objective function is the output of a stochastic simulation whose inputs follow a probability model, we show the effect of common random numbers in the reduction of overall sample complexity. We also investigate the analytical and practical pros and cons of quadratic regularization for this algorithm.

3 - On Complexity of Model-Based Derivative-Free Methods

Abraar Chaudhry, Georgia Tech, Atlanta, GA, United States, Katya Scheinberg

In many applications of mathematical optimization, one may wish to optimize an objective function without access to its derivatives. These situations call for derivative-free optimization (DFO) methods. Among the most successful approaches in practice are model-based trust-region methods, such as those pioneered by M.J.D Powell. While relatively complex to implement, these methods are now available in standard scientific computing platforms, including MATLAB and SciPy. However, theoretical analysis of their computational complexity lags behind practice. In particular, it is important to bound the number of function evaluations required to achieve a desired level of accuracy. In this paper we systematically derive complexity bounds for classical model-based trust-region methods and their modern variations. We establish, for the first time, that these methods can have the same worst case complexity than any other known DFO method.

4 - Projection versus Sketching in Derivative-free Randomized Subspace Methods

Scholar Sun, University of Waterloo, Toronto, ON, Canada, Katya Scheinberg

Scalability to higher dimensions remains a major challenge in derivative-free optimization. In the model-based paradigm, one approach is to optimize over a randomly generated, low-dimensional subspace in each iteration, thereby reducing the number of function evaluations needed to construct the model. Recently, a randomized subspace trust-region method with a geometry-correcting subroutine was shown to attain an oracle complexity of O(n/\epsilon^2), improving upon the known complexity in the deterministic setting.

This talk presents the first implementation of such an algorithm and reports numerical results on a suite of test functions, investigating the impact of subspace dimension on performance. In particular, we compare sketching-based and projection-based subspace strategies, highlighting their respective trade-offs in high-dimensional derivative-free optimization.

SatC04

Plaza II

Optimization Techniques for Large language Models Training

Invited Session

Nonlinear Optimization

Chair: Mingyi Hong, Mingyi Hong, WAYZATA, United States

Co-Chair: Jiaxiang Li, Meta, Santa Clara, CA, United States

1 - Two Perspectives on Muon for Deep Learning Optimization

Weijie Su, Wharton School, University of Pennsylvnania, Philadelphia, PA, United States

Introduced in December 2024, Muon is an optimization method for training language models that updates the weight along the direction of an orthogonalized gradient. The superiority of Muon has been quickly recognized, as demonstrated on industry-scale models; for example, it has been successfully used to train a trillion-parameter frontier language model. In this talk, we offer two perspectives to shed light on this matrix-gradient method. First, we introduce a unifying framework that precisely distinguishes between preconditioning for curvature anisotropy (like Adam) and gradient anisotropy (like Muon). This perspective not only offers new insights into Adam's instabilities and Muon's accelerated convergence but also leads to a new extension, such as PolarGrad. Next, we introduce a second perspective based on an isotropic curvature model. We derive this model by assuming isotropy of curvature (including Hessian and higher-order terms) across all perturbation directions. We show that under a general growth condition, the optimal update is one that makes the gradient's spectrum more homogeneous; that is, making its singular values closer in ratio. We then show that the orthogonalized gradient becomes optimal for this model when the curvature exhibits a phase transition in growth. Taken together, these results suggest that the gradient orthogonalization employed in Muon is directionally correct but may not be strictly optimal, and we will discuss how to leverage this model for designing new optimization methods. This talk is based on arXiv:2505.21799 and 2511.00674.

2 - A Minimalist Optimizer Design for LLM Pretraining

Athanasios Glentis, University of Minnesota, Minneapolis, MN, United States, Jiaxiang Li, Andi Han, Mingyi Hong

Training large language models (LLMs) typically relies on adaptive optimizers such as Adam, which introduce extra operations and require significant more memory to maintain first- and second-order moments than SGD. While recent works such as GaLore, Fira and APOLLO have proposed state-compressed variants to reduce memory consumption, a fundamental question remains: *What are the minimum modifications to plain SGD needed to match state-of-the-art pretraining performance?* We systematically investigate this question using a bottom-up approach, and identify two simple yet highly (memory- and compute-) efficient techniques: (1) column-wise gradient normalization (normalizing the gradient along the output dimension), which boosts SGD performance without momentum; and (2) applying first-order momentum only to the output layer, where gradient variance is highest. Combining these two techniques lead to SCALE (Stochastic Column-normAlized Last-layer momEntum), a simple optimizer for memory efficient pretraining. Across multiple LLaMA models (60M–1B), SCALE matches or exceeds the performance of Adam while using only 35–45% of the total memory. It also consistently outperforms memory-efficient optimizers such as GaLore, Fira and APOLLO, making it a strong candidate for large-scale pretraining under memory constraints. For LLaMA 7B model, SCALE outperforms the state-of-the-art memory-efficient method APOLLO in terms of both perplexity and memory consumption.

3 - Fantastic Pretraining Optimizers and Where to Find Them

Kaiyue Wen, Stanford University, Stanford, CA, United States, David Hall, Tengyu Ma, Percy Liang

AdamW has long been the dominant optimizer in language-model pretraining, despite numerous claims that alternative optimizers offer 1.4×–2× speedup. We posit that two methodological shortcomings have obscured fair comparisons and hindered practical adoption: (i) unequal hyperparameter tuning and (ii) limited or misleading evaluation setups. To address these issues, we conduct a systematic study of ten deep-learning optimizers across four model scales (0.1B–1.2B parameters) and data-to-model ratios (1–8× the Chinchilla optimum). We find that fair and informative comparisons require rigorous hyperparameter tuning and evaluations across a range of model scales and data-to-model ratios, performed at the end of training. First, optimal hyperparameters for one optimizer may be suboptimal for another, making blind hyperparameter transfer unfair. Second, the actual speedup of many proposed optimizers over well-tuned baselines is lower than claimed and decreases with model size — only ~1.1× for 1.2B-parameter models. Third, comparing intermediate checkpoints before reaching the target training budget can be misleading, as rankings between two optimizers may flip during training due to learning-rate decay. Through our thorough investigation, we find that all the fastest optimizers (e.g. Muon, SOAP) use matrix-based preconditioners — multiplying gradients with matrices rather than entry-wise scalars. However, the speedup of such matrix-based optimizers is inversely proportional to model scale, decreasing from ~1.4× over AdamW for 0.1B-parameter models to ~1.1× for 1.2B-parameter models.

4 - Scaling Learning-Rate-Free and Schedule-Free Learning to Large Language Models

Aaron Defazio, FAIR, Meta, Los Angeles, NY, United States

: I will explore a number of recent theory-driven advances in our understanding of learning rates and learning rate schedules, and show how they can be applied to training modern large scale language models.

5 - Clarifying Shampoo: Adapting Spectral Descent to Stochasticity and the Parameter Trajectory

Hao-Jun Shi, Meta Platforms, Inc., Redwood City, CA, United States, Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee

Optimizers leveraging the matrix structure in neural networks, such as Shampoo and Muon, are more data-efficient than element-wise algorithms like Adam and Signum. While in specific settings, Shampoo and Muon reduce to spectral descent analogous to how Adam and Signum reduce to sign descent, their general relationship and relative data efficiency under controlled settings remain unclear. Through extensive experiments on language models, we demonstrate that Shampoo achieves higher token efficiency than Muon, mirroring Adam's advantage over Signum. We show that Shampoo's update applied to weight matrices can be decomposed into an adapted Muon update. Consistent with this, Shampoo's benefits can be exclusively attributed to its application to weight matrices, challenging interpretations agnostic to parameter shapes. This admits a new perspective that also avoids shortcomings of related interpretations based on variance adaptation and whitening: rather than enforcing semi-orthogonality as in spectral descent, Shampoo's updates are time-averaged semi-orthogonal in expectation.

SatC05

Plaza III

Advances in Dynamic and Data-driven Optimization under Uncertainty

Contributed Session

Optimization under Uncertainty

Chair: Alexandre Jacquillat, MIT Sloan School of Management, Cambridge, MA, United States

1 - Data-Driven Sequential Search

Cagin Uru, Duke University, Durham, NC, United States, David Brown

How can one search effectively with limited prior knowledge on the distribution of alternative values? This paper addresses this question in the context of sequential search. In our data-driven setting, traditional methods that rely on a known value distribution no longer apply and feasible search policies are based solely on the history of explored alternative values. We seek to identify a policy that maximizes the worst-case ratio of expected reward compared to an oracle (referred to as Pandora) with full knowledge of the value distribution. We first consider a general setting where alternative values may follow any arbitrary distribution. We design simple policies that statically commit to a prespecified number of explorations. We show that these static policies guarantee a competitive ratio of at least 1/e ≈ 37% of the Pandora benchmark for any arbitrary value distribution. Our approach involves studying nature's problem to select a distribution to counter a policy and identifying worst-case distributions. Moreover, we study how the structure of the unknown value distribution influences achievable performance guarantees by considering a setting where feasible distributions belong to the class of monotone hazard rate distributions. In this case, we improve our guarantee to (e/(e+1))^2 ≈ 53% of the Pandora benchmark. In both settings, we derive performance limits for feasible policies to complement our performance guarantee results. We illustrate the efficacy of our policies through numerical examples based on price search. Our results provide insights to the decision makers with no prior knowledge of value distributions across many search applications.

2 - Overworking or Outsourcing: Integrated Prediction-Optimization for Overtime Control with Application in Last-Mile Delivery

Haoran Guo, Tsinghua University, Beijing, China, People's Republic of, Meng Qi, Wei Qi

Effective overtime control is essential for ensuring both operational sustainability and workforce well-being under workload fluctuations and uncertainty. In operational settings such as last-mile delivery, there are often external options to outsource workload with a higher cost, and the key trade-off is between overworking and the cost of outsourcing. To navigate this trade-off under workload uncertainty, prediction models are often introduced. However, it is common yet insufficient to simply make the downstream optimization stage overtime-aware. This insufficiency arises from a fundamental misalignment: standard prediction models are trained with statistical objectives that disregard downstream decision quality, leading to suboptimal decisions. In the overtime control setting, this can result in severe overworking and compromised employee well-being. In this paper, we formulate this problem as a contextual stochastic optimization (CSO) and propose an integrated prediction-optimization (IPO) framework. Leveraging the problem structure, we constructively prove the sufficiency of point prediction models despite the nonlinear objective. Specifically, we design a decision-aware prediction using the α-superlevel set to identify tractable geometric cases, and theoretically show this constructed prediction has a bounded excess decision risk. Guided by the theoretical insights, we develop efficient algorithms for practical implementation. Numerical experiments demonstrate our approach significantly outperforms standard predict-then-optimize (PTO) methods. Challenging a prevalent assumption, we find that simple, decision-focused predictions can outperform complex, decision-blind neural networks in decision quality. Furthermore, a case study in urban last-mile delivery reveals our approach has an escalating advantage as the overtime penalty increases, suggesting that analytical tools should align with strategic welfare goals.

3 - Relative Monte Carlo for Reinforcement Learning.

Sebastien Martin, Kellogg School of Management, Northwestern University, Evanston, IL, United States, Audrey Bazerghi

We propose a new policy gradient algorithm for reinforcement learning (RL), relative Monte Carlo (rMC), that estimates policy gradients using relative returns between a root and (coupled) counterfactual simulated paths, each instantiated by taking different actions from the root. The resulting gradient estimate is both unbiased and has low variance. rMC is compatible with any differentiable policy, including neural networks, and is guaranteed to converge even for infinite-horizon tasks.

The method exploits common random number coupling of the simulated paths to reduce variance and increase the likelihood that paths merge, thereby reducing simulation complexity.
It is particularly well-suited for operational control problems where actions have a "local'" effect, such as queueing, supply chain, or ride-hailing problems. Indeed, we show that it has provably low complexity for a family of inventory control problems. Numerical tests on an inventory and fulfillment problem show that compared to alternative RL approaches, rMC converges in far fewer iterations (lower variance), has better policy performance (unbiased), and requires minimal hyperparameter tuning.

4 - An Iterative Network Flow Algorithm for Online Pickup and Delivery

Jason Luo, Massachusetts Institute of Technology, Cambridge, MA, United States, Alexandre Jacquillat

The pickup and delivery problem involves the routing of vehicles to transport goods between pickup and delivery locations. Our work addresses two common difficulties: large numbers of vehicles and limited information regarding future demand. First, we formalize a new variant of the online pickup and delivery with no demand backlog and a distance minimization objective. We then show a network flow equivalency of the offline problem, under a constraint stating that all vehicles' fixed starting and ending locations cover all nodes. We embed the network flow into polynomial-time online algorithms for the online pickup and delivery problem. We show that the online algorithms are asymptotically 2-competitive, whereas a myopic integer programming baseline is not c-competitive for any c>0. Computational experiments show that the online algorithm returns solutions within 0.5 percent of the integer programming benchmark at less than a thousandth of the runtime.

SatC06

Director's and Lounge

Distributionally Robust Optimization and Machine Learning

Invited Session

Global Optimization

Chair: Weijun Xie, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA, 30332, United States

Co-Chair: Siyuan Chen, Georgia Institute of Technology, Atlanta, GA, United States

1 - On the Global Convergence of Noisy Gradient Descent for ResNets in the Feature Learning Regime

Zhen Yang, University of Texas at Austin, Austin, TX, United States, Rui Gao, Shuang Li

Residual networks are a foundational component of modern deep learning architectures. Despite their empirical success and widespread use, theoretical guarantees for their global convergence are limited. Existing analyses are typically asymptotic or confined to regimes that do not fully capture the crucial phenomenon of feature learning, where the network learns to build meaningful representations from the data. This paper establishes the global geometric convergence of noisy gradient descent (NGD) for training deep residual networks. Unlike prior work, our analysis holds for networks with a finite number of layers and operates within the feature learning regime.

2 - An Error Bound-based Convergence Analysis Framework for a Class of Randomized Algorithms

Man-Chung Yue, The University of Hong Kong, Hong Kong, China, People's Republic of, Zhichun Yang, Li Jiang, Tianxiang Liu

Existing error-bound-based analysis frameworks typically only provide overly conservative conclusions when applied to a class of important stochastic algorithms that exhibit certain descent properties, such as randomized coordinate descent and randomized projection methods. To address this gap, we propose a new abstract framework for analyzing such stochastic methods under on a unified error bound (UEB) condition. The proposed UEB condition subsumes many mainstream error bound conditions across various settings, including optimization, convex feasibility, and common fixed-point problems. Under the global UEB condition, we establish non-asymptotic in-expectation and asymptotic almost-sure convergence rates. Moreover, under a local UEB condition, we show asymptotic almost-sure convergence guarantees. We demonstrate the strength and versatility of our framework through several applications. For smooth unconstrained minimization, we present novel convergence guarantees for the randomized subspace descent method, which encompasses randomized coordinate descent and randomized block coordinate descent, under global and local Kurdyka-Lojasiewicz conditions. We further discuss new convergence rates for the randomized alternating Krasnoselskii-Mann method under consistent error bound conditions.

3 - Wasserstein Distributionally Robust Regret Optimization for Reinforcement Learning with Human Feedback

Jose Blanchet, Stanford University, Stanford, CA, United States, Lukas Fiechtner, Shang Liu

Distributionally Robust Optimization (DRO) is a widely adopted framework for decision-making under uncertainty, but it may lead to overly conservative solutions. Ex-ante Distributionally Robust Regret Optimization (DRRO) has been raised to overcome this over-pessimism, where the goal is to minimize the gap between the ex-ante optimal solution and the decision. While even structured problems can lead to NP-hard formulations in Wasserstein DRRO, we identify formulations that are tractable, applicable in important practical problems and that lead to solutions that can be aggressive in the presence of uncertainty. For example, we examine several examples motivated by applications in bipartite matching, shortest-path problems, and other fields such as Reinforcement Learning from Human Feedback (RLHF).

4 - Distributionally Robust Universal Classification

Siyuan Chen, Georgia Institute of Technology, Atlanta, GA, United States, Weijun Xie

The Universal Classification (UC) problem aims to find an optimal classifier from a universal policy space that minimizes the expected 0-1 loss, also known as the misclassification risk. However, conventional empirical risk minimization often results in overfitting and poor out-of-sample performance. To address this limitation, we propose the Distributionally Robust Universal Classification (DRUC) formulation, which enhances generalization by incorporating distributional robustness through a Wasserstein distance-based ambiguity set centered at the empirical distribution. To manage the infinite-dimensional nature of the DRUC policy space, we develop its in-sample DRUC counterpart, which allows for a more tractable reformulation while preserving robustness properties. We prove that, asymptotically, the in-sample DRUC formulation converges to the original UC formulation and is equivalent to the DRUC formulation. Under mild conditions, we provide non-asymptotic finite-sample performance guarantees. Furthermore, we derive a mixed-integer linear programming (MILP) reformulation to obtain the optimal in-sample DRUC policy and propose an efficient 2-approximation algorithm. Our numerical experiments show the efficiency of the proposed approximation algorithm and demonstrate the superior out-of-sample performance of the in-sample DRUC formulation.

SatC07

Congress Room

Recent Advances in Mixed-Integer Programming

Invited Session

Discrete Optimization

Chair: Shaoning Han, National University of Singapore, Singapore, Singapore

1 - On Probing Techniques in Mixed-Binary Convex Optimization Problems

Guanyi Wang, National University of Singapore, Singapore, Singapore, Shaoning Han, Haozhe Tan

Probing is an important technique for domain reduction in mixed-integer programming. This paper introduces a novel framework on constructing probing rules based on the sensitivity analysis of value functions for general mixed-binary convex optimization problems. In contrast to existing methods that mainly rely on computationally expensive reoptimization or on Lagrangian dual multipliers derived from a convex relaxation, our proposed method offers greater flexibility and achieves a balance between computational efficiency and probing strength. Computational results also demonstrate the strong performance of our new methods.

2 - Warm-starting Sequences of Perturbed Mixed-Integer Programs

Anna Deza, ISyE Georgia Tech, Atlanta, GA, United States, Santanu Dey, Fabian Badilla Mera

Schrage and Wolsey (1985) propose a sensitivity analysis for integer programs that provides dual bounds parameterized by right-hand side perturbations utilizing the structure of branch-and-bound trees. Taking a similar branch-and-bound tree-based perspective, we show how to develop dual bounds that consider simultaneous perturbations in right-hand sides, objective coefficients, and constraint matrices. We then demonstrate how these insights enable effective warm starts for sequences of perturbed mixed-integer programs, relevant in time-sensitive settings where models must be re-solved as operational conditions evolve. We present numerical results for unit commitment problems, where frequent successive data updates and tight decision timelines make rapid re-optimization highly beneficial.

3 - Convexification of Classes of Mixed-Integer Sets with L-Natural-Convexity

Qimeng Yu, Universite de Montreal, Montreal, QC, Canada, Simge Kucukyavuz

$L^\natural$(natural)-convex functions form a broad class of nonlinear functions over general integer domains arising from wide-ranging real-world applications. We explore the minimization of $L^\natural$-convex functions, of multiple $L^\natural$-convex functions with common variables, and of a mixed-integer extension of $L^\natural$-convex functions---functions defined over a mixed-integer domain with properties that resemble $L^\natural$-convexity. For each of these families of minimization problems, we propose valid linear inequalities and provide convex hull descriptions for the corresponding epigraphs. For all classes of proposed inequalities, we discuss their facet conditions, develop exact separation methods, and analyze the complexity of the separation problem. We discover hidden $L^\natural$-convexity in well-known mixed-integer structures in the integer programming literature, namely the mixing set and the continuous mixing set. We show that our findings subsume the existing polyhedral results for these sets and establish new results for the multi-capacity variant of the continuous mixing set.

4 - Polynomial-Time Solvability of a Class of Concave Minimization Problems

Shaoning Han, National University of Singapore, Singapore, Singapore

---

Concave minimization (or convex maximization) has long been a central topic in optimization, valued for its rich mathematical structure and wide range of applications. Although general concave minimization is NP-hard, we show that the problem becomes polynomially solvable when the feasible region satisfies a certain order-based property and the objective function has fixed rank. When applied to sparse PCA and matroid-constrained problems, our result recovers and strengthens existing complexity results in the literature. Beyond its theoretical significance, this finding also offers useful insights for algorithmic developement.

SatC08

Committee Room

Optimization for Decision-Making

Invited Session

Emerging Applications of Optimization

Chair: Emily Zhang, Massachusetts Institute of Technology, Cambridge, MA, United States

1 - From Trees to Treewidth: Inventory Management in Complex Supply Chain Networks

Andre Calmon, Georgia Tech Scheller College of Business, Atlanta, GA, United States, Georgina Hall, Philippe Blaettchen, Mohit Tawarmalani

We propose an exact solution approach to the Guaranteed Service Model (GSM), one of the most widely applied models for optimizing safety stock placement in supply chain networks. Based on linear programming (LP), our approach handles any directed acyclic network and any cost function that depends on a stage's incoming and outgoing service times. It scales polynomially in the number of nodes n in the network, and exponentially in its *treewidth*, which quantifies how "tree-like" a network is, and can be much smaller than n. This contrasts with existing approaches, which scale exponentially in n. The proof of exactness of our approach relies crucially on showing that the join of transportation-like polytopes remains integral, and it is more broadly applicable to other Operations Management problems. The use of linear programming makes for straightforward implementation, including when incorporating additional operational constraints. It also enables sensitivity analyses and the construction of principled bounds on the GSM's optimal value. Finally, it allows for the use of standard LP optimization software, resulting in considerable gains in solving time. In particular, we demonstrate that, on real-world data from Willems (2008), we achieve consistent and significant optimization speed-ups compared to the state-of-the-art approach for the GSM and commercial all-purpose solvers. Overall, our approach builds a new bridge between Operations Management and Computer Science, thereby providing new theoretical foundations and practical tools for managing safety stocks in complex modern supply chain networks.

2 - Near-Optimal Monitoring in Delegated Clinical Trials

Jinze Cui, Duke University, Durham, NC, United States, Peng Sun

Delegating clinical trials to contract research organizations (CROs) improves efficiency but may lead to shirking. Sponsors can deter shirking through monitoring, but investigations are costly. We study how a sponsor can induce the CRO to exert effort on every trial participant at minimal cost. Using a mechanism design framework with costly state verification, we propose a simple yet effective policy, the *uniform-pass* contract: the sponsor investigates each participant with an equal probability and compensates the CRO only if no shirking is detected. We prove the contract is near-optimal. As the trial size n increases, the total cost converges to the theoretical minimum at the rate $O(1/\sqrt{n})$, while the expected number of investigations grows in $O(\log n)$. Calibrated to Phase 3 clinical trials with 500 to 3,000 participants, the total cost under the uniform-pass contract is only 1.5% to 7.6% above the theoretical minimum, and the sponsor needs to investigate only 28 to 54 participants in expectation. These results provide both theoretical support and an operationally simple solution for scalable monitoring in delegated clinical trials.

3 - Approximation Algorithms for Inventory Problems with Decomposable Submodular Ordering Costs

Emily Zhang, Massachusetts Institute of Technology, Cambridge, MA, United States, Retsef Levi, Georgia Perakis

We develop a constant-factor approximation algorithm for the submodular joint replenishment problem (SJRP) under a broad family of setup cost functions. We consider a structured class of setup cost functions that decompose over k item categories, where the cost depends on weighted aggregate quantities within each category and allows for arbitrary interactions across categories through a joint cost function. Our approach is based on rounding the solution to a linear programming relaxation. We introduce a novel water-filling procedure that partitions the fractional solution into a hierarchy of nested regions based on marginal costs. We then select a subset to order from each region, producing a feasible integral schedule. The resulting algorithm achieves an O(k)-approximation. When the number of categories k is constant, this yields the first constant-factor guarantee for this broad spectrum of submodular setup costs, significantly expanding the class of cost functions for which such guarantees are known.

4 - Optimizing Inventory for Impact: Large-Scale Inventory Routing to Distribute Clean Cooking Devices in Kenya

Maximilian Schiffer, Technical University of Munich, Munchen, Germany, Dan Iancu, Simon Thomae

Access to clean cooking technologies remains an issue in developing economies. Around 2.1 billion people around the globe rely on polluting fuels for cooking, leading to more than 3 million premature deaths annually. In Kenya, where most households rely on polluting fuels, bio-ethanol stove suppliers use carbon credits to subsidize stove prices. Still, affordability remains a challenge due to high distribution costs caused by highly stochastic and heterogeneous demand distributed over several thousand small retailers in Nairobi alone. We address this challenge by formulating a large-scale Inventory Routing Problem with stochastic demand and propose a new two-step approach to solve it. First, we relax the problem by taking long-term averages to compute target policies and valid lower bounds. Building on these target policies, we derive dynamic routing and replenishment policies that scale well to real-world instances with several thousand stores. Our numerical experiments show that the proposed methods significantly improve over the status quo, leading to significant cost savings

SatC09

Cabinet Room

Discrete Optimization for Machine Learning and Data Analytics

Invited Session

Discrete Optimization

Chair: Jiaqi Wang, Georgia Institute of Techonology, Atlanta, GA, United States

1 - D-Optimal Design with Conditional Treatment Effect

Jiaqi Wang, Georgia Institute of Techonology, Atlanta, GA, United States, Weijun Xie, Ilya Ryzhov

We study a variant of the D-optimal design problem in which treatment effects depend on the covariates, leading to a modified information matrix. We show that the generalized D-optimality criterion remains submodular, and we develop efficient algorithms to obtain high-quality solutions.

2 - Learning to Disable Globally-Valid Cuts

Zixuan Feng, University of Florida, Gainesville, FL, United States, Aleksandr Kazachkov, Kausthubh Konuru, Ambareesh Vaidya

Cutting planes, or cuts, are a relaxation tightening technique employed by mixed-integer programming solvers, due to the fact it improves the average solving time. However, this is not always the case. For some instances, adding cuts can actually slow down the solving process for reasons including but not limited to the inequalities created are too dense, or too many cuts are added. Our research focuses on the branch-and-cut method and aims to predict whether a problem will benefit from disabling globally-valid cuts with the help of machine learning methods. Specifically, we identify a set of features that are critical for the predictive model. Furthermore, we evaluate the impact of making this decision at different stages within the root node. Our study is tested with the MIPLIB 2017 benchmark dataset and showed promising results on improving the solving time by making adaptive choices.

3 - From Majorization to Scaling: Advancing Convex Relaxations of Maximum-Entropy Sampling Problem

Lingqing Shen, Carnegie Mellon University, Pittsburgh, PA, United States, Fatma Kılınc-Karzan

In this paper, we study the maximum-entropy sampling problem (MESP) and its variants. MESP seeks to identify a small subset of variables that maximizes the determinant of a covariance submatrix, serving as a fundamental model for optimal experimental design and information acquisition. Although the problem is combinatorial and NP-hard, continuous relaxations, most notably linx and Γ factorization, have provided

tractable approximations for MESP, though often with limited tightness. We introduce a novel double-scaling technique that generalizes the existing scaling techniques. We establish its theoretical properties, proving its dominance over previous scaling bounds, and develop an alternative characterization of Γ factorization-type relaxation that, for the first time, reveals formal dominance relations between the quality of linx- and Γ factorization-type formulations, as well as their scaling-strengthened versions. Our preliminary numerical study illustrates the effectiveness of our double-scaling method applied to linx relaxation against the other scaling techniques and relaxations from the literature.

4 - A Rawlsian Mixed-Integer-Programming Approach for Fair Classification

William Yang, Lawrence Livermore National Lab, Livermore, CA, United States, Ryan Lin, Chaoyue Zhao, Shuai Huang

Binary classification is a common machine learning methodology that aims to partition data into two classes as accurately as possible. One major drawback of these methods is that they are prone to producing unfair results. One real-life example of this is the Correctional Offender Management Profiling for Alternative Sanctions algorithm, which was designed to label individuals as "low" or "high" risk for recidivism. The algorithm produced predictions that had significantly worse accuracy for individuals in one demographic group than in others. While many methods have been proposed to address this, they are known to sacrifice overall prediction accuracy for the whole population. We present a Rawlsian fairness approach that prioritizes algorithmic performance for the worst-off demographic group without sacrificing overall performance. We couple our Rawlsian fairness approach with a scoring function using a mixed-integer-programming (MIP) formulation. Our approach assigns a score to each individual and assigns them to a class if it is above or below a specific threshold. By using an MIP approach, we can solve for the optimal threshold rather than relying on the user to identify the threshold before executing the algorithm. Our MIP approach is also flexible, as it provides the user with multiple ways to obtain sparse or integer solutions. Our results yield fair outcomes even when a particular group comprises a small proportion of the entire dataset and provide insight into how our methodology compares to other fairness approaches.

SatC10

Caucus Room

Computational Primitives for Modern Optimization

Invited Session

Computational Optimization

Chair: Adrian Vladu, CNRS, Université Paris Cité, Paris, France

1 - Linear Interior-Point Methods

Bento Natura, Columbia University, New York City, NY, United States

We study algorithms for linear programs of the form min c·x, Ax = b, x ≥ 0, that access the right-hand side b and objective c only through linear comparisons. An algorithm that operates under this restriction is called a linear algorithm. Such an algorithm can query an oracle which, given vectors u and v and scalars z and y, returns the signs of u·b + z and v·c + y. The simplex method, as well as several known strongly polynomial combinatorial algorithms, can be implemented in this model. We show that interior-point methods can likewise be adapted to run as linear algorithms, which yields faster algorithms for several classes of linear programs and, through the framework of Norton, Plotkin, and Tardos, extends the family of problems that admit strongly polynomial algorithms.

2 - Data Structures for Linear Algebra

Jan van den Brand, Georgia Tech, Atlanta, GA, United States

Optimization algorithms generally rely on iterative steps that repeatedly solve linear algebra subproblems, such as computing gradients, solving linear systems, maintaining inverses, or performing repeated of matrix-vector products. This talk presents data structures that efficiently compute and maintain solutions to these linear algebra subproblems from one iteration to the next.

3 - Entrywise Approximate Linear System Solving with Applications to Markov Chains

Mehrdad Ghadiri, MIT, Cambridge, MA, United States

Quantities such as escape probabilities in Markov chains can be recovered by solving linear systems in diagonally dominant matrices arising from graph Laplacians. However, these probabilities may vary exponentially across different nodes of the graph. As a result, under standard norm-based error guarantees, accurately recovering small entries requires exponentially small error parameters, leading to prohibitively large running times. For example, a direct application of existing Laplacian solvers or fast matrix multiplication algorithms requires $\Omega(mn^2)$ and $\Omega(n^{\omega+1})$ bit operations, respectively, where m denotes the number of nonzero entries, n is the matrix dimension, and ω is the matrix multiplication exponent.

In this work, we study faster algorithms for solving such linear systems with entrywise approximation guarantees. In particular, we present an almost-linear-time algorithm for solving symmetric diagonally dominant M-matrix (SDDM) linear systems with entrywise approximation guarantees.

4 - Fast Algorithms for Regression Via Iteratively Reweighted Least Squares

Adrian Vladu, CNRS, University Paris City, Paris, France

The iteratively reweighted least squares (IRLS) method is a widely used practical approach to solving regression problems. When the quality of fit is measured by the mean squared error, the problem reduces to ordinary least squares, and hence to solving a linear system. When other measures of error are used, such as the ℓ_∞ norm of the vector of prediction errors, the problem becomes substantially harder and, at high accuracy, is equivalent to linear programming. IRLS addresses these settings by solving a sequence of weighted least squares subproblems with adaptively chosen weights. However, many variants lack robust convergence guarantees, and existing analyses often fail to explain their empirical performance.

We present a unified class of IRLS algorithms for ℓ_p regression for all $p \in [1, \infty]$. Our methods match the best known theoretical guarantees, are simple to implement, and perform well in practice. We further extend this framework to overconstrained regression with quasi self-concordant penalties, which capture a broad family of loss functions, and obtain iteration bounds that depend only on the feature dimension rather than the number of data points.

Based on joint work with Alina Ene and Ta Duy Nguyen.

SatC11

Charter Room

Computation and Analysis of Fréchet Regression Problems

Invited Session

Network Optimization

1 - End-to-End Deep Learning for Predicting Metric Space-Valued Outputs

Yidong Zhou, Unviersity of California, Davis, Davis, CA, United States, SuI Iao, Hans-Georg Müller

Many modern applications involve predicting structured, non-Euclidean outputs such as probability distributions, networks, and symmetric positive-definite matrices. These outputs are naturally modeled as elements of general metric spaces, where classical regression techniques that rely on vector space structure no longer apply. We introduce E2M (End-to-End Metric regression), a deep learning framework for predicting metric space-valued outputs. E2M performs prediction via a weighted Fréchet means over training outputs, where the weights are learned by a neural network conditioned on the input. This construction provides a principled mechanism for geometry-aware prediction that avoids surrogate embeddings and restrictive parametric assumptions, while fully preserving the intrinsic geometry of the output space. We establish theoretical guarantees, including a universal approximation theorem that characterizes the expressive capacity of the model and a convergence analysis of the entropy-regularized training objective. Through extensive simulations involving probability distributions, networks, and symmetric positive-definite matrices, we show that E2M consistently achieves state-of-the-art performance, with its advantages becoming more pronounced at larger sample sizes. Applications to human mortality distributions and New York City taxi networks further demonstrate the flexibility and practical utility of the framework.

2 - On the Computation of Fréchet Regressors on $\mathbb{S}^n$

Yamin Zhou, Rice University, Houston, TX, United States, Cesar A. Uribe

This work investigates Fréchet regression and related barycenter-type optimization problems in general metric and Riemannian settings, with particular emphasis on the sphere. Fréchet regression models the relationship between scalar or vector-valued predictors and responses in a metric space by defining the regression output as the minimizer of an affine combination of distance-based loss functions. The weights in this affine combination are determined by the data, while the functions are constructed from the underlying metric. Existing formulations predominantly specify such cost functions but provide limited guidance on how to compute the minimizers or analyze the associated optimization problems, especially in non-Euclidean geometries.

In this study, the Fréchet regression loss is formulated as an affine combination of convex functions. Depending on the structure of the weights, the resulting problem may remain convex or become a genuine difference-of-convex (DC) problem. In the convex regime, established tools from convex optimization on manifolds yield direct solution methods. In the nonconvex regime, including signed-weight barycenter problems on hemispheres, the loss is represented as a DC function and the Difference-of-Convex Algorithm (DCA) on manifolds is employed. A rigorous existence theory is developed, establishing conditions under which minimizers of the regression loss exist in the manifold setting. In addition, a detailed and solid convergence analysis of manifold DCA is provided given the objective function, extending Euclidean convergence guarantees and connecting them with real-analytic structure and Kurdyka–Łojasiewicz-type inequalities. Together, these results offer a comprehensive theoretical and algorithmic framework for Fréchet regression in general positive constant curvature metric spaces.

3 - The Signed Wasserstein Barycenter Problem

Matt Jacobs, UCSB, Isla Vista, CA, United States

Barycenter problems encode important geometric information about a metric space. While these problems are typically studied with positive weight coefficients associated to each distance term, more general signed Wasserstein barycenter problems have recently drawn a great deal of interest. These mixed sign problems have appeared in statistical inference setting as a way to generalize least squares regression to measure valued outputs and have appeared in numerical methods to improve the accuracy of Wasserstein gradient flow solvers. Unfortunately, the presence of negatively weighted distance terms destroys the L^2 convexity of the unsigned problem, resulting in a much more challenging optimization task. In this talk, I will discuss some theoretical properties of these mixed sign barycenter problems, focusing on sufficient conditions to guarantee the global optimality and uniqueness of a critical point.

Joint work with Bohan Zhou.

4 - Fréchet Regression on the Bures-Wasserstein Manifold: Existence, Optimization, and Applications

Duc Toan Nguyen, Rice University, Houston, TX, United States, Cesar A. Uribe

Fréchet regression is a flexible framework for modeling relationships in which the response variables are complex non-Euclidean objects, such as networks or probability distributions. In this work, we focus on the problem of estimating the conditional Fréchet mean on the Bures-Wasserstein manifold, the space of positive-definite matrices equipped with the Wasserstein metric. This estimation task leads to an optimization problem closely related to the Wasserstein barycenter problem, but with the notable extension of allowing negative weights. We establish a sufficient condition that guarantees the existence of a minimizer for this problem. To compute the estimator, we develop a Riemannian gradient descent algorithm and prove that it converges at a sublinear rate. Finally, we validate our approach through empirical experiments in a network regression setting.

March 21, 2026, 4:30 PM - 6:00 PM

SatD02

Georgian Room

First Order Methods for Machine Learning and Data Science

Invited Session

Optimization in Data Science

Chair: Yao Yao, University of Minnesota, Minneapolis, MN, United States

Co-Chair: Yankun Huang, Arizona State University, 101 W 5th St Apt 3043, Tempe, AZ, 85281, United States

1 - Feature Learning for High Dimensional PDE with Deep Ritz Method

Yao Yao, University of Minnesota, Minneapolis, MN, United States, Yulong Lu, Gilad Lerman

This paper investigates feature learning within the framework of the deep Ritz method for solving the stationary Schrödinger equation with Neumann boundary conditions. We first analyze the convergence of Riemannian gradient descent in an agnostic setting, where the hypothesis function is restricted to a single-index model while the PDE solution is arbitrary. We prove that gradient descent reaches an approximate global minimum: after $T = O(\log(1/\varepsilon))$ iterations, the loss is within ε of a constant multiple of the optimal loss. We then examine the loss landscape when the source term of the PDE itself follows a single-index model, considering hypothesis functions given by either a single-index model or a two-neuron multi-index model. In the single-index case, we show that the minimum Ritz energy is attained at the feature vector aligned with that of the source term. In the two-neuron case, we study the landscape of regularized Ritz losses and characterize how a second feature emerges, given that the first feature is aligned with the source, as the regularization parameter varies. Finally, numerical experiments are presented to validate the feature emergence theory in the two-neuron setting.

2 - A Difference-of-Convex Optimization Approach for Structured Bilevel Problems with Applications to Distributionally Fair Stackelberg Games

Yutian He, Department of Mathematics, University of Iowa, Iowa City, IA, United States, Beste Basciftci, Qihang Lin

A Stackelberg game is a sequential game in which a leader acts first, anticipating the responses of one or more followers who, in turn, optimize their own objectives given the leader's decision. This framework naturally gives rise to bilevel optimization problems. While classical studies of Stackelberg games primarily emphasize utility maximization, fairness considerations have received little attention.

In this work, we consider a setting with multiple groups whose utilities depend jointly on the decisions of both the leader and the followers. These group utilities may differ from those of the leader and followers themselves. We focus on a distributionally fair Stackelberg game in which the leader seeks to minimize unfairness—quantified by the Wasserstein distances among the utility distributions of different groups—subject to maintaining its own utility above a prescribed threshold.

We reformulate the resulting nonsmooth bilevel problem as a single-level nonsmooth difference-of-convex (DC) program with DC constraints. We then develop efficient lower-bounding schemes and propose tailored DC algorithms to solve these problems effectively. Numerical experiments on capacity planning and wholesaler–retailer games illustrate the value of incorporating fairness and demonstrate the efficiency of the proposed approach.

3 - Stochastic Primal-Dual Double Block-Coordinate for Two-way Partial AUC Maximization

Linli Zhou, Texas A&M University , College Station, TX, United States, Bokun Wang, My Thai, Tianbao Yang

We propose two stochastic primal-dual double block-coordinate algorithms for two-way partial AUC (TPAUC) maximization, addressing limitations of existing methods. Our approach improves convergence rates and achieves superior empirical performance on benchmark datasets for imbalanced classification.

4 - Inexact Moreau Envelope Lagrangian Method for Non-Convex Constrained Optimization under Local Error Bound Conditions on Constraint Functions

Yankun Huang, Arizona State University, Tempe, AZ, United States, Qihang Lin, Yangyang Xu

In this paper, we study the inexact Moreau envelope Lagrangian (iMELa) method for solving smooth non-convex optimization problems over a simple polytope with additional convex inequality constraints. By incorporating a proximal term into the traditional Lagrangian function, the iMELa method approximately solves a convex optimization subproblem over the polyhedral set at each main iteration. Under the assumption of a local error bound condition for subsets of the feasible set defined by subsets of the constraints, we establish that the iMELa method can find an ϵ-Karush-Kuhn-Tucker point with $\tilde O(\epsilon^{-2})$ gradient oracle complexity.

SatD03

Plaza I

Recent Advances in Derivative-Free Optimization III

Invited Session

Nonlinear Optimization

Chair: Zikai Xiong, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Abraar Chaudhry, Georgia Tech, Atlanta, GA, United States

Co-Chair: Katya Scheinberg, Georgia Tech, Atlanta, GA, United States

1 - Noise-Aware Scalable Stochastic Trust-Region via Adaptive Random Subspaces

Kwassi Joseph Dzahini, Argonne National Laboratory, Lemont, IL, United States, Jeffrey Larson, Matt Menickelly, Stefan Wild

We introduce ANASTAARS, a noise-aware stochastic trust-region algorithm with adaptive random subspace strategies that is effective from low to potentially large dimensions. The method achieves scalability by optimizing random models restricted to low-dimensional affine subspaces defined via Johnson--Lindenstrauss transforms, thereby sharply reducing function evaluation cost per iteration. Unlike earlier random subspace approaches with fixed dimensions and fully rebuilt poised sets, ANASTAARS adaptively selects the subspace dimension and updates the model by adding only a few new interpolation points while reusing past evaluations in a way that preserves poisedness, yielding substantially more efficient progress. To handle noisy objectives, it incorporates a noise-aware trust-region mechanism based on local estimates of the noise level, and its effectiveness is demonstrated on problems arising from the quantum approximate optimization algorithm (QAOA) framework.

2 - CatMADS: Categorical Variables with the MADS Algorithm

Edward Halle-Hannan, Polytechnique Montreal, Montreal, QC, Canada, Charles Audet, Youssef Diouane, Sebastien Le Digabel, Christophe Tribes

Solving optimization problems in which functions are blackboxes and variables involve different types poses significant theoretical and algorithmic challenges. Nevertheless, such settings frequently occur in simulation-based engineering design and machine learning. This paper extends the Mesh Adaptive Direct Search (MADS) algorithm to address mixed-variable problems with categorical, integer and continuous variables. MADS is a robust derivative-free optimization framework with a well-established convergence analysis for constrained quantitative problems. CatMADS generalizes MADS by incorporating categorical variables through distance-induced neighborhoods. A detailed convergence analysis of CatMADS is provided, with flexible choices balancing computational cost and local optimality strength. Four types of mixed-variable local minima are introduced, corresponding to progressively stronger notions of local optimality. CatMADS integrates the progressive barrier strategy for handling constraints, and ensures Clarke stationarity. An instance of CatMADS employs cross-validation to construct problem-specific categorical distances. This instance is compared to state-of-the-art solvers on 60 mixed-variable problems, half of which are constrained. Data profiles show that CatMADS achieves the best results, demonstrating that the framework is empirically efficient in addition to having strong theoretical foundations.

3 - Computer-aided Analysis of DFO for Noisy Smooth Convex Function with Few Gradient Oracles

Jingye Xu, Georgia Institute of Technology, Atlanta, GA, United States, Katya Scheinberg

Derivative-free optimization methods are essential when gradient information is unavailable or unreliable, but their performance deteriorates in the presence of noisy function evaluations. In particular, finite-difference gradient approximations face a fundamental dilemma: small step sizes yield accurate gradients but amplify noise, while large step sizes improve robustness at the cost of bias. This work investigates whether access to a limited number of exact gradient oracle calls can fundamentally mitigate this trade-off for smooth convex optimization. Leveraging global interpolation properties of smooth convex functions, we show how exact gradients can be used to refine noisy function evaluations via computable upper and lower bounds. To rigorously assess their worst-case performance, we formulate the analysis as a Performance Estimation Program which enables systematic, quantitative comparison of oracle allocation strategies and provides new insights into the value of sparse gradient information in noisy derivative-free optimization.

4 - Non-monotone Methods for Stochastic Derivative-free Optimization

Anjie Ding, Lehigh University, Bethlehem, PA, United States, Trang Tran, Luis Nunes Vicente

In derivative-free optimization (DFO) one minimizes functions for which the gradient is unavailable or expensive to compute. In many applications, objective function values and gradients are noisy due to simulations or system randomness. Current DFO algorithms accept trial points when a certain monotone decrease of the objection function is achieved. However, when applied to complex landscapes, such a requirement may trap the algorithm in a neighborhood of sub-optimal solutions. Non-monotone techniques allow the acceptance of trial points with temporary increases in the objective value, but have only been developed in the deterministic case. The innovation in this study is the development of non-monotone DFO techniques in the stochastic case where evaluations are noisy.

We have been investigating various non-monotone strategies, including those that replace the estimated function value at the trial point with a convex combination of past and present estimates, or with the maximum of previous observed function values. In this talk, we will present a numerical comparison of the various non-monotone techniques for both linesearch and direct-search methods. Our results demonstrate that non-monotone techniques exhibit superior performance compared to their monotone counterparts. This empirical evidence motivates our current theoretical investigation of the rate of convergence of non-monotone stochastic DFO algorithms, and we will also present our current findings.

SatD04

Plaza II

Recent Advances in Bilevel Optimization and Applications

Invited Session

Nonlinear Optimization

Chair: Mingyi Hong, Mingyi Hong, WAYZATA, United States

Co-Chair: Jiaxiang Li, Meta, Santa Clara, CA, United States

1 - A Decomposition Framework for Nonlinear Nonconvex Two-Stage Optimization

Yuchen Lou, Northwestern University, Evanston, IL, United States, Andreas Waechter, Ermin Wei, Xinyi Luo

We propose a new decomposition framework for continuous nonlinear constrained two-stage optimization, where both first- and second-stage problems can be nonconvex. A smoothing technique based on an interior-point formulation renders the optimal solution of the second-stage problem differentiable with respect to the first-stage parameters. As a consequence, efficient off-the-shelf optimization packages can be utilized. We show that the solution of the nonconvex second-stage problem behaves locally like a differentiable function so that existing proofs can be applied for the global convergence of the first-stage. We also prove fast local convergence of the algorithm as the barrier parameter is driven to zero. Numerical experiments for large-scale instances demonstrate the computational advantages of the decomposition framework.

2 - First-order Methods for Bilevel Optimization

Sanyou Mei, The Hong Kong University of Science and Technology, Hong Kong, China, People's Republic of, Zhaosong Lu

Bilevel optimization, also known as two-level optimization, is an important branch of mathematical optimization. It has found applications across various domains, including economics, logistics, supply chain, transportation, engineering design, and machine learning. In this talk, we will present first-order methods for solving a class of bilevel optimization problems using either single or sequential minimax optimization schemes. We will also discuss the first-order operation complexity of these methods and present preliminary numerical results to illustrate their performance.

3 - A Correspondence-Driven Approach for Bilevel Decision-making with Nonconvex Lower-Level Problems

Xiaotian Jiang, University of Minnesota, Minneapolis, MN, United States, Jiaxiang Li, Jiawen Bi, Mingyi Hong, Shuzhong Zhang

We consider bilevel optimization problems with general nonconvex lower-level objectives and show that the classical hyperfunction-based formulation is unsettled, since the global minimizer of the lower-level problem is generally unattainable. To address this issue, we propose a correspondence-driven hyperfunction ϕ^{cd}. In this formulation, the follower is modeled not as a rational agent always attaining a global minimizer, but as an algorithm-based bounded rational agent whose decisions are produced by a fixed algorithm with initialization and step size. Since ϕ^{cd} is generally discontinuous, we apply Gaussian smoothing to obtain a smooth approximation ϕ^{cd}_ξ, then show that its value and gradient converge to those of ϕ^{cd}. In the nonconvex setting, we identify that bifurcation phenomena, which arise when $g(x,\cdot)$ has a degenerate stationary point, pose a key challenge for hyperfunction-based methods. This is especially the case when ϕ^{cd}_ξ is solved using gradient methods. To overcome this challenge, we analyze the geometric structure of the bifurcation set under some weak assumptions. Building on these results, we design a biased projected SGD-based algorithm to solve ϕ^{cd}_ξ with a cubic-regularized Newton lower-level solver. We also provide convergence guarantees and oracle complexity bounds for the upper level.

4 - Contextual Stochastic Bilevel Optimization

Jie Wang, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, People's Republic of, Yifan Hu, Yao Xie, Andreas Krause, Daniel Kuhn

We introduce contextual stochastic bilevel optimization (CSBO) -- a stochastic bilevel optimization framework with the lower-level problem minimizing an expectation conditioned on some contextual information and the upper-level decision variable. This framework extends classical stochastic bilevel optimization when the lower-level decision maker responds optimally not only to the decision of the upper-level decision maker but also to some side information and when there are multiple or even infinite many followers. It captures important applications such as meta-learning, personalized federated learning, end-to-end learning, and Wasserstein distributionally robust optimization with side information (WDRO-SI). Due to the presence of contextual information, existing single-loop methods for classical stochastic bilevel optimization are unable to converge. To overcome this challenge, we introduce an efficient double-loop gradient method based on the Multilevel Monte-Carlo (MLMC) technique and establish its sample and computational complexities. When specialized to stochastic nonconvex optimization, our method matches existing lower bounds. For meta-learning, the complexity of our method does not depend on the number of tasks. Numerical experiments further validate our theoretical results. We also discuss how to extend CSBO to the case where the lower-level problems are infinite-dimensional functional optimization over probabiltiy space. The new problem finds applications in Sinkhorn distributionally robust optimization and meta learning for diffusion models.

SatD05

Plaza III

Risk-Aware Optimization for Power and Energy: From Algorithms to Resilient Grids

Invited Session

Optimization under Uncertainty

Chair: Robert Mieth, Rutgers University, Piscataway, NJ, United States

Co-Chair: Gejia Zhang, Rutgers University, Piscataway, NJ, United States

1 - Flow Matching for Stochastic Policies in Reinforcement Learning

Grant Ruan, MIT, Boston, MA, United States

In electric grid and many engineering systems, there are extensive coordinated decision-making problems. AI models are applied in these systems to automate certain decision steps, and it is often expensive or even impractical to dynamically update these AI models after deployment. There is a promising solution called inference-time steering, which applies lightweight adaptation at inference time so that no retraining/fine-tuning will be needed. I consider this setting in reinforcement learning, allowing the test environment deviate from the training. I leverage generative AI (flow matching) to establish a stochastic policy for multi-modal action generation (multi-peak distributions). It can further work with extra steering steps to adapt to various kinds of environmental shifts. More broadly, I will introduce the latest advances of flow matching and diffusion policies in reinforcement learning in this talk, and demonstrate the computational challenges that motivate the need of fast inference.

2 - Scalable Power Distribution System Restoration with Uncertainty-Aware Power Flow Surrogate

Lusha Wang, University of Alabama, Tuscaloosa, AL, United States

Restoring power distribution systems after major outages requires rapid coordination and reliable decision-making under uncertainty. This talk introduces a scalable restoration framework that combines advanced optimization with fast data-driven system awareness to support these challenging operations. The approach integrates efficient crew routing and network reconfiguration strategies with a lightweight surrogate for power-flow assessment, enabling near–real-time evaluation of restoration actions without relying on slow traditional simulations. Together, these components provide a practical and scalable tool that improves restoration speed, situational awareness, and resilience. Case studies on large distribution feeders illustrate how the framework can support more informed operational decisions and enhance the grid's ability to recover from extreme events.

3 - Graph-Based Reliability Modelling For Power Distribution System

Gejia Zhang, Rutgers University, Piscataway, NJ, United States, Robert Mieth

Rising demand and electrification raise concerns about whether future power network can remain adequate and dependable under heavier and more variable loads. Most customer interruptions originate within distribution networks, and outage risk is closely tied to system loading and environmental conditions, which implies that reliability depends not only on external stressors but also on operational choices. We address the need for an operational reliability metric by focusing on radial distribution systems. Viewing the grid as a graph enables the use of spectral tools. Prior work has employed the graph Laplacian and effective resistance to represent connectivity, but connectivity alone does not capture the physics of power flow. We propose a reliability metric grounded in graph theory that builds on effective resistance while embedding network operation through AC-OPF. The metric quantifies system reliability in a way that reflects both topology and operating conditions, providing a basis for incorporating reliability into routine decision making.

4 - Nonstationary Satisficing Bandits

Yixuan Zhang, University of Wisconsin-Madison, Madison, WI, United States, Ruihao Zhu, Qiaomin Xie

Motivated by the notion of satisficing in decision making, we study satisficing-regret minimization in nonstationary K-armed bandits. We establish an optimal worst-case regret of $\Theta(\log T)$; under additional structure, this can be sharpened to $\Theta(\log\log T)$ and, in some cases, $\Theta(1)$.

5 - Initial Noise Optimisation for Diffusion Inverse Problem solving

Guanyang Wang, Rutgers University, Piscataway, NJ, United States

We study how to choose or adjust the initial noise in diffusion models to improve inference for inverse problems. We introduce a noise optimization framework that treats the starting latent variable as a controllable input. By differentiating the loss through the full reverse diffusion process, we obtain a tractable gradient for the initial noise. This gives an efficient way to steer the model toward reconstructions with smaller error and more stable uncertainty estimates.

SatD06

Director's and Lounge

Heuristic and Hybrid Methods in Global Optimization

Invited Session

Global Optimization

Chair: Rohit Kannan, Virginia Tech, Blacksburg, VA, United States

Co-Chair: Dominic Flocco, University of Maryland, 1301 M ST NW APT 513, Washington, 20005, United States

1 - Hybrid Metaheuristic-Multiparametric Optimization Strategy for Bilevel Programming

Styliani Avraamidou, University of Wisconsin Madison, Madison, WI, United States, Meng-Lin Tsai

Bilevel optimization can model hierarchical optimization problems for a multitude of applications, including environmental policy [1]. Bilevel optimization problems are very challenging to solve - they are NP-hard even for the linear continuous case [2]. Common analytical approaches include KKT-based reformulations [3], branch-and-bound/cut, and multi-parametric reformulation [4]. These methods are either computationally inefficient and struggle with medium to large scale problems, or cannot be used for bilevel problems with lower-level discrete variables.

In contrast, heuristic approaches (e.g. DOMINO [1]) use a hybrid strategy: a deterministic solver handles the lower-level problem, which is then treated as a black box by the upper-level metaheuristic solver. These methods guarantee bilevel feasibility for some classes of problems, but offer no finite-time guarantees regarding convergence to either a global or even a local optimum. We propose a critical region-evaluation strategy, adapted from multi-parametric optimization, to enhance black box methods and improve their local search efficiency and solution quality.

We validate the method, comparing its performance against both deterministic multi-parametric reformulation and established global heuristic optimization algorithms. These results demonstrate that combining global search heuristics and local parameter space analysis offers a balance of the metaheuristic and multiparametric approach solving challenging classes of bilevel optimization problems, opening new pathways for applications in engineering.

1. Beykal B., et al. Journal of Global Optimization, 2020.

2. Hansen B.P., et al. SIAM Journal on Scientific and Statistical Computing, 1992.

3. Fortuny-Amat J., McCarl B. Journal of the operational Research Society, 1981.

4. Avraamidou S., Pistikopoulos E.N. Computers & Chemical Engineering, 2019.

2 - Accelerating AC Unit Commitment via SOC Relaxations and Learning-based Commitment Prediction

Berkay Becu, Georgia Institute of Technology, Atlanta, GA, United States, Santanu Dey

Alternating Current Unit Commitment (AC-UC) is a challenging large-scale mixed-integer nonlinear problem that has binary variables for generator commitment decisions with full AC network physics. Solving AC-UC directly is computationally expensive due to its non-convex power-flow equations and binary variables. In this work, we considered the Second-Order Cone (SOC) relaxation to Optimal Power Flow (OPF) problem and extended it to UC. Next, we structured methods to use historical data to decide on some set of commitment decisions to accelerate the solution of the full AC-UC problem. The SOC-UC model provides tight convex relaxations, yielding high-quality lower bounds. Using SOC-UC solutions from a collection of perturbed PGLib benchmark instances, we built a library of historical commitments and associated features describing system-level demand and generator costs. Two learning-assisted approaches are developed to predict commitment decisions for new AC-UC instances. The first, based on k-Nearest Neighbor (kNN) method, identifies the most similar SOC-UC training cases and selects their commitment schedules through majority voting. The second uses Support Vector Machines (SVMs) to learn hyperplanes that classify each binary variable into on/off or temporally consistent states based on demand and cost features. For both methods, predicted binary variables are fixed before solving AC-UC, reducing the decision space. Experiments on 50 training and 5 test instances demonstrate that kNN and SVM both recover high-quality commitment patterns while reducing AC-UC solve times. These results illustrate the potential of integrating convex relaxations with learning-based prediction to enable scalable global optimization for power-system scheduling.

3 - Diving on Continuous Variables in Mixed-Integer Nonlinear Optimization

Sourabh Kumar Choudhary, Georgia Institute of Technology, Atlanta, GA, United States, Santanu Dey, Nikolaos Sahinidis

Diving heuristics are widely used in modern solvers for Mixed Integer Programming (MIP) problems. However, general-purpose solvers restrict diving to integer variables and do not dive on continuous variables. Motivated by early success on urban-water-network optimization instances, we introduce the so-called Continuous-Variable-Diving (CVD) for general Mixed Integer Nonlinear Programming (MINLP) problems. Central to CVD is creating a MIP approximation of the MINLP. The MIP solution guides subsequent bound contraction of the continuous variables. Experimental results on MIQCQPs from QPLIB show significant gap reductions when our heuristic is compared with SCIP.

4 - A Successive Proximal DC Penalty Method with Application to Mathematical Programs with Complementarity Constraints

Dominic Flocco, University of Maryland, Washington, DC, United States, Steven Gabriel, Trine Boomsma, Martin Schmidt, Miguel Lejeune

We develop a successive proximal difference-of-convex function (DC) penalty method for solving DC programs with DC constraints. The proposed approach relies on a DC penalty function that measures violation of constraints and leads to a penalty reformulation sharing the same solution set as the original problem. The resulting penalty problem is a DC program with convex constraints and is solved using a proximal DC algorithm (DCA) with successive DC decomposition. We establish convergence to a stationary point of the penalty problem under milder constraint qualification assumptions than those typically required in the DC programming literature. We then specialize the framework to mathematical programs with complementarity constraints (MPCCs), which cover a large class of optimization and equilibrium problems arising naturally in engineering, economics, and game theory. For this application, we introduce a novel DC penalty function that penalizes violations of the complementarity constraints and establish a correspondence between S-stationary points of the MPCC and stationary points of the resulting DC penalty reformulation. Numerical experiments on MPCCs arising from bilevel optimization demonstrate that the proposed method, referred to as SPDCA, efficiently computes high-quality solutions, particularly for instances with DC, quadratic upper-level objectives, outperforming conventional mixed-integer formulations in both runtime and solution quality.

SatD07

Congress Room

Recent Advances in Mixed Integer Programming II

Invited Session

Discrete Optimization

Chair: Soroosh Shafiee, Cornell University, Ithaca, NY, United States

1 - MINLP Relaxations Exploiting Computational Geometry Tools

Mohit Tawarmalani, Purdue University, West Lafayette, IN, United States, Haisheng Zhu, Taotao He

In this talk, we propose new polyhedral relaxations for factorable programming that leverage computational geometry tools and close significant gap relative to state-of-the-art solvers. We show that our relaxations are tighter than the traditional factorable programming relaxation and the recently proposed composite relaxations when inner functions are univariate. Our proposed relaxations exploit function and domain structure to simultaneously convexify function graphs over arbitrary domains. Using computational geometry tools, we implement our relaxations for low-dimensional functions. We share our preliminary computational experience which demonstrates that our relaxations close over 40% gap on 15% of the instances from minlplib.org.

2 - GPU-Accelerated First-Order Method with Randomized Sampling for Binary Integer Programs

Ningji Wei, Texas Tech University, Lubbock, TX, United States, Jiaming Liang

We present GFORS, a GPU-accelerated framework for large binary integer programs. It couples a first-order (PDHG-style) routine that guides the search in the continuous relaxation with a randomized, feasibility-aware sampling module that generates batched binary candidates. Both components are designed to run end-to-end on GPUs with minimal CPU-GPU synchronization. The framework establishes near-stationary-point guarantees for the first-order routine and probabilistic bounds on the feasibility and quality of sampled solutions, while not providing global optimality certificates. To improve sampling effectiveness, we introduce techniques such as total-unimodular reformulation, customized sampling design, and monotone relaxation. On classic benchmarks (set cover, knapsack, max cut, 3D assignment, facility location), baseline state-of-the-art exact solvers remain stronger on small-medium instances, while GFORS attains high-quality incumbents within seconds; on large instances, GFORS yields substantially shorter runtimes, with solution quality often comparable to—or better than—the baseline under the same time limit. These results suggest that GFORS can complement exact solvers by delivering scalable, GPU-native search when problem size and response time are the primary constraints.

3 - Solving Convex Quadratic Optimization with Indicators over Graphs with Small Treewidth

Salar Fattahi, University of Michigan, Ann Arbor, MI, United States

We study convex quadratic optimization problems where each continuous variable is coupled with an indicator variable. Our focus is on the structured case in which the matrix defining the quadratic term is positive definite, with a sparsity pattern corresponding to the adjacency matrix of a graph with small treewidth. We introduce a dynamic programming algorithm that solves this problem in linear time, under certain assumptions on the underlying graph. Computational experiments on synthetic data demonstrate that our algorithm significantly outperforms state-of-the-art mixed-integer quadratic programming solvers on instances with over 20,000 indicator variables.

4 - Strong Formulation for Mixed-Integer Quadratic Optimization via Dynamic Programming

Andres Gomez, University of Southern California, Los Angeles, CA, United States

We consider a broad class of mixed-integer quadratic optimization problems that are amenable to dynamic programming algorithms. We propose a general dynamic programming algorithm to solve problems considered. While exponential in the worst case, we show that the proposed method runs in polynomial or pseudo-polynomial time in various settings of interest, including when the matrix defining the quadratic term is low rank or sparse. We show how to derive ideal formulations from the dynamic programs, that can then be combined with general branch-and-bound algorithms.

SatD08

Committee Room

Applications of Optimization for Strategic and Data-Driven Decision Making

Invited Session

Emerging Applications of Optimization

Chair: Apurv Shukla, University of Michigan, Ann Arbor, MI, United States

1 - Graphon Mechanism Design for Large-Scale Networked Systems

Apurv Shukla, University of Michigan, Ann Arbor, MI, United States

We introduce a general mechanism-design framework for large networked systems whose interaction

structure converges to a graphon.
Starting from a finite network of strategically interacting agents with convex coupling constraints, we show that as the network grows dense, the dual variables associated with global and networked resources converge to functional shadow-price fields.
These dual fields induce graphon price kernels that internalize heterogeneous externalities across the network and characterize the welfare-maximizing allocation in the graphon mean-field limit. Building on this structure, we construct a graphon mechanism that is (i) dominant-strategy incentive compatible, (ii) efficient, and (iii) strongly budget balanced in the nonatomic limit. A key feature of the mechanism is superimposability: the payment rule can be applied directly on top of any distributed primal–dual algorithm used to solve the planner's problem, without modifying its internal updates. Thus, the mechanism integrates seamlessly with existing optimization architectures.
We extend the framework to dynamic environments in which agents' private types evolve according to controlled Markov processes.
The resulting dynamic mechanism remains sequentially incentive compatible, individually rational, and superimposable at every stage.
A worked example on thermal coupling in large data centers illustrates how graphon-induced externalities and functional shadow prices naturally arise in practice. Our results generalize the Large-Scale NUM mechanism from complete networks to arbitrary dense interaction topologies, yielding the first fully structure-aware, truthful, and distributed mechanism design framework for graphon-coupled systems.

2 - Online Learning for Equilibrium Pricing in Markets under Incomplete Information

Navid Azizan, Massachusetts Institute of Technology, Cambridge, MA, United States, Devansh Jalota, Haoyuan Sun

The computation of equilibrium prices at which the supply of goods matches their demand typically relies on complete information on agents' private attributes, e.g., suppliers' cost functions, which are often unavailable in practice. Motivated by this practical consideration, we consider the problem of learning equilibrium prices over a horizon of T periods in the incomplete information setting wherein a market operator seeks to satisfy the customer demand for a commodity by purchasing it from competing suppliers with cost functions unknown to the operator. We first consider the setting when suppliers' cost functions are fixed and develop algorithms that, on three pertinent regret metrics, simultaneously achieve a regret of $O(1)$ when the customer demand is constant over time, and $O(\sqrt{T})$ when the demand varies over time. In the setting when the suppliers' cost functions vary over time, we demonstrate that, in general, no online algorithm can achieve sublinear regret on all three metrics. Thus, we consider an augmented setting wherein the operator has access to hints/contexts that reflect the variation in the cost functions and propose an algorithm with sublinear regret in this augmented setting. Finally, we present numerical experiments that validate our results and discuss various model extensions.

3 - k-SVD with Gradient Descent

Yassir Jedra, Imperial College London, London, United Kingdom, Devavrat Shah

The emergence of modern compute infrastructure for iterative optimization has led to great interest in developing optimization-based approaches for a scalable computation of k-SVD, i.e., the $k \geq 1$ largest singular values and corresponding vectors of a matrix of rank $d \geq 1$. Despite lots of exciting recent works, all prior works fall short in this pursuit. Specifically, the existing results are either for the *exact-parameterized* (i.e., $k = d$) and *over-parameterized* (i.e., $k > d$) settings; or only establish local convergence guarantees; or use a step-size that requires problem-instance-specific oracle-provided information. In this work, we complete this pursuit by providing a gradient-descent method with a simple, universal rule for step-size selection (akin to pre-conditioning), that provably finds k-SVD for a matrix of any rank $d \geq 1$. We establish that the gradient method with random initialization enjoys global linear convergence for any $k, d \geq 1$. Our convergence analysis reveals that the gradient method has an attractive region, and within this attractive region, the method behaves like Heron's method (a.k.a. the Babylonian method). Our analytic results about the said attractive region imply that the gradient method can be enhanced by means of Nesterov's momentum-based acceleration technique. The resulting improved convergence rates match those of rather complicated methods typically relying on Lanczos iterations or variants thereof. We provide an empirical study to validate the theoretical results.

4 - Multi-stage Stochastic Programming with Decision-dependent Uncertainty: An Application in Hurricane Evacuation Planning Problem

Yongjia Song, Clemson University, Clemson, SC, United States, Merve Bodur, Margarita Castro

In this talk, we will discuss multi-stage stochastic programming (MSP) models and solution approaches for hurricane evacuation planning problems. We consider logistics decision-making such as the relief item prepositioning, evacuation shelter planning, and contingency modality selection over multiple periods prior to the landfall of an impending hurricane. Using stochastic models that capture the rolling forecast information about the hurricane's attributes over time, as well as hurricane evacuation behaviorial analysis on how the likehood of evacuation may depend on the level of hurricane evacuation order (e.g., mandatory vs. voluntary), we propose MSP models which provide optimal adaptive logistics decision policies. These MSP models feature decision-dependent uncertainty due to the interactions between the evacuation modality selection and evacuation demand uncertainty. Our numerical results and sensitivity analyses demonstrate the value of MSP for hurricane evacuation planning, as well as the trade-offs between policy flexibility, solution quality, and computational effort.

SatD09

Cabinet Room

Constraint Learning: When a Neural Network Walks into an Optimization Model

Invited Session

Discrete Optimization

Chair: Thiago Serra, University of Iowa, Iowa City, IA, United States

1 - Feasibility-Guaranteed Constraint Learning Using Conformal Prediction

Daniel Ovalle Varela, Carnegie Mellon University, Pittsburgh, PA, United States, Lorenz Biegler, Ignacio Grossmann, Carl Laird, Mateo Dulce

Many optimization problems in engineering rely on constraints derived from complex physical models, simulations, or empirical observations. In many cases, these relationships are unknown, too expensive to evaluate repeatedly, or impossible to express analytically, motivating the use of constraint learning, where machine learning (ML) models approximate feasible regions directly from data. While this approach enables decision-making in settings where first-principles formulations are unavailable, standard mixed-integer constraint learning (MICL) can produce solutions that satisfy the learned constraints yet violate the true system limitations, since prediction error is not explicitly handled by the optimization model.

To address this gap, we propose Conformal Mixed-Integer Constraint Learning (C-MICL), a framework that embeds rigorous statistical guarantees into learned constraints. Instead of optimizing against point predictions, C-MICL applies conformal prediction to construct data-driven uncertainty regions around the ML outputs. Any decision that satisfies these regions is guaranteed to be feasible with probability at least $1-\alpha$, under a conditional independence assumption. The approach supports regression and classification settings and is compatible with any MIP-representable surrogate, including ReLU neural networks. Experimental results across several applications show that C-MICL reliably achieves target coverage levels, maintains competitive objective performance, and provides a scalable pathway for integrating uncertainty-aware constraint learning into optimization with formal statistical guarantees.

2 - Choosing the Right Neural Network for Optimization Models: A Systematic Trade-off Study

Jiaxiao Fang, University of Iowa, Iowa City, IA, United States, Thiago Serra

Embedding neural networks into optimization models allows us to approximate complex system components while maintaining tractability. However, larger networks with more accurate predictions often make the optimization problem that includes the neural network much harder to solve, resulting in significantly slower optimization solver performance. Although previous studies have discussed individual factors that influence this trade-off, the insights are scattered and lack a unified evaluation.

In this work, we systematically study how different neural network design choices, including the number of neurons, pruning levels, and other architectural parameters, affect both prediction accuracy and the computational performance of the resulting optimization model. Using the same dataset, we train a collection of neural networks with varying configurations and construct an accuracy-size frontier. We then measure how these configurations impact optimization runtime and the quality of solutions relative to ground truth.

Our results highlight cases where smaller or more heavily pruned networks can preserve solution quality while reducing computational cost, as well as cases where additional complexity improves accuracy but slows down the optimization process. This study provides a clearer understanding of the accuracy-complexity trade-offs involved in embedding neural networks in optimization models and offers practical guidance for selecting appropriate neural network architectures under different computational budgets.

3 - Chance-Constrained Optimization with Neural Networks

Madeline Colbert, University of Iowa, Iowa City, IA, United States, Shunyu Yao, Justin Dumouchelle, Beste Basciftci, Thiago Serra

Chance-constrained optimization problems require reformulations to obtain their computationally tractable representations. Scenario-based formulations are commonly considered to reformulate these problems, despite the high computational burden they require with the introduction of big-M constraints and auxiliary binary variables, leading to large-scale mixed-integer linear programs. To address this issue, we consider embedding quantile-regressor neural network (QRNN) as an alternative representation of the chance constraints to serve as a proxy for the large-scale scenario-based formulations. However, one issue with using this formulation is the introduction of Type I errors and feasibility issues arising from the violation of the chance constraints due to potentially imprecise QRNN predictions, a challenge that few existing studies address. This work proposes a method to mitigate these issues generated by QRNNs, compares the predictive performance and computational efficiency of scenario-based formulations with those of QRNN-based formulations to demonstrate the improvements and tradeoffs, and explores the effects of differing architectural choices on the computational and predictive performance of the resulting formulation.

4 - Optimization over Trained (and Sparse) Neural Networks: A Surrogate within a Surrogate

Thiago Serra, University of Iowa, Iowa City, IA, United States

We can approximate a constraint or an objective function that is uncertain or nonlinear with a neural network that we embed in the optimization model. This approach, which is known as constraint learning, faces the challenge that optimization models with neural network surrogates are harder to solve. Such difficulties have motivated studies on model reformulation, specialized optimization algorithms, and - to a lesser extent - pruning of the embedded networks. In this work, we double down on the use of surrogates by applying network pruning to produce a surrogate of the neural network itself. In the context of using a Mixed-Integer Linear Programming (MILP) solver to verify neural networks, we obtained faster adversarial perturbations for dense neural networks by using sparse surrogates, especially - and surprisingly - if not taking the time to finetune the sparse network to make up for the loss in accuracy. In other words, we show that a pruned network with bad classification performance can still be a good - and more efficient - surrogate.

SatD10

Caucus Room

Parallel Computing for Optimization: Methods and Applications

Invited Session

Computational Optimization

Chair: Bernard Knueven, National Renewable Energy Laboratory, Golden, CO, United States

1 - New Progressive Hedging Acceleration Heuristics in mpi-sppy

Bernard Knueven, National Renewable Energy Laboratory, Golden, CO, United States

The mpi-sppy meta-solver library facilitates scalable solutions to large-scale stochastic programming problems through the asynchronous execution of primary algorithms, such as Progressive Hedging, and assisting algorithms that compute objective value bounds or accelerate convergence of the primary algorithm. In this talk, we present recent enhancements to mpi-sppy, including newly developed techniques aimed at improving both primal and dual convergence of Progressive Hedging. We evaluate the effectiveness of these heuristic strategies on a challenging, large-scale power grid capacity expansion model.

2 - Solving Non-Convex Stochastic MINLPs via Progressive Hedging with Domain Reduction and Dynamic Gradient Information

Ali Asger, Carnegie Mellon University, Pittsburgh, PA, United States, Carl Laird

Large-scale non-convex mixed integer nonlinear programs (MINLPs) with a block-angular structure arise naturally in multi-scenario optimization contexts, including product and process family design, infrastructure planning, and other stochastic programming applications. While such problems are intractable at scale using off-the-shelf optimization solvers, their structure can be exploited through decomposition strategies to leverage parallel computation. These strategies transform the large-scale MINLPs into smaller subproblems while introducing non-anticipativity (coupling) constraints to ensure consistency of first stage (shared) decisions across all subproblems. To solve the decomposed problem, we employ progressive hedging (PH), implemented through the open-source package *mpi-sppy*, which enables efficient parallel solution of subproblems across distributed computing architectures. Although progressive hedging provides convergence guarantees for convex problems, its application to non-convex problems remains heuristic. To address this challenge, we introduce a tailored algorithmic design that incorporates dynamic gradient information at each PH iteration to adaptively adjust penalty parameters and accelerate convergence. Furthermore, we integrate domain reduction techniques into the algorithmic framework to tighten bounds on first-stage variables, substantially improving convergence in practice. Through computational experiments on process family design problems, we demonstrate the improvements in using our tailored algorithmic design compared to off-the-shelf solvers.

3 - A Partitioning Algorithm for Constructing Sets of Diverse Alternative Optima

Ignacio Andres Aravena Solis, Lawrence Livermore National Laboratory, Livermore, CA, United States, Jisun Lee, Jean-Paul Watson, Alper Atamturk

A highlighted feature of quantum computing algorithms for mixed integer programs is that they are able to produce multiple solutions for a given problem, providing users with alternatives they can evaluate against downstream processes. However, the set of solutions produced by those algorithms has not been shown to be unbiased or to cover significant portions of a set of epsilon-suboptimal acceptable solutions. Performing that assessment requires techniques that can efficiently generate sets of diverse alternative optima. In this talk, we present a novel enumeration approach for such purpose, aiming at achieving the maximum diversity among generated alternative optima within a set wall clock time limit. Instead of selecting and removing points from the acceptable solution set one-by-one, our approach removes them in pairs, selecting the most distant solutions at each iteration. This two-solution removal step is followed by a careful (non-intersecting) partition of the remainder of solution space. The two-solution removal is then recursively carried out for each remaining component of the partition, a process that we organize using a queue, and later parallelize using high-performance computing. We discuss (integer programming) strength properties of our approach in comparison with alternatives and present details of its implementation. Finally, we present a numerical comparison of our approach against approaches from the literature using standard MIP test problems and industrial-size unit commitment instances. These results show how the proposed approach is able to produce more diverse sets of solutions at a similar or smaller computational cost than prior methods.

4 - Advancing the State-of-the-Art in Stochastic Network Interdiction

Jean-Paul Watson, Lawrence Livermore National Laboratory, Livermore, CA, United States, Andrew Mastin

Stochastic network interdiction - both of the defender-attack-defender (DAD) and attacker-defender (AD) varieties - has not been analyzed in the context of recent significant advances in parallel decomposition methods for large-scale stochastic mixed-integer programs, e.g., as implemented in the open-source mpi-sppy library. We consider stochastic formulations of both DAD and AD in the context of the recently introduced Best Response Intersection (BRI) network interdiction algorithm. We analyze scalability leveraging parallel compute resources, and examine the structure of solutions to stochastic power grid interdiction problems.

SatD11

Charter Room

Compositional Optimization on Networks

Invited Session

Network Optimization

Chair: Matthew Klawonn

Co-Chair: Matthew Hale, Georgia Tech, Atlanta, GA, United States

1 - Heterogeneous Multi-Agent Multi-Target Tracking

Joana Bou, University of Florida, Gainesville, FL, United States, Tyler Hanks, Cristian F. Nino, Austin Copeland, Warren Dixon, James Fairbanks

Multi-agent target tracking in the presence of nonlinear dynamics and agent heterogeneity, where state-space dimensions may differ, is a challenging problem that traditional graph Laplacian methods cannot easily address. This work leverages the framework of cellular sheaves, a mathematical generalization of graph theory, to natively model such heterogeneous systems. While existing coordination sheaf frameworks focus on cooperative problems like consensus, this work extends them to the non-cooperative target-tracking problem. The tracking of multiple, unknown targets is formulated as a harmonic extension problem on a cellular sheaf, accommodating nonlinear dynamics and external disturbances for all agents. A decentralized control law is developed using the sheaf Laplacian, and a corresponding Lyapunov-based stability analysis is provided to guarantee tracking error convergence, with results validated by simulation.

2 - Scalable Co-Design via Linear Design Problems: Theory and Decomposition Algorithm

Yujun Huang, Massachusetts Institute of Technology, Cambridge, MA, United States, Yubo Cai, Gioele Zardini

The design of complex engineering systems necessitates a rigorous methodology to manage intricate trade-offs between component capabilities and resource constraints. While the monotone theory of co-design provides a compositional framework for synthesizing system level behavior, the specialization of these problems to specific domains remains underexplored. In this work, we present a formal mathematical framework for Linear Design Problems (LDPs), a class of co-design problems where feasible sets are represented by polyhedra within vector spaces ordered by pointed convex cones. We establish that the class of LDPs is closed under essential compositional operations, including series, parallel, and feedback connections, allowing complex systems to be modeled as wiring diagrams of modular components. Furthermore, we investigate the computational complexity and scalability of solving composed LDPs with two methods: The first adopts a compositional strategy, where solutions of individual Design Problems (DPs) are computed as Multiple objective Linear Programmings (MOLPs) and then composed to form the system-level solution. The second approach formulates the entire co-design diagram as a single MOLP to be solved globally. We provide a comparative analysis of these algorithms, evaluating their computational efficiency under various problem conditions. Finally, the framework is applied to the co-design of a task-driven Unmanned Aerial Vehicle (UAV), showing LDPss can serve as efficiently computable warm-starters for more sophisticated DP solvers.

3 - Categories and Sheaves for Optimization: From Compositional to Distributed

Hans Riess, Georgia Tech, Atlanta, GA, United States

Optimization is a foundational pillar of modern data science and control theory. In this talk, we argue that enriched category theory and sheaf theory provide a powerful lens to structure, analyze, and solve complex optimization tasks. We first show that compositional, multi-stage optimization problems (like MPC) find a natural home in the enriched category of convex bifunctions, as introduced by Rockafellar. We then extend this compositional framework beyond the convex setting to "tame" optimization, where non-convex objectives and tame feasible sets can be formally composed. Having addressed sequential composition, we shift our focus to distributed coordination over networks. We introduce “coordination sheaves" to model multi-agent systems, where tasks are assigned to nodes and consistency constraints are encoded on edges. The sheaf Laplacian, a generalization of the graph Laplacian, emerges as a natural distributed operator for solving these problems. If time allows, we recall recent results on asynchronous sheaf diffusion, demonstrating when these sheaf-based algorithms converge even in challenging asynchronous settings.

4 - Compositional Uncertainty and Adaptive Learning in Systems Co-Design

Meshal Alharbi, Massachusetts Institute of Technology, Cambridge, MA, United States, Yujun Huang, Gioele Zardini

Designing complex systems composed of many interacting components requires balancing multiple, often conflicting objectives while satisfying system-level constraints. Decisions about individual components cannot be made in isolation, as local changes can significantly impact overall system performance. Co-design has emerged as a powerful approach to solve these challenges by jointly optimizing interconnected choices, while preserving modularity and compositional structure. However, existing co-design methods typically handle uncertainty using only conservative worst-case bounds, which obscures important questions about risk tolerance. In the first part of this talk, we present a compositional framework for uncertainty in co-design that accommodates richer probabilistic descriptions and models. This framework enables decision-makers to compare designs under varying confidence levels and to understand how uncertainty in individual components influences overall system performance. In the second part, we address co-design problems in which some components can only be evaluated through costly numerical simulations. Current approaches rely on fixed grids over the design space, which scale poorly and ignore the problem’s structure. Instead, we introduce adaptive sampling methods that exploit the compositional structure of co-design to propagate beliefs about the utility of untested options across the dependency graph. Together, these contributions enable richer reasoning about co-design problems while maintaining computational tractability.

March 22, 2026, 8:00 AM - 9:30 AM

SunA01

Grand Ballroom

Adaptive Methods for Nonlinear and Stochastic Optimization II

Invited Session

Optimization in Data Science

Chair: Tianjiao Li, Massachusetts Institute of Technology, Cambridge, MA, United States

1 - The Sample Complexity of Parameter-free Optimization

Oliver Hinder, University of Pittsburgh, Pittsburgh, PA, United States, Jared Lawrence, Ari Kalinsky, Hannah Bradfield, Yair Carmon

We study the sample complexity of stochastic convex optimization when problem parameters, e.g., the distance to optimality, are unknown. We pursue two strategies. First, we develop a reliable model selection method that avoids overfitting the validation set. This method allows us to generically tune the learning rate of stochastic optimization methods
to match the optimal known-parameter sample complexity up to $\log\log$ factors.
Second, we develop a regularization-based method that is specialized to the case that only the distance to optimality is unknown. This method provides perfect adaptability to unknown distance to optimality, demonstrating a separation between the sample and computational complexity of parameter-free stochastic convex optimization.
Combining these two methods allows us to simultaneously adapt to multiple problem structures.

Experiments performing few-shot learning on CIFAR-10 by fine-tuning CLIP models and prompt engineering Gemini to count shapes indicate that our reliable model selection method can help mitigate overfitting to small validation sets.

2 - An Adaptive and Parameter-Free Nesterov's Accelerated Gradient Method for Convex Optimization

Jaewook J. Suh, Rice Universiry, Houston, TX, United States, Shiqian Ma

We propose AdaNAG, an adaptive accelerated gradient method based on Nesterov's accelerated gradient method. AdaNAG is line-search-free, parameter-free, and achieves the accelerated convergence rates $f_k - f_\star = O(1/k^2)$ and $\min_{i\in\{1, ..., k\}} \|\nabla f(x_i)\|^2 = O(1/k^3)$ for L-smooth convex function f. We provide a Lyapunov analysis for the convergence proof of AdaNAG, which additionally enables us to propose a novel adaptive gradient descent (GD) method, AdaGD. AdaGD achieves the non-ergodic convergence rate $f_k - f_\star = O(1/k)$, like the original GD. The analysis of AdaGD also motivated us to propose a generalized AdaNAG that includes practically useful variants of AdaNAG. Numerical results demonstrate that our methods outperform some other recent adaptive methods for representative applications.

3 - Parameter-Free Gradient Sliding Methods

Yan Wu, Clemson University, Clemson, SC, United States, Yuyuan Ouyang, Jimmy Zhang, Qi Luo

The gradient sliding methods were developed for problems that have separable structures: sum of convex smooth and nonsmooth functions, convex smooth objective and constraint functions, etc. The core idea of gradient sliding-type methods is that the complexity of first-order algorithms should be adaptable to different problem structures. While several gradient sliding methods have been proposed in the literature, few of them are parameter-free; most sliding algorithms require the knowledge of Lipschitz constants. In this talk, we show that gradient sliding methods can be designed without any knowledge of Lipschitz constants. Our results covers optimization problems with structure of the sum of either two convex smooth functions, or a convex smooth

function and a convex nonsmooth function. In addition, we also propose novel gradient sliding algorithms for convex functional constrained minimization and monotone variational inequalities with convex functional constraints.

4 - Instance-Optimal Stochastic Convex Optimization: Can We Improve upon Sample-Average and Robust Stochastic Approximation?

Liwei Jiang, Purdue University, West Lafayette, IN, United States, Ashwin Pananjady

We study the unconstrained optimization of a smooth and strongly convex population loss function under a stochastic oracle that introduces both additive and multiplicative noise.
We begin by showing that standard approaches such as sample average approximation and robust (or averaged) stochastic approximation can lead to suboptimal --- and in some cases arbitrarily poor --- performance with realistic finite sample sizes. In contrast, we demonstrate that a carefully designed variance reduction strategy can significantly outperform these approaches while using the same sample size. Our upper bounds are complemented by finite-sample, information-theoretic local minimax lower bounds, which highlight fundamental, instance-dependent factors that govern the performance of any estimator. Overall, our characterization shows that our variance-reduced algorithm is instance-optimal, achieving the best possible sample complexity up to logarithmic factors. Our findings are supported by several careful numerical studies.

SunA02

Georgian Room

Interactions of Optimization with Statistics, Games, and Diffusion I

Invited Session

Optimization in Data Science

Chair: Mateo Diaz, Johns Hopkins University, Baltimore, MD, United States

Co-Chair: Pedro Izquierdo Lehmann, Johns Hopkins University - Applied Math & Statistics, Baltimore, MD, United States

1 - Communication in Multi-player Games Matters: A Federated Learning Framework for Equilibrium Computation

Nicolas Loizou, Johns Hopkins University, Baltimore, MD, United States, TaeHo Yoon

Multi-player games play a prominent role in statistics, economics, finance, and engineering. Such problems have recently received significant attention, especially in the machine learning community, where adversarial training of neural networks, multi-agent reinforcement learning, and distributionally robust learning are formulated as structured multi-player games. Popular algorithms for solving games, such as the gradient method and the extragradient algorithm, are analyzed and implemented with per-iteration exchanges among players, without explicitly accounting for communication costs. However, as data sizes and the number of players in games grow, communication between players becomes the central bottleneck. To alleviate this issue, in this talk, we introduce the multi-player federated learning (MpFL) framework, a novel setting that models the clients in the FL environment as players in a game-theoretic context, aiming to reach an equilibrium. We focus on Per-Player Local Gradient Methods for MpFL and provide convergence guarantees for the new methods, showing that they can reach equilibrium with less communication. Experiments verifying our theoretical findings will also be presented.

2 - Noisy M-estimation for User-level Differential Privacy

Marco Avella Medina, Columbia University, New York, NY, United States

We propose a general optimization-based framework for computing differentially private M-estimators. We first show that robust statistics can be used in conjunction with noisy gradient descent to obtain optimal private estimators with global linear convergence. We establish local and global convergence guarantees, under local strong convexity, showing that our private estimators converge with high probability to a near-optimal neighborhood of the nonprivate M-estimators. We then extend this optimization framework to the setting where each user possesses multiple data points. We develop user-level differentially private algorithms that seek to protect the entire set of data points belonging to a user. Our main algorithm is a noisy gradient descent algorithm, combined with a user-level DP mean estimation procedure to privately compute the average gradient across users at each step. We will highlight the challenges associated with guaranteeing user-level DP and present finite sample global linear convergence guarantees for the iterates of our algorithm.

3 - Consensus-Based Bi-Level Optimization and its Application to Robust Federated Learning

Sixu Li, University of Wisconsin-Madison, Madison, WI, United States, Nicolas Garcia Trillos, Aditya Kumar Akash, Konstantin Riedl, Yuhua Zhu

Bi-level optimization problems, where one seeks the global minimizer of an upper-level objective over the globally optimal solution set of a lower-level objective, arise in various applications across science, engineering, and machine learning.

In this talk, I will introduce Consensus-Based Bi-level Optimization (CB^{2}O), a multi-particle, derivative-free method for solving bi-level optimization problems with potentially nonconvex objectives. CB^{2}O leverages a quantile-based particle selection on the lower-level objective and a Laplace principle-type approximation for the upper-level objective, ensuring intrinsic preservation of the bi-level structure over the optimization process. In the mean-field regime, we establish exponential convergence of CB^{2}O dynamics to the unique solution under suitable hyperparameter choices. I will then present CB^{2}O's application to robust federated learning against adversarial attacks, where the training is formulated as a bi-level optimization problem. I will provide a global convergence analysis under adversarial settings, demonstrating its provable robustness against a variety of attacks. Empirically, I will show CB^{2}O's effectiveness in defending against label-flipping attacks in decentralized federated learning. This talk is based on joint work with Nicolás García Trillos, Aditya Kumar Akash, Konstantin Riedl, and Yuhua Zhu.

4 -

Anytime-Feasible First-Order Methods for Constrained Nonconvex Optimization

Mahyar Fazlyab, Johns Hopkins University, Baltimore, MD, United States

Constrained nonconvex optimization is central to robotics, optimal control, and machine learning, yet many first-order methods suffer from feasibility violations and brittle penalty or stepsize tuning. In this talk, I present Safe Sequential Quadratically Constrained Quadratic Programming (SS-QCQP), a first-order method for smooth constrained nonconvex optimization that guarantees feasibility at every iteration while ensuring monotonic objective descent. SS-QCQP is derived from a control-theoretic continuous-time dynamical system whose vector field is obtained by solving a convex QCQP that enforces forward invariance of the feasible set and descent of the objective. Under standard constraint qualification conditions, the resulting dynamics achieve an (O(1/t)) convergence rate to first-order stationary points. I then introduce a safeguarded Euler discretization with adaptive step-size selection that preserves these guarantees in discrete time, yielding an (O(1/k)) rate. To improve scalability, I present an active-set variant (SS-QCQP-AS) that enforces only near-active constraints. Experiments on a multi-agent nonlinear optimal control problem demonstrate strong feasibility, predicted convergence behavior, and solution quality comparable to SQP and IPOPT.

SunA03

Plaza I

Stochastic Primal–Dual Methods with Applications to Decentralized Learning and Optimization

Invited Session

Nonlinear Optimization

Chair: Necdet Serhat Aybat, Penn State University, University Park, PA, United States

Co-Chair: Mert Gurbuzbalaban, Rutgers University, Piscataway, NJ, United States

1 - A Stochastic GDA Method with Backtracking for Solving Nonconvex-concave Minimax Problems

Necdet Serhat Aybat, Penn State University, University Park, PA, United States, Qiushui Xu, Xuan Zhang, Mert Gurbuzbalaban

We propose a stochastic GDA method with backtracking (SGDA-B) to solve nonconvex-(strongly) concave minimax problems. SGDA-B is agnostic to the Lipschitz constant L, concavity modulus $\mu \geq 0$, and the variance bound of the unbiased stochastic gradient estimator. Within $O(1/\varepsilon^4 \log(1/p)$ and $O(1/\varepsilon^7 \log(1/p))$ stochastic gradient calls, SGDA-B can compute an ε-stationary point in terms of the gradient-map norm and the gradient norm of the Moreau envelope evaluated at the random output point with probability 1-p when $\mu>0$ and $\mu=0$, respectively.

2 - High-probability Complexity Bounds for Stochastic Nonconvex Minimax Optimization

Mert Gurbuzbalaban, Rutgers University, Piscataway, NJ, United States, Yassine Laguel, Yasa Syed, Serhat Aybat

Stochastic smooth nonconvex minimax problems are prevalent in machine learning, e.g., GAN training, fair classification, and distributionally robust learning. Stochastic gradient descent ascent (GDA)-type methods are popular in practice due to their simplicity and single-loop nature. However, there is a significant gap between the theory and practice regarding high-probability complexity guarantees for these methods on stochastic nonconvex minimax problems. Existing high-probability bounds for GDA-type single-loop methods only apply to convex/concave minimax problems and to particular non-monotone variational inequality problems under some restrictive assumptions. In this work, we address this gap by providing the first high-probability complexity guarantees for nonconvex/PL minimax problems corresponding to a smooth function that satisfies the PL-condition in the dual variable. Specifically, we show that when the stochastic gradients are light-tailed, the smoothed alternating GDA method can compute an ε-stationary point within $O(\ell\varkappa^2 \delta^2/ \varepsilon^4 + \varkappa(\ell+\delta^2\log(1/q)) / \varepsilon^2)$ stochastic gradient calls with probability at least $1-q$ for any $q \in (0,1)$, where μ is the PL constant, ℓ is the Lipschitz constant of the gradient, $\varkappa=\ell/\mu$ is the condition number, and δ^2 denotes a bound on the variance of stochastic gradients. We also present numerical results on a nonconvex/PL problem with synthetic data and on distributionally robust optimization problems with real data, illustrating our theoretical findings.

3 - A Parameter-free Primal-dual Algorithm for Decentralized Constrained Convex Optimization

Qiushui Xu, Penn State University, State College, PA, United States, Necdet Serhat Aybat, Mert Gurbuzbalaban

We propose a fully distributed accelerated primal-dual method with backtracking (D-APDB) for cooperative multi-agent constrained consensus optimization problems over an undirected network of agents, where only those agents connected by an edge can directly communicate. The objective is to minimize the sum of agent-specific composite convex functions over agent-specific private nonlinear convex constraints onto which projections are not cheap to compute. Unlike existing decentralized primal-dual methods that require knowledge of the Lipschitz constants, D-APDB automatically adapts to local smoothness by employing a distributed backtracking step-size search. Each agent relies only on first-order oracles associated with its own objective and constraint functions and on local communications with the neighboring agents, without any prior knowledge of Lipschitz constants. We establish O(1/K) convergence guarantees for sub-optimality, infeasibility and consensus violation, under standard assumptions on smoothness and on the connectivity of the communication graph. To our knowledge, when nodes have nonlinear constraints, D-APDB is the first distributed method with backtracking that achieves the optimal convergence rate for the class of composite convex optimization problems subject to functional convex constraints. We also provide numerical results for D-APDB on a distributed QCQP problem illustrating the potential performance gains that can be achieved by D-APDB.

4 - DIGing Stochastic Gradient Langevin Dynamics for Bayesian Decentralized Learning on Time-varying Networks

Mustafa Kutbay, Rutgers, The State University of New Jersey, Newark, NJ, United States, Waheed Bajwa, Mert Gurbuzbalaban, Lingjiong Zhu, Muhammad Zulqarnain

This paper introduces DIGing-SGLD, a novel *decentralized stochastic gradient Langevin dynamics* algorithm designed for scalable Bayesian learning over *time-varying communication networks*. Existing decentralized SGLD methods are limited to static network topologies or suffer from steady-state sampling bias caused by network effects, even when using full data batches. DIGing-SGLD overcomes these limitations by integrating distributed inexact gradient tracking—originally developed for decentralized optimization—with Langevin-based sampling. This integration enables unbiased and communication-efficient Bayesian inference across dynamically changing network structures without requiring a central coordinator. To our knowledge, our analysis establishes the first finite-time non-asymptotic Wasserstein convergence guarantees for decentralized Langevin sampling over time-varying networks with explicit constants. Under standard strong convexity and smoothness assumptions, the proposed method achieves geometric convergence to an O(\sqrt{\eta}) neighborhood of the target distribution where \eta is the stepsize, matching the best-known rates for centralized and static-network SGLD algorithms. Numerical experiments on Bayesian linear and logistic regression tasks validate the theoretical results and showcase the performance of our method under dynamic network conditions.

SunA04

Plaza II

Information Value in Optimization

Invited Session

Optimization under Uncertainty

Chair: Amine Bennouna, Kellogg, Northwestern University, Evanston, IL, United States

Co-Chair: Aras Selvi, Princeton University, Princeton, NJ, United States

1 - Data Informativeness in Linear Optimization Under Uncertainty

Amine Bennouna, Kellogg, Northwestern University, Evanston, IL, United States, Omar Bennouna, Saurabh Amin, Asuman Ozdaglar

We study the fundamental question of how informative a dataset is for solving an optimization problem under uncertainty. In our setting, the dataset provides partial information
about an unknown parameter of the problem. Focusing on linear programs,
we characterize when a dataset is sufficient to recover an optimal decision, given an
uncertainty set on the cost vector. Our main contribution is a sharp geometric characterization that identifies the directions of the cost vector that matter for optimality,
relative to the task constraints and uncertainty set.

2 - Controlling Inventory and Information in Supply Chains

Rene Caldentey, The University of Chicago, Chicago, IL, United States

Successful supply chain operations depend on the effective coordination of inventory management strategies and the information required for their implementation. This presentation explores the integration of time-series forecasting with inventory management to assess the value of information sharing in a two-tier supply

chain. We analyze how different inventory policies influence supply chain costs under two information-sharing scenarios: (a) full information sharing and (b) no information sharing. We formulate an infinite-dimensional optimization problem where the decision variables are the MA coefficients characterizing a stationary ordering policy. By leveraging the canonical Smirnov-Beurling inner-outer factorization, we derive a fundamental mathematical identity that quantifies the value of information sharing, by viewing an inventory policy as an element of the Hardy space H2. Additionally, we show that an optimal ordering policy inherently incorporates the efficient delay of inventory replenishment in response to market demand. This delay is necessary to gradually release demand information, making the orders more predictable for upstream suppliers. We study how this delay is encoded in an inventory replenishment policy and how to optimally control it to minimize supply chain inventory costs.

3 - Estimate-Then-Optimize Versus Integrated-Estimation-Optimization Versus Sample Average Approximation: A Stochastic Dominance Perspective

Adam Elmachtoub, Columbia University, NYC, NY, United States, Henry Lam, Haofeng Zhang, Yunfan Zhao

In data-driven stochastic optimization, model parameters of the underlying distribution need to be estimated from data in addition to the optimization task. Recent literature considers integrating the estimation and optimization processes by selecting model parameters that lead to the best empirical objective performance. This integrated approach, which we call integrated-estimation-optimization (IEO), can be readily shown to outperform simple estimate-then-optimize (ETO) when the model is misspecified. In this paper, we show that a reverse behavior appears when the model class is well-specified and there is sufficient data. Specifically, for a general class of nonlinear stochastic optimization problems, we show that simple ETO outperforms IEO asymptotically when the model class covers the ground truth, in the strong sense of stochastic dominance of the regret. Namely, the entire distribution of the regret, not only its mean or other moments, is always better for ETO compared to IEO. Our results also apply to constrained, contextual optimization problems where the decision depends on observed features. Whenever applicable, we also demonstrate how standard sample average approximation (SAA) performs the worst when the model class is well-specified in terms of regret, and best when it is misspecified. Finally, we provide experimental results to support our theoretical comparisons and illustrate when our insights hold in finite-sample regimes and under various degrees of misspecification.

4 - Distributional Robustness with Domain Knowledge

Aras Selvi, Princeton University, Princeton, NJ, United States

We demonstrate that standard data-driven distributionally robust optimization methods can produce overly conservative decisions when domain knowledge rules out certain worst-case data generating distributions. To address this limitation, we develop a framework that systematically incorporates domain-specific information into a range of data-driven robust decision making models. Our experiments demonstrate that we can efficiently reduce the conservatism of such techniques while maintaining the same level of robustness guarantees.

SunA05

Plaza III

Dynamic Decisions under Uncertainty: Formulations, Algorithms, and Analysis I

Invited Session

Optimization under Uncertainty

Chair: Johannes Milz, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Alexander Shapiro, ISyE Georgia Tech, Atlanta, GA, United States

Co-Chair: Enlu Zhou, ISyE Georgia Tech, Atlanta, GA, United States

1 - Computational Hardness of Static Distributionally Robust Markov Decision Processes

Yan Li, Texas A&M University, College Station, TX, United States

We present some hardness results on finding the optimal policy for the static formulation of distributionally robust Markov decision processes. We construct problem instances such that when the considered policy class is Markovian and non-randomized, finding the optimal policy is NP-hard, and when the considered policy class is Markovian and randomized, the robust value function possesses sub-optimal strict local minima. The considered hard instances involve an ambiguity set with only two transition kernels.

2 - Sequential Optimization of Dynamically Augmented CVaR

Eugene Feinberg, Stony Brook University, Stony Brook, NY, United States

This talk discusses optimization of Conditional Value-at-Risk (CVaR) for Markov Decision Processes (MDPs) with finite state and action sets. We introduce and investigate the Dynamically augmented CVaR (DCVaR) risk objective function. The analysis of DCVaR is based on studying a specially defined Robust MDP (RMDP), in which the state space is augmented with the tail risk level. This RMDP, which we call the Dynamically augmented RMDP (DRMDP), was introduced to the literature for calculations of optimal CVaR values by value iteration more than ten years ago, but, as was understood later, these value iterations compute lower bounds of minimal static CVaRs. DCVaR is defined as a time consistent version of the static CVaR, and it is a lower bound of the static CVaR. It also can be considered as a dynamic version of the nested CVaR. This talks describes an algorithm for constructing a policy optimizing DCVaR of total discounted costs. Correctness of this algorithm is proved by studying a special mass transfer problem.

3 - Higher Order Approximations to Mean Field Stochastic Control Models

Sirui Lin, Stanford University, Stanford, CA, United States, Jose Blanchet, Peter Glynn

We consider a wide range of mean-field stochastic control problems which encompass a class of transport control models and restless bandits as particular cases. We develop the first analysis that enables O(1/N) approximations under Lipschitz dynamics, improving prior work requiring smooth conditions which are often not satisfied in common linearly constrained dynamics. The finite sample results that we obtain, which are supplied with matching lower bounds, rely on optimal transport approximations of high-dimensional truncated Gaussian distributions which are of independent interest.

4 - Dynamic Programming Decompositions of Static Risk Measures in Markov Decision Processes

Marek Petrik, University of New Hampshire, Durham, NH, United States, Gersi Doko, Erick Delage, Esther Derman, Mohammad Ghavamzadeh, Jia Lin Hau

Optimizing static risk-averse objectives in Markov decision processes is difficult because they do not admit standard dynamic programming equations that underlie common algorithms. Dynamic programming decompositions that augment the state space with discrete risk levels have recently gained popularity in the RL community. We propose and study dynamic programming decompositions of standard static risk measures, including VaR, CVaR, and EVaR. We also propose algorithms for discretizing augmented state spaces and algorithms that provide optimality guarantees. We demonstrate how dynamic program decompositions enable popular reinforcement learning algorithms, such as Q-learning, and how they relate to distributional reinforcement learning.

SunA06

Director’s and Lounge

Advances in Solution Methods for MINLPs (Part 2)

Invited Session

Global Optimization

Chair: Dahye Han, Georgia Institute of Technology, Atlanta, GA, United States

1 - Range Value-at-Risk in Two-Stage Stochastic Programs

Nan Jiang, Hong Kong University of Science and Technology, Hong Kong, Hong Kong, Zhiping Li

We study two-stage stochastic programs incorporating Range Value-at-Risk (RVaR) with bi-linear recourse. Motivated by the equivalence between the distributionally robust Value-at-Risk (VaR) minimization problem and the Range Value-at-Risk (RVaR) formulation, we show that RVaR naturally arises as the analytical counterpart of the worst-case VaR under Wasserstein ambiguity. This connection offers a new perspective for modeling decision-making under uncertainty, while the integration of RVaR into two-stage stochastic programs remains largely unexplored, particularly in settings with non-convex recourse. To address this gap, we develop a novel ranking-based reformulation of the empirical RVaR, which reveals a difference-of-convex structure amenable to efficient solution. Building on this insight, we incorporate a Tikhonov-perturbed scheme to handle bi-linear recourse, enabling implementable active-set updates and tractable subproblems. We establish subsequential convergence guarantees for the proposed algorithms under mild assumptions, and show that in the bi-linear recourse, accumulation points of the iterates correspond to critical solutions of the unregularized problem. Extensive numerical experiments demonstrate that our methods consistently deliver high-quality solutions with significantly reduced computation times compared to benchmark approaches, while exhibiting stable performance and scalability.

2 - Convexification of a Class of Bilinearly Constrained Sets Sharing a Common Variable

Hyun-Ju Oh, Clemson University, Clemson, SC, United States, Margaret Wiecek, Boshi Yang

In this talk, we study the convexification of sets defined by box constraints and bilinear inequality constraints in which all bilinear terms share a common variable. We present a complete characterization of the extreme points and the facet-defining inequalities of the resulting convex hull. In addition, we develop a separation algorithm for identifying the facet-defining inequalities.

3 - Reinforcement Learning–Guided Variable Partitioning Policy for Global Optimization of QCQPs

Agustin Castellano, Johns Hopkins University, Baltimore, MD, United States, Harsha Nagarajan

Although Reinforcement Learning (RL) has recently shown promise as a substitute for hand-crafted heuristics in combinatorial optimization, prior efforts have focused mostly on MILPs. In this work, we introduce an end-to-end RL framework that learns dynamic variable partitioning strategies that tighten MIP relaxations of parametric non-convex QCQPs. While many sequences of partition choices can lead to improved MIP relaxations and eventual closure of the optimality gap, we aim to learn a near-optimal minimal sequence of such partitions via a data-driven partitioning policy. Our approach leverages a Graph Convolutional Neural Network to extract meaningful features of the QCQP and extends the well-known Soft Actor-Critic (SAC) architecture. Entropy regularization in SAC balances variable-domain exploration and partition efficiency through a sparsity-enforcing action parameterization. This enables the RL agent to output sparse partition sets that effectively close the optimality gap. The resulting policy can be trained efficiently, can be transferred across problems of different size and performs inference in under 10ms, showing potential for integration into modern optimization pipelines. Experiments on QCQP families of more than 100 quadratic terms show median relative gaps of 6% and 1% after one and two partitioning points per variable, respectively.

SunA07

Congress Room

Recent Advances in Mixed Integer Programming III

Invited Session

Discrete Optimization

Chair: Aleksandr Kazachkov, University of Florida, Gainesville, United States

1 - Optimizing for Fairness in Generalized Kidney Exchange: Theory and Computations

Claire Chang, Cornell University, Ithaca, NY, United States, Arin Khare, David Shmoys

The seminal work of Roth, Sonmez, & Unver shows that the Edmonds-Gallai structure theorem for non-bipartite matching can be leveraged to yield a randomized algorithm to match patient-donor pairs in kidney exchange with extraordinarily strong properties. This breakthrough led to randomized polynomial-time algorithms to find maximum cardinality matchings for which individual fairness, measured by the probability that nodes are matched, for example, has maximized Nash social welfare. The societal impact of these results can not be overstated. But the algorithms actually used go beyond cardinality matching, generalizing to weighted variants, as well as also allowing structures such as paths and 3-cycles to cover patient-donor pairs. We first show that strongly polynomial algorithms that guarantee the same fairness properties can be obtained in the weighted setting. The 3-cycle extension, even just for maximum cardinality coverage is NP-hard. We provide a general result that shows that any optimization subroutine for the coverage problem (for whichever structure is allowed) can be bootstrapped using a polynomial number of calls to yield a mechanism that has analogous fairness properties to those obtained for matching. We complement these theoretical results with computational results, both on well-studied synthetic data-sets and on samples drawn from real data, that demonstrate the striking advantages of adding these fairness elements to more general kidney-exchange mechanisms.

2 - k-Coverage: Improved Heuristics for the k-Center and k-Max-Min-Dispersion Problems

Jonathan Bodine, UC Berkeley, Berkeley, CA, United States, Dorit Hochbaum

Facility location problems are a core class of discrete optimization problems with applications well beyond physical `facilities'. In this paper, we focus on the k-Center (minimizing the maximum distance) and k-Max-Min-Dispersion (maximizing the minimum separation) problems which have applications for machine learning tasks such as active learning. The current empirical literature relies on the greedy algorithm of Gonzalez that simultaneously approximates both the k-Center and k-Max-Min-Dispersion problems with best possible approximation bounds of 2 and one-half respectively. In this paper, we propose two novel heuristics specialized for either customer fulfillment or facility dispersion, which enable us to achieve a reduction in the optimality gap of more than 50% for k-Center and over 80% for k-Max-Min-Dispersion on 200 synthetic problems with known optimal solutions, and improve the objectives for large instances where the true optimum is not known. Our solutions to the k-Center problem are less impacted by outliers, resulting in solutions of much better quality for the k-Median objective. This improvement in alternative facility objectives is a significant advantage of our class of heuristics; with only minor modifications, they can prioritize different facility objectives or even mixes of objectives.

3 - A Tour-Length Estimation and Heuristics for Heterogeneous Capacitated Vehicle Routing Problem

Hyungjoo Cha, Korea University, Seoul, Korea, Republic of, Taesu Cheong, Woonghee Huh

In this talk, we study a heterogeneous capacitated vehicle routing problem in which an infinite-capacity main vehicle and a finite-capacity sub vehicle cooperate to serve customers in a planar region. The main vehicle acts as a mobile depot, while the sub vehicle repeatedly departs, collects goods, and returns to rendezvous under a one-way cargo-transfer rule. We first develop a non-asymptotic upper bound on the expected routing cost under a simple traversing-strip policy, avoiding the regression-based tour-length estimators commonly used in the literature. The bound is explicit and captures both vehicle capacity and relative operating costs. Building on this analysis, we design two constructive heuristics: a TSP-based scheme that operates on grid cells and a Hamiltonian-path scheme that reduces redundant depot returns. Computational experiments on large synthetic instances show that both heuristics produce feasible tours whose costs consistently lie below our analytical bound and an adaptation of Daganzo (1984a)'s classical approximation, with the Hamiltonian-based approach achieving the strongest performance and clear economies of scale.

4 - Speeding Up Mixed-Integer Solvers with Sparse Learning for Branching

Selin Bayramoglu, Georgia Institute of Technology, Atlanta, GA, United States, George Nemhauser, Nikolaos Sahinidis

Machine learning is being increasingly used to accelerate branch-and-bound algorithms for mixed integer optimization. Studies in the literature demonstrate success in approximating branching criteria based on strong branching, a local and expensive strategy which empirically leads to small search trees. The state-of-the-art learning methods for this problem are based on neural networks, with many parameters and a need for a large amount of data for training. In this work, we demonstrate the effectiveness of sparse models in approximating strong branching scores, showing that our models remain simple yet powerful even when trained on limited data. Moreover, our approach achieves significant solver speedups without requiring GPU hardware. We demonstrate that our methodology speeds up SCIP when applied to commonly studied mixed-integer linear problems, and a challenging mixed-integer nonlinear problem from the energy sector.

5 - When Solvers Get Moody: Quantifying and Predicting Performance Variability in Mixed Integer Programming

Mourchid Adegbindin, Cornell University, Ithaca, NY, United States, Brenda Dietrich, Jamol Pender, Andrea Lodi

Mixed-integer programming (MIP) solvers are widely used in research and practice, yet their performance can vary substantially even when solving the same instance under seemingly neutral perturbations. Such variability complicates benchmarking, planning, and the interpretation of computational studies. We introduce the Weighted Variability Score (WVS), a robust multi-component metric that aggregates dispersion in normalized runtime, final optimality gap, LP iterations, and branch-and-bound nodes across repeated runs of the same instance. On a large subset of MIPLIB 2017 instances, each solved in 100 randomized runs under a four-hour limit, WVS reveals a right-skewed distribution with a substantial upper tail: many instances are essentially stable, but some exhibit large run-to-run variability that is not explained by size or difficulty labels alone. Using a compact set of static descriptors and root-level features, a simple tree-based predictor with a gated tail uplift achieves a mean absolute error of 0.06 and Spearman rank correlation of 0.65 on a sealed test split. Feature-importance analysis of the final predictor highlights root degeneracy, matrix structure and coefficient ranges, and presolve mix as the main drivers of high WVS. By showing that variability is not only systematic but also partially predictable from inexpensive early-run features, our findings offer a practical tool for preemptive tuning and solver configuration and lay a foundation for fairer evaluation of new algorithms.

SunA08

Committee Room

Optimization and Learning for Better Healthcare Decisions

Invited Session

Emerging Applications of Optimization

Chair: Holly Wiberg, Carnegie Mellon University, Pittsburgh, PA, United States

Co-Chair: Hao Hao, Carnegie Mellon University, Pittsburgh, PA, United States

1 - ***Cancelled***Fairness-promoting Integer Programming Approaches for Medical Resident Rotation Scheduling

Karmel Shehadeh, University of Southern California, Los Angeles, CA, United States, Shutian Li, Frank E. Curtis, Beth Hochman

We introduce new integer programming (IP) approaches for the *r*esident-to-*r*otation *a*ssignment *p*roblem (RRAP). Given sets of residents, resident classes, and departments, a block structure for each class, staffing needs, rotation requirement, program rules, and resident vacation requests, the RRAP involves finding a feasible year-long rotation schedule that specifies resident assignments to rotations and vacation times. We first present an IP formulation for the RRAP, which can be easily implemented and efficiently solved using off-the-shelf solvers. However, it can lead to disparities in satisfying vacation requests among residents. To mitigate such disparities, we derive a fairness-promoting counterpart that finds an optimal rotation schedule, maximizing the number of satisfied vacation requests while minimizing a measure of disparity in satisfying these requests. Then, we design a computationally efficient Pareto Search Algorithm that finds the complete set of Pareto optimal solutions to the fairness-promoting IP quickly. Additionally, we present a user-friendly tool that implements the proposed models. Finally, we construct divers instances based on data from our collaborator and conduct extensive experiments to illustrate the potential practical benefits of our proposed approaches. Our results demonstrate the computational efficiency and implementability of our approaches and underscore their potential to enhance fairness rotation scheduling.

2 - Nonstationary Bandit Learning with Power Constraints under Habituation and Recovery Dynamics

Fengxu Li, University of Wisconsin–Madison, Madison, WI, United States, Yonatan Mintz

Multiarmed bandits (MABs) are a widely used framework for sequential decision-making under uncertainty, with significant applications in personalized healthcare. While traditional MAB methods typically assume stationary reward distributions, real-world settings are often nonstationary. For instance, in personalized interventions, treatment effectiveness may decline due to habituation and later recover, creating complex nonstationary behaviors that standard methods fail to capture.

In addition to adapting to these evolving reward structures, a critical challenge—especially in clinical trial contexts—is preserving statistical power for reliable inference. Existing bandit algorithms generally optimize cumulative rewards but often neglect adequate exploration which is necessary for detecting meaningful treatment effects.

We propose a novel Posterior Sampling algorithm explicitly designed to tackle this dual challenge. Our method adapts to nonstationary environments characterized by habituation and recovery dynamics, while simultaneously satisfying specified power constraints through controlled exploration.

We provide theoretical guarantees via sublinear regret bounds and validate the approach on two real-world datasets by simulation: a bipolar disorder clinical trial and a microrandomized trial promoting workplace standing. Results demonstrate the method's effectiveness in personalizing interventions while maintaining accurate statistical inference.

3 - Neural Index Policies for Restless Multi-Action Bandits with Heterogeneous Budgets

Himadri Pandey, Georgia Institute of Technology, Atlanta, GA, United States

Restless multi-armed bandits (RMABs) provide a scalable framework for sequential decision-making under uncertainty, but classical formulations assume binary actions and a single global budget. Real-world settings, such as healthcare, often involve multiple interventions with heterogeneous costs and constraints, where such assumptions break down. We introduce a Neural Index Policy (NIP) for multi-action RMABs with heterogeneous budget constraints. Our approach learns to assign budget-aware indices to arm--action pairs using a neural network, and converts them into feasible allocations via a differentiable knapsack layer formulated as an entropy-regularized optimal transport (OT) problem. The resulting model unifies index prediction and constrained optimization in a single end-to-end differentiable framework, enabling gradient-based training directly on decision quality. The network is optimized to align its induced occupancy measure with the theoretical upper bound from a linear programming relaxation, bridging asymptotic RMAB theory with practical learning. Empirically, NIP achieves near-optimal performance within 5% of the oracle occupancy-measure policy while strictly enforcing heterogeneous budgets and scaling to hundreds of arms. This work establishes a general, theoretically grounded, and scalable framework for learning index-based policies in complex resource-constrained environments

4 - Optimizing Invasive Species Management: A Mathematical Framework for Cost-Effective Control of Chinese Privet

Maheshi G Walawwe, University of Tennessee at Chattanooga, Chattanooga, TN, United States, Lakmali Weerasena, JIn Wang

Effective invasive species management requires strategies that balance ecological impact with limited resources. We present a discrete-time mathematical framework to model the spatiotemporal spread of Chinese privet (Ligustrum spp.), incorporating seed- and root-mediated dispersal across heterogeneous landscapes. The model captures nonlinear spatial interactions, dynamic invasion fronts, and varying habitat suitability. To inform practical management, we integrate the mathematical model with combinatorial optimization techniques, identifying adaptive, cost-effective removal strategies that minimize reinfestation risk while adhering to operational and budget constraints. Our approach accounts for treatment thresholds, spatial prioritization, and multi-period planning, enabling managers to assess trade-offs between short-term actions and long-term ecological outcomes. By coupling invasion dynamics with optimization-based decision-making, this framework provides a rigorous, data-driven tool for prioritizing interventions and designing scalable, ecologically effective invasive-species management plans.

5 - A Data-driven Optimization Approach to Designing Parsimonious Treatment Guidelines

Gian-Gabriel Garcia, University of Washington, Seattle, WA, United States, Sun Ju Lee

Evidence-based guidelines play an important role in informing how chronic diseases should be managed, as these recommendations are widely disseminated and widely implemented. However, they are often one-size-fits-most, failing to account for patient-to-patient differences. Personalized medicine has shown significant potential to improve health outcomes over clinical practice guidelines. However, the implementation of personalized medicine may be challenging, resulting in unwanted practice variation and suboptimal patient care. To find the right balance between personalized medicine and clinical guidelines, we propose the treatment guideline design problem. We develop a framework to design optimal treatment guidelines that are stratified across G groups for a population of patients wherein each person is modeled according to their own Markov Decision Process parameters. We characterize the structural properties of this problem and propose several exact and heuristic methods to solve the problem, including an exact mixed-integer linear program formulation, an exact policy-based branch-and-bound algorithm, and a heuristic that assigns the G-best policies from a set of candidate policies. Our simulation analysis shows that the G-best heuristic best scales with lager problem sizes. Applying this method to a case study on hypertension treatment planning demonstrates that with only a small number of subgroups, we can develop clinical practice guidelines which perform almost as well as personalized treatment policies.

SunA09

Cabinet Room

(Inverse) Mixed Integer Optimization

Invited Session

Discrete Optimization

Chair: Fatemeh Nosrat, University of Florida, Gainesville, FL, United States

1 - Inverse Mixed-Integer Programming: Polyhedral Approaches and Algorithms

Sam Garvin, Rice University , Houston, TX, United States, Fatemeh Nosrat, Andrew Schaefer

We examine the polyhedral structure of the inverse-feasible region (IFR) of a linear program, relating it to a certain linear subspace using orthogonal complements. When the forward-feasible region is full-dimensional, we use its binding facets to represent the extreme rays of the IFR. We also give a polyhedron with the same IFR as the forward-feasible region with fewer constraints, using the maximum weight matching problem for non-bipartite graphs to demonstrate the difference. Using this formulation, we give a cutting-plane algorithm for general inverse mixed-integer programs (MIPs). Under certain norms, we improve this algorithm by using Chvátal inequalities to generate a strictly decreasing upper bound on the optimal value of the inverse MIP which finitely converges to the true optimal value.

2 - Inverse of the Gomory Corner Relaxation of Integer Programs

Fatemeh Nosrat, University of Florida, Gainesville, FL, United States, George Lyu, Andrew Schaefer

We explore the inverse of integer programs (IPs) by studying the inverse of their Gomory corner relaxations (GCRs). We show that solving a set of inverse GCR problems always yields an upper bound on the optimal value of the inverse IP that is at least as tight as the optimal value of the inverse of the linear program (LP) relaxation. We provide conditions under which solving a set of inverse GCR problems exactly solves the inverse IP. We propose an LP formulation for solving the inverse GCR under the $L$$_1$ and L_∞ norms by reformulating the inverse GCR as the inverse of a shortest path problem.

3 - On Integral Inverse Optimization Problems

Eva Ley, TU Braunschweig, Braunschweig, Germany, Maximilian Merkert

Inverse optimization problems are bilevel problems in which the leader modifies the follower's objective such that a prescribed feasible solution becomes an optimal solution of the follower. They capture hierarchical decision-making problems like parameter estimation tasks or situations where a planner wants to steer an agent's choice.
In integral inverse optimization problems, the leader's cost modifications must be integral. While adding this integrality constraint for the cost modifications changes some inverse optimization problems like the inverse knapsack problem, others like the inverse shortest path problem or the inverse min cut problem have an optimal integral solution without loss of generality if all original weights are integer.
We prove that the integral inverse mixed-integer problem lies in coNP, similar to the (continuous) inverse mixed-integer problem. Thus, for example, the integral inverse knapsack problem is coNP-complete.

4 - Finite Sample Inverse Optimization

Kimia Ghobadi, Johns Hopkins University, Baltimore, MD, United States, Fardin Ganjkhanloo, Farzin Ahmadi

This talk introduces a finite-sample framework for parameter estimation in inverse optimization. Using concentration inequalities under sub-Gaussian noise, it builds high-probability error bounds and confidence regions from repeated noisy observations, then strengthens estimation by enforcing KKT-consistency with the forward convex model. It also highlights practical implications for data collection and robustness. The same optimality-checking intuition can potentially point to extensions for MILP forward problems, using branch-and-bound to generate competing integer solutions and add cuts iteratively.

SunA10

Caucus Room

Methods for Multi-Stage and Sequential Decision Making

Invited Session

Computational Optimization

Chair: Yifan Hu, Rutgers, Piscataway, NJ, United States

Co-Chair: Ilyas Fatkhullin, ETH Zürich, Fliederstrasse 12, Zurich, 8006, Switzerland

1 - Efficient Online Mirror Descent Stochastic Approximation for Multi-Stage Stochastic Programming

Junhui Zhang, MIT, Cambridge, MA, United States, Patrick Jaillet

We study the unconstrained and the minimax saddle point variants of the convex multi-stage stochastic programming problem, where consecutive decisions are coupled through the objective functions, rather than through the constraints. Based on the analysis of deterministic mirror descent algorithms with inexact gradients, we introduce the idea of \textit{stochastic conditional gradient oracles}, a multi-stage analog of the stochastic gradient oracles used in (classical) stochastic programming. We show one approach to construct such oracles and prove the convergence of the (accelerated) mirror descent stochastic approximation, both in expectation and with high probability. To further reduce the oracle complexity, we view the problem from a \textit{semi-online} perspective, where the stage t decision variables are constructed s stages in advance, instead of before stage 1. We show that the delay in decision making allows an asynchronous implementation of the mirror descent stochastic approximation algorithms. By avoiding computing solutions for scenarios that are inconsistent with information available during stage t, the complexity is reduced from exponential to linear in the number of stages.

2 - Multi-Time-Scale Stochastic Approximation as a Tool for Multi-Agent Learning and Distributed Optimization

Thinh Doan, UT Austin, Austin, TX, United States

Multi-time-scale stochastic approximation (SA) is a powerful generalization of the classic SA method for finding roots (or fixed points) of coupled nonlinear operators. It has attracted considerable attention due to its broad applications in multi-agent learning, control, and optimization. In this framework, multiple iterates are updated simultaneously but with different step sizes, whose ratios loosely define their time-scale separation. Empirical studies and theoretical insights have shown that such heterogeneous step sizes can lead to improved performance compared to single-time-scale (or classical) SA schemes. However, despite these advantages, existing results indicate that multi-time-scale SA typically achieves only a suboptimal convergence rate, slower than the optimal rate attainable by its single-time-scale counterpart.

In this talk, I will present our recent work on characterizing the convergence complexity of multi-time-scale SA. We develop a novel variant of this method and establish new finite-sample guarantees that achieves the optimal (O(1/k)) convergence rate. Building upon these results, I will also discuss how these advances enable the design of efficient algorithms for key problems in multi-agent learning and distributed optimization over networks.

3 -

Learning and Optimization with Multi-Fidelity Data

Swati Gupta, Sloan School of Management, MIT, Cambridge, MA, United States, Diego Duvall, Tianjiao Li, Madeleine Pollack

In many real-world decision-making problems, data arrives from sources that differ widely in fidelity and acquisition cost. Effectively leveraging such heterogeneous observations is crucial for both rapid learning and optimization. On one hand, presence of low fidelity data lead to suboptimal decisions, and on the other hand, the cost of high-fidelity data may not justify its impact on convergence. In this talk, I will present two results illustrating how incorporating data quality yields provable and practical gains.

First, in a stochastic optimization setting, we consider two gradient oracles of different costs and noise levels. For e.g., the cost of querying foundational models depends on the number of tokens specified in their input. We develop an adaptive algorithm that decides which oracle to query based on the state of the optimization process. Under a fixed budget, we show how to obtain strictly faster convergence rates compared to strategies that rely on a single oracle throughout.

Next, in a multi-armed bandit setting, we consider the feedback from N arms, which is dependent on K mechanisms used to query the arms. For example, in a call center, sales depend not only on customer features (arms), but also on the quality of the agents (mechanisms) who make the calls. We first show that ignoring the mechanism quality can lead to linear regret. We next show how to optimally use low-fidelity data samples in this framework, and discuss the impact on a real-world case study at Blackthread Analytics.

4 - Can SGD Handle Heavy-Tailed Noise?

Ilyas Fatkhullin, ETH Zurich, Zurich, Switzerland

In this work, we consider constrained stochastic optimization problems under hidden convexity, i.e., those that admit a convex reformulation via a non-linear (but invertible) map. A number of non-convex problems, ranging from optimal control, revenue and inventory management, to convex reinforcement learning, all admit such a hidden convex structure. Unfortunately, in the majority of applications considered, the map is unavailable or implicit; therefore, directly solving the convex reformulation is not possible. On the other hand, the stochastic gradients with respect to the original variable are often easy to obtain. Motivated by these observations, we examine the basic projected stochastic (sub-)gradient methods for solving such problems under hidden convexity. We provide the first sample complexity guarantees for global convergence in smooth and non-smooth settings, and discuss the extensions to more general settings with non-convex functional constraints.

SunA11

Charter Room

Recent Advances in Distributed Optimization and Learning

Invited Session

Network Optimization

Chair: Marie Maros, Texas A&M University, College Station, TX, United States

1 - A New Decomposition Paradigm for Graph-Structured Nonlinear Programs via Message Passing

Kuangyu Ding, Purdue University, West Lafayette, IN, United States, Marie Maros, Gesualdo Scutari

We study finite-sum nonlinear programs with localized interactions among decision variables, encoded by a (hyper)graph. We propose a fully decentralized \emph{Gossip-Message Passing Algorithm} (GMA), a new optimization paradigm that brings min-sum message passing into optimization and couples it with a gossip (Jacobi-type) update. Starting from a fixed-point decomposition of the optimality conditions based on a graph partition, each iteration (i) performs a Jacobi/gossip update at every agent using current messages on designated intra-cluster edges and direct neighbor couplings elsewhere, then (ii) executes a single message update on those intra-cluster edges chosen to form trees per cluster. This topology- and structure-aware design avoids multi-hop replication and min–sum's non-convergence issue on loopy graphs. For strongly convex objectives we prove linear convergence, with rates explicit in graph topology (partition, intra-cluster trees, separators) and functional curvature (local and coupling parameters), guiding partitioning and scalability with the number of agents. For general convex and nonconvex cases we obtain sublinear guarantees. We also introduce structure-preserving surrogates to reduce computation and communication. For hypergraphs with heavy overlaps, a surrogate-based hyperedge splitting restores finite-time intra-cluster message passing and ensures convergence. Numerical results corroborate the theoretical results, showing that GMA matches centralized block-Jacobi while remaining decentralized and markedly outperforms gradient-type baselines in iteration and communication.

2 - Resilient Consensus with Stepsizes

Sarper Aydin, University of South Florida, Tampa, FL, United States, Stephanie Gil, Angelia Nedich

This work investigates the resilient consensus problem over undirected graphs, incorporating stepsizes to mitigate the influence of malicious agents. Using trust observations, legitimate agents identify malicious agents and exclude them from their communication. For this purpose, agents form their trusted neighbors by comparing their cumulative trust values at each time step. Compared to prior works, we introduce stepsizes into the consensus process that progressively slow the consensus updates. This is achieved by using stepsizes to combine the standard consensus update with the agent's current value, thus attenuating the malicious effect. In the absence of finite-time detection guarantees, we show that legitimate agents reach agreement in expectation. This result holds provided the stepsizes are non-summable and the limit probability of legitimate agents correctly identifying their neighbors' types converges to one. We further derive bounds on the probability of deviation from the nominal consensus process, which is the ideal case featuring no malicious agents and no stepsizes. This bound is a function of the sum of products of stepsizes and misclassification probabilities. Numerical experiments are presented to verify the theoretical results.

3 - Decentralized High-Dimensional Inference over Networks: A Unified Perspective

Marie Maros, Texas A&M University, College Station, TX, United States

We consider the problem of solving high-dimensional statistical inference problems over a network of agents (with no coordinating agent) who have exclusive access to a fraction of the total available samples. In the high-dimensional setting, the problem dimension is much larger than the total number of available samples, making the problem ill-conditioned. Despite this, we empirically observe that obtaining a statistically meaning solution is possible with many existing decentralized schemes, given that underlying parameter to estimate lies in a low dimensional subspace. Our observations challenge the existing theories in two key ways: (i) most decentralized schemes do not break down as the problem dimensionality increases, and (ii) decentralized schemes that are expected to behave like one another behave very differently in high dimensions. To understand the behavior of decentralized optimization methods in high-dimensional inference we introduce a unified framework and analysis, allowing us to develop an understanding of the mechanisms enabling dimension independent performance of decentralized schemes.

4 - Curvature-aided Client Selection in Distributed Optimization

Reza Vafaee, Boston College, Boston, MA, United States, Usman Khan

In large-scale distributed optimization, a central coordinator often interacts with a limited subset of agents or clients at each communication round. The challenge is to select these participants so that the aggregated update closely approximates the full-network descent direction despite heterogeneous and anisotropic local objectives. Existing selection methods rely solely on first-order gradients~\cite{balakrishnan2022diverse}, ignoring the curvature information that governs how gradients translate to parameter progress. We introduce a curvature-aware selection framework, which measures client similarity in a preconditioned gradient space informed by local Hessian or Fisher information matrices. The resulting optimization problem seeks the subset whose preconditioned updates provide maximal geometric coverage, bridging gradient diversity and information geometry. This second-order perspective captures client-specific curvature mismatch, improving both convergence stability and communication efficiency in heterogeneous networks. Analytical results establish approximate submodularity of the proposed criterion and constant-factor performance of a greedy policy. Preliminary numerical studies on synthetic and real datasets indicate consistent improvements over first-order diversity baselines. The framework unifies distributed optimization and experimental design principles, offering a new lens for curvature-aware coordination in decentralized learning.

March 22, 2026, 10:00 AM - 11:30 AM

SunB01

Grand Ballroom

Some Recent Advances in Optimization and Data Science

Invited Session

Optimization in Data Science

Chair: Junhui Zhang, MIT, Cambridge, MA, United States

1 - Perturbing the Derivative: Wild Refitting for Model-Free Evaluation of Machine Learning Models under Bregman Losses

Haichen Hu, MIT, Cambridge, MA, United States, David Simchi-Levi

We study the excess risk evaluation of classical penalized empirical risk minimization (ERM) with Bregman losses. We show that by leveraging the idea of wild refitting, one can efficiently upper bound the excess risk through the so-called "wild optimism," without relying on the global structure of the underlying function class. This property makes our approach inherently model-free. Unlike conventional analysis, our framework operates with just one dataset and black-box access to the training procedure. The method involves randomized Rademacher symmetrization and constructing artificially modified outputs by perturbation in the derivative space with appropriate scaling, upon which we retrain a second predictor for excess risk estimation. We establish high-probability performance guarantees both under the fixed design setting and the random design setting, demonstrating that wild refitting under Bregman losses, with an appropriately chosen wild noise scale, yields a valid upper bound on the excess risk. Thus, our work is promising for theoretically evaluating modern opaque ML models, such as deep neural networks and generative models, where the function class is too complex for classical learning theory and empirical process techniques.

2 - Differentiable Robust Optimization with Conformal Coverage

Owen Shen, MIT, Cambridge, MA, United States, Haihao Lu, Hungpo Chao, Patrick Jaillet

We develop a framework for learning the geometry of ellipsoidal uncertainty sets end-to-end from data, targeting operating reserve procurement in electricity markets. The bilevel problem—minimizing robust dispatch cost subject to coverage constraints—is profiled into a single-level objective whose gradient combines envelope terms from dual multipliers with a quantile-sensitivity correction, requiring no differentiation through the solver. Split conformal prediction provides distribution-free coverage via a four-way data split.

3 - Hidden Convexity in Queueing Models

Minda Zhao, Georgia Institute of Technology, Atlanta, GA, United States, Xin Chen, Linwei Xin

We study the joint control of arrival and service rates in queueing systems with the objective of minimizing long-run expected cost minus revenue. Although the objective function is non-convex, first-order methods have been empirically observed to converge to globally optimal solutions. This paper provides a theoretical foundation for this empirical phenomenon by characterizing the optimization landscape and identifying a hidden convexity: the problem admits a convex reformulation after an appropriate change of variables. Leveraging this hidden convexity, we establish the Polyak-Łojasiewicz-Kurdyka (PŁK) condition for the original control problem, which excludes spurious local minima and ensures global convergence for first-order methods. Our analysis applies to a broad class of GI/GI/1 queueing models, including those with Gamma-distributed interarrival and service times. As a key ingredient in the proof, we establish a new convexity property of the expected queue length under a square-root transformation of the traffic intensity.

4 - Advances at the Intersection of Discrete Optimization and Interpretable Machine Learning

Brian Liu, MIT, Cambridge, MA, United States

Interpretable machine learning algorithms are prized for their ability to extract useful information from data and to audit predictions for trust and safety. Formulating interpretable models, such as sparse additive models or compact rule sets, using discrete optimization approaches leads to impressive modeling flexibility, offering exact control over sparsity and interaction structure. In this talk, we overview how recent advancements in discrete optimization frameworks for ensemble pruning and additive models make it possible to construct inherently transparent models that perform competitively with black-box ensembles. We discuss various formulations that encourage different behaviors in the constructed models, as well as specialized algorithms to solve the resulting problems efficiently.

SunB02

Georgian Room

Interactions of Optimization with Statistics, Games, and Diffusion II

Invited Session

Optimization in Data Science

Chair: Mateo Diaz, Johns Hopkins University, Baltimore, MD, United States

Co-Chair: Pedro Izquierdo Lehmann, Johns Hopkins University - Applied Math & Statistics, Baltimore, MD, United States

1 - Multiscale Replay: A Robust Algorithm for Stochastic Variational Inequalities with a Markovian Buffer

Milind Nakul, Georgia Institute of Technology, Atlanta, GA, United States, Tianjiao Li, Ashwin Pananjady

We introduce the Multiscale Experience Replay (MER) algorithm for solving stochastic variational inequalities (VIs) in settings where samples are collected in a Markovian buffer. MER utilizes a multi-scale scheme to emulate the behavior of VI algorithms designed for independent and identically distributed (i.i.d.) samples, thereby mitigating the adverse effects of data correlation. Notably, this property is achieved without requiring knowledge of the mixing time of the Markov chain. Our algorithm and theoretical analysis reveal how the so-called ``experience replay” heuristic, widely used in reinforcement learning, can be applied in a statistically efficient manner. We also discuss applications of MER, particularly in policy evaluation with temporal difference (TD) learning and in training generalized linear models under dependent sampling.

2 - Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization

Enrique Mallada, Johns Hopkins University, Baltimore, MD, United States, Ziqing Xu, Hancheng Min, Lachlan MacDonald, Jinqi Luo, Salma Tarmoun, Rene Vidal

Despite the empirical success of Low-Rank Adaptation (LoRA) in fine-tuning pretrained models, there is little theoretical understanding of how first-order methods with carefully crafted initialization adapt models to new tasks. In this work, we take the first step towards bridging this gap by theoretically analyzing the learning dynamics of LoRA for matrix factorization (MF) under gradient flow (GF), emphasizing the crucial role of initialization. For small initialization, we theoretically show that GF converges to a neighborhood of the optimal solution, with smaller initialization leading to lower final error. Our analysis shows that the final error is affected by the misalignment between the singular spaces of the pre-trained model and the target matrix, and reducing the initialization scale improves alignment. To address this misalignment, we propose a spectral initialization for LoRA in MF and theoretically prove that GF with small spectral initialization converges to the fine-tuning task with arbitrary precision. Numerical experiments from MF and image classification validate our findings.

3 - Active Set Identification and Rapid Convergence for Primal-Dual Degenerate Problems

Pedro Izquierdo Lehmann, Johns Hopkins University - Applied Math & Statistics, Baltimore, MD, United States

Primal-dual methods for solving optimization problems with functional constraints often exhibit a distinct two-stage behavior. Initially, they converge towards a solution at
a sublinear rate. Then, after a certain point, all iterates accurately identify the set of active constraints, after which their convergence accelerates to a much faster linear
rate. Theory characterizing this phenomenon spans more than three decades. However, most existing work only guarantees eventual active set identification and relies
heavily on nondegeneracy conditions, such as strict complementarity, which frequently fail to hold in practice. We introduce mild conditions on the problem geometry and
algorithm under which this phenomenon provably occurs. Our guarantees are entirely nonasymptotic and, importantly, do not require any nondegeneracy assumptions. We
demonstrate that our framework encompasses several widely used algorithms, including the proximal point method, the primal-dual hybrid gradient method, the alternating
direction method of multipliers, and the extragradient method.

4 - ***Cancelled*Certifying Any-Dimensional Inequalities Using de Finetti Theorems**

Eitan Levin, Caltech, Pasadena, CA, United States, Venkat Chandrasekaran

There is growing interest in computationally proving mathematical theorems (e.g., using AI). Many theorems can be stated as inequalities that hold for infinitely-many dimensions, including inequalities in symmetric functions and graph densities that arise in combinatorics. Taking an optimization perspective, we develop a framework for certifying any-dimensional inequalities by computing bounds on the limits of growing-sized optimization problems. Our key observation is a new and general connection between certifying any-dimensional inequalities and de Finetti-type theorems. In particular, we prove new de Finetti theorems and use them to certify new any-dimensional inequalities inaccessible to previous approaches.

SunB03

Plaza I

New Advances of Stochastic Nonlinear Optimization in Machine Learning I

Invited Session

Nonlinear Optimization

Chair: Sen Na, Georgia Institute of Technology, Atlanta, United States

Co-Chair: Miaolan Xie, Purdue University, West Lafayette, IN, United States

Co-Chair: Zhaosong Lu, University of Minnesota, Minneapolis, MN, United States

1 - Limitations of Smoothness Models in Nonconvex Optimization

Vivak Patel, University of Wisconsin - Madison, Madison, WI, United States

In nonconvex optimization, smoothness assumptions are critical to deriving convergence and complexity results. These smoothness assumptions focus on the continuity properties of the gradient function, and include assuming that the gradient is globally Lipschitz continuous or, more recently, that the Lipschitz constant grows as a polynomial function of the gradient norm. These more recent assumptions are often used in the analysis of methods applied in the training of neural networks, and have been justified through heuristic arguments on polynomials or through numerical experiments. Unfortunately, we show that such smoothness assumptions do not apply to interesting neural network training examples. We then show that (deterministic analogues of) methods whose convergence depends on such assumptions fall short under more appropriate assumptions. We conclude by introducing a method that overcomes these limitations under these more appropriate assumptions, and has superior performance to reasonable alternatives in numerical experiments.

2 - Finite-sum Coupled Compositional Optimization and its Application to Constrained Optimization with Multiple Functional Inequalities

Yang Tianbao, Texas A&M University, College Station, TX, United States

Finite-sum Coupled Compositional Optimization (FCCO), characterized by its coupled compositional objective structure, emerges as an important optimization paradigm for addressing a wide range of machine learning problems. In this paper, we focus on a challenging class of non-convex non-smooth FCCO, where the outer functions are non-smooth weakly convex or convex and the inner functions are smooth or weakly convex. Existing state-of-the-art result face two key limitations: (1) a high iteration complexity of $O(1/\epsilon^6)$ under the assumption that the stochastic inner functions are Lipschitz continuous in expectation; (2) reliance on vanilla SGD-type updates, which are not suitable for deep learning applications. Our main contributions are two fold: (i) We propose stochastic momentum methods tailored for non-smooth FCCO that come with provable convergence guarantees; (ii) We establish a new state-of-the-art iteration complexity of $O(1/\epsilon^5)$. Moreover, we apply our algorithms to multiple inequality constrained non-convex optimization problems involving smooth or weakly convex functional inequality constraints. By optimizing a smoothed hinge penalty based formulation, we achieve a new state-of-the-art complexity of $O(1/\epsilon^5)$ for finding an (nearly) ϵ-level KKT solution. Experiments on three tasks demonstrate the effectiveness of the proposed algorithms.

3 - Derivative-Free Sequential Quadratic Programming for Equality-Constrained Stochastic Optimization

Sen Na, Georgia Institute of Technology, Atlanta, GA, United States

We study nonlinear optimization problems with a stochastic objective and deterministic equality constraints, assuming that only zeroth-order information is available for both the objective and constraints. In this setting, we propose a Derivative-Free Stochastic Sequential Quadratic Programming (DF-SSQP) method that employs simultaneous perturbation stochastic approximation (SPSA) to randomly estimate the gradients and Hessians of the objective and constraints, requiring only a dimension-independent number of function evaluations--as few as eight per step. A key challenge distinguishing DF-SSQP from existing derivative-based SSQP methods is the random bias introduced into gradient and Hessian estimates, brought by stochastic zeroth-order approximations. As such, we introduce an online debiasing mechanism based on momentum-style estimators that properly aggregate all past gradient and Hessian information via a low-memory moving-average scheme. Under standard assumptions, we establish the global almost-sure convergence of DF-SSQP. We further provide a local asymptotic analysis, showing that the rescaled iterates exhibit asymptotic normality with a limiting covariance matrix analogous to the minimax-optimal covariance of derivative-based methods, albeit enlarged due to the absence of derivative information. This local result enables online statistical inference of model parameters. Numerical experiments on benchmark nonlinear problems demonstrate both the global convergence behavior and the local statistical properties of DF-SSQP.

4 - Nonlinear Optimization Problems Arising in Stability Analysis of Discrete Time Recurrent Neural Networks

Jayant Singh, NDSU, Lagrange, IL, United States

We consider a nonconvex optimization problem arising in the theory of global asymptotic stability of DiscreteTime Recurrent Neural Networks(RNN). RNN's are a special class of nonlinear systems, and its stability has been extensively investigated in literature. A novel stability criterion, called Method of Reduction of Dissipativity Domain has been proposed. This criterion is applicable to wider class of systems. Implementation of this criterion requires solving a sequence of nonlinear optimization problems for a given cost function. We derive conditions which guarantee an existence of at most one point of local maximum for such functions over every hyperplane. This nontrivial result is valid for wide range of neuron transfer functions.

5 - Neural-KKT based Optimization Agent for Nonlinear Programming

Saif Kazi, Los Alamos National Laboratory, Los Alamos, NM, United States, Harsha Nagarajan, Sachin Shivakumar, Russell Bent

Nonlinear Constrained Optimization a.k.a Nonlinear Programming (NLP) is an important tool in solving engineering problems [Optimal Control, Optimal Power Flow, Nonlinear Model Predictive Control etc.] with applications in multiple domains such as operations research, finance etc. Deterministic methods to solve NLPs are generally based on using the optimality criterion known as Karush-Kuhn-Tucker (KKT) conditions. Algorithms such as Interior-point methods (IPM) use Newton-Raphson to solve the nonlinear KKT conditions. This particular approach suffers from convergence issues due to the local convergence property of Newton's algorithm making it extremely important to provide good initial guess to the optimization algorithm for fast convergence.

In this study, we propose building an optimization agent based on mix of LLM and Neural Network architecture which first transforms NLP problems written in text format to the KKT conditions. In second step, the solution to these nonlinear KKT equations is posed as a residual minimization problem. This minimization problem is then encoded as a Neural Network (NN) and solved efficiently using the gradient descent based backpropagation algorithm. This step doesn't require training of neural network using artificial dataset and also mitigates the need for custom initialization associated with the classical Newton-Raphson method. We test our approach on benchmark NLP problems to showcase and compare it's performance with conventional NLP solvers.

SunB04

Plaza II

Data Value in Optimization under Uncertainty

Invited Session

Optimization under Uncertainty

Chair: Amine Bennouna, Kellogg, Northwestern University, Evanston, IL, United States

1 - What Data Enables Optimal Decisions? An Exact Characterization for Linear Optimization

Omar Bennouna, Massachusetts Institute of Technology, Cambridge, MA, United States, Amine Bennouna, Saurabh Amin, Asuman Ozdaglar

We study the fundamental question of how informative a dataset is for solving a given decision-making task. In our setting, the dataset provides partial information about unknown parameters that influence task outcomes. Focusing on linear programs, we characterize when a dataset is sufficient to recover an optimal decision, given an uncertainty set on the cost vector. Our main contribution is a sharp geometric characterization that identifies the directions of the cost vector that matter for optimality, relative to the task constraints and uncertainty set.
We further develop a practical algorithm that, for a given task, constructs a minimal or least-costly sufficient dataset.
Our results reveal that small, well-chosen datasets can often fully determine optimal decisions---offering a principled foundation for task-aware data selection.

2 - Online Learning with Mild Adversaries without Distributional Priors

Moise Blanchard, Georgia Tech, Somerville, MA, United States

Classical results in statistical learning typically consider two extreme data-generating models: i.i.d. instances from an unknown distribution, or fully adversarial instances, often much more challenging statistically. To bridge the gap between these models, recent works introduced several frameworks to quantify mild adversaries. For instance, in the smoothed framework, the adversary is constrained to have density bounded by $1/\sigma$ compared to some fixed base measure μ, interpolating i.i.d. and adversarial scenarios depending on the smoothness parameter σ. Alternatively, in the abstention framework, clean data is generated i.i.d. from some fixed distribution μ then covariates may be arbitrarily corrupted, but the learner may decide to abstain if they believe a data has been corrupted. In both scenarios, most prior works heavily relied on the prior knowledge of μ. Arguably, this is a strong assumption both in practice and in theory since distributional knowledge is not necessary with full i.i.d. data (PAC learning). We resolve this conceptual gap by proposing algorithms agnostic to this base distribution μ. We propose the first algorithm to achieve sublinear regret without well-specified assumptions. In classification, it achieves \tilde O(\sqrt{T/σ}) regret for T iterations. Similarly, in the abstention setting, we provide the first class of algorithms which provide a non-trivial positive trade-off between misclassification error and the number of abstentions when the data was not corrupted. Specifically, for a budget of T^{1-a} abstentions it achieves misclassification error \tilde O(T^{4a}). We will conclude by discussing corresponding statistical gaps between having prior distributional knowledge of μ or not.

3 - Data Attribution in High-Dimensions and without Strong Convexity

Ittai Rubinstein, Massachusetts Institute of Technology, Cambridge, MA, United States, Sam Hopkins

Data attribution methods aim to quantify how training examples shape model predictions, supporting applications in interpretability, unlearning, and robustness. The dominant tools in practice are *influence functions* (IF) and *Newton step* (NS) approximations, yet their theoretical guarantees and practical accuracy have remained poorly understood. In this talk, I will present new analytic techniques that uncover the scaling laws of the approximation error of IF and NS. Our results improve on prior analyses both by establishing asymptotically sharper bounds and by avoiding dependence on the global strong convexity parameter, which is often prohibitively small in practice. These insights not only explain long-standing empirical observations —such as why and when NS is more accurate than IF—but also guide the design of new methods. As an application, I will present *rescaled influence functions* (RIF), a simple, drop-in replacement for IF that matches the efficiency of IF while achieving the accuracy of NS. I will discuss both theoretical advances and empirical results on real-world datasets. Together, these contributions provide a first principled understanding of data attribution methods and demonstrate how to turn this understanding into more reliable tools.

4 - A Minimalist Bayesian Framework for Stochastic Optimization

Kaizheng Wang, Columbia University, New York, NY, United States

The Bayesian paradigm offers principled tools for sequential decision-making under uncertainty, but its reliance on a probabilistic model for all parameters can hinder the incorporation of complex structural constraints. We introduce a minimalist Bayesian framework that places a prior only on the component of interest, such as the location of the optimum. Nuisance parameters are eliminated via profile likelihood, which naturally handles constraints. As a direct instantiation, we develop a MINimalist Thompson Sampling (MINTS) algorithm. Our framework accommodates structured problems, including continuum-armed Lipschitz bandits and dynamic pricing. It also provides a probabilistic lens on classical convex optimization algorithms such as the center of gravity and ellipsoid methods. We further analyze MINTS for multi-armed bandits and establish near-optimal regret guarantees.

SunB05

Plaza III

Dynamic Decisions under Uncertainty: Formulations, Algorithms, and Analysis II

Invited Session

Optimization under Uncertainty

Chair: Johannes Milz, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Alexander Shapiro, ISyE Georgia Tech, Atlanta, GA, United States

Co-Chair: Enlu Zhou, ISyE Georgia Tech, Atlanta, GA, United States

1 - Distributionally Robust Control (Part I)

Dan Iancu, Stanford University, Stanford, CA, United States, Bahar Taskesen, Cagil Kocyigit, Daniel Kuhn

We study a distributionally robust generalization of the classic discrete-time, finite-horizon Linear-Quadratic-Gaussian (LQG) control problem. The state and observation noise distributions are chosen adversarially from divergence-based ambiguity sets centered at a nominal distribution that is zero-mean and Gaussian. The distribution of noise terms may have nonzero means and must have finite second moments satisfying an orthogonality condition (which reduces to uncorrelatedness when means are zero). Under a mild structural assumption on the divergence – satisfied by several relevant examples – we prove that a policy depending affinely on the history of observations is optimal, and the adversary's worst-case distribution is Gaussian, and with an additional weak assumption, we prove that the adversary sets the mean to zero and the optimal policy becomes linear. We develop a Frank-Wolfe algorithm based on solving standard LQG subproblems via Kalman filtering and dynamic programming, and we show empirically that it consistently outperforms semidefinite-programming reformulations. We show that the results hold when the nominal distribution is elliptical – rather than Gaussian – if the divergence is the 2-Wasserstein metric.

2 - Distributionally Robust Control (Part II)

Bahar Taskesen, University of Chicago, Chicago, IL, United States, Dan Iancu, Cagil Kocyigit, Daniel Kuhn

This talk presents an infinite horizon, average cost formulation of distributionally robust linear quadratic Gaussian control with time invariant system matrices and a time invariant nominal disturbance model. We study a zero sum game between a controller and an adversary who selects disturbance distributions from divergence based ambiguity sets centered at a Gaussian nominal model. Importantly, we do not impose stationarity a priori on the control policies or on all distributions in the ambiguity set, which introduces additional technical challenges and requires slightly stronger assumptions than in the finite horizon case. Under mild regularity conditions on the divergence, stabilizability and detectability that mirror classical LQG assumptions for infinite horizon models, we establish the existence of a Nash equilibrium with a tractable and interpretable structure: the controller employs an optimal stationary linear output feedback policy, while the adversary selects a worst case time invariant Gaussian distribution. The analysis relies on Fan type minimax duality and shows that classical LQG separation principles extend to this distributionally robust setting, thereby generalizing our finite horizon results and aligning the robust model with the structural insights of traditional infinite horizon LQG with known noise distributions.

3 - Central Limit Theorems for Sample Average Approximations in Stochastic Optimal Control

Johannes Milz, Georgia Institute of Technology, Atlanta, GA, United States, Alexander Shapiro

We establish central limit theorems for the Sample Average Approximation (SAA) method in discrete-time, finite-horizon Stochastic Optimal Control. Using the dynamic programming principle and backward induction, we characterize the limiting distributions of the SAA value functions. The asymptotic variance at each stage decomposes into two components: a current-stage variance arising from immediate randomness, and a propagated future variance accumulated from subsequent stages. This decomposition clarifies how statistical uncertainty propagates backward through time. Our derivation relies on a stochastic equicontinuity condition, for which we provide sufficient conditions. We illustrate the variance decomposition using the classical Linear Quadratic Regulator (LQR) problem. Although its unbounded state and control spaces violate the compactness assumptions of our framework, the LQR setting enables explicit computation and visualization of both variance components.

4 - ***Cancelled*Dynamic Risk Measure under Filtration Mismatch**

Rui Gao, The University of Texas at Austin, Austin, TX, United States, Jincheng Yang, Luhao Zhang

Dynamic risk measures are usually defined on a filtration generated by some data process, while in practice, the information structure is often approximated by a scenario tree. This creates a mismatch between the reference filtration used for modeling and the target filtration of the implemented environment---a feature largely overlooked in the literature, but crucial for concepts like time consistency. We introduce a new notion of time consistency that explicitly accounts for such mismatched filtrations and collapses to the classical definition when they coincide. We then propose a new dynamic risk measure that is time-consistent in this generalized sense and can be interpreted as reweighting and relocating conditional scenarios, simultaneously adjusting both their probabilities and support.

SunB06

Director's and Lounge

Advances in Deterministic Methods II

Invited Session

Global Optimization

Chair: Matthew Stuber, University of Connecticut, Mansfield, CT, United States

Co-Chair: Joseph Scott, Georgia Institute of Technology, Atlanta, United States

1 - A New Bounding Framework for Parametric ODE Solutions

Kamil Khan, McMaster University, Hamilton, ON, Canada

Joint work with Huiyi Cao.

In chemical engineering applications such as safety verification, nonconvex optimization problems must be solved to guaranteed global optimality, to within a specified tolerance. Typical methods for deterministic global optimization compute crucial bounding information by minimizing convex relaxations of objective functions. However, generating useful convex relaxations can be difficult if process models include embedded systems of ordinary differential equations (ODEs), since established reachable-set methods for ODEs are impressive yet still somewhat limited in scope and fidelity. These dynamic relaxation methods compute convex relaxations for ODE solutions together with interval bounds, by solving an auxiliary relaxation ODE system. By analogy, in the well-known McCormick relaxation approach for constructing convex relaxations of composite functions, these relaxations are also computed in tandem with constant interval bounds, which provide useful range estimates for intermediate variables.

We present a framework that extends state-of-the-art relaxation approaches for ODE solutions, with the ultimate goal of aiding methods for dynamic global optimization. This framework makes use of the relationship between ODE solution relaxations and their related interval bounds. We take mathematical techniques that were used to establish correctness of ODE solution relaxations, and we adapt these to develop several new types of interval bounds on ODE solutions, with guaranteed inclusion monotonicity properties. (We also derive several established methods in a new way.) For example, we obtain new effective interval bounds based on the edge-concave relaxations of Hasan (2018), which significantly outperform convex relaxations in some cases. Numerical examples are presented, and implications are discussed.

2 - Separable Edge-Concave Underestimators for High-Dimensional Signomials

Bimol Nath Roy, Texas A&M University, College Station, TX, United States, Faruque Hasan

We present a new relaxation based on separable edge-concave underestimator (SEC) for general n-dimensional signomial problems. While the vertex polyhedral convex envelope of an edge-concave underestimator is constructed with convex combination of function evaluations at all 2^n domain vertices, the SEC underestimator uses a single linear hyperplane constructed at only $(n+1)$ linearly independent vertices.

We show that there exists a single n-dimensional hyperplane passing through all 2^n vertices of any separable function defined over a box. Moreover, we provide an explicit form of the linear hyperplane representing the convex envelope of the SEC underestimator. Having an explicit form of the hyperplane leads to an efficient linear programming relaxation of the original problem with high-dimensional signomials. We test our proposed relaxation technique to obtain root node lower bounds for several test functions with dimensions up to 100. We find that our method of relaxation, on average, provides tighter and faster lower bounds for higher dimensional signomials when compared with the performance of state-of-the-art commercial solvers. Future work entails to extending the separable edge-concave underestimators for general C^2 nonconvex problems.

3 - Rotational Transformations for QCQP

Jan Kronqvist, KTH Royal Institute of Technology, Stockholm, Sweden, Yehor Blokhin

In this talk, we investigate how rotational and linear transformations can be used in conjunction with classical convexification methods to obtain stronger convex relaxations. The main ideas are applicable to bounded problems with nonconvex constraints and/or objective defined by twice differentiable functions. But, we focus on global optimization of non-convex quadratic problems with some linear constraints (not only variable bound constraints).

We introduce a novel family of relaxation methods obtained by linearly transforming the vector space to better capture the geometry of the feasible set. Preliminary experiments indicate that the presented approaches, integrated with the aBB method and McCormick envelopes, produce significantly tighter convex bounds.

4 - Scaling Deterministic Global Optimization With SIMT Architectures

Matthew Stuber, University of Connecticut, Mansfield, CT, United States, Robert Gottlieb, Dimitri Alston

Specialized computing architectures offer opportunities to dramatically boost the computational performance of algorithms for scientific computing applications. One dominant strategy is the use of graphics processing units (GPUs) for algorithms with highly repetitive floating-point operations and high data throughput, such as machine learning model training and data analysis. Such computing architectures are classified as “single-instruction, multiple-thread” or SIMT, and effectively execute a single instruction across many computing elements simultaneously. In contrast to conventional computing architectures (e.g., x86 and x86-64), which are classified as “multiple-instruction, multiple-data” or MIMD, that allow multiple independent and different instructions to be executed simultaneously on different data. Until recently, the acceleration of mathematical programming algorithms was not possible with SIMT architectures. Recently, MadNLP.jl was developed to dramatically accelerate the solution of large-scale nonlinear programs to local optimality. However, for applications where guaranteed optimality is paramount, such as in robust design, deterministic global optimization is required. In this talk, we discuss our recent advances in the development of a novel deterministic global optimization algorithm and software implementation for SIMT architectures. Our approach utilizes McCormick-based relaxations and their subgradients to construct linear programming lower-bounding problems. A novel implementation of PDLP was developed to solve LP lower-bounding problems entirely on the GPU with minimal data transfer between the host CPU and the GPU. Implemented within a spatial branch-and-bound framework, this approach achieves a significant speedup over an equivalent CPU implementation.

SunB07

Congress Room

Optimization, Game Theory, and Society

Invited Session

Discrete Optimization

Chair: Federico Bobbio, Northwestern University, Chicago, IL, United States

1 - On Path Competition among LEO Satellites

Federico Bobbio, Northwestern University, Chicago, IL, United States, Randal Berry, Michael Honig, Rakesh Vohra, Thanh Nguyen, Vijay Subramanian

We study competition between low Earth orbit (LEO) satellite constellations when firms must share a finite spatial resource and overlapping coverage regions. Operators choose how densely to deploy satellites and how to allocate limited capacity across geographic demand, taking into account that satellites move, footprints overlap, and users can subscribe only when they fall within the coverage of an available satellite. This creates congestion-type interactions both in physical space (collision and interference constraints) and in economic space (competition for users). We formulate the interaction as a leader–follower game in which an incumbent chooses its deployment first and an entrant best responds, and we characterize how footprint size, capacity, and demand heterogeneity shape equilibrium deployment, congestion, and the division of demand. The analysis highlights conditions under which unconstrained deployment can limit effective entry or reduce overall welfare, and it suggests how access, spacing, or coordination rules for LEO infrastructure could preserve competition while respecting physical and operational constraints in a congested orbital environment.

2 - Log-Optimal Portfolio Construction for Binary Options with Combinatorial Constraints

Jeff Decary, University of Connecticut, Storrs, CT, United States, David Bergman, Bin Zou

We study the problem of optimal wealth allocation across independent binary options with known payouts, aiming to maximize log utility under practical constraints. This general framework arises in settings such as prediction markets (e.g., Kalshi and Polymarket), financial event contracts (e.g., Nadex), and sports betting (e.g., Draftkings and FanDuel). The Kelly Criterion provides a classical solution for the bet sizing of a single binary option, and numerous papers have explored extensions to multiple binary options. Our work expands on this body of research by investigating how to incorporate combinatorial constraints into the model, including limits on the number of binary options to select in a portfolio, a requirement that arises in many settings. These constraints considerably increase the computational complexity of the problem, thereby necessitating advanced solution methodologies. To address this challenge, we develop a logic-based Benders decomposition algorithm that provides a scalable and computationally efficient solution framework. Although broadly applicable, we focus on sports betting due to its market scale and unique inclusion of parlay options. We also study how sportsbooks can make slight modifications to parlay pricing so that optimal allocations do not include parlay options, even though such options may remain attractive to bettors.

3 - Three Stage Stochastic Network Design and Routing for Cold Food Supply Chains Under Uncertainty

Rosemarie Santa Gonzalez, Georgia Institute of Technology, Atlanta, GA, United States, Janosch Ortmann

Access to fresh and temperature sensitive foods in regions with limited infrastructure requires distribution systems that can operate under substantial uncertainty in supply, workforce, and demand. Motivated by ongoing work with community partners and tested on data from cold food supply chains, this study develops a three-stage stochastic optimization model for the joint design and operation of such systems. In the first stage, the model determines the optimal configuration and capacity of the distribution network under uncertainty in farmer production availability, driver participation, and location-specific demand. The second stage specifies routing decisions after the first stage realization is revealed, accounting for variable driver availability, stochastic travel conditions, and cold-chain requirements. The third stage optimizes the allocation of limited perishable inventory at each site when both realized supply and the number of households present may deviate from forecasts. By integrating supply, workforce, and demand uncertainty into a unified multi-stage framework, the model provides an approach for improving reliability and operational efficiency in real world cold supply chains.

4 - Online Learning for Equilibrium Pricing in Markets under Incomplete Information

Haoyuan Sun, Massachuesetts Institute of Technology, Cambridge, MA, United States, Devansh Jalota, Navid Azizan

The computation of *equilibrium* prices at which the supply of goods matches their demand typically relies on complete information on agents' private attributes, e.g., suppliers' cost functions, which are often unavailable in practice. Motivated by this practical consideration, we consider the problem of learning equilibrium prices over a horizon of T periods in the incomplete information setting wherein a market operator seeks to satisfy the customer demand for a commodity by purchasing it from competing suppliers with cost functions unknown to the operator. We first consider the setting when suppliers' cost functions are fixed and develop algorithms that, on three pertinent regret metrics, simultaneously achieve a regret of $O(1)$ when the customer demand is constant over time, and $O(\sqrt{T})$ when the demand varies over time. In the setting when the suppliers' cost functions vary over time, we demonstrate that, in general, no online algorithm can achieve sublinear regret on all three metrics. Thus, we consider an augmented setting wherein the operator has access to hints/contexts that reflect the variation in the cost functions and propose an algorithm with sublinear regret in this augmented setting. Finally, we present numerical experiments that validate our results and discuss various model extensions.

SunB08

Committee Room

Optimization of Sustainable Logistics

Invited Session

Emerging Applications of Optimization

Chair: Gina Villafane, University of Waterloo, 200 University Avenue West, Waterloo, N2L 3G1, Canada

1 - Evaluating Just-in-Time ship arrivals at the Panama Canal: A Model-Based Assessment of Emission-Reducing Scheduling Policies

Gabriel Fuentes, Norwegian School of Economics, Bergen, Norway, Torstein Takvam, Stein W. Wallace

The Panama Canal is a vital hub for international shipping, and the scheduling of canal operations is critical for the efficiency and sustainability. As with other maritime bottlenecks, the Panama Canal currently requires most scheduled ships to arrive hours ahead of their transit, creating queues of ships to prevent canal idle time. Conversely, ship operators can reduce emissions and costs by adjusting sailing speed and arriving Just In Time (JIT). Aligning with the Canal's goal of encouraging sustainable supply chains, JIT arrivals present an opportunity to reduce emissions. However, it remains unclear if emission-reducing policies will impact the canal's revenue, which is primarily driven by canal throughput. In this study, we propose a combined ship scheduling and speed reduction model as a mixed integer linear program (MILP) for evaluating how changes to the canal's scheduling policy can impact the sailing speed and thereby emission from ships transiting the canal. Our heuristic decomposition method ensures an acceptable computation time. We validate the model's output applying historical data from the Panama Canal Authority (ACP) and the Automatic Identification System (AIS). The results demonstrate significant emission reduction potential, consistent with prior research. Furthermore, by comparing profit and emission reducing objectives, the results suggest that a change in policy is practically feasible without a significant difference in canal throughput, which is the primary driver of revenue. With acknowledged limitations, the results indicate that the proposed model is suitable for evaluating the impact of policy changes to the ship scheduling procedures at the Panama Canal.

2 - An Integrated Approach to Container Demand Prediction and Berth Allocation

Sahian Rosales, Georgia Tech Panama Logistics Innovation & Research Center, Panama City, Panama, Gabriel Fuentes, Jorge Barnett

Accurate container throughput forecasting is essential for port capacity planning in Panama, where maritime logistics drive economic competitiveness. This study integrates machine learning-based demand forecasting with berth allocation optimization to support proactive port operations. Using AIS-derived vessel draft variations and Random Forest regression tested on 2019-2024 data from Panama Maritime Authority, we estimated container throughput and achieve high-accuracy predictions ($R^2 = 0.98$, RMSE ≈ 6,000 TEU) for ports in the vicinity of the Panama Canal. These forecasts feed a Berth Allocation Problem (BAP) formulation that assigns vessels to berths while minimizing idle times under anticipated demand conditions. By incorporating forecasted volumes, port operators gain advance visibility into capacity constraints and reduce the congestion risks. We implement the approach at Manzanillo International Terminal with preliminary validation across additional Panamanian ports, showing the methodology's generalization.

3 - A Mathematical Model for Estimating International Ammonia Demand for Maritime Transport Fuel

César Cerda, Universidad Adolfo Ibáñez, Santiago, Chile, Tito Homem-de-Mello, Frederic Babonneau, Gabriel Fuentes, Pierre Cariou

One of the main challenges that the new green hydrogen (GH2) industry is facing is to obtain and to secure off-takers. Recently, maritime transport has gained traction as potential demand for green ammonia, which can be produced from GH2. The shipping industry is a hard-to-abate sector, being responsible for nearly 3% of global CO2 emissions. Considering this, the IMO (International Maritime Organization) has been developing a globally applicable regulation, the Net Zero Framework (NZF), which would allow for the progressive penalization of vessels' CO2 emissions, thereby sending a clear signal to the various stakeholders to move forward along the sector's decarbonization. This would make it the first industry with globally uniform regulation for CO2 emissions, which could encourage the bankability of related projects. This research presents an optimization model that estimates the bunkering points' locations and the required fuel volume, using historical transit data of container ships entering and leaving the Americas. The model minimizes the total cost of the ecosystem formed by ships owners and ammonia-based bunkering distributors, subject to production constraints of potential suppliers. It provides insights into the potential impact of a price of emissions on the ammonia fuel supply and demand development, and Chile's participation as producer.

4 - On the Panama Canal Scheduling and Fresh Water Sustainability

Gina Villafane, University of Waterloo, Waterloo, ON, Canada, James H. Bookbinder

To enable interoceanic transit, the Panama Canal uses freshwater. Because of the mountainous geography, water from a lake at a higher altitude than each entrance to the canal facilitates vessel passages. With each vessel crossing, millions of gallons of freshwater are discharged to the oceans. Thus, increasing transits means greater water loss. This is the fundamental trade-off of the canal’s design.

In the past, water availability was not a concern, due to Panama’s rainforest climate. But severe droughts are risking its service reliability, creating rippling economic effects on global supply chains. The canal requires precise water conservation strategies for every scheduled vessel, to maintain its watershed and attain its business goals.

Scheduling vessels through the canal is highly constrained; not only by its operational capacity but also by opposing transit directions, pre-emptive processing, and cargo-related navigational restrictions.

To find operational efficiencies while considering conflicting goals, we propose a multi-objective bidirectional hybrid flow shop model that considers two water conservation modes. We define a Pareto frontier of non-dominated solutions that balances makespan and water utilization. As an alternative to drastically restricting transits during droughts, the Panama Canal Authority's watershed-level policy can be represented in this model, and water conservation proactively implemented.

5 - Optimization and Simulation for Crowd-Based Food Delivery in Dublin

Javier Faulin, Public University of Navarre, Pamplona, Spain, Adrian Serrano-Hernandez, Luis Cadarso, Peter Keenan

Crowd-based bike delivery systems offer a scalable, cost-effective alternative to traditional restaurant delivery models, which rely on costly owned-fleet vehicles. This research evaluates the viability of such systems in Dublin through an agent-based simulation model integrating population characteristics, biker behavior, and restaurant operations.The framework combines a computational agent-based simulation with multiple optimization modules to guide agent interactions. A biased-randomization savings heuristic for Vehicle Routing Problems, adapted from traditional fleet-based systems, optimizes delivery routes within the simulation. The model explores diverse scenarios varying demand rates, biker adoption, and operational parameters. Results demonstrate promising performance across key performance indicators including economic efficiency, delivery reliability, and customer satisfaction. This work contributes to understanding the viability of sharing-economy delivery models and provides practical optimization insights for urban logistics networks operating at the intersection of data-driven decision-making and sustainable mobility solutions.

6 - Electric Vehicle Scheduling and Vehicle-to-Grid Integration in Micro-grids

Nathan Cho, Cornell University, New York, NY, United States, Andrea Lodi, Anna Scaglione

The logistical challenges and significant costs associated with procuring and transporting fuel to micro-grids, for example those created ad hoc in case of emergency responses or in military bases, underscore the need for sustainable and resilient energy solutions. Integrating renewable energy sources and electric vehicles into those micro-grids offers a promising approach to enhance energy security and operational readiness. This paper explores the optimization of travel, charging, and discharging schedules for a fleet of electric trucks, aiming to minimize reliance on fuel-generated electricity while ensuring that mission-critical transportation needs are met. We extend the classical Vehicle Scheduling Problem by incorporating Electric Vehicle Scheduling Problem dynamics and Vehicle-to-Grid capabilities, developing a comprehensive optimization model that addresses both logistical and energy demands within a micro-grid context. Utilizing a column generation approach, we efficiently solve large-scale instances and demonstrate significant improvements in fuel efficiency, renewable energy utilization, and overall operational cost. Computational experiments using realistic demand and solar generation data illustrate that integration of vehicle-to-grid enabled electric vehicles substantially reduces fuel consumption and can generate surplus energy returned to the grid. The results indicate that while battery constraints may require an increased fleet size, strategic scheduling of charging and discharging yields considerable economic and operational benefits. Our findings provide valuable insights for planners aiming to optimize energy use, reduce dependence on traditional fuel sources, and enhance operational resilience in remote environments.

SunB09

Cabinet Room

Optimization for Illicit Trafficking Networks

Invited Session

Emerging Applications of Optimization

Chair: Xiaowei Guo, Clemson University, CLEMSON, United States

Co-Chair: Thomas Sharkey, Clemson University, Clemson, SC, United States

1 - Intervening into Sex Trafficking Networks with Forced Criminality and Quotas

Xiaowei Guo, Clemson University, CLEMSON, SC, United States, Thomas Sharkey

We study interventions into sex trafficking networks where traffickers provide their victims quotas to meet through participating in multiple illegal activities. The interaction between the trafficker and their victims is modeled as a bilevel optimization problem: the trafficker (upper-level) assigns quotas to the victims (lower-level) while aiming to minimize their total risk, whereas victims aim to fulfill their quotas quotas with minimal harm. The intervener into the trafficking networks seek to maximize the total risk faced by the traffickers and is, therefore, interdicting a bilevel optimization problem.
We propose approaches to reformulate the intervener's problem into (i) a max-min mixed-integer program by examining the optimality conditions of each victim's optimization problem and (ii) a single-level mixed-integer program by examining the optimality conditions of each trafficker's bilevel optimization problem. These approaches can be applied to both the optimistic and pessimistic assumptions for the bilevel optimization problems of the traffickers.
Experimental results show that our single-level reformulation consistently outperforms the max-min MIP reformulation, especially when integrating additional theoretical results to reduce the considered solution space.

2 - A Bayesian Nonparametric Modeling for Child Trafficking Vulnerabilities in Sierra Leone

Wenyu Gao, UNC Charlotte, Charlotte, NC, United States

Child trafficking remains a critical social challenge marked by complex spatial and contextual heterogeneity. This study employs a Bayesian nonparametric modeling framework to investigate trafficking vulnerability patterns across Sierra Leonean chiefdoms, administrative units analogous to cities in the United States, within three high-burden districts. Incorporating areal information from the chiefdoms as random effects, a generalized linear mixed model (GLMM) is developed to assess the relationship between trafficking prevalence and potential risk factors. By applying a nonparametric Bayesian clustering approach to the random effects, the study innovatively identifies latent risk structures and spatial dependencies. Specifically, covariate-dependent models such as the Weighted Dirichlet Process Mixture (WDPM) and Dependent Dirichlet Process (DDP) models are leveraged to integrate socioeconomic and cultural covariates, thereby enhancing clustering precision and estimation accuracy. Unlike traditional clustering techniques, the Bayesian nonparametric framework adaptively determines cluster configurations without requiring a predefined number of groups. Clustering outcomes are rigorously validated through community-level hypothesis testing to ensure robustness and reliability. This novel modeling framework provides a comprehensive and data-driven approach to understanding trafficking risks in resource-limited settings. By uncovering hidden vulnerability patterns, the study delivers actionable insights for designing targeted, context-specific interventions. The findings underscore the potential of advanced Bayesian modeling to inform evidence-based policy-making and optimize resource allocation in efforts to combat child trafficking effectively.

3 - Encoding Spatiotemporal Patterns for Counter-trafficking Intelligence

Nickolas Freeman, University of Alabama, Tuscaloosa, AL, United States, Roya Shomali, Gregory Bott

Adult Service Websites (ASWs) host advertisements for commercial sex services and have been linked to several cases of sexual exploitation and sex trafficking. Given the connection to sex trafficking, ASW data is viewed as an important source of intelligence for counter-trafficking operations. One area where ASW data can provide insight is the geographical patterns of the individuals represented in the data. This is important because it is widely believed that sex traffickers move their victims to make it difficult for law enforcement to track and recover victims and to isolate victims from familiar people or places. Although previous research has investigated geographical patterns in ASW data, these studies are limited by small sample sizes and inferior methods for linking ads at the individual level. This research leverages recent advances in processing and linking ASW data, along with techniques from network science and artificial intelligence, to develop new methods for detecting geographic patterns in ASW data. Moreover, we demonstrate several use cases for the methods: 1) understanding geographical communities individuals move within, 2) ranking major traffic corridors by prevalence of use, and 3) identifying similar entities by spatiotemporal travel patterns.

4 - IMBWatch -- A Spatio-Temporal Graph Neural Network Approach to Detect Illicit Massage Business

Swetha Varadarajan, University of Cape Town , Cape Town , South Africa, Abhishek Ray, Lumina Albert

Illicit Massage Businesses (IMBs) are a covert and persistent form of organized exploitation, operating under the guise of wellness services while facilitating human trafficking, sexual exploitation, and coerced labor. Detecting IMBs is challenging because traffickers rely on coded digital advertisements, rapidly rotating staff, shifting business locations, and shared infrastructure such as phone numbers, addresses, and aliases. Traditional detection strategies—community tips, regulatory inspections, and manual reviews—are reactive and struggle to uncover the broader operational networks that enable IMBs to persist.

To address these limitations, we introduce **IMBWatch**, a spatio-temporal graph neural network (ST-GNN) framework designed to identify IMB activity at scale. IMBWatch builds dynamic graphs from open-source intelligence, where nodes represent entities such as businesses, individuals, phone numbers, and aliases, and edges encode relational, spatial, and temporal dependencies. The model integrates graph convolutional operations with temporal attention mechanisms to capture how IMB networks evolve across time and geography, enabling the discovery of subtle but consistent behavioral patterns.

Empirical evaluations on real-world data from multiple U.S. cities show that IMBWatch outperforms baseline approaches, achieving higher accuracy. The system produces actionable insights that support more targeted and proactive investigative strategies. IMBWatch is scalable, adaptable to other trafficking-related domains, and will be made publicly available with anonymized datasets and open-source code. This makes it a practical, data-driven tool for strengthening anti-trafficking efforts and advancing research on illicit network behavior in complex, evolving environments.

SunB10

Caucus Room

Cut to Fit: Task-Tailored Optimization under Uncertainty

Invited Session

Computational Optimization

Chair: Aras Selvi, Princeton University, Princeton, NJ, United States

1 - Prior-independent Bidding Strategies for First-Price Auctions

Omar Mouchtaki, New York University, New York, NY, United States, Rachitesh Kumar

First-price auctions are one of the most popular mechanisms for selling goods and services, with applications ranging from display advertising to timber sales. Unlike their close cousin, the second-price auction, first-price auctions do not admit a dominant strategy. Instead, each buyer must design a bidding strategy that maps values to bids—a task that is often challenging due to the lack of prior knowledge about competing bids. To address this challenge, we conduct a principled analysis of prior-independent bidding strategies for first-price auctions using worst-case regret as the performance measure. First, we develop a technique to evaluate the worst-case regret for (almost) any given value distribution and bidding strategy, reducing the complex task of ascertaining the worst-case competing-bid distribution to a simple line search. Next, building on our evaluation technique, we minimize worst-case regret and characterize a minimax-optimal bidding strategy for every value distribution. We achieve it by explicitly constructing a bidding strategy as a solution to an ordinary differential equation, and by proving its optimality for the intricate infinite-dimensional minimax problem underlying worst-case regret minimization. Our construction provides a computationally-tractable procedure for deriving minimax-optimal bidding strategies. When the value distribution is continuous, it yields a deterministic strategy that maps each value to a single bid. Importantly, our result allows us to quantify, through minimax regret, the performance loss due to a lack of knowledge about competing bids. We leverage this to analyze the impact of the value distribution on the performance loss.

2 - Stopping Rules for Stochastic Gradient Descent via Anytime-Valid Confidence Sequences

Liviu Aolaritei, UC Berkeley, Berkeley, CA, United States

We study a basic but unresolved question in stochastic optimization: when should stochastic gradient descent (SGD) be stopped based only on its observed trajectory? We develop anytime-valid confidence sequences for stochastic gradient methods that remain valid under continuous monitoring and directly yield statistically valid stopping rules. In convex optimization, they certify weighted suboptimality under general stepsize schedules; in nonconvex optimization, they certify weighted first-order stationarity. The result is a unified framework for online stopping of SGD with provable complexity guarantees

3 - Achieving First-order Statistical Improvements in Data-Driven Optimization

Tianyu Wang, Columbia University, New York, NY, United States, Henry Lam

Recent proliferation of data-optimization integration has led to a range of methods that aim to improve the statistical performance of data-driven optimization decisions. However, while many of these methods are motivated intuitively from a robustness or regularization perspective, their resulting statistical benefits are often less clear and, even if available, are argued in a case-by-case fashion. We provide a systematic dissection of data-driven optimization formulations using the view of "directionally perturbed" empirical optimization, which demonstrably covers most of the existing formulations. On the negative side, we argue that under mild smoothness conditions, any such formulations can result in at best second-order improvements. On the positive side, we show that in the presence of auxiliary information such as the availability of additional unsupervised data, we can construct a principled methodology, by building connections to the concept of Monte Carlo control variate, to achieve general first-order improvements in terms of excess risk.

4 - Emergence of Operationally-Interpretable Features in AI Optimization

X.Y. Han, Chicago Booth, Chicago, IL, United States

While recent work in OR/OM increasingly explores /how/ to better train AI algorithms for decision-making, less attention has been given to a complementary question: once AI predictors are successfully trained, what rules has the AI actually learned? This talk demonstrates how the geometry of a network’s learned embeddings can shed light on this issue. Using representative OR decision-making problems, such as multi-class scheduling, we show that after an AI learns to solve these problems, its internal vector representations can be mapped to classically understood models. This mapping allows for interpretability in the AI's decision-making procedure.

5 - Beyond Maximum Likelihood: Variational Inequality Estimation for Generalized Linear Models

Yao Xie, Georgia Institute of Technology, Atlanta, GA, United States, Linglingzhi Zhu, Jonghyeok Li

Generalized linear models (GLMs) are fundamental tools for statistical modeling, with maximum likelihood estimation (MLE) serving as the classical method for parameter inference. While MLE performs well in canonical GLMs, it can become computationally inefficient near the true parameter value. In more general settings with non-canonical or fully general link functions, the resulting optimization landscape is often non-

convex, non-smooth, and numerically unstable. To address these challenges, we investigate an alternative estimator based on solving the variational inequality (VI) formulation of the GLM likelihood equations, originally proposed by Juditsky and Nemirovski \cite{juditsky2019signal} as an alternative for solving nonlinear least-squares problems. Unlike their focus on algorithmic convergence in monotone settings, we analyze the VI approach from a statistical perspective, comparing it systematically with the MLE. We also extend the theory of VI estimators to a broader class of link functions, including non-monotone cases satisfying a strong Minty condition, and show that it admits weaker smoothness requirements than MLE, enabling faster, more stable, and less locally trapped optimization. Theoretically, we establish both non-asymptotic estimation error bounds and asymptotic normality for the VI estimator, and further provide convergence guarantees for fixed-point and stochastic approximation algorithms.
Numerical experiments show that the VI framework preserves the statistical efficiency of MLE while substantially extending its applicability to more challenging GLM settings.

SunB11

Charter Room

Quantum Computing

Invited Session

Computational Optimization

Chair: David Bernal Neira, Purdue University, West Lafayette, United States

Co-Chair: Mohammadhossein Mohammadisiahroudi, University of Maryland Baltimore County, Ellicott City, MD, United States

1 - Model Sparsification via Quantum-inspired Hamiltonian Descent

Yuxiang Peng, Purdue University, West Lafayette, IN, United States, Zhiyuan Jia, Pengyu Liu, Yu-Hsuan Wu, Tabish Shaik, Jianhao Ma, Jiaqi Leng, Xiaodi Wu, Chaoyue Zhao

Large language model (LLM) inference incurs substantial computational cost, driven primarily by dense matrix multiplications. Model sparsification aims to reduce this cost by reconstructing model behavior using sparse weight matrices. Building on SparseGPT, the sparsification process can be formulated as a set of mixed-integer quadratic programming (MIQP) problems. However, the scale and volume of these optimization tasks make conventional CPU-based solvers impractical. To address this challenge, we employ a GPU-native metaheuristic optimization method, quantum-inspired Hamiltonian descent (QIHD). QIHD produces near-optimal sparse reconstructions with significantly lower deviation from the original dense models than existing approaches. For small-scale LLMs, QIHD completes post-training sparsification within tens of GPU hours, demonstrating its practicality for accurate and scalable model pruning.

2 - Gradient Flows in Quantum Optimization

Jiaqi Leng, University of California, Berkeley, Berkeley, CA, United States

First-order methods are the workhorse of modern machine learning, but they are notoriously inefficient in non-convex landscapes, frequently becoming trapped by local minima and saddle points. These classical limitations make them a prime target for quantum acceleration. In this talk, I will first establish a fundamental connection between gradient flows (i.e., the continuous-time limit of gradient descent) and the evolution of a continuous-variable quantum system under a corresponding Hamiltonian. This mapping yields a new class of quantum algorithms that not only inherit the efficiency of gradient descent but also leverage quantum tunneling to escape saddle points and local minima. I will present results demonstrating this quantum advantage, including provable speedups and superior empirical performance in finding global solutions. Finally, I will briefly survey some recent advances in generalizing this quantum dynamical approach to higher-order optimization methods.

3 - Limitations to Decoded Quantum Interferometry for Max Cut

Ojas Parekh, Sandia National Labs, Albuquerque, NM, United States

Decoded Quantum Interferometry (DQI) is a framework for approximating special kinds of discrete optimization problems that relies on problem structure in a way that sets it apart from other classical or quantum approaches. We show that the instances of MaxCut on which DQI attains a nontrivial asymptotic approximation guarantee are solvable exactly by efficient classical algorithms. In this case the problem structure that enables good DQI approximations also renders the problem classically tractable.

4 - Continuous-Variable Quantum Algorithms for Dissipative Dynamics: From PDEs to Optimization

Zeguan Wu, University of Pittsburgh, Bethlehem, PA, United States, Min Chen, Yu Gan, Junyu Liu

Simulating nonlinear differential equations on quantum computers typically requires Carleman linearization and fault-tolerant digital hardware. We present an alternative analog approach using continuous-variable quantum systems, where classical fields are encoded into coherent states of bosonic modes. Leveraging the Koopman-von Neumann formalism, nonlinear dynamics are lifted to linear evolution implementable with logarithmic circuit depth. We then explore preliminary extensions to optimization, where engineered Lindblad dissipation can drive the system toward solutions of least-squares problems.

March 22, 2026, 11:45 AM - 12:30 PM

Plenary 5

Grand Ballroom

Convex (and Nonconvex) Optimization

Plenary

Plenary

Chair: Katya Scheinberg, Georgia Institute of Technology, Atlanta, GA, United States

1 - Convex (and Nonconvex) Optimization

Stephen Boyd, Stanford University, Stanford, CA, United States

Convex optimization has emerged as a useful tool for applications that include data analysis and model fitting, machine learning and statistics, resource allocation, engineering design, network design and optimization, finance, and control and signal processing. Disciplined convex programming (DCP) has become the basis for several domain specific languages for convex optimization, which allow a user to specify a convex optimization problem using a simple human readable language. If the user follows the DCP rules, their problem can be certified as convex, and automatically compiled into a standard form for a numerical solver, greatly simplifying the specification and solution of convex problems. In this talk we cover the basics of DCP and give several examples. We finish by presenting an extension of DCP to nonlinear programming, called DNLP, that extends many of the useful properties of DCP to NLP.

March 22, 2026, 1:30 PM - 3:00 PM

SunC01

Grand Ballroom

Recent Advances in Simple Bilevel Optimization

Invited Session

Optimization in Data Science

Chair: Nam Ho-Nguyen, The University of Sydney, SYDNEY, Australia

1 - Recent Advances in Convex Bi-level Optimization

Shoham Sabach, Cornell University, Ithaca, NY, United States

Bi-level optimization is a powerful framework for hierarchical decision-making, where one optimization problem is nested within another. This approach has found widespread applications in fields such as machine learning, energy systems, and economics. However, the growing scale and complexity of these systems underscore the growing need for efficient and theoretically grounded algorithms to address bi-level problems. In this talk, we focus on the class of convex bi-level optimization problems, presenting recent advancements in algorithm design and their theoretical convergence and complexity analysis. We will also address some of the inherent challenges in this domain and highlight open questions that continue to drive research in this field.

2 - Conditional Gradient Methods with Standard LMO for Stochastic Simple Bilevel Optimization

Khanh-Hung Giang-Tran, Cornell University, Ithaca, NY, United States, Soroosh Shafiee, Nam Ho-Nguyen

We propose efficient methods for solving stochastic simple bilevel optimization problems with convex inner levels, where the goal is to minimize an outer stochastic objective function subject to the solution set of an inner stochastic optimization problem.

Existing methods often rely on costly projection or linear optimization oracles over complex sets, limiting their scalability.

To overcome this, we propose an iteratively regularized conditional gradient approach that leverages linear optimization oracles exclusively over the base feasible set. Our proposed methods employ a vanishing regularization sequence that progressively emphasizes the inner problem while biasing towards desirable minimal outer objective solutions. In the one-sample stochastic setting and under standard convexity assumptions, we establish non-asymptotic convergence rates of $\tilde{O}(t^{-1/4})$ for both the outer and inner objectives. In the finite-sum setting with a mini-batch scheme, the corresponding rates become $\tilde{O}(t^{-1/2})$. When the outer objective is nonconvex, we prove non-asymptotic convergence rates of $\tilde{O}(t^{-1/7})$ for both the outer and inner objectives in the one-sample stochastic setting, and $\tilde{O}(t^{-1/4})$ in the finite-sum setting. Experimental results on over-parametrized regression and dictionary learning tasks demonstrate the practical advantages of our approach over existing methods, confirming our theoretical findings.

3 - On Incentive Mechanism Design in Federated Learning: Guarantees for Federated Averaging with Noncooperative Data Participation

Fateme Maleki, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States, Krishnan Raghavan, Farzad Yousefian

Federated learning enables distributed model training without sharing raw data. In this framework, keeping the clients engaged remains challenging since participants may receive uneven benefits from their contributions. We consider nonconvex and L-smooth local objective functions for the clients. We present an Incentive-enabled Federated Averaging method that rewards participants based on their level of participation in each round, through allowing clients to control the size of their training data dynamically in each round of communication. In each communication round, clients exchange both local parameters with the server and their updated training dataset sizes, which are dynamically adjusted using a Nash equilibrium seeking update rule. Subsequently, the server transmits updated global parameters back to clients to initiate the following round. We theoretically analyze the proposed method, obtaining explicit iteration and communication complexity guarantees. We plan to conduct numerical experiments on MNIST and CIFAR-10 datasets to evaluate the effectiveness of our approach, with a focus on examining the convergence behavior of the loss function.

4 - Bilevel Mixed-Integer Linear Program with Binary Tender

Bo Zhou, University of Michigan, Ann Arbor, MI, United States, Ruiwei Jiang, Siqian Shen

Bilevel programs model sequential decision interactions between two sets of players and find wide applications in real-world complex systems. In this paper, we consider a bilevel mixed-integer linear program with binary tender, wherein the upper and lower levels are linked via binary decision variables and both levels may involve additional mixed-integer decisions. We recast this bilevel program as a single-level formulation through a value function for the lower-level problem and then propose valid inequalities to replace and iteratively approximate the value function. We first derive a family of Lagrangian-based valid inequalities that give a complete description of the value function, providing a baseline method to obtain exact solutions for the considered class of bilevel programs. To enhance the strength of this approach, we further investigate another two types of valid inequalities. First, when the lower-level value function has intrinsic special properties such as supermodularity or submodularity, we exploit such properties to separate the Lagrangian-based inequalities quickly. Second, we derive decision rule-based valid inequalities, where linear decision rules and learning techniques are explored respectively. We demonstrate the effectiveness and efficiency of the proposed methods in extensive numerical experiments, including instances of general bilevel mixed-integer programs and those of a facility location interdiction problem.

5 - Network Flow-Based Cuts for Binary Bilevel Programs

Sebastian Vasquez, Carnegie Mellon University, Pittsburgh, PA, United States, Leonardo Lozano, Willem-Jan van Hoeve

Solving binary bilevel programs remains challenging due to the absence of strong yet tractable relaxations. Recent advances have embedded network flow polytopes into the high-point relaxation to obtain exact extended reformulations and valid dual bounds, but the resulting network size often limits scalability. We propose a new network flow-based framework for binary bilevel programs that uses networks as oracles to derive cuts directly in the original variable space. Building on existing approaches, we develop new exact reformulations and introduce mechanisms to strengthen structured cuts through tilting. Moreover, we design an extended network encoding that enables the derivation of stronger cuts and the tilting of any valid cut. Computational experiments demonstrate the effectiveness of the proposed approach.

SunC02

Georgian Room

Advances in Optimization for Data Science

Contributed Session

Optimization in Data Science

Chair: Minseo Lee, VIRGINIA POLYTECHNIC INSTITUTE, Blacksburg, United States

1 - MINLP for Regularized Symbolic Regression with Applications to Data-driven Discovery of Physical Laws

Dimitrios Fardis, Georgia Institute of Technology, Atlanta, GA, United States, Radhakrishna Tumbalam Gooty, Nikolaos Sahinidis

In recent years, symbolic regression has been approached through MINLP formulations, which provide deterministic optimality guarantee in contrast to the traditional approach of genetic programming. We explore various regularization metrics in the context of symbolic regression that aim to balance the accuracy and parsimony of the models built. We present extensive computational experimentation with the different metrics as we evaluate them for solvability of the resulting MINLP, and we assess which metrics facilitate the discovery of physical laws from ground-truth datasets.

2 - A Mathematical Programming Approach to Self-Attention in Transformer Models

Minseo Lee, Virginia Polytechnic Institute , Blacksburg, VA, United States, Esra Buyuktahtakin Toy

The integration of neural networks into mathematical optimization is well established for simple architectures but remains challenging for complex models such as the Transformer. The main difficulty stems from the self-attention mechanism, whose quadratic interactions and non-convex softmax function complicate direct optimization. This paper presents the first rigorous mathematical programming formulation that captures the computational structure of a single self-attention head.

We develop a reduced, abstract representation that explicitly models the key nonlinearities of self-attention. To ensure tractability, we introduce systematic piecewise-linear approximations for the softmax and interaction terms, enabling the formulation to balance fidelity and computational efficiency. The resulting model preserves the expressive behavior of attention while keeping the problem size manageable for modern solvers.

This formulation provides a critical building block for embedding high-fidelity, pre-trained Transformer components within Mixed-Integer Nonlinear Programming (MINLP) frameworks. By enabling optimization models to directly incorporate attention-based surrogate functions, this work advances the integration of deep learning and mathematical programming and supports the development of new hybrid AI–optimization methodologies.

3 - Bayesian Optimization for Automated Tuning of Predictive Control in Type 1 Diabetes Management

Marzia Cescon, University of Houston, Houston, TX, United States

In control theory, tuning control parameters after synthesizing a control law is often performed through trial-and-error. This process is limited by practical constraints such as time, computational burden, and lack of insight into the achievable optimal performance. Additionally, managing uncertainty remains a challenge—some approaches rely on zone
control, which alters the model or reference trajectory entirely, while others retune model or controller parameters. Artificial Pancreas systems for Type 1 Diabetes (T1D) management exemplify these challenges, compounded by strict safety requirements and the scarcity of real-world data, making optimization expensive. In this study, we investigate the use of Bayesian Optimization (BO) with Expected Improvement (EI) and Latin Hypercube Sampling (LHS) for offline autotuning of a predictive controller in an Artificial Pancreas setting. Our results show a significant improvement in glycemic control. Specifically, the average Time In Range (TIR) increased from 46.13% ± 7.07 (manual tuning) to 82.81% ± 12.03 using LHS-based BO-GP-EI. To address safety concerns, we constrained the maximum insulin input daily
patient-specific basal dose, reducing Time Below Range (TBR) from 5.46%±6.76 to 0%, while maintaining a TIR of 69.45%±17.29. These findings demonstrate the effectiveness of BO-based autotuning in optimizing control parameters and enforcing safety constraints.

4 - Novel Gradient-informed Acquisition Functions for Efficient Bayesian Optimization

Lea Prade Njoua Dongmo, University of Houston, Houston, TX, United States, Marzia Cescon

This work explores the use of novel acquisition functions in Expensive Black-Box Optimization(EBBO) problems under strict data limitations. While standard acquisition functions such as, e.g., Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB)are widely used in this setting, they often under perform in highly nonlinear or sharply featured response landscapes. To address their limitations, we introduce two novel acquisition functions, namely, Best Value Gradient Exploitation with smart exploration (BVGEE) and its enhanced variant BVGEE+, that combine empirical gradient estimation with adaptive sampling. Benchmarking the proposed acquisition functions on four synthetic examples shows that BVGEE+ consistently delivers low-variance performance with minimal computational overhead, making it a strong candidate for EBBO in scenarios characterized by limited data availability and complex dynamics.

5 - Optimizing Large language Models: A Roadmap for Sustainable and Ethical AI

Meetu Malhotra, Harrisburg University of Science and Technology, Harrisburg, PA, United States

In an era where technology and sustainability must coexist, large language models (LLMs) have emerged as transformative tools across various industries, from healthcare to finance, education to entertainment. This presentation delves into the multifaceted landscape of LLMs, beginning with an introduction to their capabilities and diverse applications. As we explore the innovative use cases that highlight the profound impact of LLMs, we also confront the pressing challenges they pose, particularly regarding energy consumption, carbon emissions, and the inherent biases present in these models. Addressing bias is crucial for ensuring fairness and inclusivity in AI applications, while establishing robust guardrails is essential to mitigate risks associated with misuse. With the increasing deployment of LLMs, understanding their environmental footprint and ethical implications becomes necessary. Through this presentation, I aim to foster a deeper awareness of these issues, including the critical need to address bias in LLMs, and outline a roadmap for optimizing their performance and efficiency, minimizing their environmental impact, and maximizing their potential while ensuring responsible usage

SunC03

Plaza I

New Advances of Stochastic Nonlinear Optimization in Machine Learning II

Invited Session

Nonlinear Optimization

Chair: Sen Na, Georgia Institute of Technology, Atlanta, United States

Co-Chair: Miaolan Xie, Purdue University, West Lafayette, IN, United States

Co-Chair: Zhaosong Lu, University of Minnesota, Minneapolis, MN, United States

1 - Variance-reduced First-order Methods for Stochastic and Finite-sum Optimization

Zhaosong Lu, University of Minnesota, Minneapolis, MN, United States, Sanyou Mei, Yifeng Xiao

We consider stochastic and finite-sum optimization problems with deterministic constraints. Existing methods typically focus on finding an approximate stochastic solution that ensures the expected constraint violations and optimality conditions meet a prescribed accuracy. However, such an approximate solution can possibly lead to significant constraint violations in practice. To address this issue, we propose variance-reduced first-order methods that treat the objective and constraints differently. Under suitable assumptions, our methods can find stronger approximate stochastic solutions with complexity guarantees, offering more reliable constraint satisfaction than existing approaches. This is joint work with Sanyou Mei (HKUST) and Yifeng Xiao (UMN).

2 - Stochastic Adaptive Optimization with Unreliable Inputs: A Unified Framework for High-Probability Complexity Analysis

Miaolan Xie, Purdue University, West Lafayette, IN, United States, Katya Scheinberg

We consider an unconstrained continuous optimization problem where, in each iteration, gradient estimates may be arbitrarily corrupted with a probability greater than 1/2. Additionally, function value estimates may exhibit heavy-tailed noise throughout the algorithm's execution. This framework accommodates challenging scenarios where both gradient and function value estimates can have heavy-tailed noise, making it applicable to many real-world problems with outliers and data anomalies, particularly relevant to stochastic and derivative-free optimization. We introduce an algorithmic and analytical framework that provides high-probability bounds on iteration complexity for this setting where the gradient and function value estimates are unreliable. The analysis offers a unified approach, encompassing methods such as line search and trust region.

3 - Accelerated Stochastic First-order Method for Convex Optimization under Heavy-tailed Noise

Chuan He, Linkoping University, Linkoping, Sweden, Zhaosong Lu

We study convex composite optimization problems where the objective is the sum of a prox-friendly function and a convex function whose subgradients are estimated under heavy-tailed noise. We first show that an accelerated vanilla stochastic proximal subgradient method---without clipping or normalization---achieves universally optimal first-order oracle complexity for smooth, weakly smooth, and nonsmooth convex problems, as well as for stochastic convex optimization under heavy-tailed noise.

4 - Understanding and Improving Shampoo and SOAP via Kullback-Leibler Minimization

Wu Lin, Vector Institute, Toronto, ON, Canada, Scott Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, Roger Grosse

Shampoo and its efficient variant, SOAP, employ structured second-moment estimations and have shown strong performance for training neural networks (NNs). In practice, however, Shampoo typically requires step-size grafting with Adam to be competitive, and SOAP mitigates this by applying Adam in Shampoo's eigenbasis -- at the cost of additional memory overhead from Adam in both methods. Prior analyses have largely relied on the Frobenius norm to motivate these estimation schemes. We instead recast their estimation procedures as covariance estimation under Kullback-Leibler (KL) divergence minimization, revealing a previously overlooked theoretical limitation and motivating principled redesigns. Building on this perspective, we develop and , practical schemes that match or exceed the performance of Shampoo and SOAP in NN pre-training while achieving SOAP-level per-iteration runtime. Notably, KL-Shampoo does not rely on Adam to attain competitive performance, eliminating the memory overhead introduced by Adam. Across our experiments, KL-Shampoo consistently outperforms SOAP, Shampoo, and even KL-SOAP, establishing the KL-based approach as a compelling foundation for designing structured methods in NN optimization.

5 - Stochastic Galerkin Method and Hierarchical Preconditioning for PDE-constrained Optimization

Akwum Onwunta, Lehigh University, Bethlehem, PA, United States

e develop efficient hierarchical preconditioners for optimal control problems governed by partial differential equations with uncertain coefficients. Adopting a discretize-then-optimize framework that integrates finite element discretization, stochastic Galerkin approximation, and advanced time-discretization schemes, the approach addresses the challenge of large-scale, ill-conditioned linear systems arising in uncertainty quantification. By exploiting the sparsity inherent in generalized polynomial chaos expansions, we derive hierarchical preconditioners based on truncated stochastic expansion that strike an effective balance between computational cost and preconditioning quality. Numerical experiments demonstrate that the proposed preconditioners significantly accelerate the convergence of iterative solvers compared to existing methods, providing robust and efficient solvers for both steady-state
and time-dependent optimal control applications under uncertainty.

SunC04

Plaza II

Structured Nonconvex Constrained Optimization: Methods and Applications

Invited Session

Nonlinear Optimization

Chair: Linglingzhi Zhu, Georgia Institute of Technology

1 - A Smoothing Moving Balls Approximation Method

Ting-Kei Pong, Hong Kong Polytechnic University, Hong Kong, Hong Kong

We consider the problem of minimizing a difference-of-convex objective over a nonlinear conic constraint. We assume that the cone is closed, convex and pointed with a nonempty interior, and the support function of a compact base of its dual cone admits a so-called majorizing smoothing approximation. These conditions are satisfied by widely studied cones such as the nonnegative orthants and the cone of positive semidefinite matrices. We reformulate the conic constraint equivalently as a single inequality constraint involving the aforementioned support function, and adapt the moving balls approximation (MBA) method for its solution. In essence, in each iteration of our algorithm, we approximate the support function by a smooth approximation function and apply one MBA step. The subproblems that arise in our algorithm always involve only one single inequality constraint. We design explicit rules to evolve the smooth approximation functions from iteration to iteration and establish the global iteration complexity for obtaining an approximate Karush-Kuhn-Tucker point. In addition, in the convex setting, we establish convergence of the sequence generated, and study its local convergence rate under a standard Hölderian growth condition.

This is a joint work with Nung-sing Sze and Jiefeng Xu.

2 - First Order Algorithms for Constrained Bilevel Optimization Problems

Jiawei Zhang, University of Wisconsin–Madison, Madison, WI, United States

In this talk, I will discuss how to use first-order algorithms to find an (ϵ, δ)-Goldstein stationary point of a constrained bilevel optimization problem. We will rely purely on first-order updates, and the assumptions we adopt are all practical or mild.

3 - Primal-dual Methods for Nonsmooth Nonconvex Optimization with Orthogonality Constraints

Linglingzhi Zhu, Georgia Institute of Technology, Atlanta, GA, United States

Recent advancements in data science have significantly elevated the importance of orthogonally constrained optimization problems. The Riemannian approach has become a popular technique for addressing these problems due to the advantageous computational and analytical properties of the Stiefel manifold. Nonetheless, the interplay of nonsmoothness alongside orthogonality constraints introduces substantial challenges to current Riemannian methods, including scalability, complicated subproblems, and cumulative numerical errors that threaten feasibility. In this paper, we take a retraction-free primal-dual approach and propose a linearized smoothing augmented Lagrangian method specifically designed for nonsmooth and nonconvex optimization with orthogonality constraints. Our proposed method is single-loop and free of subproblem solving. We establish its iteration complexity of $\mathcal{O}(\epsilon^{-3})$ for finding ϵ-KKT points, matching the best-known results in the Riemannian optimization literature. Additionally, by invoking the standard Kurdyka-\L ojasiewicz (K\L) property, we demonstrate asymptotic sequential convergence of the proposed algorithm. The key challenge lies in balancing the primal-dual updates to ensure feasibility, due to the nonconvex nature of the orthogonality constraints. This reflects a broader difficulty in applying primal-dual methods to nonconvex constrained problems and our method represents the first use of classical dual updates for solving orthogonality constrained problems with convergence guarantees, even in the smooth setting. Empirically, we demonstrate the superior computational efficiency and scalability of our proposed method on both smooth and nonsmooth orthogonal constrained problems compared with state-of-the-art algorithms.

4 - The Latest Developments in the Knitro Optimization Solver

Youssouf Emine, Artelys, Montreal, QC, Canada

Knitro was originally developed in the 1990s as an interior-point algorithm for general nonlinear, non-convex optimization. Over the years, Knitro evolved into a more general optimization toolbox that includes an Active-Set LP-based solver, a Sequential Quadratic Programming (SQP) solver, specialized LP, QP, and SOCP solvers, a branch-and-bound solver for mixed integer programming, and multi-start heuristics for global optimization. To add to this toolbox of algorithms, we have recently started developing first-order methods that do not require any matrix factorizations. The hope is that these might be able to provide useful solutions to extremely large-scale models where the factorizations in interior-point methods become too expensive. In this talk we will present some of this work, primarily based on Augmented Lagrangian (AL) type methods.

5 - ***Cancelled*Bi-level Optimization Method for Non-smooth Non-convex Lower-level Problem via Uniform Smoothing**

Dung Vo, Wayne State University, Detroit, MI, United States, Dat Tran, Prashant Khanduri

Bi-level optimization (BLO) refers to an optimization framework with two levels of hierarchy: an upper-level and a lower-level problem. In recent years, BLO has emerged as a fundamental technique for addressing hierarchical structures in modern machine learning (ML) problems. Most existing algorithms for BLO rely on specific assumptions about the lower-level objective, such as (strong) convexity, Polyak--Łojasiewicz (PL) conditions, and/or smoothness, to ensure tractable analysis and convergence guarantees. In this work, we consider a broader class of BLO problems, where the lower-level objective satisfies the Kurdyka--Łojasiewicz (KL) condition and, in general, may be non-convex and non-smooth. First, we develop tractable (smoothing-based) reformulations of the original BLO problem and establish equivalence between the developed reformulations. Then, we utilize regularization-based techniques to solve the reformulated problem and develop both gradient-free and first-order algorithms for solving the general BLO problems. Importantly, we establish finite-time convergence guarantees for the developed algorithms. Finally, we corroborate the theoretical results with experiments on modern ML problems.

SunC05

Plaza III

Stochastic Optimization for Reinforcement Learning

Invited Session

Optimization under Uncertainty

Chair: Caleb Ju, Georgia Institute of Technology, 755 Ferst Dr NW, Atlanta, GA, 30332, United States

1 - Exploration from a Primal-Dual Optimization Lens in Reinforcement Learning

Bo Dai, Georgia Tech, Atlanta, GA, United States, Tong Yang, Yuejie Chi, Lin Xiao

Online reinforcement learning (RL) with complex function approximations such as transformers and deep neural networks plays a significant role in the modern practice of artificial intelligence. Despite its popularity and importance, balancing the fundamental trade-off between exploration and exploitation remains a long-standing challenge; in particular, we are still in lack of efficient and practical schemes that are backed by theoretical performance guarantees. Motivated by recent developments in exploration via optimistic regularization, this paper provides an interpretation of the principle of optimism through the lens of primal-dual optimization. From this fresh perspective, we set forth a new value-incentivized actor-critic (VAC) method, which optimizes a single easy-to-optimize objective integrating exploration and exploitation -- it promotes state-action and policy estimates that are both consistent with collected data transitions and result in higher value functions. Theoretically, the proposed VAC method has near-optimal regret guarantees under linear Markov decision processes (MDPs) in both finite-horizon and infinite-horizon settings, which can be extended to the general function approximation setting under appropriate assumptions.

2 - The Lagrangian Method for Safe Learning in Dynamic Games

Soham Das, The University of Tennessee, Knoxville, TN, United States, Santiago Paternain, Luiz Chamon, Ceyhun Eksin

We introduce the concept of a *Lagrangian game* as a general framework for solving constrained Markov games (CMGs). CMGs model multiagent systems operating in dynamic environments under global or local feasibility constraints. These constraints encode safety, resource, and regulatory requirements and serve as the foundation for safe multiagent reinforcement learning (safe MARL). In our approach, agents repeatedly construct and solve an unconstrained Lagrangian game based on their current multipliers, simulate finite state-action trajectories to estimate cumulative rewards and constraint violations, and update the multipliers using a gradient step. This iterative process generates a sequence of Lagrangian games and corresponding policy profiles. We show that the resulting sequence of policies forms a nonstationary, approximately Nash solution for the original CMG and achieves constraint feasibility with controlled accuracy. Our analysis covers both general-sum and cooperative CMGs, including constrained multiagent MDPs, where the method yields nonstationary solutions that outperform any stationary policy up to a quantifiable error. We further extend the framework to locally constrained Markov games, where agents have individual private constraints and update multipliers using local information, while still converging to feasible, nonstationary approximate Nash equilibria.

3 - Value Mirror Descent for Reinforcement Learning

Zhichao Jia, Georgia Institute of Technology, Atlanta, GA, United States, Guanghui Lan

Value-based methods have been extensively studied for computing an ε-optimal policy of a discounted Markov decision process (DMDP). In particular, given a DMDP with state space S, action space A, discount factor $\gamma \in (0,1)$ and rewards in [0,1], the algorithm developed in ``Near-Optimal Time and Sample Complexities for Solving Discounted Markov Decision Process with a Generative Model'' by Sidford et al. finds an ε-optimal policy with high probability using $O(|S||A|(1-\gamma)^{-3}\varepsilon^{-2})$ samples, which nearly matches the lower complexity bound. In this talk, we introduce a novel value mirror descent method. Our proposed method also achieves nearly optimal sample complexity for solving a DMDP with high probability. Moreover, in the presence of a strongly convex regularizer, our method attains an ε-optimal policy for a regularized DMDP with high probability using $O(|S||A|(1-\gamma)^{-5}\varepsilon^{-1})$ samples. To the best of our knowledge, this is the first value-based method achieving an $O(\varepsilon^{-1})$ dependence in the strongly convex regularized setting.

4 - Auto-explorative Model-free Reinforcement Learning

Caleb Ju, Georgia Institute of Technology, Atlanta, GA, United States, Guanghui Lan

The exploration-exploitation dilemma in reinforcement learning is a fundamental challenge to the design of efficient algorithms. Existing policy gradient methods address this by either positing or explicitly injecting exploration, which can yield sub-optimal performance and result in non-implementable methods. This work introduces a new class of auto-explorative methods. These are methods that automatically explore both state and action spaces without any exploration strategies while adapting to the true, unknown problem parameters like the mixing time of the induced Markov chain. The resulting parameter-free methods do not estimate these unknown parameters and instead utilize a novel and easy-to-compute certificate. For finite state and action Markov decision processes, we show these methods achieve the same worst-case complexity as when the parameters are known. Moreover, we extend these auto-explorative methods to the case where linear function approximators are used and demonstrate a similar complexity. Such results can potentially be applied to general state and action spaces as well.

SunC06

Director's and Lounge

Advances in Global Optimization: Algorithms and Applications

Contributed Session

Global Optimization

Chair: Rohit Kannan, Massachusetts Institute of Technology, Cambridge, MA, United States

Co-Chair: Harsha Nagarajan, Los Alamos National Laboratory, Los Alamos, NM, United States

1 - Surrogate-Based Global Flowsheet Optimization Using Unit-Level Global Optimal Designs

Sai Tarun Ganapavarapu, Georgia Institute of Technology, Atlanta, GA, United States, Joseph Scott

Physics-embedded chemical process flowsheets exhibit inherent nonconvexities, trapping local solvers and rendering global optimization computationally intractable. Traditional surrogates built using simulation data are often difficult to optimize and fail to ensure material and energy balance closure.

We propose an alternative framework, first solving the cost optimization problem for individual unit operations, generating optimal designs across diverse feed conditions and product targets. A surrogate model is then constructed to map these inputs (feed conditions, product targets) directly to the optimal cost. We compare surrogate types (e.g., polynomial regression, linear decision trees, neural networks), analyzing their accuracy versus the dimensionality of the resulting surrogate-embedded optimization. Furthermore, multi-unit flowsheets contain various connectivity constraints to direct material and energy flows. We exploit dependent connectivity constraints to reduce the degrees of freedom (DOF). A new surrogate using DOF as inputs predicts interconnecting stream properties. These predictions feed the unit-level cost surrogates, directly linking flowsheet connectivity DOF to total cost. This approach significantly reduces the optimization problem's dimensionality, enabling the solution of larger-scale flowsheets.

This problem-agnostic hierarchical decomposition methodology provides a powerful framework for a broad class of nonconvex network optimization problems, such as water distribution networks, gas networks, and supply chain problems.

2 - A Novel Branch-and-Bound Algorithm to Solve the Best Subset Selection Problem for High-dimensional Data

Vikram Singh, Texas A&M- Texarkana, Texarkana, TX, United States, Min Sun

Over the last decade, there has been growing interest among researchers in studying sparse methods for high-dimensional data. At the heart of this task lies the Best Subset Selection (BSS) in linear regression, a classical problem in statistics that exemplifies the computational difficulty of solving sparsity-constrained optimization problems. The BSS problem is known to be NP-hard, and finding the global optimum becomes increasingly difficult as the problem dimension grows. Branch-and-bound (BB) methods provide a general framework for solving BSS. However, in the underdetermined case the bound-based pruning is less likely to occur early in the search tree of any BB method, making it more computationally extensive. We propose a novel BB method that employs a specialized enumeration to generate candidate solutions and guarantee convergence to the global optimum. As with any BB method, its efficiency critically depends on how memory usage is controlled and how strong candidate solutions are identified during the search. Our method leverages the echelon form of each parent node to efficiently compute lower bounds for a sequence of child nodes in lower dimensions, making it particularly effective in the underdetermined case. We compare our method with an existing BB method and with GUROBI's quadratic mixed-integer solver.

3 - Specialized Spatial Branch-and-Bound for Global Optimization of Gaussian Process Posteriors

Wei-Ting Tang, University of Wisconsin-Madison, Madison, WI, United States, Akshay Kudva, Yilin Xie, Calvin Tsay, Joel Paulson

Gaussian process (GP) models are widely used nonparametric surrogates across science and engineering due to their flexibility, ability to represent uncertainty in a calibrated way, and rigorous treatment of noisy observations. Many surrogate-based optimization methods embed GP posterior mean and variance terms, leading to challenging global problems, especially when minimizing the popular lower confidence bound (LCB) acquisition function (which systematically balances exploration and exploration). Spatial branch-and-bound (BnB) methods offer a principled framework for solving such problems deterministically, but their performance hinges on constructing relaxations that are both tight and computationally tractable. Existing approaches, such as those based on McCormick relaxations, often yield loose bounds for GP posteriors and scale poorly with the number of training points. We introduce a specialized BnB algorithm that combines closed-form analytic bounds with piecewise-linear relaxations for selected kernel contributions, resulting in mixed-integer (quadratic) relaxations that can be solved efficiently using modern MIP solvers. Our method applies to a broad class of stationary kernels and accommodates both the O(N) posterior mean and $O(N^2)$ variance terms. Benchmark results in 4-10 dimensions demonstrate significant reductions in CPU time and BnB tree size relative to BARON, SCIP, and MAiNGO.

4 - Adaptive Levenberg-Marquardt Third-Order Newton's Method

Yubo Cai, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, United States, Gioele Zardini

This paper proposes the Adaptive Levenberg-Marquardt Third-Order Newton's Method (ALMTON) for unconstrained nonconvex optimization. While higher-order methods theoretically offer superior convergence rates, existing third-order approaches face a significant trade-off between practical utility and global convergence. Standard third-order adaptive regularization (AR3) employs fourth-order regularization to ensure global convergence, but this yields computationally challenging quartic subproblems. In the contrast, while the unregularized third-order Taylor model (a cubic polynomial) is efficiently solvable via semidefinite programming (SDP) [Ahmadi and Zhang, 2022], the corresponding Newton method is only locally convergent.

ALMTON bridges this gap, becoming the first globally convergent variant of the unregularized third-order Newton's method. It achieves global convergence by integrating unregularized steps with an adaptive Levenberg-Marquardt (LM) regularization framework. Critically, ALMTON uses a (second-order) LM term instead of fourth-order regularization, preserving the cubic polynomial structure of the subproblem. This allows ALMTON to leverage a unified SDP template for both regularized and unregularized steps. The algorithm strategically prioritizes efficient, unregularized steps, using the LM-regularized model only as needed to ensure globalization.

We establish a worst-case complexity bound of $O(\epsilon^{-2})$ for finding an ϵ-approximate stationary point. We benchmark ALMTON against state-of-the-art ARC and AR3 variants in [Cartis et al., 2025] on standard test problems. Empirical results illustrate that ALMTON can comprehensively outperform these algorithms in terms of objective evaluations, computational efficiency, and the number of subproblem solves, demonstrating its potential as an efficient, globally convergent high-order optimization method.

5 - Hexaly, Hybrid Optimization Solver

Fred Gardi, Hexaly, Brooklyn, NY, United States

Hexaly is an innovative hybrid optimization solver based on a nonlinear, set-based modeling formalism. This formalism unifies and extends modeling paradigms from mixed-integer linear programming, nonlinear programming, constraint programming, and black-box optimization. Internally, Hexaly integrates a broad range of exact and heuristic optimization techniques, including branch-and-bound, column and cut generation, constraint propagation, and local search, among others. From a computational standpoint, Hexaly demonstrates competitive performance with state-of-the-art solvers such as Gurobi, CPLEX, and OR-Tools, delivering efficient and scalable solutions to routing, scheduling, packing, clustering, and location problems. This presentation introduces the so-called set-based modeling formalism and demonstrates its scalability for solving large-scale problems. It further examines how the solver exploits this formalism to automatically orchestrate advanced exact and heuristic solution methods drawn from the current state of the art.

SunC07

Congress Room

Novel Applications of Integer Programming

Invited Session

Discrete Optimization

Chair: Diego Moran, Rensselaer Polytechnic Institute, Troy, NY, United States

1 - Optimization Approaches for Designing Experiments

Zulal Isler Ardic, Rensselaer Polytechnic Institute, Troy, NY, United States, Diego Moran, Burak Kocuk

Designing experiments is essential for drawing reliable conclusions in research. A key challenge lies in dividing participants into groups that are balanced and representative of the population. Traditional methods, such as randomization and re-randomization, are widely used to reduce bias, but they do not guarantee optimal group balance, particularly in studies with small samples or heterogeneous, multi-attribute populations. Optimization-based methods have emerged as a promising alternative, enabling researchers to directly target balance criteria across several covariates simultaneously. In this talk, we present new multi-objective optimization models that integrate diverse discrepancy metrics, including moment differences, Wasserstein distance, variation distance, and a new bin-based metric. Through computational experiments, we evaluate the strengths and limitations of each method in achieving group similarity and empirically show that our multi-objective approaches demonstrate promising performance.

2 - New Methodologies for Solving the Close Enough TSP

Xiangnan Zhang, Rensselaer Polytechnic Institute, Troy, NY, United States, James Bailey, Bahar Cavdar, Diego Moran

The Close-Enough Traveling Salesman Problem (CETSP) generalizes the classical Traveling Salesman Problem by relaxing the requirement of visiting each customer at an exact location. Instead, the salesman may visit any point within a specified neighborhood of each customer, typically modeled as a disk. In this talk, we present new approaches for solving the CETSP based on a two-stage formulation: in the first-stage, representative points are selected within each neighborhood, and in the second-stage, a classical TSP is solved on these points. We propose algorithms that leverage the structure of this formulation in different ways, including the use of Minimum Spanning Trees to approximate the second-stage TSP. Finally, we present computational experiments on benchmark datasets to evaluate the performance of our methods.

3 - Robust Admission via Two-Stage Stable Matching under Ranking Uncertainty

Gustavo Angulo, Pontificia Universidad Católica de Chile, Santiago, Chile, Matías Giddings, Pablo Marshall

The Bachillerato Inicia UC program offers a pathway for students from Chilean technical high schools to articulate into undergraduate programs at Pontificia Universidad Católica de Chile. Upon applying, candidates rank up to three preferred programs. However, articulation is determined only after one year, based on their academic ranking within the cohort - a value unknown at the time of admission. The challenge is to select a feasible subset of students to admit, respecting the order of applicant scores, while guaranteeing that, under any realization of final rankings, each admitted student can be matched to one of their declared preferences, respecting program capacities. We formalize this as a two-stage stable matching problem under ranking uncertainty and design an algorithm that characterizes the set of robustly admissible candidates. The model ensures articulation guarantees that are consistent with declared preferences and capacity constraints. We present implementation results from the 2024 and 2025 admission cycles.

4 - On the Convexification of a Class of Mixed-Integer Conic Sets

Rui Chen, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, People's Republic of, Guxin Du, Linchuan Wei

We investigate mixed-integer second-order conic (SOC) sets with a nonlinear right-hand side in the SOC constraint, a structure frequently arising in mixed-integer quadratically constrained programming (MIQCP). Under mild assumptions, we show that the convex hull can be exactly described by replacing the right-hand side with its concave envelope. This characterization enables strong relaxations for MIQCPs via reformulations and cutting planes. Computational experiments on distributionally robust chance-constrained knapsack variants demonstrate the efficacy of our reformulation techniques.

5 - Randomized Augmentation Algorithms for Total Variation-Regularized Integer Programs

Dominic Yang, Argonne National Laboratory, Lemont, IL, United States, Sven Leyffer, Miles Bakenhus

We address a class of integer programs on graphs with total variation regularization. Total variation regularizers are common in problems such as image denoising and topology optimization for their ability to smooth solutions while still preserving sharp edges. Our methods are centered around augmentation algorithms which maintain a set of moves which repeatedly "augment" a given feasible point until it cannot be improved any further using such moves. In the context of total variation-regularized problems, we show the moves which update the solution on a connected subgraph are a Graver basis, meaning they act as an optimality certificate, and the augmentation algorithm produces a globally optimal solution in certain cases. Unfortunately, this set of moves is so large as to be computationally intractable to search completely, so we substitute an exhaustive search with a randomized approach which samples connected subgraphs. We demonstrate how to randomize the simplex algorithm applied to an associated linear program to produce random connected subgraphs and how to bias the sampler to produce moves which are more likely to improve the current iterate's objective. We then apply our method to some problems from topology optimization and show it is competitive with the standard approaches.

SunC08

Committee Room

Risk-Aware Methods for Decision-Making Under Uncertainty

Invited Session

Emerging Applications of Optimization

Chair: William Yang, Lawrence Livermore National Lab, Livermore, CA, United States

Co-Chair: Shixiang Woody Zhu, Carnegie Mellon University, Pittsburgh, PA, United States

1 - A Risk-Averse Stochastic Optimization Approach for Reliably Operating Energy Storage Technology Supply Chains

Michael Grappone, Lawrence Livermore National Laboratory, Livermore, CA, United States, William Yang, Elizabeth Glista, Emily Tucker, Mengyao Yuan, Jean-Paul Watson

There exist unique challenges to reliably managing supply chains electric vehicle (EV) batteries. They require assembly from multiple raw materials sourced from across the world. Furthermore, raw materials are often processed at locations that are different than the final production stage, which adds more complexity. There is a high degree of uncertainty in the price for these technologies because of the ever-changing landscape of the energy sector. Moreover, the uncertainty from market and geopolitical factors impacts the ability to ship materials from one location to another and is exacerbated by the highly interconnected nature of battery supply chains. To address these challenges, we develop a risk-averse stochastic optimization methodology to provide secure operation of the energy storage technology supply chain. Our model considers not only the raw material flow, but also the intermediate processing and the final production stage. Our model also allows for each stage to take place in different locations. We additionally incorporate stochastic components such as uncertainties caused by pricing. Finally, our model optimizes the Conditional Value at Risk metric, which allows us to produce results that are more reliable compared to traditional stochastic optimization methods. We present experimental results that demonstrate our methodology's efficiency and its ability to reliably provide cost-effective solutions. A major consequence of our methodology is that it can allow EV producers to securely obtain necessary technologies.

2 - Multi-Stage Adaptive Robust Optimization for Long-term Grid Resilience Planning

Shixiang Woody Zhu, Carnegie Mellon University, Pittsburgh, PA, United States

Long-term grid resilience planning requires making a sequence of interdependent investments under deep uncertainty in future hazards, climate trajectories, and system evolution. Classical robust and stochastic planning frameworks either assume a fixed uncertainty set or rely on scenario enumeration that cannot capture evolving risk profiles over multi-decadal horizons. This paper presents a multi-stage adaptive robust optimization (MS-ARO) framework for long-term grid resilience planning that dynamically reallocates investments as new information unfolds. The proposed formulation couples (i) a stage-wise uncertainty model that captures the progressive revelation of extreme-weather risk, infrastructure aging, and load growth; (ii) adaptive decision policies that allow mid-course corrections while respecting a global budget; and (iii) a tractable decomposition algorithm that exploits the nested structure of the uncertainty sets. Computational experiments, using large-scale outage and infrastructure data from U.S. utilities, demonstrate that MS-ARO substantially outperforms static and two-stage robust models by enabling earlier identification of high-risk regions, avoiding myopic over-hardening, and yielding more reliable long-term investment portfolios. Our results highlight the importance of adaptive, risk-aware optimization for grid resilience in the face of compounding climate and operational uncertainties.

3 - Adaptive Two-stage Expansion Planning Problems

Valerie Thomas, ISyE Georgia Tech, Atlanta, GA, United States, Yuang Chen, Beste Basciftci

We introduce an adaptive two-stage stochastic optimization model for energy infrastructure expansion planning under uncertain demand. Unlike traditional two-stage models that are too rigid or multi-stage models that are overly flexible, this new model balances commitment and flexibility by allowing investment decisions to be revised at specific adaptation times. The model's performance is evaluated using two metrics and compared to conventional approaches. To improve its practicality, this paper also presents five strategies for sharing adaptation times among related investments, which further reduces the number of policy revisions. A case study on Rwanda's electrification plan demonstrates that this approach can lead to cost savings of up to 3.44% compared to the two-stage model and significantly fewer adaptations compared to the multi-stage model. The model results also provide valuable policy implications, such as the optimal adaptation frequency and timing, as well as the ideal planning horizon length.

4 - An Interval-based Stochastic Dominance Approach for Pruning Uncertain Solutions in Uncertain Multi-objective Optimization Problems

Chathuri Aththanayake, University of Tennessee at Chattanooga, Chattanooga, TN, United States, Lakmali Weerasena

Many real-world problems in engineering, management, and agriculture involve multiple conflicting objectives, along with uncertainty in input data and model outcomes. While Pareto optimality is fundamental for identifying efficient solutions in Multi-objective Optimization Problems (MOPs), it does not capture the uncertainty and often results in a large solution set, complicating decision-making. In Uncertain Multi-objective Optimization Problems (UMOPs), the solution pool becomes even larger due to the need to represent uncertainty, making the selection of the most appropriate solution even more challenging. To address this issue, we propose a newly defined index for interval-based uncertain solutions and apply it within a constrained similarity-based clustering model to group solutions with similar characteristics. Additionally, we utilize the proposed index to introduce a new stochastic dominance rule, which identifies the Stochastic Pareto Set. Furthermore, we provide a computational analysis of the proposed index to evaluate its effectiveness in pruning solutions with different density levels among interval-based stochastic non-dominated solutions. This framework provides decision-makers with a more tractable and informative solution set for UMOPs, enhancing decision support in uncertain multi-objective environments.

5 - A General Framework for Finite-Sample Evaluation of CVaR Optimization

Kimiya Jozani, Virginia Tech, Blacksburg, VA, United States, Sajad Khodadadian, Esra Toy

Certifying the quality of solutions to risk-averse stochastic optimization problems remains a central challenge in data-driven decision making. This difficulty is especially pronounced in Conditional Value-at-Risk (CVaR) models, where the tail-focused nature of the objective makes standard asymptotic SAA (Sample Average Approximation) guidance unreliable for determining when a computed solution is sufficiently accurate. To address this limitation, we introduce a comprehensive finite-sample certification framework that provides explicit, high-confidence optimality guarantees for CVaR-based SAA solutions. The framework generates rigorous upper bounds on the optimality gap that can be incorporated into existing SAA workflows without altering model structure. To corroborate the effectiveness of this approach, we apply the method to risk-aware epidemic control problems, where mitigation policies must be selected under significant structural uncertainty. Our finite-sample certificates enable decision makers to reduce sampling effort while maintaining conservative performance guarantees, which is critical in time-sensitive public health contexts where rapid, reliable assessment of risk-averse policies is needed. Numerical studies demonstrate that the approach reduces sampling effort while maintaining conservative guarantees, thereby enabling reliable risk-averse decision support in time-sensitive epidemic settings.

SunC09

Cabinet Room

Advances in Methodologies for Transportation and Supply Chain Problems

Invited Session

Emerging Applications of Optimization

Chair: Weiqing(Lynn) Xu, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Haoyun Deng, Georgia Institute of Technology, Atlanta, GA, United States

1 - Bilevel Risk-Averse Routing and Resource Allocation for Urban Air Mobility

Haoyun Deng, Georgia Institute of Technology, Atlanta, GA, United States, Weijun Xie, Peng Wei

Urban air mobility (UAM) provides a promising solution for alleviating ground traffic congestion and achieving zero direct aviation emissions. The successful and reliable implementation of UAM requires careful coordination between regulatory authorities, such as the Federal Aviation Administration (FAA), and airline operators, particularly in uncertain conditions of passenger demand and airspace capacity. In this paper, we introduce a bilevel risk-averse optimization framework that integrates airline routing decisions with FAA capacity management and delay assignment at critical airspace bottlenecks. To address the computational complexity of this integrated routing and resource allocation problem, we propose and evaluate both heuristic and exact optimization methods. Numerical experiments demonstrate the effectiveness of the proposed approaches and provide valuable insights for stakeholders seeking efficient, equitable, and robust UAM operations in uncertain environments.

2 - A Two-Layer Dynamic Discretization Discovery Algorithm for Time-Spaced Network Design with Operational Constraints

Weiqing(Lynn) Xu, Georgia Institute of Technology, Atlanta, GA, United States, Alan Erera, Oktay Gunluk

Solving large-scale service network design problems is a central challenge in transportation logistics. While Dynamic Discretization Discovery (DDD) has proven effective when using time-expanded models, existing methods struggle with complex operational constraints, such as maximum wait times at terminals and First-In-First-Out (FIFO) loading models. We introduce a new DDD methodology based on a two-layer relaxation graph that explicitly separates arrival and departure events into distinct layers. This novel structure allows us to systematically identify and repair three distinct types of temporal infeasibilities—short travel time arcs, long waiting arcs, and long dispatch arcs—that arise due to the sparse time discretization. Our approach iteratively refines this sparse time-expanded network, guaranteeing convergence to a feasible operational solution without necessarily constructing the full model. The framework is applied within a two-stage planning process, translating frequency-based weekly delivery plans into detailed, executable schedules. Initial numerical results show our method is effective, demonstrating its potential to solve previously intractable, high-fidelity consolidation scheduling problems.

3 - Submodularity and Supermodularity in Min-Cost Flow Interdiction

Ruiwei Jiang, University of Michigan, Ann Arbor, MI, United States, Bo Zhou, Siqian Shen

We consider the min-cost flow (MCF) interdiction problem, which is a bilevel program with an attacker interdicting (attacking) certain components of a network and a defender sending a MCF in the ensuing network. To solve the bilevel program effectively, we investigate when the defender's minimum cost is submodular or supermodular with respect to the attacker's interdiction (attacking) decisions. In numerical experiments, we demonstrate an order-of-magnitude speed up brought by exploiting these properties to generate valid inequalities for solving MCF interdiction.

4 - Hospital-at-Home: Unifying Patient Selection and Distributed Supply-Chain Optimization under Uncertainty

Peiyu Wang, Carnegie Mellon University, Pittsburgh, PA, United States

Hospital-at-Home (HaH) reframes care as a distributed healthcare system. It eases patients from the pain of traveling to hospitals while saving millions of dollar for hospitals. HaH operations hinge on two coupled decisions: which patients to enroll and how to coordinate last-mile resources (nurses, depots, supplies). Today, for most of the hospitals, HaH enrollment still largely depends on clinical decisions by empirical judgement from the docter. The logistics of HaH are not systematically optimized. In this background, we propose a framework that jointly makes both decisions and explicitly accounts for uncertain demands, to improve the efficiency and safety for HaH programs.

This talk presents a concise blueprint that treats uncertainty in patient demands as a first-class design variable and turns it into robust operational choices in both patient selection and the downstream supply chain. We combine distributionally robust optimization (DRO) and submodular optimization to capture cohort-level risk pooling and tradeoff in logistics and supply cost. We will outline the end-to-end flow, from patient features to uncertainty-aware enrollment and depot assignment, and show how the DRO + submodular pairing yields robust policies that scales to realistic constraints (coverage, geography, staffing), which significantly improves the distributed health supply chains. The goal is a practical path from models to deployment: fewer assumptions, clearer trade-offs, and decisions that remain reliable for HaH patients.

5 - Dantzig-Wolfe and Arc-Flow Reformulations: A Systematic Comparison

Daniel Yamin, Carnegie Mellon University, Pittsburgh, PA, United States, Willem-Jan van Hoeve, Ted Ralphs

Dantzig-Wolfe and Arc-Flow reformulations are two of the most widely used approaches for solving large-

scale integer programs. Although they are closely connected, they are typically studied in isolation. Arc-Flow formulations, in particular, can be derived from dynamic programming models, decision diagrams, or other problem-specific constructions, which makes a generic comparison with Dantzig-Wolfe formulations nontrivial. Building on a unified formulation and notation, this study clarifies the theoretical connections and computational trade-offs between the two reformulations. We establish conditions under which valid inequalities in the original, Dantzig-Wolfe, or Arc-Flow spaces can be translated across reformulations without loss of strength, and reinterpret iterative strengthening methods, such as decremental state-space relaxation and column elimination, through the lens of cutting planes. To empirically assess these insights, we benchmark both reformulations under identical conditions on the vehicle routing problem with time windows using state-of-the-art column- and cut-generation techniques. The results identify clear contrasts: the Arc-Flow reformulation benefits from path recombination and performs best when subproblems are highly relaxed or low-dimensional, whereas the Dantzig-Wolfe reformulation is more efficient when the master problem remains compact. Overall, our study provides a unified perspective and practical guidelines for choosing between Dantzig-Wolfe and Arc-Flow reformulations in large-scale integer programming.

SunC10

Caucus Room

Learning for Optimization, Optimization for Learning, and Experimentation

Invited Session

Computational Optimization

Chair: Peter Zhang, Carnegie Mellon University, Pittsburgh, PA, United States

1 - A Deep Learning Approach to Multistage Stochastic Programming

Yizhe Huang, University of Texas at Austin, Austin, TX, United States

Multistage stochastic programming is a powerful framework for sequential decision-making under uncertainty with complex constraints, but solving large-scale instances remains challenging due to the curse of horizon. We address this issue using deep learning. Unlike most existing value-based methods that use neural networks to approximate the value function, we adopt a policy-based approach that parameterizes the policy with neural networks and handles constraints via a Lagrangian formulation. We establish convergence guarantees for the proposed method, drawing on tools from Wasserstein gradient flows.

2 - Adaptive Neyman Allocation

Jinglong Zhao, Boston University, Boston, MA, United States

In the experimental design literature, Neyman allocation refers to the practice of allocating units into treated and control groups, potentially in unequal numbers proportional to their respective standard deviations, with the objective of minimizing the variance of the treatment effect estimator. This widely recognized approach increases statistical power in scenarios where the treated and control groups have different standard deviations, as is often the case in social experiments, clinical trials, marketing research, and online A/B testing. However, Neyman allocation cannot be implemented unless the standard deviations are known in advance. Fortunately, the multi-stage nature of the aforementioned applications allows the use of earlier stage observations to estimate the standard deviations, which further guide allocation decisions in later stages. In this paper, we introduce a competitive analysis framework to study this multi-stage experimental design problem. We propose a simple adaptive Neyman allocation algorithm, which almost matches the information-theoretic limit of conducting experiments. We provide theory for estimation and inference using data collected from our adaptive Neyman allocation algorithm. We demonstrate the effectiveness of our adaptive Neyman allocation algorithm using both online A/B testing data from a social media site and synthetic data.

3 - Robust Paths: Geometry and Computation

Hao Hao, Carnegie Mellon University, Pittsburgh, PA, United States, Peter Zhang

Applying robust optimization often requires selecting an appropriate uncertainty set both in shape and size, a choice that directly affects the trade-off between average-case and worst-case performances. In practice, this calibration is usually done via trial-and-error: solving the robust optimization problem many times with different uncertainty set shapes and sizes, and examining their performance trade-off. This process is computationally expensive and ad hoc. In this work, we take a principled approach to study this issue for robust optimization problems with linear objective functions, convex feasible regions, and convex uncertainty sets. We introduce and study the robust path: a set of robust solutions obtained by varying the uncertainty set's parameters. Our central geometric insight is that a robust path can be characterized as a Bregman projection of a curve (whose geometry is defined by the uncertainty set) onto the feasible region. This leads to a surprising discovery that the robust path can be approximated via the trajectories of standard optimization algorithms, such as the proximal point method, of the deterministic counterpart problem. We give a sharp approximation error bound and show it depends on the geometry of the feasible region and the uncertainty set. We also illustrate two special cases where the approximation error is zero: the feasible region is polyhedrally monotone (e.g., a simplex feasible region under an ellipsoidal uncertainty set), or the feasible region and the uncertainty set follow a dual relationship. We demonstrate the practical impact of this approach in two settings: portfolio optimization and adversarial deep learning.

4 - Incentivizing Desirable Effort in Strategic Classification: An Optimization Approach

Diptangshu Sen, Georgia Institute of Technology, Atlanta, GA, United States, Vasiliki Efthymiou, Chara Podimata, Juba Ziani

Machine learning tools like classifiers are now ubiquitous in decision-making for many high-stakes applications like loan approvals, granting bail, admissions and hiring, etc. When agents receive unfavorable outcomes from the classifier, it naturally creates incentives for them to modify their features strategically to improve future outcomes. Such modifications may either be genuine improvement or gaming or a combination of both. However, this strategic behavior is problematic: the classifier is now deployed on data that does not match the training-time distribution, and this can significantly hurt its accuracy. The emerging area of ``**strategic classification**" studies this problem, with the goal of designing classifiers that are robust to distribution shifts induced by strategic behavior. However, an equally important yet under-explored goal is to ensure that the design of the classifier also incentivizes agents to perform *desirable* modifications in the first place. Optimization theory provides natural tools to model, formalize, and tackle such design problems. In this work, we borrow heavily from these tools to explore the complexity of designing "good" classifiers---accurate classifiers that also incentivize desirable agent behavior---, under varying levels of modification power and classifier information available to the agents. Our work demonstrates the enormous power of optimization in addressing emerging challenges in other disciplines.

5 - Hierarchical Policy Optimization for Sequential Decision Making

Jiangweizhi Peng, University of Minnesota, Minneapolis, MN, United States, Mingyi Hong

For long-horizon sequential decision-making problems, standard policy-gradient methods typically operate at a single temporal scale, which leads to high variance and weak credit assignment, especially when returns are driven by delayed consequences of earlier choices. We study policy optimization for such problems through the lens of the options framework, a hierarchical architecture that introduces temporally extended decisions on top of primitive actions. Building on this structure, we derive coupled gradient estimators that decompose returns into high- and low-level advantages, where the high-level advantage captures coarse progress over extended time intervals and the low-level advantage refines primitive decisions toward the goal. We analyze the resulting gradient estimators from an optimization perspective, highlighting how temporal abstraction improves the conditioning of the effective objective and reduces estimator variance. Across environments, we observe more stable training dynamics and improved sample efficiency compared to flat actor–critic baselines with standard generalized advantage estimation. Overall, the results indicate that our proposed gradient estimators, instantiated under the hierarchical options framework, provide a principled and effective way to exploit temporal abstraction in long-horizon settings and suggest a path toward more scalable training schemes for complex decision-making systems.

SunC11

Charter Room

Network Optimization

Contributed Session

Network Optimization

Chair: Mert Gurbuzbalaban, Rutgers University, Piscataway, NJ, United States

1 - Optimal Transmission Switching with De-energization using Benders Decomposition

Benoit Jeanson, TU Delft, Delft, Netherlands, Mathieu Tanneau, Simon Tindemans

This paper proposes a Benders decomposition framework for the Optimal Transmission Switching with De-energization (OTSD) problem, which ensures secure power grid operations by preventively disconnecting power lines. A key obstacle for solving OTSD at scale is the representation of potential loss of network connectivity ---and the resulting localized blackouts--- after individual line outages. State-of-the-art MILP solvers struggle to tackle the problem's scale and underlying combinatorial explosion, which has motivated the development of expert heuristics, albeit without optimality guarantees.

To overcome these limitations, we propose a scalable Benders decomposition framework for OTSD, combined with several performance improvements. At each iteration, a security analysis identifies the set of sub-problems that yield violated cuts, thus reducing the computational burden of generating Benders cuts. In addition, we improve the granularity of feasibility cuts by introducing additional variables in the master problem. Finally, using graph-theoretic arguments, we present a closed-form formula for separating optimality cuts without solving a linear problem.

The proposed approach is evaluated on test systems representative of medium-sized sub-transmission grids encountered in real-world operations. Numerical experiments demonstrate that, when combined with a local search mechanism, the proposed method improves best-known solutions by up to 30% within minutes of runtime, significantly outperforming standard MILP solvers. These results highlight the potential of decomposition-based techniques to address large-scale OTSD instances and advance decision-making for secure power system operations.

2 - A Lagrangian Relaxation Matheuristic for the Selective One-to-One Pickup-and-Delivery

Eli Olinick, Southern Methodist University, Dallas, TX, United States, Yuanyuan Dong

We address the Selective One-to-One Pickup-and-Delivery Problem (SOPDP). The single-vehicle SOPDP, which is NP-hard, requires simultaneously solving two problems: (1) determining how to route an empty delivery vehicle back from its current location to its depot by a scheduled arrival time, and (2) selecting a profit-maximizing subset of spot-market delivery requests along the route subject to the vehicle’s capacity. For finding provably optimal solutions to the multi-vehicle SOPDP, we propose a compact mixed integer programming (MIP) formulation and simple, but effective, preprocessing routines. We also propose a Lagrangian Relaxation Matheuristic (LRM) that takes advantage of multithreading to find high quality solutions within practical time limits. Results from computational experiments on a set of 300 problem instances demonstrate the effectiveness of the LRM. For example, the minimum, median, mean, and maximum running times for Gurobi to solve the MIP for the largest instances are 5 seconds, 56 minutes, 5 hours, and 10 hours respectively. The corresponding times for the LRM are 1, 5, 8, and 30 minutes, respectively; and the minimum, median, and maximum gaps are 0%, 2.98%, and 4.82%, respectively.

3 - Learning Dynamics in Network Games under Partial Visibility

Sepideh Mohammadi, University of Tennessee, Knoxville, TN, United States, Jim Ostrowski, Soham Das

We study a variant of stochastic fictitious play under incomplete information (partial visibility). We hypothesize that, for any network game, this variant of fictitious play converges to an approximate Nash equilibrium for some bounds of observation level. We show that, in the graph coloring game on cycles, stochastic fictitious play consistently reaches approximate equilibrium under partial visibility, while also demonstrating that such convergence does not hold in full visibility settings. Thus, full information may be suboptimal for learning agents in some network games to reach equilibrium.

4 - Black-Box Bayesian Optimization for Traffic Microsimulation Calibration

Abhilasha Saroj, Oak Ridge National Laboratory, Knoxville, TN, United States, Shaked Regev, Guanhao Xu, Jinghui Yuan, Roy Luo, Ross Wang

Real-Twin is an open-source platform for semi-automated, consistent traffic microsimulation scenario generation. This paper presents its new automated calibration capability based on a refined Black-Box Bayesian Optimization (BBBO) framework. We apply BBBO to calibrate traffic volumes and driving behavior parameters of a traffic simulation model in SUMO, a microscopic traffic simulation software. We then compare its performance against state-of-the-art black-box optimization algorithms commonly used for traffic simulation calibration, Genetic Algorithm, Tabu Search, and Simulated Annealing. A key contribution is the implementation of an efficient, lightweight BBBO-based calibration module within Real-Twin that is explicitly designed for constrained computational budgets. The explainability of BBBO is leveraged for feature engineering, including dimensionality reduction, to focus the search on the most influential parameters and reduce calibration costs. We evaluate all methods in terms of effectiveness, efficiency, and robustness across multiple trials. BBBO consistently and quickly meets the calibration targets across repetitions, while some other methods can stagnate and occasionally fail. This demonstrates BBBO's superior performance and makes it a strong candidate for scalable, real-world digital twin calibration workflows.

March 22, 2026, 3:30 PM - 5:00 PM

SunD01

Grand Ballroom

Recent Advances in Optimization for Data Science

Contributed Session

Optimization in Data Science

Chair: Mingxi Zhu, Georgia Institute of Technology, Decatur, GA, United States

1 - Recent Advances in Stochastic Optimization under Heavy-Tailed Noise

Zijian Liu, New York University, New York City, NY, United States

Recently, the study of heavy-tailed noise in first-order stochastic optimization has garnered significant attention, as it has been recognized as a more realistic condition, supported by numerous empirical observations. Specifically, the stochastic noise (difference between the stochastic and true gradient) is considered to have only a finite p-th moment where $1<p\leq 2$, instead of assuming it always satisfies the classical finite variance assumption. To address this more challenging setting, people have proposed various algorithms and demonstrated that they converge at optimal rates. In this talk, I will discuss some of my recent

advances in this topic. 1. For nonconvex optimization (whether smooth or not), I will show that the widely adopted mechanism of gradient clipping is unnecessary by providing new convergence rates for algorithms that do not contain clipping. Moreover, I also establish matched lower bounds demonstrating the optimality of the analysis. 2. For (strongly) convex optimization (w/wo smoothness), I will show that, under the bounded domain assumption, the gradient clipping is also unnecessary and classical algorithms (e.g., projected SGD) are sufficient to guarantee optimal convergence. The proof technique provably generalizes to harder online convex optimization beyond stochastic convex optimization. Additionally, I will talk about how to ensure convergence without requiring any problem-dependent parameters under heavy-tailed noise (e.g., the tail index p). 3. Finally, I will discuss the shortcomings of the existing heavy-tailed assumption and propose a new one to characterize the noise behavior. Under this new condition, I will provide new upper and lower bounds faster than the existing ones.

2 - Memory/Query Complexity Tradeoffs for High-accuracy Feasibility Problems

Jialin Yu, Georgia Institute of Technology, Atlanta, GA, United States, Moise Blanchard

We investigate tradeoffs between convergence rates of optimization algorithms and their memory usage for finding a feasible point within a ball of size \epsilon>0 with access to a separation oracle in dimension d. This feasibility problem is a foundational problem in optimization and naturally generalizes first-order convex optimization. Prior works identified three main convergence rate regimes for this problem. In low-accuracy regimes, memory does not play a critical role since gradient descent is already optimal; in moderate-accuracy regimes, prior works showed exponential lower bound tradeoffs between memory and query complexity; but in high-accuracy regimes much less is known. In particular, there are exponential gaps between (1) known memory-constrained algorithms which achieve (d\log(1/\epsilon))^d query complexity and (2) the classical query lower bound d\log(1/\epsilon) by Nemirovski and Yudin. Further, it is unknown whether memory plays any substantial role for solving feasibility problems in this regime. In this work, we show the first memory/query complexity lower bound tradeoff for high-accuracy regimes by proving that any algorithm with less than quadratic memory must have at least d (\log(1/\epsilon))^2 oracle complexity.

3 - Turning Data Heterogeneity into an Advantage through Primal–Dual Methods for Machine Learning

Mingxi Zhu, Georgia Institute of Technology, Decatur, GA, United States, Yinyu Ye

Data heterogeneity is a fundamental challenge in modern machine learning. Many gradient-based training algorithms are known to suffer from slow convergence when agents are heterogeneous. This paper is the first to show that, for machine learning problems with separable loss functions and multiple local agents each holding heterogeneous data, this negative effect can turn into an advantage for primal–dual optimization methods. Specifically, while primal-only optimization algorithms (e.g., multi-block PCG, distributed Newton) suffer from slow convergence under heterogeneity, primal–dual algorithms (e.g., EXTRA, consensus multi-block ADMM) can in fact benefit from data heterogeneity. By enforcing consensus constraints and leveraging dual updates, these methods extract additional “momentum” from heterogeneity and accelerate convergence. We further extend our analysis to federated learning, where data heterogeneity is typically harmful to primal-based methods such as FedAvg. In contrast, we demonstrate that heterogeneity can also accelerate primal–dual FL algorithms (e.g., Fed-ADMM). Our results suggest that the central challenge of heterogeneity in multi-agent learning is not to eliminate it, but to exploit it through the design of primal–dual methods, highlighting their potential to turn data heterogeneity from a liability into an advantage in distributed machine learning.

4 - Safe Screening Rules for Sparse Index-tracking Portfolio Optimization

Nanari Wada, University of Tsukuba, Tsukuba, Japan, Yuichi Takano

Sparse index tracking aims to replicate the performance of a benchmark index using a small subset of assets and is formulated as a mixed-integer quadratic optimization problem with a cardinality constraint for limiting the number of invested assets. This problem is NP-hard, making it computationally challenging to obtain a globally optimal solution to large-scale problems. Safe screening rules are known as an effective technique for accelerating sparse estimation by pre-eliminating irrelevant variables while ensuring global optimality. However, existing screening rules cannot be directly applied to budget-constrained portfolio optimization, where the sum of investment weights must be one. In this talk, we propose novel safe screening rules for index-tracking problems with both cardinality and budget constraints. We construct a convex perspective relaxation of the problem and derive screening scores by analyzing its Fenchel dual. Based on these scores, we establish theoretically guaranteed conditions for identifying assets that should be included in or excluded from the optimal portfolio. Computational results demonstrate that our screening rules can significantly speed up the solution of large-scale index-tracking problems without compromising global optimality.

5 - A CPU/GPU Implementation of Randomized Rounding for Symbolic Regression

Xingjian Tao, Georgia Institute of Technology, Atlanta, GA, United States, Nikolaos Sahinidis

Symbolic regression aims to uncover interpretable, closed-form expressions that map input variables to observed outputs. While state-of-the-art solvers—such as Operon and PySR—rely on genetic programming, recent work has approached symbolic regression as a mixed-integer nonlinear program (MINLP) and introduced STAR, a randomized rounding algorithm to obtain high quality solutions of the MINLP. In this work, we develop a novel GPU-accelerated implementation of STAR. We describe the algorithms utilized, including candidate generation, redundancy elimination, and constant optimization. Extensive experiments on SRBench demonstrate that our approach achieves higher symbolic solution rates than most top-performing SRBench methods.

SunD02

Georgian Room

Recent Progress on Theoretical Foundations of LLM Optimizers

Invited Session

Computational Optimization

Chair: Yiping Lu, Northwestern University, Evanston, IL, United States

1 - On the Width Scaling of Neural Optimizers Under Matrix Operator Norms I: Making Operators Play Nice Together

2 - On Neural Optimizers Under Matrix Operator Norms II: Convergence Rates, Dimension Dependence, and Batch-Size Effects

3 - Structured preconditioners in adaptive optimization: A unified analysis

SunD03

Plaza I

Optimization Techniques for Large language Models Training II

Invited Session

Nonlinear Optimization

Chair: Jiaxiang Li, Meta, Santa Clara, CA, United States

Co-Chair: Mingyi Hong, Mingyi Hong, WAYZATA, United States

1 - MuonBP: Faster Muon via Block-Periodic Orthogonalization

Kaan Ozkara, Amazon, Santa Clara, CA, United States, Mingyi Hong

Gradient orthogonalization is a simple strategy that shows great utility in speeding up gradient descent. The Muon optimizer (Jordan, Jin, et al., 2024) combines gradient orthogonalization with first-order momentum and achieves significant improvement in data efficiency over Adam/AdamW (Loshchilov and Hutter, 2019) for language model training. However, when using model parallelism, gradient orthogonalization introduces additional overhead compared to coordinate-wise optimizers (such as AdamW) due to additional gather and scatter operations on gradient matrix shards from different devices. This additional communication can amount to a throughput hit of 5%-10% compared to Adam/AdamW. To remedy this, we propose Muon with Block-Periodic Orthogonalization (MuonBP), which applies orthogonalization independently to matrix shards on each device and periodically performs full orthogonalization to maintain training stability at scale. We show how to adjust the learning rate from the baseline to MuonBP and give convergence guarantees for this algorithm. Crucially, our theory dictates that we use two stepsizes: one for the blockwise orthogonalization steps, and one for the full orthogonalization steps. Our method is simple, requires minimal hyperparameter adjustments, and achieves competitive iteration complexity compared with baseline Muon while providing per-iteration throughput comparable to coordinate-wise methods such as AdamW. When training an 8B model with eight-way tensor parallelism and ZeRO optimizer state sharding, MuonBP achieves 8% throughput increase compared to Muon with no degradation in performance.

2 - Hybrid Preconditioning for Training of Large Language Models

Tuo Zhao, ISyE Georgia Tech, Atlanta, GA, United States

Training Large Language Models (LLMs) typically relies on monolithic preconditioning strategies, where a single update rule—whether based on curvature approximation or spectral normalization—is applied uniformly across all parameters. In this talk, we demonstrate that a **Hybrid Preconditioning** approach, which applies distinct conditioners to different components of the parameter space, offers a more effective path to convergence.

We illustrate this concept through two recent algorithms: **Cosmos** and **NoMuon**.

First, we discuss **Cosmos** (arXiv:2502.17410), which operates via an orthogonal decoupling of the update direction. Instead of treating the entire space equally, Cosmos applies **SOAP** (a refined second-order preconditioner) specifically to the **Principal Subspace** to capture dominant curvature information, while utilizing **Muon** (a coarse second-order preconditioner) for the **Complementary Subspace**. This combination naturally balances aggressive curvature exploitation with spectral regularization.

Second, we introduce **NoMuon** (arXiv:2510.05491), which extends this hybrid philosophy to tensor structures. It applies **Muon** in the matrix space, while simultaneously utilizing **neuron-wise adaptive scaling**.

We argue that these combinations are complementary, and leverage the strengths of different conditioners where they are most needed. Our empirical analysis confirms that this hybrid strategy consistently outperforms methods relying on a single type of preconditioning, achieving superior efficiency in training LLMs.

3 - PiKE: Adaptive Data Mixing for Large-Scale Multi-Task Learning Under Low Gradient Conflicts.

Meisam Razaviyayn, University of Southern California, Los Angeles, CA, United States, Zeman Li, Yuan Deng, Pelin Zhong, Vahab Mirrokni

Modern foundation models are trained on diverse datasets to enhance generalization across tasks and domains. A central challenge in this process is determining how to effectively mix and sample data from multiple sources. This naturally leads to a multi-task learning (MTL) perspective. While prior work in MTL has emphasized mitigating gradient conflicts, we observe that large-scale pretraining scenarios—such as multilingual or multi-domain training—often exhibit little to no gradient conflict. Motivated by this observation, we propose PiKE (Positive gradient interaction-based K-task weights Estimator), an adaptive data mixing algorithm that dynamically adjusts sampling weights during training. PiKE exploits non-conflicting gradient interactions to minimize a near-tight upper bound on the average loss decrease at each step, while incurring negligible computational overhead. We provide theoretical convergence guarantees and show that PiKE outperforms static and non-adaptive mixing baselines. Furthermore, we extend PiKE to promote balanced learning across tasks. Extensive experiments on large-scale language model pretraining confirm that PiKE achieves faster convergence and improved downstream performance compared to existing approaches.

4 - Lions and Muons: Optimization via Stochastic Frank-Wolfe

Maria-Eleni Sfyraki, University of California San Diego, San Diego, CA, United States, Jun-Kun Wang

Stochastic Frank-Wolfe is a classical optimization method for solving constrained optimization problems. On the other hand, recent optimizers such as Lion and Muon have gained quite significant popularity in deep learning. In this work, we provide a unifying perspective by interpreting these seemingly disparate methods through the lens of Stochastic Frank-Wolfe. Specifically, we show that Lion and Muon with weight decay can be viewed as special instances of a Stochastic Frank-Wolfe, and we establish their convergence guarantees in terms of the Frank-Wolfe gap, a standard stationarity measure in non-convex optimization for Frank-Wolfe methods. We further find that convergence to this gap implies convergence to a KKT point of the original problem under a norm constraint for Lion and Muon. Moreover, motivated by recent empirical findings that stochastic gradients in modern machine learning tasks often exhibit heavy-tailed distributions, we extend Stochastic Frank-Wolfe to settings with heavy-tailed noise by developing two robust variants with strong theoretical guarantees, which in turn yields new variants of Lion and Muon.

SunD04

Plaza II

Recent Advances in Nonlinear Stochastic Optimization and Related Applications

Invited Session

Nonlinear Optimization

Chair: Baoyu Zhou, Arizona State University, Tempe, AZ, United States

Co-Chair: Haoming Shen, University of Arkansas

1 - On the Convergence and Complexity in Nonsmooth Stochastic Optimization with Sample Average Approximation

Jingyi Wang, LLNL, Livermore, CA, United States

This talk consists of three new developments in nonsmooth stochastic optimization. First, we will present our latest results on the stochastic estimate error of Clake subgradient using sample average approximation (SAA). In particular, we consider Lipschitz continuous and regular functions. Next, we design a novel line search process for a sequential quadratic programming algorithm that ensures convergence for weakly concave, hence nonsmooth, objective functions. Finally, using the SAA estimate, we derive iteration and subgradient evaluation complexity for constrained problems with weakly concave objective functions.

2 - Sequential Quadratic Optimization for Solving Expectation Equality Constrained Stochastic Optimization Problems

Haoming Shen, University of Arkansas, Fayetteville, AR, United States, Yang Zeng, Baoyu Zhou

A sequential quadratic programming method is designed for solving general smooth nonlinear stochastic optimization problems subject to expectation equality constraints. We consider the setting where the objective and constraint function values, as well as their derivatives, are not directly available. The algorithm applies an adaptive step size policy and only relies on objective gradient estimates, constraint function estimates, and constraint derivative estimates to update iterates. Both asymptotic and non-asymptotic convergence properties of the algorithm are analyzed. Under reasonable assumptions, the algorithm generates a sequence of iterates whose first-order stationary measure diminishes in expectation. In addition, we identify the iteration and sample complexity for obtaining a first-order ε-stationary iterate in expectation. The results of numerical experiments demonstrate the efficiency and efficacy of our proposed algorithm compared to a penalty method and an augmented Lagrangian method.

3 - Sequential Quadratic Programming for Physics-Informed Neural Networks

Kexin Li, University of Arkansas, Fayetteville, AR, United States

Physics-informed neural networks (PINNs) provide a powerful framework for solving inverse and forward problems governed by physical laws, but their training remains challenging due to severe ill-conditioning and the presence of complex constraints. In this work, we investigate a stochastic sequential quadratic programming (SQP) framework for efficiently solving constrained PINN optimization problems. The proposed approach aims to exploit the structure of the physical constraints through iterative quadratic subproblems, providing improved convergence and stability compared with standard gradient-based training. Beyond synthetic PDE benchmarks, we also explore the use of real-world data—such as California wildfire aerosol optical depth (AOD) measurements—to assess the method's practical effectiveness in data- and physics-driven modeling.

4 - A Nonmonotone Line Search Method for Stochastic Optimization

Yunfan Zeng, Arizona State University, Tempe, AZ, United States

We propose a nonmonotone line search method for solving general smooth, unconstrained, stochastic optimization problems. Unlike traditional stochastic line searches that require monotonic decrease of the noisy function values, our method allows temporary increases within a nonmonotone acceptance window, thereby avoiding overly conservative step-size reductions. Under mild assumptions on the stochastic oracles, our proposed algorithm achieves global convergence as well as advanced numerical performance.

SunD05

Plaza III

New Developments in MDPs and POMDPs

Invited Session

Optimization under Uncertainty

Chair: Hao Zhang, University of British Columbia, Vancouver, BC, Canada

1 - A Faster Exact Algorithm for POMDPs via Backward Induction

Hao Zhang, University of British Columbia, Vancouver, BC, Canada, Yining Wang

The state-of-the-art method for solving POMDPs exactly is "incremental pruning" and its variations ("Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes" by Cassandra, Littman & Zhang, UAI 1997). In this paper, we introduce a new algorithm that delivers an order-of-magnitude speedup for dense POMDPs (those that produce exponentially many value hyperplanes or extreme points even with small state, action, and observation sets). Our approach builds on the dual POMDP framework developed in "Partially Observable Markov Decision Processes: A Geometric Technique and Analysis" (Zhang, 2010, *Operations Research*).

2 - Landscape of Policy Optimization for Finite Horizon MDPs with General State and Action

Xin Chen, Georgia Institute of Technology, Atlanta, GA, United States, Yifan Hu, Minda Zhao

Policy gradient methods are widely used in reinforcement learning. Yet, the nonconvexity of policy optimization imposes significant challenges in understanding the global convergence of policy gradient methods. For a class of finite-horizon Markov Decision Processes (MDPs) with general state and action spaces, we develop a framework that provides a set of easily verifiable assumptions to ensure the Polyak-Łojasiewicz-Kurdyka (PŁK) condition of the policy optimization. Leveraging the PŁK condition, policy gradient methods converge to the globally optimal policy with a non-asymptomatic rate despite nonconvexity. Our results find applications in various control and operations models, including entropy-regularized tabular MDPs, Linear Quadratic Regulator (LQR) problems, stochastic inventory models, and stochastic cash balance problems, for which we show an ϵ-optimal policy can be obtained using a sample size \tilde{O}(\epsilon^{-1}) and polynomial in terms of the planning horizon by stochastic policy gradient methods. Our result establishes the first sample complexity for multi-period inventory systems with Markov-modulated demands and stochastic cash balance problems in the literature.

3 - Reuse Historical Trajectories in Natural Policy Gradient via Importance Sampling

Yuhao Wang, Georgia Institute of Technology, Atlanta, GA, United States, Enlu Zhou

Reinforcement learning provides a mathematical framework for learning-based control, whose success largely depends on the amount of data it can utilize. The efficient utilization of historical trajectories obtained from previous policies is essential for expediting policy optimization. Empirical evidence has shown that policy gradient methods based on importance sampling work well. However, existing literature often neglect the interdependence between trajectories from different iterations, and the good empirical performance lacks a rigorous theoretical justification. In this paper, we study a variant of the natural policy gradient method with reusing historical trajectories via importance sampling. We show that the bias of the proposed estimator of the gradient is asymptotically negligible, the resultant algorithm is convergent, and reusing past trajectories helps improve the convergence rate. We further apply the proposed estimator to popular policy optimization algorithms such as trust region policy optimization. Our theoretical results are verified on classical benchmarks.

4 - Statistical Tractability of Off-policy Evaluation of History-dependent Policies in POMDPs

Yuheng Zhang, University of Illinois Urbana-Champaign, Champaign, IL, United States, Nan Jiang

We investigate off-policy evaluation (OPE), a central and fundamental problem in reinforcement learning (RL), in the challenging setting of Partially Observable Markov Decision Processes (POMDPs) with large observation spaces. Recent works of Uehara et al. (2023a); Zhang & Jiang (2024) developed a model-free framework and identified important coverage assumptions (called belief and outcome coverage) that enable accurate OPE of memoryless policies with polynomial sample complexities, but handling more general target policies that depend on the entire observable history remained an open problem. In this work, we prove information-theoretic hardness for model-free OPE of history-dependent policies in several settings, characterized by additional assumptions imposed on the behavior policy (memoryless vs. history-dependent) and/or the state-revealing property of the POMDP (single-step vs. multi-step revealing). We further show that some hardness can be circumvented by a natural model-based algorithm — whose analysis has surprisingly eluded the literature despite the algorithm's simplicity — demonstrating provable separation between model-free and model-based OPE in POMDPs.

SunD06

Director’s and Lounge

Innovations in Optimization Under Uncertainty and Complex Systems

Contributed Session

Optimization under Uncertainty

Chair: Francesco Lugani, N/A

1 - Continuous-support Wasserstein Set Through the Lens of Gaussian Mixture Distributions and Distributionally Robust Chance-constrained Programming

Shibshankar Dey, IEMS, Northwestern University, Evanston, IL, United States, Sanjay Mehrotra

We study linear chance-constrained optimization problems under a continuous-support Wasserstein ambiguity set through the lens of Gaussian mixture model (GMM). We cast the Wasserstein distance over the support of Gaussian distribution parameters using the Bures–Wasserstein metric, rather than the classical Euclidean metric–based form. The resulting GMM-based distributionally robust (DR) chance-constrained program (GMM-DR-CCP) is an infinite-dimensional problem. We establish duality of the worst-case chance satisfaction constraint using the newly cast ambiguity set that allows us to present the infinite-dimensional GMM-DR-CCP equivalently as a finite-dimensional semi-infinite problem. A cutting-surface algorithm is used to solve the semi-infinite problem where a piecewise-linear (PWL) approximation is employed to approximate the standard Gaussian CDF \Phi. While the master problem leverages our earlier results to relate approximation accuracy with the master-problem optimal objective under compact decision space, analogous properties for the subproblem follow from the continuous differentiability of a fractional composite map of \Phi on a compact joint mean–covariance space. The cutting-surface algorithm is shown to attain any given optimality gap in finite iterations using a finite number of linear segments for PWL approximation. Numerical experiments with electric-vehicle charging station capacity planning and portfolio optimization demonstrate that a continuous-support DR model yields diversified and stable decisions without sacrificing objective values compared with its discrete-support counterpart. The computational result also addresses the limitations of discrete-support models based on the out-of-sample performance analysis, and their bias toward observed or sampled data. The EV application with a charging demand dataset further demonstrates the intrinsic robustness of GMM distribution representation.

2 - Data-driven Policies For Two-stage Stochastic Linear Programs

Harsha Gangammanavar, Southern Methodist University, Dallas, TX, United States, Chhavi Sharma

A stochastic program typically involves several parameters, including deterministic first-stage parameters and stochastic second-stage elements that serve as input data. These programs are re-solved whenever there is a change in any of the input parameters. However, in practical applications, quick decision-making is necessary, and solving a stochastic program from scratch for every change in input data can be computationally costly. This work addresses this challenge for two-stage stochastic linear programs (2-SLPs) with varying right-hand sides of the first-stage constraints. We construct a Piecewise Linear Difference-of-Convex (PLDC) policy by leveraging optimal bases from previous solves. This PLDC policy retains optimal solutions for previously encountered parameters and provides high-quality solutions for new right-hand side vectors whose bases are already discovered. Our proposed policy directly applies to the extensive form of the 2-SLP. When stage decomposition algorithms, such as the L-shaped and Stochastic Decomposition, are applied to solve the 2-SLPs, we develop the L-Shaped (and Stochastic Decomposition) guided sequential procedure, which iteratively tracks the policy quality and feeds new basis information to the policy. We assess the performance of our policy through analytical and numerical techniques. Our compelling experimental results show that a good percentage of test instances are feasible and optimal at the solutions returned by the policy.

3 - Solving Stochastic Programs With Decision-Dependent Uncertainty by Discretization

Francesco Lugani, University of Copenhagen, Copenhagen, Denmark, Giovanni Pantuso

In two-stage stochastic programming, it is typically assumed that the behavior of the uncertainty is independent of past decisions. However, in many real-world systems, current choices actively shape the probability of future outcomes. In this work, we relax this assumption by allowing the distribution of second-stage uncertain parameters to depend on first-stage decisions through a general distribution map. We propose an approximation scheme in which this distribution map is replaced by a piecewise-constant approximation over a partition of the first-stage feasible set. Under mild assumptions, we show that the original stochastic program arises as the limit of these approximations as the partition resolution increases. These convergence results establish the theoretical foundation for a practical algorithm that computes approximate solutions to decision-dependent stochastic programs within a prescribed error tolerance.

4 - Tessera: A Novel Zero-Data Digital Twin for Urban Climate Modeling

Ajay Vinjamuri, Rock Ridge High School, Ashburn, VA, United States

Cities worldwide face escalating climate risks, such as extreme heat and flooding, demanding innovative tools to proactively design resilient urban environments. Current urban digital twins often focus on single hazards or require extensive real-world data, limiting their use for proactive, multi-hazard scenario planning, especially in data-scarce regions. This project introduces Tessera, a novel digital twin framework that integrates Urban Heat Island (UHI) modeling, flood risk simulation, and synthetic city generation to create a virtual laboratory for climate policy planning. Tessera's core is a hybrid ConvGRU model that predicts spatiotemporal UHI effects by fusing dynamic climate data with static urban features. A key innovation is Tessera's use of procedurally generated synthetic cities, which enables robust climate impact analysis and the testing of design interventions where detailed city data is unavailable.

The framework's capability to rapidly simulate multi-hazard climate scenarios is validated by strong predictive performance, achieving a Root Mean Squared Error of 1.44°C for UHI prediction and an AUROC score of 0.73 for flood zone identification. This proven accuracy allows Tessera to serve as a scalable decision-support tool for urban planners and policymakers to stress-test future climate scenarios, representing a critical step towards the proactive, data-driven urban planning needed to navigate the complexities of climate change.

5 - ***Cancelled*A Dive-and-fix Approach for Large-scale Pooling Problems**

Ke-Ni Xiang, Beijing Institute of Technology, Beijing, China, People's Republic of, Zhi-Wei Wei, Xin Sun, Yan-Ru Wang, Yan-Ming Cao, Wei-Kun Chen, Yu-Hong Dai

The pooling problem is an NP-hard nonconvex bilinear programming problem that has many industrial applications, including petrochemical refining, wastewater treatment, natural gas transportation, open-pit mining, and animal feed problems. In this talk, we develop a dive-and-fix algorithm which starts with the all zeros trivial solution and then iteratively augments the current solution by involving a new terminal node until all terminal nodes are considered. Two key advantages of the proposed dive-and-fix algorithm that make it particularly suitable for solving large-scale pooling problems are as follows. First, it only requires solving linear programming subproblems, each of which considers a single terminal node. Second, it guarantees to return a feasible solution (usually with high solution quality). As a byproduct of analysis, we prove that checking whether the optimal value of a pooling problem is zero can be accomplished in polynomial time, which closes an open question in the literature. Finally, computational results demonstrate the effectiveness and efficiency of the proposed dive-and-fix algorithm over the direct use of an off-the-shelf global solver and a state-of-the-art greedy construction heuristic algorithm in the literature.

SunD07

Congress Room

Computation for Sequential and Mixed-Integer Nonlinear Optimization

Invited Session

Discrete Optimization

Chair: Yongzheng Dai, Georgia Institute of Technology, Atlanta, GA, United States

1 - Outer-Inner Cutting-Plane Approximation for AC Unit Commitment

Yongzheng Dai, The Ohio State University, Columbus, OH, United States, Antonio Conejo, Feng Qiu

AC Unit Commitment (ACUC) is a challenging mixed-integer nonlinear programming problem in power systems. Thought it can be convexified as a mixed-integer second-order-conic programing (MISOCP), such MISOCPs keep hard to scale by state-of-art solvers. To address the computational challenges in solving ACUC, we propose a problem-specific outer-inner cutting-plane approximation algorithm to solve the MISOCP exactly with a series of mixed-integer linear programs. We apply our algorithm to solve ACUCs, which is a foundmental but the most time-consuming subroutine in decomposition methods for robust grid optimizations. Our experiments demonstrate that our algorithm can produce solutions within a standard optimality gap in a reasonable time for moderately large instances.

2 - Learning to Optimize for Mixed-Integer Non-linear Programming with Feasibility Guarantees

Bo Tang, Massachusetts Institute of Technology, Boston, MA, United States, Elias B. Khalil, Jan Drgona

Mixed-integer nonlinear programs (MINLPs) arise in domains as diverse as energy systems and transportation, but are notoriously difficult to solve, particularly at scale. While learning-to-optimize (L2O) methods have been successful at continuous optimization, extending them to MINLPs is challenging due to integer constraints. To overcome this, we propose a novel L2O approach with two integer correction layers to ensure the integrality of the solution and a projection step to ensure the feasibility of the solution. We prove that the projection step converges, providing a theoretical guarantee for our method. Our experiments show that our methods efficiently solve MINLPs with up to tens of thousands of variables, providing high-quality solutions within milliseconds, even for problems where traditional solvers and heuristics fail. This is the first general L2O method for parametric MINLPs, finding solutions to some of the largest instances reported to date.

3 - ML-Guided Primal Heuristics for Mixed Binary Quadratic Programs

Weimin Huang, University of Southern California, Los Angeles, CA, United States, Natalie Isenberg, Jan Drgona, Draguna Vrabie, Bistra Dilkina

Mixed Binary Quadratic Programs (MBQPs) are an important and complex set of problems in combinatorial optimization. As solving large-scale combinatorial optimization problems is challenging, primal heuristics have been developed to quickly identify high-quality solutions within a short amount of time. Recently, a growing body of research has used machine learning to accelerate solution methods for challenging combinatorial optimization problems. Despite the increasing popularity of these ML-guided methods, a large body of work has focused on Mixed-Integer Linear Programs (MILPs). MBQPs are challenging to solve due to the combinatorial complexity coupled with nonlinearities. We present ML-guided primal heuristics for Mixed Binary Quadratic Programs (MBQPs) by adapting and extending existing work on ML-guided MILP solution prediction to MBQPs. We introduce a new neural network architecture for MBQP solution prediction and a new training data collection procedure. Moreover, we extend existing loss functions in solution prediction and propose to combine contrastive weighted cross-entropy losses. We evaluate the methods on standard and real-world MBQP benchmarks and show that the developed ML-guided methods significantly outperform existing primal heuristics and state-of-the-art solvers. Furthermore, models trained with our proposed loss function outperform other ML-based methods adapted from MILPs and improve generalization in cross-regional inference on a real-world wind farm layout optimization problem.

4 - Maximum Load Assortment Optimization: Approximation Algorithms and Adaptivity Gaps

Marouane Ibn Brahim, Cornell Tech / Cornell University, New York, NY, United States, Omar El Housni, Danny Segev

Motivated by modern-day applications such as attended home delivery and preference-based group scheduling, where decision makers wish to steer customers toward choosing the same alternative, we introduce a novel class of assortment optimization problems, referred to as *maximum load assortment optimization*. Given a universe of products and a stream of customers whose decisions are governed by the multinomial logit choice model, we define the random *load* of any product as the total number of customers who select it. Our objective is to offer an assortment of products to each customer so that the expected maximum load across all products is maximized. We consider both static and dynamic formulations. In the static setting, a single assortment is carried throughout the process of customer arrivals, whereas in the dynamic setting, the decision maker offers a personalized assortment to each customer, based on the entire information available at that time. Both formulations present a wide range of computational challenges and analytical questions. The main contribution of this paper resides in proposing efficient algorithmic approaches for computing near-optimal static and dynamic assortment policies. In particular, we develop a polynomial time approximation scheme for the static problem, and demonstrate that an elegant policy utilizing weight-ordered assortments yields a 1/2 approximation. Concurrently, we prove that such policies provide a 1/4 approximation of the dynamic formulation, establishing a constant factor bound on its adaptivity gap. Finally, we design an adaptive policy whose expected maximum load is within factor 1-eps of optimal, admitting a quasi-polynomial time implementation.

5 - Approximate Solution of Infinite-horizon Risk-Sensitive Markov Decision Processes

M. Soheil Hemmati, University of Oklahoma, Norman, OK, United States, Daihan Zhang, Saumya Sinha, Andrew Schaefer

Infinite-horizon risk-sensitive Markov decision processes (MDPs) with discounted costs are difficult to solve because optimal policies may be non-stationary. Standard approaches use state discretization or value function approximation to approximately solve the risk-sensitive problem using (risk-neutral) MDP techniques. For these methods, only limited performance guarantees vis-à-vis explicit stopping criteria or approximation error bounds exist. In this work, we develop variants of value iteration, linear programming, and policy iteration that operate directly on the risk-sensitive MDP. Our analysis establishes that the approximate value functions produced by value iteration and linear programming converge to the optimal risk-sensitive value function, and the resulting policies are provably near-optimal. We further show that the policies generated by our policy-iteration method converge in value to an optimal policy, despite the inherent non-stationarity of risk-sensitive optimal solutions. In addition to asymptotic guarantees, we derive explicit stopping conditions that ensure arbitrarily accurate approximations of the optimal value function.

SunD08

Committee Room

Novel Applications of Optimization Methods

Contributed Session

Emerging Applications of Optimization

Chair: Seung Jin Choi, Virginia Tech, Blacksburg, VA, 24060, United States

1 - Recursive Partitioning and Batching for Network Design with Service Time Guarantees at Massive Scale

Myungeun Eom, Georgia Institute of Technology, Atlanta, GA, United States, Alan Erera, Alejandro Toriello

Motivated by the parcel delivery industry, we study a network design problem with service time guarantees at industrial scale. This tactical service network design problem determines primary paths and delivery schedules for commodities to minimize transportation and handling costs while ensuring committed service times. To construct a solution for a real-world instance with over 1,000 nodes, one million arcs, and 40,000 commodities, we propose a recursive graph partitioning and batching method. This method partitions the network into smaller regions and solves the problem for each region, considering only commodities whose origin and destination are in the same region. For commodities crossing regions, we first determine an appropriate sub-network, then solve a restricted model on this sub-network. To handle the large number of commodities, we divide these into batches based on schedule slack (intuitively, how flexibly a commodity can be scheduled) and solve the problem sequentially over each batch. Finally, we reoptimize commodities in low-utilization trailers. We demonstrate the scalability and efficiency of our approach through computational studies on real-world instances from an industry partner. Our method finds high-quality solutions after several hours of computing time, while a commercial solver is unable to even build a model for a much smaller instance.

2 - Cooperative Integer Programming Games: Core Stability and Optimal Coalition Structures

Hyunwoo Lee, Virginia Tech, Blacksburg, VA, United States, Robert Hildebrand, Esra Buyuktahtakin Toy, Jian Yang

In cooperative game theory, a characteristic-function game is defined by a set of agents N and a characteristic function V that assigns to each coalition $C \subseteq N$ a real value $V(C)$, i.e., $V: 2^{|N|} \rightarrow \mathbb{R}$. Traditionally, $V(C)$ is specified by problem-specific formulas or rules.
We introduce a new concept—cooperative integer programming games (CIPGs)—in which (i) each player has an individual integer programming model; (ii) each coalition's value is obtained by pooling the constraint sets of its members to form a coalition-level integer program; and (iii) the characteristic function V maps every coalition C to the optimal value of the corresponding pooled or individual integer program.
Our primary goal is to identify an optimal coalition structure (OCS) and an OCS satisfying stability (OCSS).
We derive a \textit{stability inequality} that guarantees stability within each coalition—meaning each coalition is in the Core with respect to itself. Building on this inequality, we develop a cutting-plane algorithm to compute OCSS, in which the initial agent set is partitioned into coalitions that satisfy the stability condition. Computational experiments on cooperative knapsack games demonstrate the effectiveness of the proposed method in finding OCSS solutions.

3 - Safety-Constrained Reinforcement Learning for Naval Warfare Reconnaissance with an Intelligent Target

Seung Jin Choi, Virginia Tech, Blacksburg, VA, United States, Elson Cibaku, Anna Svirsko, Daphne Skipper, Esra Buyuktahtakin Toy

Modern naval warfare requires autonomous systems capable of persistent maritime reconnaissance and search under uncertainty, resource constraints, and adversarial behavior. We present a safety-constrained, heterogeneous multi-agent reinforcement learning (MARL) framework that jointly optimizes agents' tactical decisions and refueling logistics under partial observability. The method combines centralized training with decentralized execution and Proximal Policy Optimization, augmented by Lagrangian constrained policy optimization to ensure safety feasibility and mission continuity under strict fuel limitations.

A central feature is an intelligent, adaptive target whose motion induces evolving spatial uncertainty. To capture this, we introduce an Area-of-Uncertainty (AOU) field that updates dynamically with agents' sensing and target behavior. The AOU provides an information-theoretic reward for reducing uncertainty, enabling coordinated exploration and pursuit driven by information value rather than static search heuristics.

We evaluate the framework in a maritime reconnaissance scenario with stochastic, adaptive target motion, partial observability, and operational constraints. AOU-guided policies achieve high mission completion, robust constraint satisfaction, and efficient fuel and sensing usage. This work advances optimization-informed MARL for dynamic, adversarial naval environments and supports next-generation autonomous systems for contested maritime operations. This research was funded by the Office of Naval Research Mathematical and Resource Optimization program under Grants N000142412280 and N0001425GI01857.

4 - Who Gets to Ride? Comparing Transportation Eligibility Policies for Saanich Public Schools

Olivia Phillips, Georgia Tech, Atlanta, GA, United States, Arthur Delarue, Sebastien Martin

In public school transportation, eligibility policies are often changed in response to community feedback without a full understanding of the impact of these changes on operation costs. Saanich School District in British Columbia, Canada, services nearly 8000 students and currently faces such policy decisions without a clear way to compare costs and benefits of various policies.

This is not an issue unique to Saanich; almost every public school system provides transportation to its students. With limited funding, schools are forced to make difficult budgeting decisions every year, and transportation is a major portion of this annual budget. Each school must therefore choose an eligibility policy that correctly weighs the costs of increasing transportation availability with the benefits to students, their families, and the wider community.

In order to find a better policy, we have begun designing optimization models that maximize the utility to all parties involved while keeping costs below a given threshold. By adding nuance to this model and varying its parameters, we will be able to provide Saanich with a tool to more easily understand the impact on various policy decisions and ultimately decide on the policy that works best for their district.

5 - ***Cancelled***Optimizing Patient-Specific Intermittent Androgen Deprivation Therapy: A Mixed-Integer Nonlinear Programming Approach

Stefan Faulkner, Georgia Tech, Atlanta, GA, United States, David Goldsman, Edwin Romeijn

Prostate cancer represents the second leading cause of cancer mortality in American men, with androgen deprivation therapy (ADT) serving as standard treatment for biochemically recurrent disease. While intermittent ADT (IADT) can reduce cumulative treatment exposure and toxicity compared to continuous therapy, current protocols rely on fixed, heuristic rules derived from clinical trials. These one-size-fits-all approaches fail to account for patient-specific tumor dynamics and mechanisms of treatment resistance. We introduce a novel mixed-integer nonlinear programming (MINLP) framework to derive optimal, personalized IADT schedules that maximize time to progression.

We formulate the optimization problem using a biologically motivated model of prostate-specific antigen (PSA) dynamics that incorporates cancer stem-like cell self-renewal as a mechanism underlying resistance evolution. The model detects clinical progression through a multi-day consecutive PSA increase detection framework using binary indicator variables. We develop both a baseline simulation model replicating the Bruchovsky et al. clinical trial protocol and an optimization model treating key protocol parameters as decision variables rather than fixed values.

Using patient archetypes parameterized with literature-derived cancer stem cell self-renewal rates, we demonstrate the optimization framework's ability to derive patient-specific treatment schedules. Computational experiments suggest that personalized optimization of protocol parameters can extend time to progression compared to fixed clinical trial protocols. This work exemplifies optimization at the interface of clinical data, treatment decisions, and improved patient outcomes in oncology.

SunD09

Cabinet Room

Optimization for Generative Models

Invited Session

Emerging Applications of Optimization

Chair: Yao Xie, Georgia Institute of Technology, Atlanta, GA, United States

Co-Chair: Linglingzhi Zhu, Georgia Institute of Technology

1 - Score Matching for Discrete Diffusion Models

Yuchen Wu, Cornell University, Ithaca, NY, United States

Existing statistical guarantees for score estimation have largely focused on continuous diffusion models, leaving the discrete setting significantly under-explored. This is a critical gap, as discrete models often outperform continuous ones on discrete target distributions. We address this by deriving a minimax lower bound for the score entropy loss and proposing a thresholding estimator that provably attains this bound up to poly-logarithmic factors.

2 - Unlock Diffusion Models for Generative Optimization

Minshuo Chen, Northwestern University, Evanston, IL, United States

Diffusion models have achieved state-of-the-art performance in generative modeling, with their power significantly enhanced by fine-tuning toward task-specific objectives. However, existing fine-tuning methods often lack a principled foundation and offer limited performance guarantees. In this talk, we present a mathematical framework, termed as generative optimization, for understanding guidance-based diffusion model fine-tuning, providing a systematic perspective on its optimization properties and algorithmic design. We abstract task-specific objectives as a reward function and fine-tuned diffusion models aim to maximize the reward by generating solutions. In the offline (one-step) setting, we show that guidance enables high-quality sample generation, achieving optimality akin to off-policy bandit algorithms. In the online (iterative) setting with objective function queries, we establish a strong connection between guided diffusion and numerical optimization.

3 - Discrete Guidance Matching: Exact Guidance for Discrete Flow Matching

Liyan Xie, University of Minnesota, Minneapolis, MN, United States, Zhengyan Wan, Yidong Ouyang, Fang Fang, Hongyuan Zha, Guang Cheng

Guidance provides a simple and effective framework for posterior sampling by steering the generation process towards the desired distribution. When modeling discrete data, existing approaches mostly focus on guidance with the first-order Taylor approximation to improve the sampling efficiency, which may introduce non-negligible approximation errors. A novel guidance framework for discrete data is proposed to address this problem: We derive the exact transition rate for the desired distribution given a learned discrete flow matching model, leading to guidance that only requires a single forward pass in each sampling step, significantly improving efficiency. This unified novel framework is general enough, encompassing existing guidance methods as special cases, and it can also be seamlessly applied to the masked diffusion model. We demonstrate the effectiveness of our proposed guidance on energy-guided simulations and preference alignment on text-to-image generation and multimodal understanding tasks.

4 - No-Regret Generative Modeling via Parabolic Monge-PDE

Nabarun Deb, University of Chicago, Chicago, IL, United States, Tengyuan Liang

We introduce a novel generative modeling framework based on a discretized parabolic Monge-Amp\`{e}re PDE, which emerges as a continuous limit of the Sinkhorn algorithm commonly used in optimal transport. Our method performs iterative refinement in the space of Brenier maps using a mirror gradient descent step. We establish theoretical guarantees for generative modeling through the lens of no-regret analysis, demonstrating that the iterates converge to the optimal Brenier map under a variety of step-size schedules. As a technical contribution, we derive a new Evolution Variational Inequality tailored to the parabolic Monge-Amp\`{e}re PDE, connecting geometry, transportation cost, and regret. Our framework accommodates non-log-concave target distributions, constructs an optimal sampling process via the Brenier map, and integrates favorable learning techniques from generative adversarial networks and score-based diffusion models. As direct applications, we illustrate how our theory paves new pathways for generative modeling and variational inference.

5 - Bellman Optimality of Average-Reward Robust Markov Decision Processes with a Constant Gain

Shengbo Wang, University of Southern California, Los Angeles, CA, United States, NIAN SI

Learning and optimal control under robust Markov decision processes (MDPs) have received increasing attention, yet most existing theory, algorithms, and applications focus on finite-horizon or discounted models. The average-reward formulation, while natural in many operations research and management contexts, remains underexplored. This is primarily because the dynamic programming foundations are technically challenging and only partially understood, with several fundamental questions remaining open. This paper steps toward a general framework for average-reward robust MDPs by analyzing the constant-gain setting. We study the average-reward robust control problem with possible information asymmetries between the controller and an S-rectangular adversary. Our analysis centers on the constant-gain robust Bellman equation, examining both the existence of solutions and their relationship to the optimal average reward. Specifically, we identify when solutions to the robust Bellman equation characterize the optimal average reward and stationary policies, and we provide sufficient conditions ensuring solutions' existence. These findings expand the dynamic programming theory for average-reward robust MDPs and lay a foundation for robust dynamic decision making under long-run average criteria in operational environments.

6 - Worst-case Generation via Minimax Optimization in Wasserstein Space

Xiuyuan Cheng, Duke University, Durham, NC, United States, Yao Xie, Linglingzhi Zhu, Yunqin Zhu

Worst-case generation plays a critical role in evaluating robustness and stress-testing systems under distribution shifts, in applications ranging from machine learning models to power grids and medical prediction systems. We develop a generative modeling framework based on min-max optimization over continuous probability distributions, namely the Wasserstein space. Unlike traditional discrete distributionally robust optimization approaches, which often suffer from scalability issues, limited generalization, and costly worst-case inference, our framework exploits Brenier's theorem to characterize the least favorable distribution as the pushforward of a transport map from a continuous reference measure. Based on the min-max formulation, we propose a Gradient Descent Ascent (GDA)-type scheme that updates the decision model and the transport map alternately, establishing global convergence guarantees without requiring convexity-concavity type assumptions. We also propose to parameterize the transport map using a neural network that can be trained simultaneously with the GDA iterations by matching the transported training samples, thereby achieving a simulation-free approach. The efficiency of the proposed method as a worst-case generator is validated by numerical experiments on synthetic and image data.

SunD10

Caucus Room

Neural Network and Reinforcement Learning for Stochastic Optimal Controls and Mean-field Games

Invited Session

Computational Optimization

Chair: Wei Cai, Southern Methodist University, 17515 Woods Edge Dr, 75287

Co-Chair: Thaleia Zariphopoulou, UT Austin

1 - Martingale Deep Neural Networks for Stochastic Optimal Controls

Wei Cai, Southern Methodist University, Dallas, TX, United States

In this talk, we will present a highly parallel and derivative-free martingale neural network method, based on

the probability theory of Varadhan's martingale formulation of PDEs, to solve Hamilton-Jacobi-Bellman (HJB) equations arising from stochastic optimal control problems (SOCPs), as well as general quasilinear parabolic partial differential equations (PDEs). In both cases, the PDEs are reformulated into a martingale problem such that loss functions will not require the computation of the gradient or Hessian matrix of the PDE solution, and can be computed in parallel in both time and spatial domains. Moreover, the martingale conditions for the PDEs are enforced using a Galerkin method realized with adversarial learning techniques, eliminating the need for direct computation of the conditional expectations associated with the martingale property. For SOCPs, a derivative-free implementation of the maximum principle for optimal controls is also introduced. The numerical results demonstrate the effectiveness and efficiency of the proposed method, which is capable of solving HJB and quasilinear parabolic PDEs accurately and fast in dimensions as high as 10,000.

2 - Learning Mean-Field Games through Mean-Field Actor-Critic Flow

Mo Zhou, University of California, Los Angeles, Los Angeles, CA, United States, Ruimeng Hu, Haosheng Zhou

We propose the Mean-Field Actor-Critic (MFAC) flow, a continuous-time learning dynamics for solving mean-field games (MFGs), combining techniques from reinforcement learning and optimal transport. The MFAC framework jointly evolves the control (actor), value function (critic), and distribution components through coupled gradient-based updates governed by partial differential equations (PDEs). A central innovation is the Optimal Transport Geodesic Picard (OTGP) flow, which drives the distribution toward equilibrium along Wasserstein-2 geodesics. We conduct a rigorous convergence analysis using Lyapunov functionals and establish global exponential convergence of the MFAC flow under a suitable timescale. Our results highlight the algorithmic interplay among actor, critic, and distribution components. Numerical experiments illustrate the theoretical findings and demonstrate the effectiveness of the MFAC framework in computing MFG equilibria.

3 - Neural Network Training and Fine-tuning of Diffusion Models

Yinbin Han, Stanford University, Stanford, CA, United States

Diffusion models have emerged as powerful tools for generative modeling. In this talk, I will present theoretical results on both learning score functions and fine-tuning diffusion dynamics. First, I will introduce a framework for analyzing score estimation with neural networks trained by gradient descent. With a connection between score matching and mean squared regression, we establish the first algorithmic generalization and sample-complexity guarantees for learning score functions via early stopping. In the second part of the talk, I will discuss a KL-regularized stochastic control formulation of fine-tuning. A policy iteration algorithm is proposed with global linear convergence and provable regularity throughout training.

4 - When SOC Meets CTMC: From Discrete Neural Sampler Training to Diffusion LLM Finetuning

Yuchen Zhu, Georgia Institute of Technology, Atlanta, GA, United States

Discrete diffusion models have emerged as a powerful modeling method for categorical data on finite state spaces. Driven by Continuous-time Markov Chains (CTMC), a pure jump process natively defined on the state space, discrete diffusion models face unique challenges in Reinforcement Learning (RL) post-training due to a lack of trajectory differentiability. In this talk, I will introduce an effective, scalable framework for addressing this challenge and performing RL finetuning with discrete diffusion models in a theoretically-grounded way, leveraging Stochastic Optimal Control (SOC) techniques applied to CTMC. I will discuss two applications of this framework.

In the first part of the talk, I will introduce the Masked Diffusion Neural Sampler (MDNS), an effective neural sampler training strategy for high-dimensional discrete distributions, which achieves state-of-the-art performance in learning sophisticated distributions, such as Ising and Potts models.

In the second part of the talk, I will present Distribution Matching Policy Optimization (DMPO), a novel RL strategy tailored to finetuning diffusion Large Language Models (diffusion LLMs) and grounded in SOC techniques, which significantly outperforms previous baselines on multiple math reasoning and planning benchmarks.

SunD11

Charter Room

Networked Systems: Analysis, Control and Optimization

Invited Session

Network Optimization

Chair: Sebin Gracy, South Dakota School of Mines, Rapid City, SD, United States

1 - Networked Competitive Bivirus SIS spread with Higher Order Interactions

Sebin Gracy, South Dakota School of Mines, Rapid City, SD, United States, Brian Anderson, Mengbin Ye, Cesar Uribe

The paper studies the simultaneous spread of two competing viruses over a network of population nodes with higher- order interactions (HOI), using a continuous-time time-invariant competitive bivirus networked susceptible-infected- susceptible (SIS) system. In this paper, by HOI, we mean interactions among group sizes of no more than three nodes. The first key contribution is to establish several important general properties for generic systems. Namely, there are a finite number of equilibria, each equilibrium is nondegenerate, and the system is a strongly monotone dynamical system. Put together, we establish that for almost all initial conditions, the system will converge to a stable equilibrium (of which there may be many). We then turn our focus to characterizing the existence and stability of the equilibria of this system, which are i) the disease-free equilibrium (DFE), ii) single-virus endemic equilibria, and iii) coexistence equilibria (where both viruses are present). We present a range of conditions on the existence or nonexistence of various equilibria. Two key features underpin our results: First, we substantially relax the connectivity conditions of the network relative to existing literature. More specifically, for securing several important general properties for generic systems, we do not require strong connectivity of the standard pairwise interaction graph. Second, we identify dynamical phenomena, including multiple stable equilibria, which are known to be impossible without HOI. The latter illustrates the novel insights that are obtained by including HOI into models of epidemic spread. Finally, we illustrate our results using a real-world large-scale network.

2 - The Defense of Networked Targets in General Lotto Games

Keith Paarporn, University of Colorado, Colorado Springs, Colorado Springs, CO, United States, Adel Aghajan, Jason Marden

Ensuring the security of networked systems is a significant problem due to the susceptibility of modern infrastructures to adversarial interference. A central component of this problem is how defensive resources should be allocated to mitigate the severity of potential attacks on the system. In this paper, we consider a General Lotto game where a defender and attacker deploy resources on the nodes of a network, and the objective is to secure as many links as possible. The defender secures a link only if it out-competes the attacker on both of its associated nodes. For bipartite networks, we completely characterize equilibrium payoffs and strategies for both the defender and attacker. Surprisingly, the resulting payoffs are the same for any bipartite graph. On arbitrary networks, we provide lower and upper bounds on the defender's max-min value. Notably, the equilibrium payoff from bipartite networks serves as the lower bound. This suggests that more connected networks are easier to defend against attacks. These findings are confirmed through numerical solutions of a convex optimization problem geared towards obtaining tighter upper bounds for arbitrary graph structures.

3 - Distributed Asynchronous Multi-Agent State Estimation

Matthew Hale, Georgia Tech, Atlanta, GA, United States, Adam Pooley

Distributed trajectory estimation arises in many applications across autonomy and robotics. However, existing implementations typically do not consider asynchrony in agents' communications and computations, even though this asynchrony commonly arises in practice. Therefore, we propose an asynchronous block coordinate descent algorithm for distributed trajectory estimation of multi-robot systems. We consider a team of agents that observes a team of robots, and the agents communicate and jointly compute estimates of the robots' states over time. Then, we derive an approximation of the maximum a posteriori estimation problem, and we show that it introduces only negligible errors in optima. Then we show how to use the structure of a given estimation problem to determine the minimal communication topology that agents must use. We apply that result to show our approximation promotes sparsity of agents' communication topology. Next, we show that the approximated objective function is strongly convex and prove that agents' iterates converge exponentially fast to its unique minimizer. In simulation we compare our algorithm against a state-of-the-art algorithm and show that our algorithm attains a solution with up to 63% less error. Experiments show the accuracy of our approach when sensing real robots.

4 - Optimal Distributed Mitigation in Layered Epidemic Processes

Jose Caiza, Purdue University , West Lafayette, IN, United States

Most models of networked epidemic processes focus exclusively on person-to-person transmission, and existing mitigation strategies typically rely on centralized optimization methods such as spectral optimization or minimizing the basic reproduction number. However, many diseases also spread through underlying infrastructure systems—including transportation networks, water distribution networks, and hospitals—which can significantly influence epidemic dynamics. In this work, we introduce a **layered epidemic model** that couples a population network with an infrastructure network carrying a pathogen-borne disease. We propose a new metric, the local basic reproduction number (LBRN), which provides sufficient local conditions for predicting outbreaks. Since LBRNs depend only on local information, they naturally enable distributed decision-making. Leveraging this property, we develop an LBRN-based optimization framework for both budget and rate constrained optimal allocation problems with preventive and corrective control actions. Under mild assumptions imposed on the allocation cost functions, we show that these problems can be formulated as convex geometric programs. The nature of the LBRNs allows the resulting optimization to be solved through standard distributed dual-decomposition algorithms. We conclude with numerical experiments demonstrating the scalability of our method and comparing its performance with existing centralized approaches as the network size increases.