Learning Optimal Controllers by Policy Gradient: Global Optimality via Convex Parameterization

Yue Sun and Maryam Fazel

***Abstract*— Common reinforcement learning methods seek optimal controllers for unknown dynamical systems by searching in the "policy" space directly. A recent line of research, starting with [1], aims to provide theoretical guarantees for such direct policy-update methods by exploring their performance in classical control settings, such as the infinite horizon linear quadratic regulator (LQR) problem. A key property these analyses rely on is that the LQR cost function satisfies the "gradient dominance" property with respect to the policy parameters. Gradient dominance helps guarantee that the optimal controller can be found by running gradient-based algorithms on the LQR cost. The gradient dominance property has so far been verified *on a case-by-case basis* for several control problems including continuous/discrete time LQR, LQR with decentralized controller, $\mathcal{H}_2/\mathcal{H}_\infty$ robust control.**

In this paper, we make a connection between this line of work and classical convex parameterizations based on linear matrix inequalities (LMIs). Using this, we propose a unified framework for showing that gradient dominance indeed holds for a broad class of control problems, such as continuous- and discrete-time LQR, minimizing the L_2 gain, and problems using system-level parameterization. Our unified framework provides insights into the landscape of the cost function as a function of the policy, and enables extending convergence results for policy gradient descent to a much larger class of problems.

I. Introduction

Linear quadratic regulator (LQR) is one of the most well studied optimal control problems for decades [2]. Consider the continuous time linear time-invariant dynamical system,

$$\dot{x} = Ax + Bu, \;\; x(0) = x_0, \tag{1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^p$ is the input, and A, B are constant matrices describing the dynamics. The goal of optimal control is to determine the input series $u(t)$ that minimizes some cost function that typically depends on state and input. In the infinite horizon LQR problem, with constant matrices $Q \in \mathbf{S}^n_{++}, R \in \mathbf{S}^p_{++}$, one minimizes

$$\text{loss}(u(t)) := \mathbf{E}_{x_0} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t)) dt \tag{2}$$

It is known that the optimal controller is linear in the state, referred to as static state feedback, and can be described as $u(t) = Kx(t)$ for a constant $K \in \mathbb{R}^{p\times n}$ [2]. This can be obtained by solving the algebraic Riccati equation (ARE) [3], [4]. A large number of works have studied the solution of ARE, including approaches based on iterative algorithms [5], algebraic solution methods [6], and semidefinite programming [7]. However, this approach is in sharp contrast to how one would typically minimize a cost function through gradient descent on K, usually used in reinforcement learning settings.

In many practical cases, the system dynamics is unknown, and among the optimal control algorithms, there are two major types. The first type is model based methods, when the system is first identified and then a controller is constructed based on the identified system. System identification has a long history, as reviewed [8]. Recently [9] gave sample complexity bounds for state-observed system. [10]–[13] describe the joint system identification and optimal control approaches.

Another type of method is model free method, when the controller is directly trained by observing the cost (or loss) function, without characterizing the dynamics. Here one does not necessarily estimate the system parameters A, B. [14] is a review of reinforcement learning area and optimal control and studies some fixed point type dynamic programming methods. Q-learning is a typical model free method for reinforcement learning, and it is applied to LQR as in [10], [15], [16].

This paper mainly focus on another model-free method called policy gradient descent. It calls for an estimate of the cost (2) as well as its gradient with respect to controller K when $u = Kx$. One hopes that gradient descent with respect to K converges to the optimal controller K^*. The policy gradient descent is more recently reviewed by [17], [18]. [1] provides a counterexample showing that minimizing the quadratic LQR cost as a function of K is not convex, quasi-convex or star-convex.

Recently people have witnessed the *empirical* success of first order methods in solving nonconvex reinforcement learning problems. [19, Ch. 3] proposes the gradient based method for optimal control and extends to decentralized control. [20] studies feedback control with dynamical controllers, and observes that gradient descent with Youla parameterization is robust within the set of stabilizing controllers while other parameterizations are not. On *theoretical* side, [1] gives the first result by proving the coercivity and gradient dominance property of $\mathcal{L}(K)$ for the discrete time LQR. Based on this, [1] shows the linear convergence of gradient based method. Later [21] shows a similar result for the continuous time case, [22], [23] give a more detailed analysis for both discrete and continuous time LQR. [24] and [25] shows similar results for two settings of zero-sum LQ games. [26] studies the convergence of gradient descent on $\mathcal{H}_2$ control with $\mathcal{H}_\infty$ constraint and shows that gradient descent implicitly makes the controller robust. [27] shows the convergence for finite-horizon distributed control under the quadratic invariance assumption. Those papers all show convergence of policy gradient descent by gradient dominance property,

Y. Sun (yuesun@uw.edu) and M. Fazel (mfazel@uw.edu) are with Department of Electrical and Computer Engineering, University of Washington, Seattle, USA.

but investigate different control problems and the proofs are given case by case.

Traditionally, convex parameterization (convexification) such as Youla parametrization, Q-parameterization, or the more recent System Level Synthesis (SLS) have allowed the reformulation of certain control design problems as semidefinite programs. In this paper, we are interested to see if these methods can help us distill the essence of the gradient dominance property of the original control problem that is nonconvex in K.

Control for nonlinear systems is far more difficult, typically via dynamic programming, solving Bellman equations [28], or recent deep RL that led to empirical success in control of complex systems. Yet it is still mysterious how deep learning models work in this context, and recent theoretical studies have focused on linear systems in hope of providing insights into more complex cases.

Contributions: In this paper, we will build a bridge between nonconvex policy gradient descent and known convex parameterization methods, which provides insight into why convergence to the optimal solution happens despite nonconvexity in all the problems cited above. We use a mapping between the landscape of convex and nonconvex objectives, and use this mapping to prove the gradient dominance property of the nonconvex objective under reasonable assumptions.

Our result is quite general—we show that continuous time LQR is a special case that our theorem applies to, and we generalize the guarantees provided by this method to a range of other control problems including instances of optimal control, robust control, mixed design and system level synthesis (some are in appendix [29] due to space limit). Thus for all these problems, if one wants to understand whether the (nonconvex) loss with respect to controller parameter K can be minimized by policy gradient descent (first-order optimization methods that update K), one can directly check if it is covered by our theorem, avoiding a case-by-case analysis. Also, as discussed in [1], theoretical guarantees for first-order methods naturally lead to guarantees for the more practical zeroth-order optimization or sampling-based methods, which do not need access to the gradient of the cost with respect to K.

The rest of this paper is structured as follows. Sec. II reviews the continuous-time LQR problem. Sec. III presents our main result on the gradient dominance property for the nonconvex loss. Sec. IV lists more examples of control problems covered by the main theorem. Sec. V gives a proof sketch with intuitive connections between the nonconvex and convex formulations. The detailed proof appears in [29].

II. Review of convex parameterization for continuous time LQR

Convexification method (e.g., solving optimal control by linear matrix inequalities (LMI) in [30]) is widely used in optimal control problems, and here we discuss its application for continuous time LQR [21]. Define a continuous time linear time invariant system (1) where x is state and u is input signal, and x_0 is the initial state. We assume that $\mathbf{E}(x_0 x_0^\top) = \Sigma \succ 0$. This is a commonly used setup such as in [22, §3.3], [19, Paper 3].

One can then consider minimizing the linear quadratic (LQ) loss (2) as a function of $u(t)$ where Q, R are positive definite matrices. It is known [2] that, the input signal that minimizes the loss function $\mathrm{loss}(u)$ is given by a static state feedback controller, denoted by $u(t) = K^* x(t)$. K^* can be obtained by solving linear equations, called riccati equations. Note that once we know the optimal state feedback controller is static, we can write loss as $\mathcal{L}(K)$ which is a function of K instead, and search only in static state feedback controllers.

An alternative approach is reparameterizing to obtain a convex formulation, as used in [21], which we will review here, starting from the Lyapunov equation. Suppose the initial state satisfies $\mathbf{E}(x_0 x_0^\top) = \Sigma \succ 0$, and $\dot{x} = Ax$. Then with a matrix $P \in \mathbf{S}_{++}^{n\times n}$ (P is a positive definite matrix) as the variable, the Lyapunov equation is written as

$$AP + PA^\top + \Sigma = 0 \tag{3}$$

In our setup (1), we use a state feedback controller $u = Kx$, thus we have $\dot{x} = (A+BK)x$. We denote the set of stabilizing controllers as $\mathcal{S}_{K,\mathrm{sta}}$, which is defined as

$$\mathcal{S}_{K,\mathrm{sta}} = \{K : \ \mathrm{Re}(\lambda_i(A+BK)) < 0, \ i = 1, ..., n\}.$$

If a state feedback controller is applied, the loss is only bounded when $K \in \mathcal{S}_{K,\mathrm{sta}}$ and is coersive in $\mathcal{S}_{K,\mathrm{sta}}$ [23]. Replace A by the closed loop system matrix $A + BK$ in the Lyapunov equation, and let $L = KP \in \mathbb{R}^{p\times n}$, we get

$$AP + PA^\top + BL + L^\top B^\top + \Sigma = 0$$

Let $\mathcal{A}(P) = AP + PA^\top, \ \mathcal{B}(L) = BL + L^\top B^\top$, which are often referred to as Lyapunov maps. Assume $\mathcal{A}$ is invertible, then we have the relation

$$\mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0. \tag{4}$$

Indeed, once we fix the system and any stabilizing controller A, B, K, the matrices P as well as $L = KP$ are uniquely determined. P is the Grammian matrix

$$P = \int_0^\infty e^{t(A+BK)} \ \Sigma \ e^{t(A+BK)^\top} \ dt. \tag{5}$$

P is positive definite if $\Sigma \succ 0$. We are interested in the loss function $\mathcal{L}(K)$ when[1] $K \in \mathcal{S}_{K,\mathrm{sta}}$, which corresponds to (2) by inserting $u(t) = Kx(t)$.

$$\mathcal{L}(K) = \mathbf{Tr}((Q + K^\top RK)P). \tag{6}$$

One can construct a bijection from P, L to K, and prove that, if we minimize $f(L, P)$ subject to (4), the optimizer P^*, L^* will map to the optimal K^*, and this minimization problem is convex, so we can solve it by convex optimization algorithms.

Convex reparameterization for Continuous time LQR: Suppose the dynamics and loss are (1) and (2), and let $\mathbf{E}(x_0 x_0^\top) = \Sigma \succ 0$. Denote the (static) state feedback

[1] If K is not a stabilizing controller, we define $\mathcal{L}(K) = +\infty$.

controller by K, so that $u(t) = Kx(t)$. The optimal control problem then is

$$\min_K \ \mathcal{L}(K), \quad \text{s.t.} \quad K \in \mathcal{S}_{K,\text{sta}} \tag{7}$$

where $\mathcal{L}(K)$ is the cost in (2) with $u = Kx$. This problem can be expressed as the following equivalent convex problem,

$$\min_{L,P,Z} \ f(L,P,Z) := \mathbf{Tr}(QP) + \mathbf{Tr}(ZR) \tag{8a}$$

$$\text{s.t. } \mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0, \ P \succ 0, \tag{8b}$$

$$\begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0 \tag{8c}$$

The connection between the two problems is distilled in Sec. III. For all feasible (L, P, Z) triplets in (8), we can take the first two elements (L, P), and they form a bijection with all stabilizing controllers K in (7). The loss function values are equal under the bijection. So we can solve for L^*, P^*, and $K^* = L^*(P^*)^{-1}$.

III. Main result

Motivated by methods that use gradient descent in the policy space, we ask whether running a gradient-based algorithm and getting $\nabla_K \mathcal{L}(K) = 0$ for some K in fact gives the globally optimum K^*. [1], [21] show the coercivity and gradient dominance property of $\mathcal{L}(K)$ for the discrete time and continuous time LQR respectively. In this paper, we generalize these results from the special case of continuous-time LQR to a much broader set of control problems, showing the gradient dominance property of the nonconvex losses as functions of policy.

We present our main result in Theorem 1. It is described as a pair of problems satisfying Assumptions 1, 3, which covers problems extending beyond continuous time LQR. In Sec. IV we will review more examples showing the generality of this result.

We begin by considering an abstract description of the pair of problems (7) and (8). These problem descriptions cover LQR as discussed in the last section, as well as more problems discussed in Sec. IV. Consider the problems

$$\min_K \quad \mathcal{L}(K), \quad \text{s.t. } K \in \mathcal{S}_K, \tag{9}$$

and

$$\min_{L,P,Z} \quad f(L,P,Z), \quad \text{s.t.} \quad (L,P,Z) \in \mathcal{S}, \tag{10}$$

where the sets $\mathcal{S}_K, \mathcal{S}$ capture the control constraints. They are defined differently for each specific example in Sec. IV. For example, for continuous time LQR, $\mathcal{S}_K$ is the set of all stabilizing controllers (7) and $\mathcal{S}$ is the intersection of (8b) & (8c). We allow special cases when (10) depends only on L, P,

$$\min_{L,P} \quad f(L,P), \quad \text{s.t.} \quad (L,P) \in \mathcal{S} \tag{11}$$

We distill three properties of the two problems (9) and (10) that will be critical for Theorem 1, and allow us to cover more problems as discussed in Sec. IV.

Assumption 1. *The feasible set $\mathcal{S}$ is convex in (L, P, Z). The cost function $f(L, P, Z)$ is convex, bounded, and differentiable in $(L, P, Z) \in \mathcal{S}$.*

Assumption 1 imply the second problem is convex. Next, we extract the property of the connection between (7) and (8), and give an abstract description of the assumptions for (9) and (10).

Assumption 2. *Let P be always invertible[2] in $\mathcal{S}$. Assume we can express $\mathcal{L}(K)$ as:*

$$\mathcal{L}(K) = \min_{L,P,Z} \ f(L,P,Z) \quad \text{s.t. } (L,P,Z) \in \mathcal{S}, \ LP^{-1} = K.$$

With the assumptions above, we will present the main theorem.

Theorem 1. *We consider the problems* (9) *and* (10)*, and we require Assumptions 1,2. Let K^* denote the global minimizer of $\mathcal{L}(K)$ in S_K. Then there exist constants $C_1, C_2 > 0$ independent of K,*

1) if f is convex, the gradient of $\mathcal{L}$ satisfies[3]

$$\|\nabla \mathcal{L}(K)\|_F \geq C_1(\mathcal{L}(K) - \mathcal{L}(K^*)). \tag{12}$$

2) if f is μ-strongly convex, the gradient satisfies

$$\|\nabla \mathcal{L}(K)\|_F \geq C_2(\mu(\mathcal{L}(K) - \mathcal{L}(K^*)))^{1/2}. \tag{13}$$

The constants C_1, C_2 are discussed below and in [29].

Remark 1. *The constants are case by case. We show that, for continuous time LQR, in the sublevel set where $\mathcal{L}(K) \leq a$, we define*

$$\nu = 4a\left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^2,$$
$$C_{1,1} = 2a\nu\lambda_{\min}^{-1}(\Sigma)\lambda_{\min}^{-1/2}(Q)\lambda_{\min}^{-1/2}(R),$$
$$C_{1,2} = 2a^2\nu^2\lambda_{\min}^{-2}(\Sigma)\lambda_{\min}^{-3/2}(Q)\lambda_{\min}^{-1/2}(R),$$

Then $C_1 = (\max\{C_{1,1}, C_{1,2}\})^{-1}$. [21] gives another convex formulation with strong convexity and we can get C_2 for that form, the details are in [29].

Our lower bound of the gradient, $\|\nabla \mathcal{L}(K)\|_F \gtrsim (\mathcal{L}(K) - \mathcal{L}(K^*))^\alpha$, is known as Lojasiewicz inequality [31]. When $\alpha = 1/2$, it is also called the *gradient dominance* property. If Lojasiewicz inequality holds, all local minima of the objective function are global minima, then an iterative method with $\|\nabla \mathcal{L}(K)\|_F \to 0$ makes the iterates converge to the global minimum.

Assumption 2 is a rather weak assumption. Assumption 3 is a stronger one covered by Assumption 2 that, we assume

[2]The invertibility of P guarantees a well defined map between L, P and K, which is usually true, e.g., for the instances in Sec. IV.

[3]We always consider the directional derivative of a feasible direction within descent cone.

that there is a bijection between K and (L, P). This is true for many control problems such as continuous time LQR. We emphasize the special case for Assumptions 1,3 since it is easy to illustrate in Sec. V.

Assumption 3. *1) (Bijection between two feasible sets) Let P be invertible, $K = LP^{-1}$ define a bijection*[4] *$K \leftrightarrow (L, P)$. For any such bijection $K \leftrightarrow (L, P)$, $\exists Z$, such that $(L, P, Z) \in \mathcal{S}$.*

2) (Equivalence of functions) Choose a controller $K \in \mathcal{S}_K$ with corresponding $(L, P) \in \mathcal{S}$. Then $\mathcal{L}(K) = \min_Z f(L, P, Z)$ subject to $(L, P, Z) \in \mathcal{S}$.

Our main theorem suggests that, when the original nonconvex optimization problem can be mapped to a convex optimization problem that satisfies Assumptions 1,2 or 1,3, all stationary points of the nonconvex objective are global minima. So if we can evaluate the gradient of nonconvex objective and run gradient descent algorithm, the iterates converge to the optimal controller.

IV. Optimal control problems covered by main theorem

In order to reach the conclusion, Theorem 1 requires an optimal control problem (9), its convexified form (10) and a few assumptions. This is an abstract and general description that does not need the exact continuous time LQR formulation in Sec. II. We can easily check that the continuous time LQR satisfies the Assumptions 1,3, thus we can directly apply Theorem 1 to argue that the continuous time LQR cost $\mathcal{L}(K)$ satisfies (12).

Below (more examples in [29]), we will list some examples to show that Theorem 1 covers a wide range of optimal control problems. This shows the **generality** of Theorem 1. They can be optimized by policy gradient descent.

A. Discrete time LQR

We consider a discrete time linear system

$$x(t+1) = Ax(t) + Bu(t),\ x(0) = x_0, \tag{14}$$

The goal is to find a state feedback controller K minimizing

$$\mathcal{L}(K) = \mathbf{E}_{x_0} \sum_{i=0}^{\infty} x(t)^\top Q x(t) + u(t)^\top R u(t),\ u = Kx$$

Similar to the continuous time system, one can choose the same parameterization P, L, Z and another PSD matrix $G \in \mathbb{R}^{n\times n} \succeq 0$ and solve the following problem

$$\min_{L,P,Z,G}\ f(L, P, Z, G) := \mathbf{Tr}(QP) + \mathbf{Tr}(ZR) \tag{15a}$$

$$\text{s.t. } P \succ 0,\ G - P + \Sigma = 0, \tag{15b}$$

$$\begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0,\ \begin{bmatrix} G & AP+BL \\ (AP+BL)^\top & P \end{bmatrix} \succeq 0 \tag{15c}$$

The goal is to argue that $\mathcal{L}(K)$ and (15) has the connection such that Theorem 1 applies, so that the stationary point of $\mathcal{L}(K)$ has to be the global optimum.

[4]Note that generally $K = LP^{-1}$ cannot guarantee a bijection. However bijection is possible with the extra constraint $(L, P) \in \mathcal{S}$.

Lemma 1. *The LQR problem* $\min\ \mathcal{L}(K)$ *with stabilizing K, and problem* (15)*, satisfy Assumption 1, 2.*

Proof. (15) is a convex optimization problem. Now we prove Assumption 2, i.e., we prove that $L(K)$ equals the minimum of the problem (15) with an extra constraint $K = LP^{-1}$.

- We first minimize over Z, the minimizer is $Z = LP^{-1}L^\top$. Now replace L by KP and the loss becomes $\mathbf{Tr}((Q + K^\top RK)P)$.
- Eliminate G by

$$\begin{bmatrix} P - \Sigma & AP+BL \\ (AP+BL)^\top & P \end{bmatrix} \succeq 0$$

Using Schur complement, it is equivalent to

$$(AP+BL)P^{-1}(AP+BL)^\top - P + \Sigma \preceq 0$$

Plug in $L = KP$, we have

$$(A+BK)P(A+BK)^\top - P + \Sigma \preceq 0.$$

The loss does not involve G so it does not change.

- Now, we need to prove that $\mathcal{L}(K)$ is equal to

$$\begin{aligned} &\min_P\ \mathbf{Tr}((Q + K^\top RK)P) \\ &\text{s.t. } (A+BK)P(A+BK)^\top - P + \Sigma \preceq 0. \end{aligned} \tag{16}$$

The constraint (16) can be written as

$$(A+BK)P(A+BK)^\top - P + \Theta = 0,\ \Theta \succeq \Sigma.$$

- Denote the solution to $(A+BK)P(A+BK)^\top - P + \Theta = 0$ as $P(\Theta)$. $P(\Theta)$ for all $\Theta \succeq \Sigma$ covers the feasible points of (16). $P(\Theta)$ is expressed as:

$$P(\Theta) = \sum_{t=0}^{\infty} (A+BK)^t \Theta ((A+BK)^\top)^t$$

So $P(\Theta) \succeq P(\Sigma)$, $\forall \Theta \succeq \Sigma$. Since Q and $K^\top RK$ are positive semidefinite, $\mathbf{Tr}((Q + K^\top RK)P)$ achieves the minimum at $P = P(\Sigma)$.
- At the end, $P(\Sigma)$ is the Grammian $\mathbf{E}\sum_{t=0}^{\infty} x(t)x(t)^\top$ when $\mathbf{E}x(0)x(0)^\top = \Sigma$. We studied the connection between continous time Grammian (5) and the loss (6), a similar result holds for discrete time LQR:

$$\mathbf{Tr}((Q + K^\top RK)P(\Sigma)) = \mathcal{L}(K).$$

We build the connection between minimizing $\mathcal{L}(K)$, and the convex optimization (15). We argued this pair of problems satisfies the assumptions of Theorem 1. Theorem 1 suggests that $\mathcal{L}(K)$ is gradient dominant, so we can approach K^* by gradient descent on K. This is essentially the conclusion of [1], [22]. Note that the proof of discrete time LQR [1], [22] and continuous time LQR [21], [23] cannot trivially extend to each other.

B. Minimizing L_2 gain

We quote from [30] the problem of minimizing the L_2 gain with static state feedback controller K. As discussed in [30, §6.3.2], this problem has an associated convex optimization problem and we can show it satisfies Assumption 1,2.

We consider minimizing the L_2 gain of a closed loop system. The continuous time linear dynamical system is

$$\dot{x} = Ax + Bu + B_w w,\ y = Cx + Du \tag{17}$$

For any signal z, denote

$$\|z\|_2 := \left(\int_0^\infty \|z(t)\|_2^2 dt\right)^{1/2}$$

Suppose we use a state feedback controller $u = Kx$, and aim to find the optimal controller K^* that minimizes the L_2 gain. We minimize the squared L_2 gain as

$$\min_K\ \mathcal{L}(K) := (\sup_{\|w\|_2=1} \|y\|_2)^2.$$

This problem can be further reformulated as [30, §7.5.1]

$$\begin{aligned}&\min_{L,P,\gamma}\ f(L,P,\gamma) := \gamma,\ \text{s.t.}\\ &\begin{bmatrix} AP + PA^\top + BL + L^\top B^\top + B_w B_w^\top & (CP+DL)^\top \\ CP + DL & -\gamma I\end{bmatrix}\\ &:= M(L,P,\gamma) \preceq 0.\end{aligned} \tag{18}$$

The minimum L_2 gain is $\sqrt{\gamma^*}$ and $K^* = L^* P^{*-1}$. We will show in [29] that the parameters K and (L, P, γ), with loss $\mathcal{L}(K)$ and $f(L, P, \gamma)$, satisfy Assumptions 1,2. Thus we can claim that all stationary points of $\mathcal{L}(K)$ are global minimum.

[30, §6.3.2] suggests that L_2 gain is also the $\mathcal{H}_\infty$ norm of transfer function, so it covers the instances in [26]. We discuss this further in [29].

C. System level synthesis (SLS) for finite horizon time varying discrete LQR

Different from the previous examples, we consider a time varying system in a finite horizon, where we seek to design a time varying controller. This problem and its convex parameterization are introduced in [32]. It satisfies Assumption 1,3. We consider the following linear dynamical system

$$x(t+1) = A(t)x(t) + B(t)u(t) + w(t) \tag{19}$$

over a finite horizon $0, \dots T$. Let the state be x and the input be u. Define

$$X = \begin{bmatrix} x(0) \\ ... \\ x(T)\end{bmatrix},\ U = \begin{bmatrix} u(0) \\ ... \\ u(T)\end{bmatrix},$$

$$W = \begin{bmatrix} x(0) \\ w(0) \\ ... \\ w(T-1)\end{bmatrix}, Z = \begin{bmatrix} 0 & 0 & ... & 0 & 0 \\ I & 0 & ... & 0 & 0 \\ 0 & I & ... & 0 & 0 \\ ... & & & & \\ 0 & 0 & ... & I & 0\end{bmatrix},$$

$$\mathcal{A} = \text{diag}(A(0), ..., A(T-1), 0),$$
$$\mathcal{B} = \text{diag}(B(0), ..., B(T-1), 0)$$

Now we consider the time varying controller K that links state and input as

$$u(t) = \sum_{i=0}^{t} K(t, t-i)x(i) \tag{20}$$

and let

$$\mathcal{K} = \begin{bmatrix} K(0,0) & 0 & ... & 0 \\ K(1,1) & K(1,0) & ... & 0 \\ ... & & & \\ K(T,T) & K(T,T-1) & ... & K(T,0)\end{bmatrix}$$

We will minimize some loss function with the constraint. For example, in the discrete time LQR regime, let the input be (20) and define (More examples of nonquadratic cost in [32, §2.2])

$$\mathcal{L}(\mathcal{K}) = \sum_{t=0}^{T} x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t), \tag{21}$$

here $Q(t), R(t) \succeq 0$. We will minimize $\mathcal{L}(\mathcal{K})$ where $\mathcal{K}$ is the variable.

Convex problem: The dynamics (19) can be written as

$$X = Z\mathcal{A}X + Z\mathcal{B}U + W = Z(\mathcal{A} + \mathcal{B}\mathcal{K})X + W$$

We define the mapping from W to X, U by

$$\begin{bmatrix} X \\ U\end{bmatrix} = \begin{bmatrix}\Phi_X \\ \Phi_U\end{bmatrix} W.$$

where Φ_X, Φ_U are block lower triangular. There is a constraint on Φ_X, Φ_U:

$$\begin{bmatrix} I - Z\mathcal{A} & -Z\mathcal{B}\end{bmatrix}\begin{bmatrix}\Phi_X \\ \Phi_U\end{bmatrix} = I. \tag{22}$$

It is proven in [32, Thm 2.1] that $\mathcal{K} = \Phi_U \Phi_X^{-1}$. $\mathcal{K}$ and Φ_X, Φ_U is a bijection given Φ_X, Φ_U satisfying (22).

Let $\mathcal{Q} = \text{diag}(Q(0), ..., Q(T))$, $\mathcal{R} = \text{diag}(R(0), ..., R(T))$, the LQR loss with $x(0) \sim \mathcal{N}(0, \Sigma)$ and no noise is

$$f(\Phi_X, \Phi_U) = \left\|\text{diag}(\mathcal{Q}^{1/2}, \mathcal{R}^{1/2})\begin{bmatrix}\Phi_X(:,0) \\ \Phi_U(:,0)\end{bmatrix}\Sigma^{1/2}\right\|_F^2$$

the LQR loss with $x(0), w(t)$ being i.i.d from $\mathcal{N}(0, \Sigma)$ is

$$f(\Phi_X, \Phi_U) = \left\|\text{diag}(\mathcal{Q}^{1/2}, \mathcal{R}^{1/2})\begin{bmatrix}\Phi_X \\ \Phi_U\end{bmatrix}\Sigma^{1/2}\right\|_F^2.$$

If we solve $\min_{\mathcal{K}}\ \mathcal{L}(\mathcal{K})$, $\mathcal{K}$ being lower left triangular, with the above two models of $w(t)$, both can be minimized with constraint (22):

$$\min_{\Phi_X, \Phi_U}\ f(\Phi_X, \Phi_U),\ \text{s.t.}\ \begin{bmatrix} I - Z\mathcal{A} & -Z\mathcal{B}\end{bmatrix}\begin{bmatrix}\Phi_X \\ \Phi_U\end{bmatrix} = I,$$
$$\Phi_X, \Phi_U \text{ are lower left triangular}$$

This problem is convex and satisfy Assumption 1. [32, Thm 2.1] suggests the relation between $\mathcal{L}$ and f satisfying the Assumption 3 for Theorem 1. With Theorem 1, we can argue that all stationary points of $\mathcal{L}(\mathcal{K})$ are global minimum.

V. Proof Sketch

The full proof of Theorem 1 is in [29], and this section is a sketch of the proof. We illustrate the idea in Figure 1, which, on the high level, maps the loss function in original space of controller K where the loss is nonconvex, and the parameterized space with L, P, Z where the loss is convex.

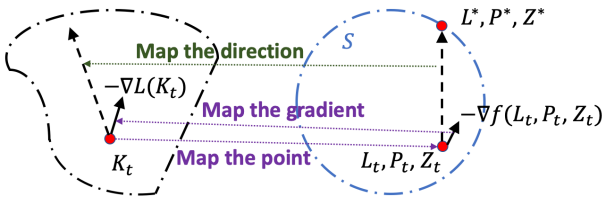

Fig. 1: Mapping between nonconvex and convex landscapes. Suppose we run gradient descent at iteration t, for any controller K, we can map it to L, P, Z in the other parameterized space. and then we map the direction $(L^*, P^*, Z^*) - (L, P, Z)$ and the gradient $\nabla f(L, P, Z)$ back to the original K space. Since in (L, P, Z) space the loss is convex, then $\langle \nabla f(L, P, Z), (L^*, P^*, Z^*) - (L, P, Z) \rangle < 0$. We prove that similar correlation holds for the nonconvex objective.

For simplicity, we sketch the proof using Assumptions 1,3. For any point K, we can find a point (L, P, Z) in the parameterized space. If it is not the optimizer, we can find the line segment linking (L, P, Z) and the optimizer (L^*, P^*, Z^*). Note that the optimization problem is convex in this space so that $\langle \nabla f(L, P, Z), (L^*, P^*, Z^*) - (L, P, Z) \rangle$ is upper bounded by $f(L^*, P^*, Z^*) - f(L, P, Z)$. Then with the help of our assumptions, we can map the directional derivative back to the original K space, and show that the directional derivative in $\mathcal{L}(K)$ is not 0.

Before concluding, we remark that the assumptions in Theorem 1 come from an optimization theory perspective, and we do not dive into the control theoretic interpretations of the constants and assumptions. Our approach has the benefit that it unifies the analysis of many control problems in a single abstract result. We leave it to future work to refine the analysis to obtain the best case-specific convergence rates, and to provide an interpretation of the associated constants in terms of control theoretic notions.

References

[1] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for linearized control problems," *arXiv preprint arXiv:1801.05039*, 2018.

[2] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[3] R. F. Stengel, *Optimal control and estimation*. Courier Corporation, 1994.

[4] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. Springer Science & Business Media, 2013, vol. 36.

[5] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[6] P. Lancaster and L. Rodman, *Algebraic riccati equations*. Clarendon press, 1995.

[7] V. Balakrishnan and L. Vandenberghe, "Semidefinite programming duality and linear time-invariant systems," *IEEE Transactions on Automatic Control*, vol. 48, no. 1, pp. 30–41, 2003.

[8] L. Ljung, "System identification: theory for the user," *PTR Prentice Hall, Upper Saddle River, NJ*, pp. 1–14, 1999.

[9] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*, 2018, pp. 439–473.

[10] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proceedings of 1994 American Control Conference-ACC'94*, vol. 3. IEEE, 1994, pp. 3475–3479.

[11] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

[12] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 5582–5588.

[13] H. Mania, S. Tu, and B. Recht, "Certainty equivalent control of LQR is efficient," *arXiv preprint arXiv:1902.07826*, 2019.

[14] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[15] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.

[16] D. Lee and J. Hu, "Primal-dual q-learning framework for LQR design," *IEEE Transactions on Automatic Control*, vol. 64, no. 9, pp. 3756–3763, 2018.

[17] S. M. Kakade, "A natural policy gradient," in *Advances in neural information processing systems*, 2002, pp. 1531–1538.

[18] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade, "Towards generalization and simplicity in continuous control," in *Advances in Neural Information Processing Systems*, 2017, pp. 6550–6561.

[19] K. Mårtensson, "Gradient methods for large-scale and distributed linear quadratic control," Ph.D. dissertation, Lund University, 2012.

[20] J. W. Roberts, I. R. Manchester, and R. Tedrake, "Feedback controller parameterizations for reinforcement learning," in *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 2011, pp. 310–317.

[21] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 7474–7479.

[22] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[23] J. Bu, A. Mesbahi, and M. Mesbahi, "Policy gradient-based algorithms for continuous-time linear quadratic control," *arXiv preprint arXiv:2006.09178*, 2020.

[24] J. Bu, L. J. Ratliff, and M. Mesbahi, "Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games," *arXiv preprint arXiv:1911.04672*, 2019.

[25] K. Zhang, Z. Yang, and T. Basar, "Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games," *Advances in Neural Information Processing Systems*, vol. 32, pp. 11 602–11 614, 2019.

[26] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," in *Learning for Dynamics and Control*, 2020, pp. 179–190.

[27] L. Furieri, Y. Zheng, and M. Kamgarpour, "Learning the globally optimal distributed lq regulator," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 287–297.

[28] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[29] Y. Sun and M. Fazel, "Analysis of policy gradient descent for control: Global optimality via convex parameterization." [Online]. Available: https://github.com/sunyue93/Nonconvex-optimization-meets-control/blob/master/convexify.pdf

[30] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[31] S. Lojasiewicz, "A topological property of real analytic subsets," *Coll. du CNRS, Les équations aux dérivées partielles*, vol. 117, pp. 87–89, 1963.

[32] J. Anderson, J. C. Doyle, S. H. Low, and N. Matni, "System level synthesis," *Annual Reviews in Control*, vol. 47, pp. 364–393, 2019.