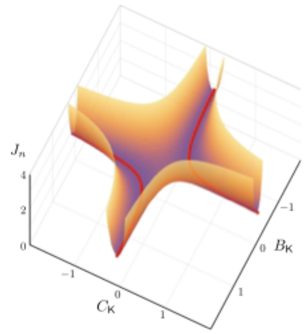
Toward a Theoretical Foundation of Policy Optimization for Control

Policy optimization (PO) methods allow agents to learn optimal behaviors by finding the best policy for the current state of the environment. PO has been a workhorse of Reinforcement Learning (RL), and a key ingredient in breakthroughs such as beating humans at sophisticated games and complex robotic manipulation. However, theoretical understanding of why and when it ``works" remains limited.

Earlier work by IFDS faculty [1] began to investigate PO, focusing on the Linear Quadratic Regulator (LQR)—a canonical control problem that served as a starting point to study gradient-based PO methods. We gave the first statistical and computational guarantees for PO applied to the LQR problem, showing its convergence to the *globally optimal* policy despite nonconvexity. In addition, motivated by applications where gradients are not available and one can at best access only samples of the cost function, we showed that sample-based PO converges to the optimal solution under mild assumptions, with sample complexity and iteration complexity that are polynomial in relevant problem-dependent quantities.

Encouragingly, our work has sparked interest in the broader controls community—our approach relying on *coercivity* and the *gradient dominance* property (a case of the Polyak-Lojasiewisz inequality) was extended to more complex control design problems (robust state-feedback, risk-sensitive control, LQ games). This has led to two recent trends: (1) using control problems as benchmarks for less-understood RL algorithms, and (2) theoretically-sound use of RL-style methods in control.

We next investigated the harder problem of analyzing PO for *partially observed* systems (e.g., output-feedback synthesis) where even for linear dynamics, the optimal policy will itself be dynamic (history dependent) unlike the case with LQR (Figure shows an example of non-isolated, disconnected globally optimal controllers). To this end, we characterize the optimization landscape and the saddle-points for the Linear Quadratic Guassian problem, leveraging convex parameterizations to make promising progress towards a full understanding [2,4].



This research thrust makes connections between control and RL to leverage the traditional strengths of the two viewpoints, as well as bridge the two research communities. Our survey paper on this topic [3] gives a unified view of recent results and the future outlook, and our tutorials/workshops at flagship conferences in 2023 (full-day workshop at ACC, half-day at L4DC, and minisymposia at SIAM Optimization conference) aim to engage control, optimization, and machine learning communities.

Credits. Supported in part by NSF DMS 2023166, CCF 2212261, CCF 1740551, Darpa Lagrange FA8650-18-2-7836.

Publications. [1] M. Fazel, R. Ge, S. Kakade, M. Mesbahi, Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator, Proc. of Intl. Conf. on Machine Learning (ICML), July 2018.
[2] Y. Zheng, Y. Sun, M. Fazel, N. Li, Escaping High-order Saddles in Policy Optimization for Linear Quadratic Gaussian (LQG) Control, Proc. of IEEE Conf. on Decision and Control (CDC), Dec 2022.
[3] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, T. Başar, Towards a Theoretical Foundation of Policy Optimization for Learning Control Policies, Annual Review of Control, Robotics, and Autonomous Systems, May 2023.
[4] Z. Ren, Y. Zheng, M. Fazel, N. Li, On Controller Reduction in Linear Gaussian Control with Performance Bounds, Learning for Dynamics and Control Conf. (L4DC), June 2023.