**AFRL-RQ-WP-TR-2020-0005**

# CONTROL AND LEARNING OF UNCERTAIN DYNAMICAL SYSTEMS: OPTIMIZATION, SAMPLING, AND REGRET

**Maryam Fazel, Sham Kakade, and Mehran Mesbahi**

**University of Washington**

**NOVEMBER 2019**
**Final Report**

**DISTRIBUTION STATEMENT A. Approved for public release.**
**Distribution is unlimited.**

**STINFO COPY**

**AIR FORCE RESEARCH LABORATORY**
**AEROSPACE SYSTEMS DIRECTORATE**
**WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7542**
**AIR FORCE MATERIEL COMMAND**
**UNITED STATES AIR FORCE**

# NOTICE AND SIGNATURE PAGE

AFRL-RQ-WP-TR-2020-0005 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

*//Signature//

RICHARD D. SNYDER
Work Unit Manager
Design and Analysis Branch

//Signature//

CHARLES TYLER
Chief, Design and Analysis Branch
Aerospace Vehicles Division

//Signature//

PHILIP S. BERAN, PhD
Technical Advisor, Design and Analysis Branch
Aerospace Vehicles Division

*Disseminated copies will show “//Signature//” stamped or typed above the signature blocks.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS**.

| | | |
|---|---|---|
| **1. REPORT DATE** ***(DD-MM-YY)***<br>November 2019 | **2. REPORT TYPE**<br>Final | **3. DATES COVERED** ***(From - To)***<br>13 April 2018 – 13 October 2019 |

**4. TITLE AND SUBTITLE**

CONTROL AND LEARNING OF UNCERTAIN DYNAMICAL SYSTEMS: OPTIMIZATION, SAMPLING, AND REGRET

**5a. CONTRACT NUMBER**
FA8650-18-2-7836

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
DARPA

**6. AUTHOR(S)**

Maryam Fazel, Sham Kakade, and Mehran Mesbahi

**5d. PROJECT NUMBER**
N/A

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**
Q1YT

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

University of Washington
4333 Brooklyn Avenue NE
Seattle, WA 98195-0001

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Research Laboratory
Aerospace Systems Directorate
Wright-Patterson Air Force Base, OH 45433-7542
Air Force Materiel Command
United States Air Force

Defense Advanced Research Projects
Agency/Defense Sciences
Office (DARPA/DSO)
3701 N. Fairfax Drive
Arlington, VA 22203

**10. SPONSORING/MONITORING AGENCY ACRONYM(S)**
AFRL/RQVC

**11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)**
AFRL-RQ-WP-TR-2020-0005

**12. DISTRIBUTION/AVAILABILITY STATEMENT**

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

**13. SUPPLEMENTARY NOTES**

This report is the result of contracted fundamental research, which is deemed exempt from Public Affairs Office security and policy review in accordance with Deputy Assistant Secretary of the Air Force (Science, Technology, Engineering) (SAF/AQR) memorandum dated 10 Dec 08 and Air Force Research Laboratory Executive Director (AFRL/CA) policy clarification memorandum dated 16 Jan 09.

**14. ABSTRACT**

This report shows that first order methods can be used to provide an effective bridge between optimal control theory and sample-based reinforcement learning. The work focuses on the linear quadratic regulator problem and Markov decision processes. Some of the results include a proof that gradient descent starting from a stabilizing policy converges to the globally optimal policy and an algorithm that provides nearly tight regret bounds for the control of a linear dynamical system with adversarial disturbances.

**15. SUBJECT TERMS**

optimal control theory, reinforcement learning, linear quadratic regulator, Markov decision processes

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT:** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** (Monitor) |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | Richard D. Synder |
| Unclassified | Unclassified | Unclassified | SAR | 25 | **19b. TELEPHONE NUMBER** (Include Area Code)<br>(937) 713-7212 |

# TABLE OF CONTENTS

## LIST OF FIGURES

## PREFACE

Major advances have been made in recent years in the control of uncertain dynamical systems using reinforcement learning and data-driven approaches; examples range from allowing robots to perform more sophisticated controls tasks such as robotic hand manipulation, to sequential decision making in game domains (e.g., AlphaGo). Many of these successes have relied on sampling-based reinforcement learning algorithms, including the deep Reinforcement Learning (DeepRL) approaches, where we have little theoretical understanding of their efficiency from statistical or computational perspectives. In contrast, control theory (optimal and adaptive control) has a rich body of tools, with provable guarantees, for related sequential decision-making problems, particularly those that involve continuous control. These latter techniques are often model-based: they estimate an explicit dynamical model first (e.g., system identification) and then design optimal controllers in contrast to the direct model-free approaches, such as those in DeepRL.

The objective of this project has been to build an overarching bridge between these two lines of work, namely, between optimal control theory and sample-based reinforcement learning methods, to better connect the model-based and model-free methods, and provide rigorous theory for practical and popular sampling-based methods in applied machine learning and applied control. In particular, we report advances on the Linear Quadratic Regulator in control (from a new perspective) and on Markov Decision Processes.

# 1 SUMMARY

Direct policy gradient methods for reinforcement learning and continuous control problems are a popular approach for a variety of reasons: 1) they are easy to implement without explicit knowledge of the underlying model, 2) they are an "end-to-end" approach, directly optimizing the performance metric of interest, 3) they inherently allow for richly parameterized policies. A notable drawback is that even in the most basic continuous control problem (the linear quadratic regulator), these methods must solve a non-convex optimization problem, where little is understood about their efficiency from both computational and statistical perspectives. In contrast, system identification and model-based methods in optimal control theory have a more solid theoretical footing. This project has aimed to bridge the gap between the two communities of control theory and RL.

Our exploration has focused on two pillars of system theory and learning, namely, **Linear Quadratic Regulator (LQR)** problem on one hand, and **Markov Decision Processes (MDPs)** on the other hand. In this project, we have considered the LQR problem in both discrete and continuous time and over an infinite time horizon. When exact gradients of the cost function with respect to the control gain are available to our control algorithms, the optimal solution to the problem is well-known. In this case, we ask whether applying (several versions of) the popular policy gradient methods to this problem will give the same known solution. The reason this question is challenging is that algorithms that update the "policy" must solve a nonconvex optimization problem to find the globally optimal policy.

Our first set of results in this project had been as follows. We proved that despite the nonconvexity, gradient descent starting from a stabilizing policy *converges to the globally optimal policy*. We then extended this to sampling-based policy gradient type methods without access to exact gradients, at the cost of taking a number of samples of function value. Thus, we are able to conclude that policy gradient methods globally converge to the optimal solution, and have *sample complexity* and *computational complexity* that depends polynomially on relevant problem parameters.

The results above appeared in a paper published in ICML 2018 (International Conference on Machine Learning). This work, which obtained the first theoretical guarantees on convergence of policy gradient methods on the linear quadratic regulator problem, is receiving attention from both the control theory and the machine learning communities, contributing to the growing interest in the intersection of these two fields (the paper has already has 74 citations in just over a year according to Google Scholar).

Since the publication of our ICML paper, our team has been able to delve into deeper technical aspects of the problem setup: we identified some of the control theoretic aspects of the direct policy updates, extensions to natural gradient and quasi-Newton updates, and also pointed out conditions under which direct policy update recovers some other algorithms for solving this class of problems. As such, our work has not only provided a bridge between learning and control in the context of LQR, but also has provided new insights into reasoning about algorithm design for this important class of control synthesis problems. In fact, we have also been able to extend the setup of policy gradients/natural gradient/quasi-Newton updates to continuous LQR problems, requiring distinct analysis techniques. The main issue here is that although one can construct a map between discrete-time and continuous time representations of a dynamical systems (the so-

called bilinear transform), there is no analogous map between the corresponding feedback gains. That is, gradient policy updates for discrete-time systems cannot in general be mapped directly to the corresponding updates for continuous-time problems. However, using the overarching theme of the project, we have been able to show certain desirable properties of the continuous-time LQR, that would allow the extension of the direct policy updates.

One of the advantages of adopting direct policy updates via first-order gradient based approach to control synthesis is the ability to restrict the control structure on a subspace or convex set, and then use a "projected" gradient (when this projection is not costly). This is in sharp contrast to means of imposing structure on the control (policy) through the so-called certificates, such as the solution of the Riccati equation (that parameterizes the cost-to-go). In this latter case, one needs to impose a structure on this certificate, such that when this structured certificate is used for control synthesis, the resulting controller has the desired structure. This plan of action, however, turns out to be rather intricate as it is far from obvious what structure on the certificate leads to the desired structure on feedback control. Direct policy update circumvents this issue. However, there are a number of theoretical questions that need to be addressed to for this more direct structured synthesis approach. For example, it is not clear whether the process of policy update followed by a projection retains the stabilizing feature of the feedback gains. It is also not clear whether this projected gradient update leads to stationary point or whether the set of structured stabilizing feedback gains is in fact connected. In this project, we have obtained several results on these questions: we have shown that certain structures on the feedback gain can lead to exponentially many connected components (in the topological sense) and also derived sufficient conditions on this structure under which one can ensure that the set of stabilizing structured feedback gains has only one connected component. Furthermore, we have shown that when the controller is restricted to a subspace, projected gradient update with carefully chosen learning rate does in fact convergence to stationary point of the structured LQ problems.

In another work, we have studied the control of a linear dynamical system with adversarial disturbances (as opposed to statistical noise). The objective we consider is one of regret: we desire an online control procedure that can do nearly as well as that of a procedure that has full knowledge of the disturbances in hindsight. Our main result is an efficient algorithm that provides nearly tight regret bounds for this problem. From a technical standpoint, this work generalizes upon previous work in two main aspects: our model allows for adversarial noise in the dynamics, and allows for general convex costs.

We have also examined the problem of low-order linear system identification, using the nuclear norm of the system's Hankel matrix as a regularizer to promote a low-order solution. Model order is a measure of model complexity, and corresponds to system memory length, or the minimal number of states that are needed to describe the dynamics. While Hankel-nuclear-norm regularization is used in practice, it does not yet have a theoretical analysis that quantifies the complexity of the inputs (sample complexity analysis). We have started to obtain such sample complexity bounds for a specific system identification setup, where multiple input rollouts are allowed.

Finally, we also examine related questions in **Markov Decision Processes** (MDPs). Specifically, we focus on: 1) the convergence and approximation properties of policy gradient methods, and 2) understanding discovery of how a system can learn to explore and manipulate its environment in the absence of any reward signal.

# 2 INTRODUCTION

## 2.1 Linear Quadratic Regulator (LQR) in continuous control

As mentioned in the introduction, in this project we have adopted a multi-pronged approach to the problem of direct policy optimization for uncertain linear dynamical systems; in control theory, such models are represented in the form,

$$\dot{x}(t) = Ax(t) + Bu(t) + Ew(t)$$

or

$$x_{k+1} = Ax_k + B_k + Ew_k$$

where $A$ is the system matrix, $B$ encodes how the input effect the dynamics of the systems, and $E$ captures how disturbances influence the dynamics. The former is referred to as the continuous time model and the latter as discrete time linear time-invariant model. In control theory, one aims to design a control $u(t)$ or $u_k$ such that certain stability and performance objectives are satisfied, given some knowledge of the system matrices. In this project, our first objective has been to completely resolve the adoption of direct policy updates for the so-called linear quadratic regulator (LQR) problem, where an integral quadratic cost is minimized over an infinite horizon for a linear system of the forms above. LQR is one of the pillars of the so-called state-space approach to control synthesis. Historically, LQR has been approached from the perspective of characterizing the cost-to-go, where it is first shown that the cost-to-go from a given state assumes a quadratic form, that can be found using the Riccati equation. Although this approach is very powerful, it is also very sensitive to the assumptions about the problem structure and knowledge of the system matrices. For example, one imposes a structure on the desired feedback gain, say a sparsity pattern, or when the control is synthesized using output feedback, the entire machinery that involves solving Riccati the equation becomes problematic. Our first objective has been to explore to what extend gradient descent and its variants can be adopted for direct policy updates for LQR and then use the insights to go beyond the LQR setup. In particular in our project we have addressed the following:

a) Complete characterization of the direct policy optimization for LQR (for discrete time LTI models) for the model-based and model-free case, highlighting the convergence properties of the algorithm on the system parameters

b) Complete characterization of the direct policy optimization for LQR (for continuous time LTI models) highlighting the convergence properties of the algorithm on the system parameters

c) Proposing a quasi-Newton algorithm for policy optimization and optimal stepsize and showing its connection to other classes of iterative algorithms for LQR, including the so-called Kleinman and Hewer algorithms.

d) Complete characterization of the set of stabilizing feedback gains for discrete and continuous time system, and identifying its key topological and metrical properties and their algorithmic implications.

e) Proposing continuous flows for solving LQR-type control synthesis problems, including gradient flow, natural gradient flow, and quasi-Newton flow, both for discrete and continuous time models.

f) We also consider the questions of robustness, where the disturbances $w_t$ may be adversarial. Here, we focus on comparing to the best linear controller which knows the disturbances in advance. We provide the first low regret result, showing that this is indeed possible.

## 2.2 Introduction to Markov Decision Processes (MDPs)

A (finite) Markov Decision Process (MDP) $M = (S, A, P, r, \gamma, \rho)$ is specified by: a finite state space $S$; a finite action space $A$; a transition model $P$ where $P(s'|s, a)$ is the probability of transitioning into state $s'$ upon taking action $a$ in state $s$; a reward function $r : S \times A \rightarrow [0, 1]$ where $r(s, a)$ is the immediate reward associated with taking action $a$ in state $s$; a discount factor $\gamma \in [0, 1)$; a starting state distribution $\rho$ over $S$.

A deterministic, stationary policy $\pi : S \rightarrow A$ specifies a decision-making strategy in which the agent chooses actions adaptively based on the current state, i.e., $a_t = \pi(s_t)$. The agent may also choose actions according to a stochastic policy $\pi: S \rightarrow \Delta(A)$ (where $\Delta(A)$ is the probability simplex over $A$).

A policy induces a distribution over trajectories $\tau = (s_t, a_t, r_t)_{t=0}^{\infty}$, where $s_0$ is drawn from the starting state distribution $\rho$, and, for all subsequent timesteps $t$, $a_t \sim \pi(\cdot\ |s_t)$ and $s_{t+1} \sim P(\cdot\ |s_t, a_t)$. The value function $V^\pi : S \rightarrow R$ is defined as the discounted sum of future rewards starting at state s and executing $\pi$, i.e. $V^\pi(s)$ is the expected, discounted sum future rewards when the policy $\pi$ is executed starting from state $s$.

We further define $V^\pi(\rho)$ as the expected value under the initial state distribution $\rho$:

$$V^\pi(\rho) := E_{s\sim\rho}[V^\pi(s)].$$

For MDPs, policy gradient methods are among the most effective methods in challenging reinforcement learning problems with large state and/or action spaces. Our work focused on their most basic theoretical convergence properties, where little was known before. This included:

a) A complete characterization of how fast they converge to a globally optimal solution (say with a sufficiently rich policy class).

b) A complete characterization of how they cope with approximation error due to using a restricted class of parametric policies.

c) Bounds on their finite sample behavior. Such characterizations are important not only to compare these methods to their approximate value function counterparts (where such issues are relatively well understood, at least in the worst case), but also to help with more principled approaches to algorithm design.

We also looked at the question of pure exploration, where we seek to discover what an agent is capable of doing with a reward function. For example, suppose an agent is in a (possibly unknown) Markov Decision Process in the absence of a reward signal, what might we hope that an agent can efficiently learn to do? Our work studied a broad class of objectives that are defined solely as functions of the state-visitation frequencies that are induced by how the agent behaves. For example, one natural, intrinsically defined, objective problem is for the agent to learn a policy which induces a distribution over state space that is as uniform as possible, which can be measured in an entropic sense. Our work here provided an efficient algorithm to optimize such intrinsically defined objectives, when given access to a black box planning oracle (which is robust to function approximation).

# 3 METHODS, ASSUMPTIONS, AND PROCEDURES

## 3.1 LQR methods, direct policy optimization, and system identification

We first considered the LQR problem in discrete time and continuous time and over an infinite time horizon. This control synthesis problem assumes a linear model for the underlying dynamical system and aims to synthesis an optimal control with respect to an integral quadratic cost. The linearity assumption on the model however, is due to the fact that one aims to derive fundamental theoretic guarantees on resulting closed loop system, such as robustness to delays or model uncertainty. In practice, however, LQR is often used to synthesized a time-varying/state-dependent controller when the underlying nonlinear systems is linearized around some nominal trajectory or equilibrium point. In this latter case the knowledge of the nonlinear model again becomes crucial for any formal guarantees. A power approach for analyzing the robustness of such linear and nonlinear systems is through Lyapunov theory, that in the case of LQR, is intimately related to the cost-to-go. The cost-to-go in the case of LQR is obtained through the solution of the Riccati equation, but again, this can be done given the knowledge of the system model. Hence, adopting a direct policy update to LQR and other related problems, although seem natural, but pose a number of issues that need to be addressed, such as ensure that updated policies remain stabilizing. Our assumption in this work adopts the typical requirements for LQR synthesis: the cost functions for continuous and discrete time models are of the forms,

$$\int_0^\infty x(t)^T Q x(t) + u(t) R u(t)$$

or

$$\sum_{k=0}^{\infty} x_k^T Q x_k + u_k R u_k$$

The cost function parameters $Q$ and $R$ are positive semidefinite and positive definite respectively, and the pairs $(Q, A)$ and $(A, B)$ are detectable and stabilizable, respectively. Since we need to initialize the direct policy update from some initial controller that is already in the feedback loop, it is assumed that we have access to either an initial stabilizing feedback gain, or alternatively, that the system is open loop stable. Our team has since worked on other data-driven approaches to obtain this first stabilizing feedback gain.

The algorithms we consider are based on gradient descent on the control cost as a function of the control policy (which we have been referring to as direct policy update methods). We are interested in analyzing these methods due to their popularity in practice. We have shown strong properties (convergence to the *global* minimum) for the LQR problem. Thinking towards the next stages of the current project, we asked: What can one say more generally, beyond the favorably-structured LQR problem? Understanding of gradient descent (and more generally, first-order optimization algorithms) for nonconvex landscapes will have to resolve issues related to saddle points: how can we ensure the algorithm can progress towards a local minimum and not get stuck in a saddlepoint? To examine this more broadly, we have used tools from Riemannian geometry to understand gradient descent on a smooth manifold and its rates of escape from undesirable saddle points and convergence to local optima. Our results on this problem will be published in the proceedings of the upcoming 2019 NeurIPS conference.

A related (classical) problem in control theory is system identification. While methods for identifying a linear dynamical system given input-output observations are well-studied, understanding their sample complexity (how much data is needed for a given identification accuracy) is much more recent topic, which brings recent statistical techniques to classical control. We focus on methods that use regularization to encourage fitting *low-order* dynamical models. Model order is a measure of complexity: memory length, smallest possible dimension for (hidden) state $x_t$. Given data $u_t$, $y_t$, $t = 0, 1, \dots, T$, we would like to identify a low-order model given by the set of matrices $(A, B, C)$, or equivalently the Markov parameters $CA^{t-1}B$, $t = 1, 2, \dots$ such that the block-Hankel matrix (defined in section 4) has a low rank. We have started to examine this problem, focusing on nuclear-norm regularized least squares fitting; our work on this topic is in progress. The methods used are statistical guarantees for recovering a structured low-rank matrix from random measurements (in this case, by applying random inputs to the system).

### 3.2 MDP methods

This work studies ascent methods for the optimization problem:

$$\max V^{\pi\theta}(\rho), \theta \in \Theta$$

where $\{\pi_\theta | \theta \in \Theta\}$ is some class of parametric (stochastic) policies. We consider a number of different policy classes. One is complete in the sense that any stochastic policy can be represented in the class; specifically, we consider the standard softmax policy class. We also consider a restrictive policy class, which may not contain the optimal policy.

- *For the softmax parameterization*: For unconstrained $\theta \in R^{|S||A|}$ we have that $\pi_\theta(a|s) = \theta_{s,a}/Z\pi_\theta$ (where $Z$ is a normalizing constant). The softmax parameterization is complete.
- *Restricted parameterizations*: We also study parametric classes $\{\pi_\theta | \theta \in \Theta\}$ that may not contain all stochastic policies. Here, the best we may hope for is an agnostic result where we do as well as the best policy in this class. For example, this class may be neural network policies.

For optimization, we actually consider a different measure under a distribution $\mu$ over states (as opposed to $\rho$), which will become clear in our results section. The policy gradient algorithm we consider is:

$$\theta^{(t+1)} = \theta(t) + \eta \nabla_\theta V^{(t)}(\mu)$$

The widely used natural gradient algorithm we consider is:

$$\theta^{(t+1)} = \theta(t) + \eta F^{-1} \nabla_\theta V^{(t)}$$

where $F$ is Fisher information matrix under the state action visitation measure (as in [Kakade, ‘02]).

# 4 RESULTS AND DISCUSSION

## 4.1 Direct policy optimization with exact gradients

The cost function of the typical LQR problem is a function of the initial condition of the system, although the optimal policy (feedback gain) is independent of any initialization, $x_0$ or $x(0)$. Our first step to adopt direct policy update for LQR has been to randomize over the initial conditions (or choosing a spanning set of initial conditions) and then letting the input $u$ be state feedback form $u = -Kx$. Then the LQR cost function becomes a function of $K$, namely $f(K)$. In order to write the gradient update for LQR in the form,

$$K_{j+1} = K_j - \eta_j \nabla f(K_j)$$

One has to characterize the gradient; we have shown that this gradient is of the form,

$$\nabla f(K) = 2\big(RK - B^T X(A - BK)\big)Y$$

where $Y$ satisfies the Lyapunov equation,

$$A_K Y A_K^T - Y + \Sigma = 0$$

with $\Sigma$ representing the covariance of the random initial conditions or sum of outer products of the deterministic spanning initial conditions. Using this setup, we were then able to show a number of important properties for $f(K)$, namely, (1) $f$ is real analytic function over its domain, (2) $f$ is coercive and has a compact sublevel sets, and (3) $f$ is gradient dominated. We note that it is exactly these properties of the LQR problem that allows the application of Polyak's results on the convergence of first order methods for gradient dominated functions. Gradient dominance essentially bounds how much the function value deviates from its optimal value as a function of its gradient. This property then allows the adoption of first order method for certain classes of (nonconvex) optimization problems, and in particular LQR. That is, by correctly choosing the stepsize, the gradient update above can be used for LQR with a global convergence. This key observation has then provided springboard to consider extensions of gradient descent update, some of which further improve the convergence of the baseline update. In particular, our team also considered the so-called natural gradient update and quasi-Newton updates for LQR; some examples of performance for these algorithms are included in Figure 1 below.

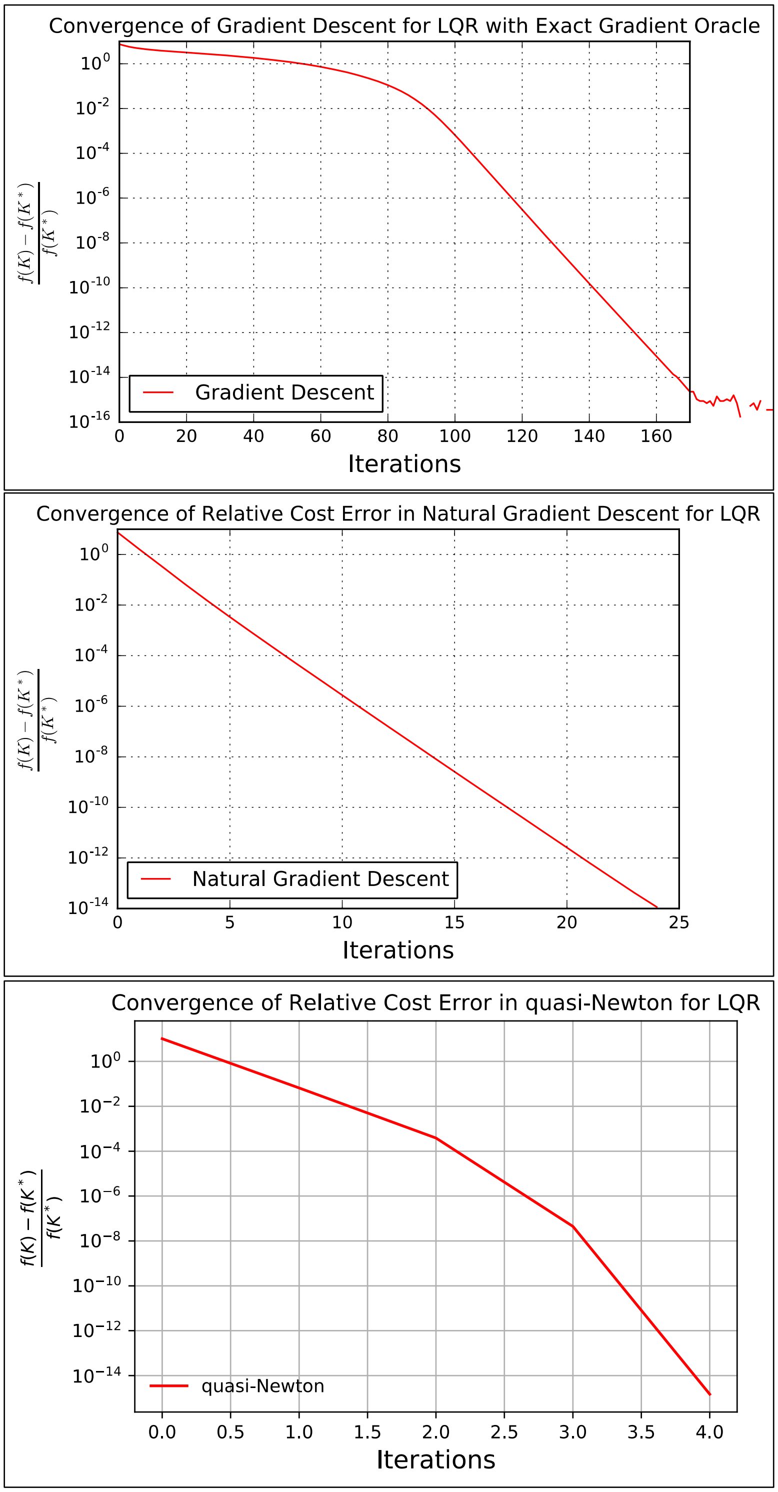


**Figure 1 Performance of gradient descent, natural gradient descent, and quasi-Newton iterations for the LQR problem**

It can also be shown that the quasi-Newton update with the optimal stepsize, in fact recovers the so-called Kleinman-Hewer algorithms for solving discrete Riccati equations. As such, the goal of connecting control theory with RL has had a number of side-benefits, including providing an overarching framework on how various algorithms for control synthesis are related to each other.

Continuous flows that mimic their discrete time implementation in terms of numerical algorithms, offer unique analytic insights for design and analysis off numerical algorithms. In such a setting, gradient descent, for example, can be examined using the machinery of gradient flows in a streamlined fashion. In our project we thus became interested to examine gradient flows for nonconvex direct policy update for LQR. Continuous flow policies can highlight some of the global characteristics of their discrete counterparts. For example, given that direct policy updates for control synthesis in general and LQR in particular is a nonconvex optimization problem (nonconvex objective function over a nonconvex set of stabilizing feedback gains), the selection of the stepsize requires intricate analysis. Gradient flows and their extensions, namely natural gradient flow and quasi-Newton flow, circumvent the issue of stepsize selection and offer a more transparent convergence analysis. We have been able to pursue this agenda to a great extent in this project, showing exponential convergence for a host of gradient flow policies, admitting easy to work with Lyapunov functions. Some of the numerical results for our work are presented in Figure 2 below.

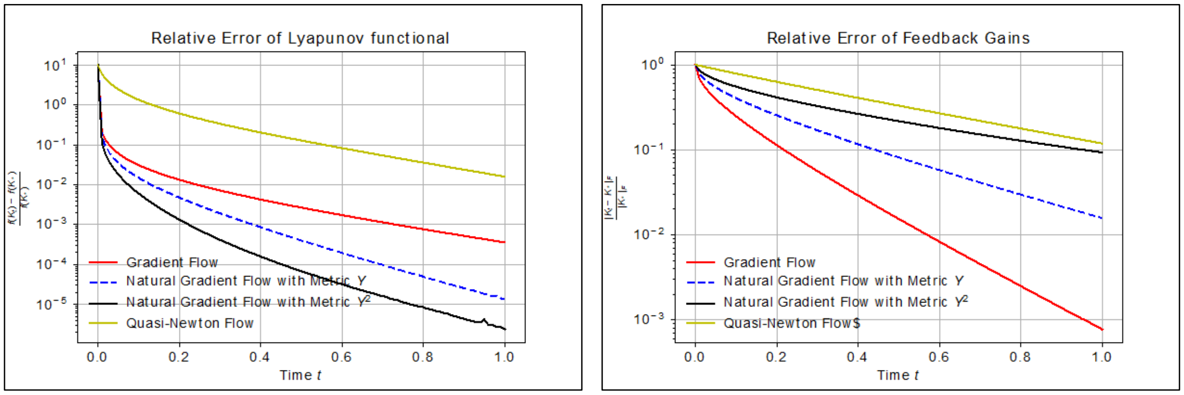


**Figure 2 Performance of gradient flow, natural gradient flow, and quasi-Newton flow for the LQR problem**

## 4.2 Direct policy optimization with sampling (inexact gradients)

When direct gradient of the LQR cost with respect to the policy are not available (e.g., model-free set up, with unknown $A$, $B$ matrices), it is common to try to estimate the cost gradients by perturbing the policy (around the current point) and observing the change in the cost. The idea is analogous to zeroth-order or derivative-free optimization. We analyze this approach for the LQR problem, and show that with enough samples (number of rollouts, times length of each rollout, times number of algorithm iteration) convergence to the global optimum happens when the initial policy $K_0$ is stabilizing, and stepsize is chosen appropriately.

## 4.3 Geometry of set of stabilizing controllers

Direct policy updates for LQR and its extension has required us to examine the metrical and topological properties of the set of stabilizing feedback policies in a more systematic way. This is due to the fact that by solving for the cost-to-go first, for example using the Riccati equation, one can only implicitly address the stabilizing features of the resulting feedback gain. More generally, such topological properties have received renewed interest in system literature as they have direct implications for adopting learning algorithms for control design. We also note that insights into topological and metrical properties of stabilizing feedback gains also reveal fundamental shortcomings in certain optimization algorithms. For example, if the set of stabilizing feedback gains has several path-connected components, the solutions of gradient-type learning algorithms will be highly dependent on the initialization process. It is thus surprising that despite the long historical interest in characterizing the set of stabilizing feedback gains, research works on its set-theoretic and topological properties are rather limited. This is potentially due to significantly more interest in characterizing the set of certificates for stabilizing controllers, e.g., in terms of linear matrix inequalities. In this work we have examined the convexity, connectedness, and topological properties of the set of stabilizing feedback gains.

Our work has characterized topological, metrical, and geometric properties of the set of stabilizing controllers for both continuous and discrete-time LTI systems. We have shown that the set of stabilizing state-feedback gains for a continuous SISO system is regular open, unbounded, in general nonconvex, and path-connected in the Euclidean topology. In the meantime, the set of stabilizing output-feedback controllers is shown to be open but not connected in general, and can be bounded or unbounded. In recent works, based on the implicit assumption that stable and unstable intervals of the feedback gain interlace, it has been stated that the set of stabilizing output feedback controllers for SISO systems can have at most $n$ connected components. If this assumption does not hold, however, the line of reasoning reported in the literature lead to the upper bounds of $2n$ and $n$, respectively. In the work supported by this project we have proved a tight bound of $n/2$ for continuous as well as discrete time LTI systems (as seen in Figure 3); moreover, all of our results are constructive (they lead to algorithms for characterizing these sets) and rely on basic topology and analytic theory of polynomials.

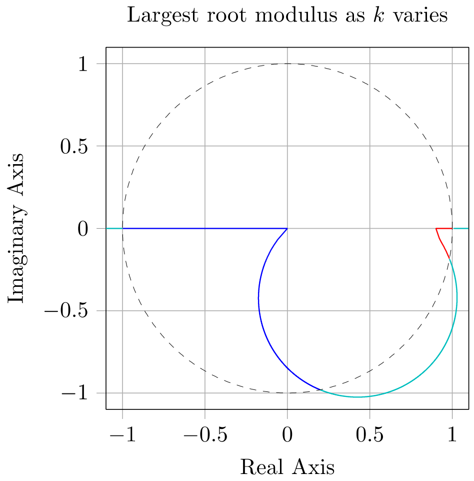


**Figure 3 Stabilizing feedback gains for an output feedback problem**

*For discrete time systems, the spectra of the closed loop system has to be in the unit disk; this example shows that for an output feedback problem on a system for n=4 states, the set of stabilizing feedback gains can show two connected components*

The separate treatment for continuous and discrete time systems has been warranted; in fact, in contrast to the folklore expectation of unified properties for continuous and discrete time systems, there are counterexamples to show that the analogies between the two are far from complete. The distinct difference between continuous and discrete LTI systems might be due to the fact that the generalized bilinear transform has poles and thus not continuous. Therefore, generalizing the proposed topological properties of the set of stabilizing feedback gains from continuous LTI systems to discrete ones is not straightforward. Nevertheless, in this project we have been able to show that the set of stabilizing state feedback gains for discrete-time LTI SISO systems enjoys some of the topological properties as its continuous counterpart, i.e., open and path connected in Euclidean topology and nonconvexity. But in contrast to the continuous case, the set of stabilizing state feedback gains is bounded. For output feedback SISO systems, the corresponding set of stabilizing gains is open, bounded and in general nonconvex, but is no longer path-connected. Accordingly, we have proved that the set can have at most $n/2$ path-connected components, which is a tight bound supported by simulation results. In this part of the project, we have also been able to propose an algorithm for determining the intervals of stabilizing feedback gains for general continuous and discrete LTI systems. This algorithm also computes the number of unstable roots in each unstable interval.

## 4.4 Direct policy updates for structured control design

Our work has also considered the extension of direct policy updates via gradient descent to the problem of designing feedback gains with an arbitrary sparsity pattern. Such structured control synthesis problems are notoriously difficult. In this project, we have proposed a formalism to set up the problem where projected gradient descent of the form,

$$K_{i+1} = \mathcal{P}_U\big(K_j - \eta\nabla f(K_j)\big)$$

can be applied to structured synthesis problems. However, in the case of structured synthesis, the LQR cost function is no longer gradient dominated and the choice of stepsize cannot be generalized from the unstructured case. In our work, we have adapted the machinery developed for the unstructured LQR for the structured synthesis: we first define the initial state independent LQR formulation and then show that the cost function can be equivalently defined as the unstructured LQR cost function restricted to the linear space defined by the information-exchange graph; as such, the cost function is smooth in the subspace topology and has a coercive property. Using this setup, we the obtain the gradient and Hessian of the cost function, leading to a natural choice of stepsize by bounding the Hessian over the initial sublevel set. We show this stepsize will guarantee a non-asymptotic sublinear convergence rate to the first-order stationary point.

Structured control synthesis is in general is an open line of research in control theory, believed to be NP-hard event for the stabilization issue. That is, given a linear time-invariant system encoded by the pair $(A, B)$, there is currently no polynomial time algorithm that can verify whether there exists a $K$ with a given sparsity pattern that $A - BK$ is stable; noting that this problem is considered difficult even there is no notion of optimality in the problem setup. For example, we have shown that the set of structured stabilizing feedback gains for a linear system can have exponentially many connected components- as such it would be difficult in general to find first order methods for structured synthesis as the underlying algorithm needs to figure out a way to navigate a disconnected feasible set. Nevertheless, we are still able to show that our proposed algorithm can find controllers in each component that satisfy first order stationary condition on each connected subset. Such insights were then adopted for networked systems, when the structure of the feedback gain is induced by a graph. In this project, we have had a number of observations and computational studies to shed light on this notoriously difficult problem in control theory—a representative numerical example is shown in Figure 4 below.

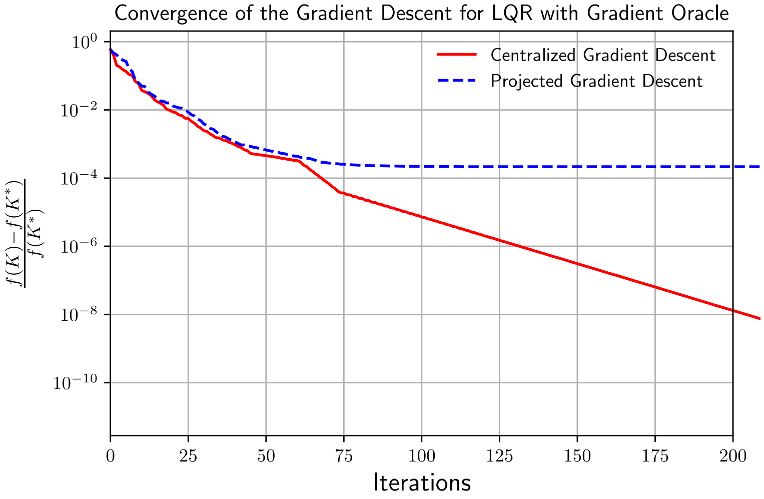


**Figure 4 Performance of Distributed Gradient Descent vs. Centralized Gradient Descent for a networked system with a specified sparsity pattern**

## 4.5 LQR with adversarial disturbances and regret

Our studies the robust control of linear dynamical systems, whereas before the linear dynamical system is governed by the dynamics equation

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

where $x_t$ is the state, $u_t$ is the control and $w_t$ is a disturbance to the system. The key differences here as follows: first, the disturbance $w_t$ may be adversarial chosen (as opposed to some known statistical model); second, at every time step $t$, the controller suffers a cost $c(x_t, u_t)$ to enforce the control (where $c$ is may be a more general convex cost as opposed to just a quadratic cost). In other words, we consider the setting of online control with arbitrary disturbances. Formally, the setting involves, at every time step $t$, an adversary selecting a convex cost function $c_t(x, u)$ and a disturbance $w_t$, and the goal of the controller is to generate a sequence of controls $u_t$ such that a sequence of convex costs $c_t(x_t, u_t)$ is minimized. This setting generalizes a fundamental problem in control theory (including the Linear Quadratic Regulator) which has been studied over several decades. However, despite the significant research literature on the problem, our generalization and results address several challenges that have remained. It is worthwhile discussing the challenges that we had to address:

*Challenge 1*. Perhaps the most important challenge we address is in dealing with arbitrary disturbances $w_t$ in the dynamics. This is a difficult problem, and so standard approaches almost exclusively assume i.i.d. Gaussian noise. Worst-case approaches in the control literature, also known as $H_\infty$-control and its variants, are overly pessimistic. Instead, we take an online (adaptive) approach to dealing with adversarial disturbances.

*Challenge 2*. Another limitation for efficient methods is the classical assumption that the costs $c(x_t, u_t)$ are quadratic, as is the case for the linear quadratic regulator. Part of the focus in the literature on the quadratic costs is due to special properties that allow for efficient computation of the best linear controller in hindsight. One of our main goals is to introduce a more general technique that allows for efficient algorithms even when faced with arbitrary convex costs.

*Our contributions*. In this work, we tackle both challenges outlined above: coping with adversarial noise, and general loss functions in an online setting. For this we turn to the time-trusted methodology of regret minimization in online learning. In the field of online learning, regret minimization is known to be more robust and general than statistical learning, and a host of convex relaxation techniques are readily available. To define the performance metric, denote for any control algorithm $A$,

$$J_T(A) = \sum_{t=1}^{T} c_t(x_t, u_t).$$

The standard comparator in control is a linear controller, which generates a control signal as a linear function of the state, i.e. $u_t = -Kx_t$. Let $J_T(K)$ denote the cost of a linear controller from a certain class $K \in Class_K$. For an algorithm $A$, we define the regret as the sub-optimality of its cost with respect to the best linear controller from a certain set

$$Regret = J_T(A) - \min_{K \in Class} J_T(K).$$

Our main result is an efficient algorithm for control which achieves regret $\mathcal{O}(\sqrt{T})$ in the setting described above. Ours is the first algorithm which achieves regret $\mathcal{O}(\sqrt{T})$ even in the presence of bounded adversarial disturbances. Previous regret bounds needed to assume that the disturbances $w_t$ are drawn from a distribution with zero mean and bounded variance. Furthermore, our regret bounds apply to any sequence of adversarially chosen convex loss functions. Previous efficient algorithms applied to convex quadratic costs only. Our results above are obtained using a host of techniques from online learning and online convex optimization, notably online learning for loss functions with memory and improper learning using convex relaxation.

The algorithm is one which can be viewed as an online optimization method, except that it takes the gradient using a memory. In fact, a notable contribution here is the use of two new proof techniques:

*Improper Policy Class*: We parameterize the policy we execute at every step as a linear function of the disturbances in the past. This leads to a convex relaxation of the problem. We avoid a linear dependence on time for the number of parameters in our policy, by additionally including a stable linear controller in our policy allowing us to effectively consider only $\mathcal{O}(\log(T))$ previous perturbations.

*A novel reduction to "online convex optimization with memory"*: The choice of the policy class with an appropriately chosen horizon $H$ allows us to reduce the problem to compete with functions with truncated memory. This naturally falls under the class of online convex optimization with memory. The key to our approach are new methods to bound the regret on truncated functions, where we use the Online Gradient Descent based approach.

## 4.6 Convergence of policy gradients and common RL algorithms in MDPs

*The softmax parameterization*: This is the most commonly used parameterization. Our work provided the first global convergence guarantees using only first-order gradient information for this widely-used parameterization. Our first result for this parameterization establishes the asymptotic convergence of the policy gradient algorithm; the analysis challenge here is that the optimal policy (which is deterministic) is attained by sending the softmax parameters to infinity. In order to establish a convergence rate to optimality for the softmax parameterization, we then consider a relative entropy regularizer and provide an iteration complexity bound that is polynomial in all relevant quantities. The use of our relative entropy regularizer is critical to avoiding collapsing gradients, an issue discussed in practice; in particular, the more general approach of entropy based regularizers is fairly common in practice.

For these aforementioned algorithms, the convergence rates depend on a certain distribution mismatch coefficient, which is the (worst case) ratio between probability that an optimal policy reaches some state $s$ in comparison to the probability of $s$ under our start state measure $\mu$. This is reason in which we seek to have a measure $\mu$ which has coverage over all the states.

We then consider the Natural Policy Gradient (NPG) algorithm [Kakade, 2002], which can be considered a quasi second-order method due to the use of its particular preconditioner, and provide an iteration complexity to achieve an $\varepsilon$-optimal policy that is polynomial in $1/\varepsilon$ and has no dependence on the number of states, the number of actions, or the distribution mismatch coefficient.

Restricted parameterizations and function approximation: We now summarize our results with regards to policy gradient methods in the setting where we work with a restricted policy class, which may not contain the optimal policy. In this sense, these methods can be viewed as approximate methods. The focus in the function approximation setting is to avoid the worst-case $L_\infty$ guarantees that are standard in approximate dynamic approaches.

We focus on average case guarantees, that support the applicability of supervised machine learning methods to solve the underlying approximation problem. This is because supervised learning methods, like classification and regression, typically only have bounds that depend on the expected error under a distribution, as opposed to worst-case guarantees over all possible inputs.

One key contribution of this work is in precisely quantifying the notion of (average case) approximation error that is relevant for policy gradient methods; for the natural gradient method, we quantify this in terms of the precisely defined regression error based on how well the policy class can approximate certain value functions, a notion related to that of compatible function approximation error [Sutton et al., 1999]. Furthermore, due to the direct nature of policy gradient methods and due to our precise quantification of approximation error, we provided finite sample and computational complexity results for the natural gradient algorithm. In particular, we provide a model-free, linear time algorithm for the natural policy gradient, requiring only simulation-based rollouts (or restarts).

Our main result is showing how the Natural Policy Gradient (NPG) has a convergence guarantee that is comparable to the Conservative Policy Iteration (CPI) [Kakade and Langford, 2002]. CPI has been the algorithm with the strongest performance guarantees to date, and our work showed how the NPG has the same guarantee. This is perhaps surprising since the NPG is the a widely used algorithm, which has not had any performance analysis. One significant advantage of NPG over CPI is that the explicit parametric policy representation in NPG (and other policy gradient methods) leads to a succinct policy representation in comparison to CPI or related boosting-style methods, where the representation complexity of the policy of the latter class of methods grows linearly in the number of iterations (since these methods add one policy to the ensemble per iteration). This increased representation complexity is likely why CPI is less widely used in practice.

### 4.7 Maximum Entropy Exploration and Curiosity Driven Learning

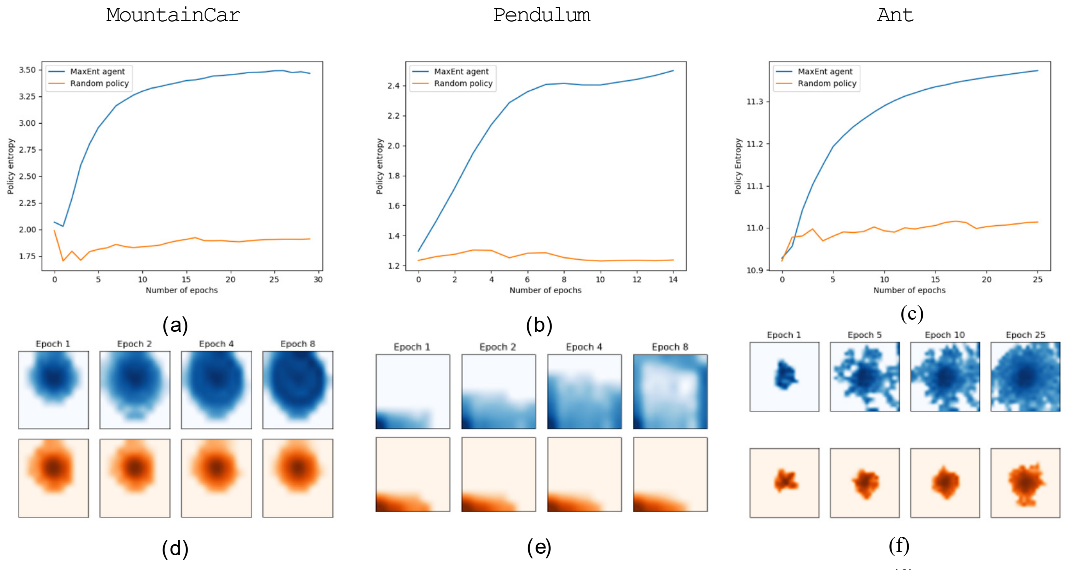
The goal here is for the agent to discover what it is capable of doing, without given any explicit reward signal. We provide an efficient algorithm to optimize such intrinsically defined objectives, when given access to a black box planning oracle (which is robust to function approximation). Furthermore, when restricted to the tabular setting where we have sample based access to the MDP, our proposed algorithm is provably efficient, both in terms of its sample and computational complexities. Key to our algorithmic methodology is utilizing the conditional gradient method (a.k.a. the Frank-Wolfe algorithm) which utilizes an approximate MDP solver.

To facilitate exploration in potentially unknown MDPs within a restricted policy class, we assume access to the environment using the following two oracles:

*Approximate planning oracle*: Given a reward function (on states) $r : S \to R$ and a sub-optimality gap $\varepsilon$, the planning oracle returns a stationary policy $\pi = ApproxPlan(r, \varepsilon)$ with the guarantee that $V(\pi) \geq \max_\pi V(\pi) - \varepsilon$, where $V(\pi)$ is the value of policy $\pi$.

*State distribution estimate oracle*: A state distribution oracle estimates the state visitation distribution (the frequencies with which a policy visits the states in the MDP), $d_\pi = DensityEst(\pi, \varepsilon)$ of any given (non-stationary) policy $\pi$, guaranteeing that $\|d_\pi - d_\pi\|_\infty \leq \varepsilon$.

Given access to these two oracles, we describe a method that provably optimizes any continuous and smooth objective over the state-visitation frequencies. Of special interest is the maximum entropy and relative entropy objectives. Our main result provides an efficient algorithm such that for any $\beta$-smooth objective function $R$, and any $\varepsilon > 0$, in $O(1/\varepsilon \log 1/\varepsilon)$ calls to *ApproxPlan* and *DensityEst*, it returns a policy $\pi$ with $R(d_\pi) \geq \max_\pi R(d_\pi) - \varepsilon$. Some sample results obtained using the maximum entropy agent are shown in Figure 5.



(a) (b) (c)

(d) (e) (f)

**Figure 5 Comparison of the maximum entropy agent to a random baseline policy for three sample problems**

*In each plot, blue represents the MaxEnt agent, and orange represents the random baseline. (a), (b), and (c) show the entropy of the policy evolving with the number of epochs. (d), (e), and (f) show the log-probability of occupancy of the two-dimensional state space. In (f), the infinite x-y grid is limited to the range [-20,20]×[-20,20]*

# 5 CONCLUSIONS

The objective of this project has been to build an overarching bridge between these two lines of work, namely, between optimal control theory and sample-based reinforcement learning methods. This has been accomplished in the context of two pillars of control theory and learning, namely, linear quadratic regulator problem and Markov decision processes. It has been shown that first order methods can effectively be used to provide a bridge between the two disciplines by clearly highlighting the interplay between modeling, data-driven decision making, statistical reasoning, and uncertainty. As such, this research has opened up a number of intriguing directions at the interface of control theory and learning that need to explored further in the coming years.

# 6 REFERENCES

1. Bu, Jingjing, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. 2019. "LQR through the Lens of First Order Methods: Discrete-Time Case," http://arxiv.org/abs/1907.08921.

2. Fazel, Maryam, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. 2018. "Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator." In Proceedings of the 35th International Conference on Machine Learning. https://doi.org/10.1089/end.2008.0274.

3. Hewer, G. 1971. "An Iterative Technique for the Computation of the Steady State Gains for the Discrete Optimal Regulator." IEEE Transactions on Automatic Control 16 (4): 382–84.

4. Polyak, B., M. Khlebnikov, and P. Shcherbakov. 2013. "An LMI Approach to Structured Sparse Feedback Design in Linear Control Systems." In 2013 European Control Conference (ECC), 833–38.

5. Bu, Jingjing, Afshin Mesbahi, and Mehran Mesbahi. 2019. "On Topological and Metrical Properties of Stabilizing Feedback Gains: The MIMO Case." arXiv [cs.SY]. arXiv. http://arxiv.org/abs/1904.02737.

6. Mesbahi, Mehran, and Magnus Egerstedt. 2015. "Graphs for Modeling Networked Interactions." Encyclopedia of Systems and Control. https://doi.org/10.1007/978-1-4471-5058-9_212.

7. Agarwal, Naman, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. 2019. "Online Control with Adversarial Disturbances." arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1902.08721.

8. Hazan, Elad, and Sham Kakade. 2019. "Revisiting the Polyak Step Size." arXiv [math.OC]. arXiv. http://arxiv.org/abs/1905.00313.

9. Jin, Chi, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. 2019. "Stochastic Gradient Descent Escapes Saddle Points Efficiently." arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1902.04811.

10. Kakade, Sham M. 2002. "A Natural Policy Gradient." In Advances in Neural Information Processing Systems 14, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani, 1531–38. MIT Press.