Exploiting Linear Models for Model-Free Nonlinear Control: A Provably Convergent Policy Gradient Approach

Guannan Qu, Chenkai Yu, Steven Low, Adam Wierman

***Abstract*— Model-free learning-based control methods have seen great success recently. However, such methods typically suffer from poor sample complexity and limited convergence guarantees. This is in sharp contrast to classical model-based control, which has a rich theory but typically requires strong modeling assumptions. In this paper, we combine the two approaches. We consider a dynamical system with both linear and non-linear components and use the linear model to define a warm start for a model-free, policy gradient method. We show this hybrid approach outperforms the model-based controller while avoiding the convergence issues associated with model-free approaches via both numerical experiments and theoretical analyses, in which we derive sufficient conditions on the non-linear component such that our approach is guaranteed to converge to the (nearly) global optimal controller.**

I. Introduction

Recent years have seen great success in using learning-based methods for the control of dynamical systems. Examples cut across a broad spectrum of applications, including robotics [1], autonomous driving [2], energy systems [3], and more. Many of these learning-based methods are model-free in nature, meaning that they do not explicitly estimate the underlying model and do not explicitly make any assumptions on the parametric form of the underlying model [4]–[6]. Examples of such methods include policy gradient methods [7]–[9] and approximate dynamic programming [10]–[12]. Because model-free methods do not explicitly assume a parametric model class, they can potentially capture hard-to-model dynamics [13], which has led to empirical success in highly complex tasks [14]–[16]. However, the theoretic understanding of model-free approaches is extremely limited, and empirically they suffer from poor sample complexity and convergence issues [17], [18].

This stands in contrast to the classical model-based control, where one first estimates a parametric form of the model (e.g. linear state space model) and then develops a controller using tools from classic control theory. This approach has a rich history, including theoretical guarantees [19], [20], and is typically more sample efficient [18]. However, one major drawback of model-based control is that the model class might fail to capture complex real-world dynamics, in which case model error makes theoretical guarantees invalid.

Given the contrasts between model-free control and model-based control, the literature that focuses on providing a theoretic understanding of the two approaches is largely distinct, with papers focusing on either model-based approaches (e.g. [20]–[23]) or model-free approaches (e.g. [7], [8], [24]). There have been recent empirical approaches suggesting that model-based and model-free approaches can be combined to achieve the benefit of both, e.g., [13], [17], [25]; however, a theoretical understanding of the interplay between the approaches, especially when the dynamical system is nonlinear, remains open.

Contribution. In this paper, we study how model-based and model-free methods can be combined to achieve the benefits of both in a particular setting where the dynamical system's state space representation is a sum of two parts: a linear part, which is the most commonly used model class in model-based control, and a non-parametric non-linear part. This form of decomposition is widely used in practice. For example, engineers often have good approximate linear models for real-world dynamical systems such as energy systems [26] and mechanical systems [27]. The difference between the linear approximation and the real dynamics is often nonlinear and nonparametric, though understood to be small.

In this context, we introduce an approach for combining model-based methods for the linear part of the system and model-free approaches for the nonlinear part. In detail, we first use a model-based approach to design a state-feedback controller based on the linear part of the model. Then, we use this controller to warm start a model-free policy search. This warm start is similar in spirit to several empirically successful methods in the recent literature, e.g. [17], [25], however, no theoretical guarantees are known for existing approaches. In contrast, we prove guarantees on the convergence of the approach to an (almost) globally optimal state-feedback linear controller. Our analysis shows that the approach combines the benefits of model-based methods and model-free methods, capturing the unmodeled dynamics ignored by the model-based control while avoiding the convergence issues often associated with model-free approaches.

The key technical contribution underlying our approach is a landscape analysis of the cost as a function of the state-feedback controller. We show that the model-based controller obtained from the linear part of the system falls inside a convex region of the cost function which also contains the (almost) global minimizer. As a result, when using a warm start from the model-based controller, our approach is guaranteed to converge to the global minimizer. To highlight the necessity of the warm start, we show examples in which the landscape is non-convex and contains spurious local minima and even has a disconnected domain. Thus, a model-

Guannan Qu, Steven Low and Adam Wierman are with Department of Computing and Mathematical Sciences, California Institute of Techonology, Pasadena, CA (emails: gqu@caltech.edu, slow@caltech.edu, adamw@caltech.edu). Chenkai Yu is with Tsinghua University, Beijing, China (email: yck17@mails.tsinghua.edu.cn).

free approach that ignores model information may fail to converge to the global minimizer.

Related Work. Our work is mostly related to the class of model-free policy search methods for the Linear Quadratic Regulator (LQR), which date back to the early work of [28], [29] and have received considerable attention recently, e.g. [7]–[9], [24], [30]–[38]. A common theme in this line of work is that the underlying dynamical system is assumed to be *linear*, under which the cost function is shown to satisfy a "gradient dominance" property [7], which implies the model-free policy search method will converge to the global optimal controller. While these results provide a theoretic understanding of model-free methods, the benefits of using model-free methods for linear systems is not clear. For example, [18] shows that when the dynamics is actually linear, model-based methods are more sample efficient than model-free approaches. On the other hand, applications where model-free approaches have seen the most success are those involving the control of nonlinear dynamics [39]. However, though there has been empirical success, an understanding of model-free approaches for nonlinear systems is lacking. Our work makes an initial step by analyzing a model-free policy search method for nonlinear systems with a particular structure.

Our work is also related to empirical approaches suggested in the literature on reinforcement learning that involve augmenting model-free reinforcement learning with model-based approaches for various goals [40], [41], such as for gradient computation [42], [43], generate trajectories for model-free training [44]. Among these, the most related to our work are [17], [25], [45], [46], which use model-based methods as a starting point for model-free policy search. However, these papers focus on empirical evaluation, and to the best of our knowledge, we are the first to provide a theoretic justification on the effectiveness of combining model-based and model-free methods.

Beyond the above, our work is also related to a variety of areas at the interface of learning and control:

Model-based LQR. When the model is linear and is known, the optimal control problem can be solved via approaches like Algebraic Ricatti Equation [19] and dynamic programming [47]. When the linear model has unknown parameters, various system identification approaches have been proposed to estimate the system parameter, e.g. classic results such as [48], [49] or more recent ones with a focus on finite sample complexity, e.g., [50]–[52]. In addition, there have been recent efforts to provide end-to-end frameworks that combine system identification and control design [20], sometimes in an online setting, e.g. [21], [22], [53]–[57].

Control of nonlinear systems. There is a vast literature on the control of nonlinear dynamical systems, see e.g. [58], [59], including techniques like feedback linearization [60]. Specifically, our model is related to a practice in nonlinear control where one first linearizes the nonlinear system and design a controller based on the linear model [58, Sec. 3.3]. Our proposed approach goes beyond this by using model-free policy search to improve the controller obtained from the linear system. In addition, our problem is also related to the Circle/Popov criterion in nonlinear control [61], which certify the stability of an interconection of a linear system and a nonlinear system with bounded sector. In contrast, our approach not only certifies stability, but also designs a stabilizing controller with optimal cost.

Robust control. The fact that our model is a summation of a linear part and a small nonlinear part can be understood from the robust control angle [62], where the linear model can be viewed as the nominal plant and the nonlinear part can be viewed as an uncertain perturbation [63]. However, robust control methods like H_∞ and H_2/H_∞ mixed design [64] seek to design controllers with worst-case guarantees against all possible perturbations [65, Sec 4], whereas our work seeks to learn the best controller for the actual instance of the perturbation (the non-linear part of the model).

II. Model

We consider a dynamical system with state $x_t \in \mathbb{R}^n$ and control input $u_t \in \mathbb{R}^p$,

$$x_{t+1} = Ax_t + Bu_t + f(x_t), \tag{1}$$

where A is n-by-n, B is n-by-p, and $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $f(0) = 0$ and is "small" compared to A and B. We focus on the class of linear controllers, $u_t = -Kx_t$ for $K \in \mathbb{R}^{p\times n}$ and we consider the following quadratic cost function $C : \mathbb{R}^{p\times n} \to \mathbb{R}$,

$$C(K) = \mathbb{E}_K \sum_{t=0}^{\infty} \left(x_t^\top Q x_t + u_t^\top R u_t\right), \tag{2}$$

where the expectation is taken with respect to x_0 that is drawn from a fixed initial state distribution $\mathcal{D}$, and the subscript K in the expectation indicates the trajectory $\{x_t\}_{t=0}^\infty$ in the expectation is generated by controller K.

The system in (1) is the sum of a linear part and a "small" non-linear part f. Such a decomposition can be found in many practical situations, as discussed in the introduction. In such settings, f represents the error of the approximated linear model, which is small if the linear approximations are accurate in practice. Alternatively, (1) can be a result of linearization of a non-linear model, with f capturing the higher order residuals.

In this paper, we assume f is unknown while some estimate $(\hat{A}, \hat{B})$ of (A, B) is known, since in various engineering domains, the linear model (A, B), or at least some approximation of (A, B), is readily available. Alternatively, $(\hat{A}, \hat{B})$ can also be the result of system identification for the unknown system.

Combining model-based and model-free control. We propose a framework that combines model-based and model-free methods to find an optimal linear state feedback controller that minimizes the cost (2). Concretely, the framework works as follows:

- Compute model-based controller $\hat{K}_{\text{lin}}$ to be the optimal LQR controller for linear system $(\hat{A}, \hat{B})$ and cost matrices (Q, R).
- Use $\hat{K}_{\text{lin}}$ as an initial point for model-free policy search. There can be many variants of policy search, including zeroth-order policy search [7] or actor-critic methods

[32]. In Section III we propose a concrete approach (Algorithm 1).

Compared with a standard model-free approach, where the initial point is unspecified and is usually obtained through trial and error, this hybrid approach makes use of model-based control to warm start the model-free policy search algorithm. This intuitive idea is powerful given the complexity of the cost landscape. To illustrate the importance of this warm start approach, we provide two examples (Examples 1 and 2) showing that, even when f is small compared to A, B, the landscape of $C(K)$ may contain spurious local minima (Example 1), and the set of stabilizing state feedback controllers may not even be connected (Example 2). As such, model-free approaches will likely fail to converge to the global minimizer. In contrast, in the examples, the model-based controller $\hat{K}_{\text{lin}}$ (when $(\hat{A}, \hat{B}) = (A, B)$) stays well within the attraction basin of the global minimizer. Hence the proposed hybrid approach with the model-based warm start converges at least when $(\hat{A}, \hat{B})$ is an accurate enough estimate of (A, B). In the next section, we formalize this intuition and provide theoretic results on the landscape of $C(K)$ as well as the convergence of the proposed approach.

Example 1 (Cost landscape may contain spurious local minima)**.** *Consider the following one-dimensional dynamics* $x_{t+1} = 0.5x_t + u_t + f(x_t)$*, and* $f(x) = 0.01x/(1 + 0.9\sin(x))$*, satisfying* $|f(x)| \le 0.1|x|$*. We set* $x_0 = 50, Q = 10, R = 1$*. When using a linear state feedback controller* $u_t = -Kx_t$*, the cost is given in Figure 1, which has many local minima. However,* $\hat{K}_{\text{lin}}$ *lies within the attraction basin of the global minimizer* K^* *and is in fact very close to* K^**.*

Example 2 (Finite-cost controllers may be disconnected)**.** *Suppose* $n = 2$ *and* $p = 1$*. Let*

$$A = 0.95\begin{bmatrix}\cos 0.2 & -\sin 0.2\\ \sin 0.2 & \cos 0.2\end{bmatrix},\ B = \begin{bmatrix}0.2\\ 0.15\end{bmatrix},$$
$$Q = \begin{bmatrix}1 & -0.999\\ -0.999 & 1\end{bmatrix},\ R = 0.5,\ x_0 = \begin{bmatrix}5\\ -6\end{bmatrix},$$

$$f(x) = \frac{0.1(\left[\begin{smallmatrix}3\\0\end{smallmatrix}\right] - x) - (A - I)x - B\begin{bmatrix}-1 & -0.2\end{bmatrix}x}{(\|x - \left[\begin{smallmatrix}3\\0\end{smallmatrix}\right]\|^2 + 1)^2} + \frac{0.7(\left[\begin{smallmatrix}4.5\\-3\end{smallmatrix}\right] - x) - (A - I)x - B\begin{bmatrix}0 & 0.7\end{bmatrix}x}{(\|x - \left[\begin{smallmatrix}4.5\\-3\end{smallmatrix}\right]\|^2 + 1)^2} + \frac{0.9(\left[\begin{smallmatrix}5\\-1\end{smallmatrix}\right] - x) - (A - I)x - B\begin{bmatrix}-0.2 & 0.5\end{bmatrix}x}{(\|x - \left[\begin{smallmatrix}5\\-1\end{smallmatrix}\right]\|^2 + 1)^2} + f_0,$$

where f_0 *is such that* $f(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]) = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$*. In this case, the set of controllers* $K = [K_1, K_2]$ *with finite cost is not connected, as shown in Figure 2. Moreover, this phenomenon exists even for very small* f*. Starting from the above values, we can simultaneously make* A *closer to* I*,* B *closer to 0, and the coefficients* $0.1, 0.7, 0.9$ *in function* f *closer to 0 (with the same factor) in order to maintain this phenomenon. A detailed explanation of this example can be found in Appendix A.*

Notation. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectrum norm for matrices, and $\|\cdot\|_F$ to denote the Frobenius norm. For matrices A, B of the same dimension, $\langle A, B\rangle = \text{Tr}(A^\top B)$ denotes the trace inner product. For symmetric matrices A, B, $A \succeq B$ means $A - B$ is positive semi-definite. Notation $\sigma_{\min}(\cdot)$ denotes the smallest eigenvalue of a symmetric square matrix. Additionally, $y_1 \lesssim y_2$ and $y_1 \asymp y_2$ mean $y_1 \le cy_2$ and $y_1 = cy_2$ respectively for some numerical constant c.

III. Main Results

Our main technical result characterizes the landscape of the cost function in order to prove the convergence of the proposed approach combining model-based and model-free techniques. For concreteness, we use a particular instance of the policy search method and show its convergence, but the approach is more general and can be extended to other methods.

Before stating our results, we discuss their assumptions. The first assumption is about the pair Q, R in the cost function and is standard [21].

Assumption 1. *Q and R are positive definite matrices satisfying $R + B^\top QB \succeq \sigma I$ for some $\sigma > 0$, and $\|Q\| \le 1$, $\|R\| \le 1$.*

The assumption $\|Q\| \le 1, \|R\| \le 1$ in Assumption 1 is for ease of calculation, and is without loss of generality as we can always rescale the cost function to guarantee it is satisfied. Our next assumption concerns the pair (A, B) and is again standard [20].

Assumption 2. *The pair (A, B) is controllable. Let K^*_{lin} be the optimal controller associated with the linear system $x_{t+1} = Ax_t + Bu_t$, and we assume $\|(A - BK^*_{\text{lin}})^t\| \le c_{\text{lin}}\rho^t_{\text{lin}}, \forall t$, for some $\rho_{\text{lin}} \in (0, 1)$ and $c_{\text{lin}} > 0$. Further, we assume $\max(\|A\|, \|B\|, \|K^*_{\text{lin}}\|, 1) \le \Gamma$ for some $\Gamma > 0$.*

The next assumption is on the initial state distribution.

Assumption 3. *The initial state distribution $\mathcal{D}$ is supported in a region with radious D_0. Further, $\mathbb{E}\, x_0x_0^\top \succeq \sigma_x I$ for some $\sigma_x > 0$.*

The requirement of bounded support is only for simplification of the proof. It can be replaced with a bound on the second and the third moment of the initial state if desired at the expense of extra complexity. Finally, we assume that f and the Jacobian of f are Lipschitz continuous or, in other words, the first and second order derivatives of f are bounded. This quantifies the "smallness" of f.

Assumption 4. *We assume f is differentiable, $f(0) = 0$, $\|f(x) - f(x')\| \le \ell\|x - x'\|$, and $\|\frac{\partial f(x)}{\partial x} - \frac{\partial f(x')}{\partial x}\| \le \ell'\|x - x'\|$ for some $\ell, \ell' > 0$, where $\frac{\partial f(x)}{\partial x}$ is the Jacobian of $f(x)$.*

Before we state our result, we must also define what we mean by the "global" domain of $C(K)$. One natural definition for the domain of C is the set of (global or local) stabilizing controllers for the nonlinear system (1). However, to the best of our knowledge, the stabilization of nonlinear systems is a challenging topic and such a set is not clearly characterized. For this reason, we consider an alternative domain $\Omega(c_0, \rho_0) = \{K : \|(A - BK)^t\| \le c_0\rho_0^t\}$ for some

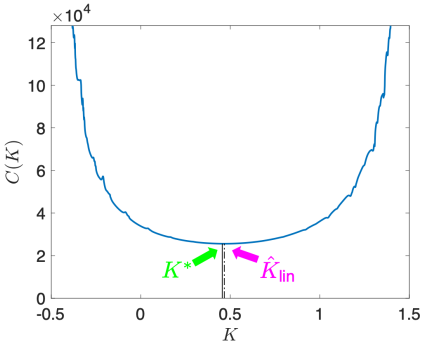


Fig. 1: Cost landscape in Example 1, where $\hat{K}_{\text{lin}}$ is computed under $(\hat{A}, \hat{B}) = (A, B)$.

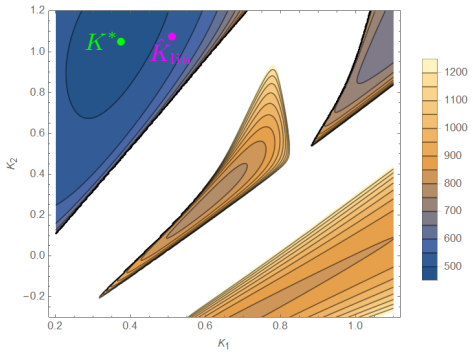


Fig. 2: Cost landscape in Example 2, where $\hat{K}_{\text{lin}}$ is computed under $(\hat{A}, \hat{B}) = (A, B)$.

$c_0 \geq 1, \rho_0 \in (0, 1)$ to be chosen later. We consider this domain since it is clearly characterized and also because when $\rho_0 \to 1$, $c_0 \to \infty$, this set captures the set of all stabilizing controllers for the linear system (A, B).[1]

We now move to our results. Our first result characterizes the landscape of the cost function. It shows that when ℓ and ℓ' (the Lipschitz constant for f and Jacobian of f respectively) are small enough, $C(K)$ achieves its global minimum inside a local neighborhood of K^*_{lin}, which as defined in Assumption 2 is the optimal LQR controller for the linear part (A, B) of the system, or in other words $\hat{K}_{\text{lin}}$ when $(\hat{A}, \hat{B}) = (A, B)$. Further, within this local neighborhood, $C(K)$ is strongly convex and smooth. Theorem 1 is our most technical result and a proof is provided in Appendix B.

Theorem 1. *For any $\rho_0 \in [\frac{\rho_{\text{lin}}+1}{2}, 1)$ and $c_0 \geq 2c_{\text{lin}}$, let $\Omega = \Omega(c_0, \rho_0)$. If $\ell \lesssim \frac{(\sigma\sigma_x)^2(1-\rho_0)^8}{\Gamma^9 c_0^{15} D_0^4}$, $\ell' \lesssim \frac{(\sigma\sigma_x)^2(1-\rho_0)^8}{\Gamma^9 c_0^{18} D_0^5}$, then:*

(a) $C(K)$ is finite in Ω and the trajectories satisfies $\|x_t\| \leq 2c_0(\frac{\rho_0+1}{2})^t\|x_0\|$ for any $x_0 \in \mathbb{R}^n, K \in \Omega$;

(b) there exists a region $\Lambda(\delta) = \{K : \|K - K^*_{\text{lin}}\|_F \leq \delta\} \subset \Omega$ with $\delta \asymp \frac{\sigma_x \sigma (1-\rho_0)^4}{\Gamma^5 c_0^7 D_0^2}$ such that $C(K)$ is μ-strongly convex and h-smooth inside $\Lambda(\delta)$, with $\mu = \sigma\sigma_x$ and $h \asymp \frac{\Gamma^4 c_0^4 D_0^2}{(1-\rho_0)^2}$;

(c) the global minimum of $C(K)$ over Ω is achieved at a point $K^* \in \Lambda(\frac{\delta}{3})$, which is also the unique stationary point of $C(K)$ inside $\Lambda(\delta)$.

We comment that, while our landscape result is a local convexity result around the global minimum K^*, we are also able to show that K^*_{lin} (which can be efficiently approximated based on $\hat{A}, \hat{B}$ provided that they are accurate enough) is within the convex region around K^* and, as such, within the attraction basin of K^*. This is different than existing landscape analysis for non-convex optimization in other contexts like deep learning, where only local convexity is shown without showing how to enter its attraction basin [66], [67].

Given the landscape result, it is perhaps not surprising that the model-free policy search method converges to the global minimizer K^* when warm starting with the model-based optimal LQR controller $\hat{K}_{\text{lin}}$ provided $(\hat{A}, \hat{B})$ is an accurate enough estimate of (A, B), because both K^* and $\hat{K}_{\text{lin}}$ lie in the same convex region of the cost function. In the following, we prove this formally by considering a version of model-free policy search algorithm - the zeroth order policy search with one point gradient estimator. The proposed algorithm is stated in Algorithm 1 with the gradient estimator subroutine given in Algorithm 2. To state the convergence results of our algorithm, we first consider the simpler case $(\hat{A}, \hat{B}) = (A, B)$ in Theorem 2. Theorem 2 shows that the landscape result in Theorem 1 ensures that when $(\hat{A}, \hat{B}) = (A, B)$, Algorithm 1 converges to the global minimum of $C(K)$ over Ω, hence outperforming the model-based controller and avoiding the non-convergence issue of model-free approaches shown before.

Theorem 2. *Suppose $(\hat{A}, \hat{B}) = (A, B)$. Under the conditions in Theorem 1, for any $\varepsilon > 0$ and $\nu \in (0, 1)$, if the step size $\eta \leq \frac{1}{h}$, the number of gradient descent steps $M \geq \frac{1}{\eta\mu}\log(\delta\sqrt{h/\varepsilon})$, and the gradient estimator parameters satisfy $r \leq \frac{1}{3h}e_{grad}$,*

$$J \geq \frac{1}{e_{grad}^2}\frac{d^3}{r^2}\log\frac{4dM}{\nu}\max(18(C(K^*)+2h\delta^2)^2, 72C_{\max}^2),$$
$$T \geq \frac{2}{1-\rho_0}\log\frac{6dC_{\max}}{e_{grad}r},$$

where $e_{grad} = \min(\frac{\mu}{2}\sqrt{\frac{\varepsilon}{h}}, \mu\frac{\delta}{3})$, $d = pn$, and $C_{\max} = \frac{40\Gamma^2 c_0^2}{1-\rho_0}D_0^2$, then with probability at least $1-\nu$, $C(K_M) - C(K^*) \leq \varepsilon$.

The proof of Theorem 2 is provided in Appendix I. The above Theorem 2 guarantees the convergence to the optimal controller when $(\hat{A}, \hat{B}) = (A, B)$. We next present Corollary 1, which shows that as long as $(\hat{A}, \hat{B})$ is an accurate enough estimate of (A, B), the same results as in Theorem 2 hold. Corollary 1 is a combination of Theorem 2 and a LQR perturbation result in [21].

Corollary 1. *There exists a perturbation constant c_{per} depending on A, B, Q, R such that when $\max(\|\hat{A}-A\|, \|\hat{B}-B\|) \leq \frac{\min(\delta,1)}{6\,\text{c}_{\text{per}}}$ where δ is the constant defined in Theorem 1,*

[1] This is because for any stabilizing controller K of linear system (A, B), there must exist $c_0 > 0, \rho_0 \in (0, 1)$ s.t. $\|(A - BK)^t\| \leq c_0\rho_0^t, \forall t \geq 0$.

the same results in Theorem 2 hold with the same parameter choice of η, M, r, J, T as Theorem 2.

The proof of Corollary 1 is provided in Appendix I. In addition to the above theoretical guarantees, the hybrid approach numerically appears to have better sample complexity even when the model-free methods do converge. We illustrate such results in the next section.

Our result shows that the proposed hybrid approach is guaranteed to converge to the global minimum only when ℓ, ℓ' are bounded. Such a requirement on ℓ, ℓ' is intuitive since when the "size" of f is much larger than the linear part (A, B), a warm start based on the linear model does not make much sense as the linear model is a poor estimation of the dynamics. There should be a threshold on the "size" of f, below which the hybrid approach will work. Our result provides a (potentially conservative) lower bound on the threshold. Tighter bounds are interesting goals for future work.

Finally, we comment that Algorithm 1 with the gradient estimator Algorithm 2 is but one of many possibilities for policy search methods, e.g. two-point gradient estimator [30], REINFORCE [6], actor-critic methods [68], and with our landscape result (Theorem 1), similar versions of our convergence result (Theorem 2) can be proven for these other types of policy search methods. Further, there are various results suggesting ways to reduce the variance of the gradient estimator [69]–[71] that could also be incorporated into the framework here.

Remark 1. *In this paper, the search space is the class of linear controllers $u_t = -Kx_t$. The focus on linear controllers is in line with a common practice in nonlinear control that first linearizes the nonlinear system and then designs a linear controller for the linearized system. Compared to this approach, we use the same class of controllers, but search for the best one considering the nonlinear residual. Those being said, for nonlinear systems, the optimal controller is in general nonlinear. The idea in this paper can also be used to search over nonlinear controllers, where one first warm-starts from the optimal LQR controller for the linear system, and then learns a nonlinear residual controller, but the analysis is much more challenging because popular nonlinear controllers include neural networks, whose analysis remains largely open.*

Remark 2. *The problem setting of this paper does not consider process noise, and the only randomness in the system arises from the initial state. The analysis and algorithm in this paper can be generalized to the case with process noise.*

IV. Numerical Experiments

To illustrate our approach, we contrast it with model-free and model-based approaches using two sets of experiments: (i) synthetic random instances and (ii) the cart inverted pendulum.

Algorithm 1: Model-Free Policy Search with Model-Based Warm Start

Input: Linear Model $(\hat{A}, \hat{B})$, cost matrix (Q, R), parameters η, M, r, J, T

$\hat{K}_{\text{lin}} \leftarrow$ OPT-LQR$(\hat{A}, \hat{B}, Q, R)$ `// Find the optimal controller for the linear system`

$K_0 \leftarrow \hat{K}_{\text{lin}}$ `// Warm start`

for $m = 0, 1, \ldots, M-1$ **do**

  $\widehat{\nabla C}(K_m) \leftarrow$ GradientEstimator(K_m, r, J, T)

  $K_{m+1} \leftarrow K_m - \eta \widehat{\nabla C}(K_m)$

return K_M

Algorithm 2: GradientEstimator

Input: Controller K, parameters r, J, T

for $j = 1, 2, \ldots, J$ **do**

  `/* Sample random direction` U_j `from sphere with radius` r `in Frobenius norm */`

  Sample $U_j \sim \text{Sphere}(r)$

  `/* Sample a trajectory under perturbed controller` $K + U_j$ `*/`

  Sample $x_0 \sim \mathcal{D}$

  for $t = 0, 1, \ldots, T$ **do**

    Set $u_t = -(K + U_j)x_t$

    Receive the next point x_{t+1} from system

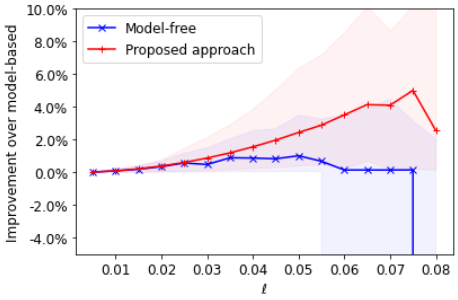
  Calculate approximate cost $\widehat{C}_j = \sum_{t=0}^{T} [x_t^\top Q x_t + u_t^\top R u_t]$

return $\widehat{\nabla C}(K) = \frac{1}{J} \sum_{j=1}^{J} \frac{d}{r^2} \widehat{C}_j U_j$ *where* $d = pn$ `// One point gradient estimator`

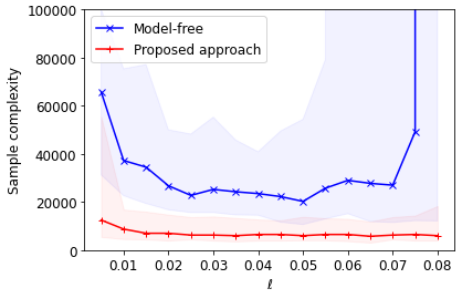
A. Synthetic experiments

Our first set of experiments focuses on random synthetic examples. We set n (the dimension of state) and p (the dimension of input) to be 2. We generate A and B randomly, with each entry drawn from a Gaussian distribution $N(0, 1)$, where A is normalized so that the spectral radius of A is 0.5. The initial state distribution $\mathcal{D}$ is a uniform distribution over a fixed set of 2 initial states, which are drawn from i.i.d. zero-mean Gaussians with norm normalized to be 2. The cost is set as $Q = 2I, R = I$. We set $f(x) = \ell x/(1 - 0.9\sin(x))$, where all operations here are understood as entry wise and ℓ is a parameter that we increase from $\ell = 0.005$ to $\ell = 0.08$. For each ℓ, we run both our hybrid approach and the model-free approach (starting from $K = 0$ as this system is open loop stable) with algorithm parameters $\eta = 0.01$, $T = 50$, $r = 0.001$, $J = 10$, and $M = 200$. We repeat the above procedures for 50 times, each time with A, B and $\mathcal{D}$ regenerated, and then plot the final cost achieved by both approaches (normalized as the improvement over the model-based LQR controller)[2] as a function of ℓ in Figure 3a. We also plot the sample complexity as a function of ℓ for both

[2]The improvement is counted as $-\infty$ if a run fails to converge to a stabilizing controller.



(a)



(b)

Fig. 3: Simulation results for synthetic experiments. Solid lines represent the median and shaded regions represent the 25% to 75% percentiles.

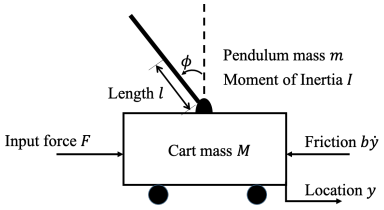


Fig. 4: Cart inverted pendulum model with $M = 0.5\,\mathrm{kg}$, $m = 0.2\,\mathrm{kg}$, $b = 0.1\,\mathrm{N\,s\,m^{-1}}$, $I = 0.006\,\mathrm{kg\,m^2}$, $l = 0.3\,\mathrm{m}$.

approaches in Figure 3b, where sample complexity is the number of state samples needed for the respective algorithm to converge.[3] The results show that both the proposed hybrid approach and the model-free approach can outperform the model-based LQR controller. Moreover, the proposed hybrid approach consistently outperforms the model-free approach in terms of the final cost achieved and the sample complexity.

B. Inverted Pendulum

Our second set of experiments focuses on the cart inverted pendulum model (cf. Figure 4), where the goal is to stabilize the pendulum in the upright position. This is a nonlinear system with a widely accepted approximated linear model, and we provide its dynamics and its linearization below in continuous time [27],

$$\begin{aligned}\begin{bmatrix}\dot{y}\\ \dot{\phi}\\ \ddot{y}\\ \ddot{\phi}\end{bmatrix} &= \begin{bmatrix}\dot{y}\\ \dot{\phi}\\ \begin{bmatrix} M+m & -ml\cos\phi \\ -ml\cos\phi & I+ml^2\end{bmatrix}^{-1}\begin{bmatrix}-b\dot{y}-ml(\dot{\phi})^2\sin\phi+F\\ mgl\sin\phi\end{bmatrix}\end{bmatrix}\\ &\approx \begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & \frac{m^2gl^2}{I(M+m)+Mml^2} & \frac{-(I+ml^2)b}{I(M+m)+Mml^2} & 0\\ 0 & \frac{mgl(M+m)}{I(M+m)+Mml^2} & \frac{-mlb}{I(M+m)+Mml^2} & 0\end{bmatrix}\begin{bmatrix}y\\ \phi\\ \dot{y}\\ \dot{\phi}\end{bmatrix}\\ &\quad + \begin{bmatrix}0\\ 0\\ \frac{I+ml^2}{I(M+m)+Mml^2}\\ \frac{ml}{I(M+m)+Mml^2}\end{bmatrix}F,\end{aligned}$$

where the "≈" is obtained by setting $\sin\phi \approx \phi, \cos\phi \approx 1$ and $(\dot{\phi})^2\sin\phi \approx 0$. We identify the state as $x = [y, \phi, \dot{y}, \dot{\phi}]^\top$ and the input as $u = F$. We discretize both the nonlinear system and the linear approximation above using forward discretization with the step size $\tau = 0.05s$ to obtain a discrete time nonlinear system and its approximation, and we set f to be the difference of the two. We also set $Q = I, R = 1$, and the initial state distribution is a Dirac distribution centered on $x_0 = [0.8, 0.8, 0.2, 0.2]^\top$.

We run the proposed approach as well as the model-free approach, where the model-free approach is initialized at $K = [k_1, k_2, k_3, k_3]$ which is generated randomly with k_1, k_2 drawn from $[-15, 0]$, and k_3, k_4 drawn from $[0, 15]$.[4] For both approaches, we set the algorithm parameters as $\eta = 0.01$, $r = 0.001$, $T = 2000$, $J = 3$, $M = 500$. We do 50 runs for both approaches, plot the learning processes in Figure 5a. We also plot the histogram of the final cost achieved by both approaches (normalized as the improvement over the model-based LQR controller) in Figure 5b.

The results show that the model-free approach fails to find a stabilizing controller in roughly 40% of the runs, whereas almost all runs of the proposed approach can find a stabilizing controller,[5] even though the model-free approach always starts from a stabilizing controller. Further, both the proposed hybrid approach and the model-free approach outperform the model-based LQR controller if they do reach a stabilizing controller. However, the proposed hybrid approach consistently achieves larger improvements than the model-free approach.

V. Conclusion

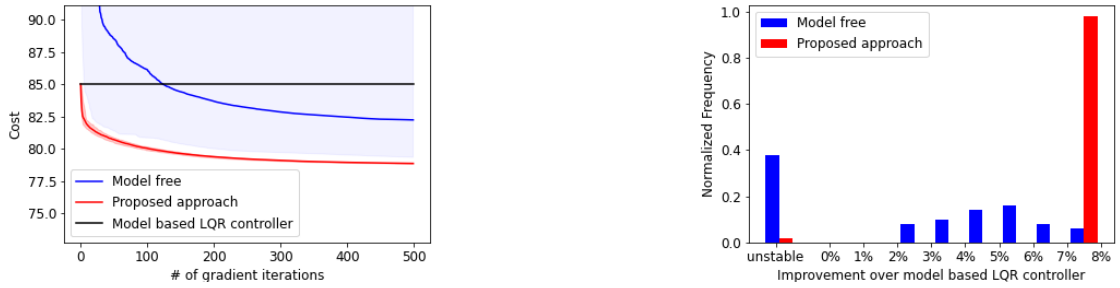
In this paper, we consider a dynamical system with a (roughly) known linear component and a (small) nonlinear component, and we propose an approach that combines model-based LQR control with model-free policy search to achieve the best of both worlds.

Our work represents an initial step towards making model-free policy search methods more reliable by exploiting known model information. An immediate next step is to relax the

[3]The sample complexity is counted as ∞ if a run doesn't converge to a stabilizing controller.

[4]Such an initialization is obtained through trial and error with the goal of ensuring a stabilizing initial controller with high probability. If the initial controller is unstable, we resample until it is stable.

[5]We use a small number of trajectories for calculating the gradient ($J = 3$). As such, the proposed approach has a small probability of not converging when this gradient estimate is poor.



(a) (b)

Fig. 5: Simulation results for the inverted pendulum example. In (a), solid lines represent the median and shaded regions represent the 10% to 90% percentiles.

bound on ℓ, ℓ' in the landscape result (Theorem 1) and understand how much "non-linearity" can be tolerated until the linear model is no longer informative for control design. Another important direction is to enlarge the search space to nonlinear controllers (Remark 1), which is a challenging task given that popular parameterizations of nonlinear controllers involve neural networks, which are hard to analyze.

REFERENCES

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," 2015.

[2] D. Li, D. Zhao, Q. Zhang, and Y. Chen, "Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 83–98, 2019.

[3] D. Wu, D. Kalathil, and L. Xie, "Deep reinforcement learning-based robust protection in electric distribution grids," *arXiv preprint arXiv:2003.02422*, 2020.

[4] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[5] D. P. Bertsekas, *Dynamic programming and optimal control 3rd edition, volume II*. Belmont, MA: Athena Scientific, 2011.

[6] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[7] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," *arXiv preprint arXiv:1801.05039*, 2018.

[8] 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.

[9] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *arXiv preprint arXiv:1912.09135*, 2019.

[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. Tu and B. Recht, "Least-squares temporal difference learning for the linear quadratic regulator," *arXiv preprint arXiv:1712.08642*, 2017.

[12] K. Krauth, S. Tu, and B. Recht, "Finite-time analysis of approximate policy iteration for the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, 2019, pp. 8512–8522.

[13] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, "Model-based reinforcement learning via meta-policy optimization," *arXiv preprint arXiv:1809.05214*, 2018.

[14] S. Levine and V. Koltun, "Guided policy search," in *International Conference on Machine Learning*, 2013, pp. 1–9.

[15] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[16] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[17] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7559–7566.

[18] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," *arXiv preprint arXiv:1812.03565*, 2018.

[19] K. Zhou, J. C. Doyle, K. Glover *et al.*, *Robust and optimal control*. Prentice hall New Jersey, 1996, vol. 40.

[20] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, pp. 1–47, 2019.

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

[22] M. Simchowitz and D. J. Foster, "Naive exploration is optimal for online LQR," *arXiv preprint arXiv:2001.09576*, 2020.

[23] M. Simchowitz, K. Singh, and E. Hazan, "Improper learning for non-stochastic control," *arXiv preprint arXiv:2001.09254*, 2020.

[24] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," *arXiv preprint arXiv:1812.08305*, 2018.

[25] T. Silver, K. Allen, J. Tenenbaum, and L. Kaelbling, "Residual policy learning," *arXiv preprint arXiv:1812.06298*, 2018.

[26] A. Benchaib, *From Small Signal to Exact Linearization of Swing Equations*. John Wiley & Sons, Ltd, 2015, ch. 3, pp. 57–86. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119135760.ch3

[27] M. Magdy, A. El Marhomy, and M. A. Attia, "Modeling of inverted pendulum system with gravitational search algorithm optimized controller," *Ain Shams Engineering Journal*, vol. 10, no. 1, pp. 129–149, 2019.

[28] T. Rautert and E. W. Sachs, "Computational design of optimal output feedback controllers," *SIAM Journal on Optimization*, vol. 7, no. 3, pp. 837–852, 1997.

[29] K. Mårtensson and A. Rantzer, "Gradient methods for iterative distributed control synthesis," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 549–554.

[30] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem," *arXiv preprint arXiv:1912.11899*, 2019.

[31] B. Gravell, P. M. Esfahani, and T. Summers, "Learning robust controllers for linear quadratic systems with multiplicative noise via policy gradient," *arXiv preprint arXiv:1905.13547*, 2019.

[32] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," *arXiv preprint arXiv:1907.06246*, 2019.

[33] K. Zhang, Z. Yang, and T. Basar, "Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games," in

Advances in Neural Information Processing Systems, 2019, pp. 11 602–11 614.

[34] K. Zhang, B. Hu, and T. Basar, "Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence," in *Learning for Dynamics and Control*, 2020, pp. 179–190.

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

[36] J. P. Jansch-Porto, B. Hu, and G. Dullerud, "Convergence guarantees of policy optimization methods for markovian jump linear systems," *arXiv preprint arXiv:2002.04090*, 2020.

[37] ——, "Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of markovian jump systems," *arXiv preprint arXiv:2006.03116*, 2020.

[38] I. Fatkhullin and B. Polyak, "Optimizing static linear feedback: Gradient method," *arXiv preprint arXiv:2004.09875*, 2020.

[39] V. Pong, S. Gu, M. Dalal, and S. Levine, "Temporal difference models: Model-free deep RL for model-based control," *arXiv preprint arXiv:1802.09081*, 2018.

[40] T. Che, Y. Lu, G. Tucker, S. Bhupatiraju, S. Gu, S. Levine, and Y. Bengio, "Combining model-based and model-free RL via multi-step control variates," 2018.

[41] T.-L. Vuong and K. Tran, "Uncertainty-aware model-based policy optimization," *arXiv preprint arXiv:1906.10717*, 2019.

[42] N. Mishra, P. Abbeel, and I. Mordatch, "Prediction and control with temporal segment models," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2459–2468.

[43] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa, "Learning continuous control policies by stochastic value gradients," in *Advances in Neural Information Processing Systems*, 2015, pp. 2944–2952.

[44] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep Q-learning with model-based acceleration," in *International Conference on Machine Learning*, 2016, pp. 2829–2838.

[45] S. Bansal, R. Calandra, K. Chua, S. Levine, and C. Tomlin, "Mbmf: Model-based priors for model-free reinforcement learning," *arXiv preprint arXiv:1709.03153*, 2017.

[46] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine, "Residual reinforcement learning for robot control," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6023–6029.

[47] D. P. Bertsekas, *Dynamic programming and optimal control, volumn I*, 3rd ed. Belmont, Mass.: Athena Scientific, 2005.

[48] L. Ljung, "System identification," *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–19, 1999.

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

[50] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," *arXiv preprint arXiv:1802.08334*, 2018.

[51] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 5655–5661.

[52] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-time system identification for partially observed LTI systems of unknown order," *arXiv preprint arXiv:1902.01848*, 2019.

[53] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[54] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time analysis of optimal adaptive policies for linear-quadratic systems," *arXiv preprint arXiv:1711.07230*, 2017.

[55] Y. Ouyang, M. Gagrani, and R. Jain, "Learning-based control of unknown linear systems with thompson sampling," *arXiv preprint arXiv:1709.04047*, 2017.

[56] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, 2018, pp. 4188–4197.

[57] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," *arXiv preprint arXiv:1902.06223*, 2019.

[58] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[59] A. Isidori, *Nonlinear Control Systems Design 1989: Selected Papers from the IFAC Symposium, Capri, Italy, 14-16 June 1989*. Elsevier, 2014.

[60] T. Westenbroek, D. Fridovich-Keil, E. Mazumdar, S. Arora, V. Prabhu, S. S. Sastry, and C. J. Tomlin, "Feedback linearization for unknown systems via reinforcement learning," 2019.

[61] W. M. Haddad and V. Chellaboina, *Nonlinear dynamical systems and control: a Lyapunov-based approach*. Princeton university press, 2011.

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

[63] I. R. Petersen and R. Tempo, "Robust control of uncertain systems: Classical results and recent developments," *Automatica*, vol. 50, no. 5, pp. 1315–1335, 2014.

[64] P. P. Khargonekar and M. A. Rotea, "Mixed H_2/H_∞ control: a convex optimization approach," *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 824–837, 1991.

[65] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback control theory*. Courier Corporation, 2013.

[66] S. Oymak and M. Soltanolkotabi, "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 4951–4960. [Online]. Available: http://proceedings.mlr.press/v97/oymak19a.html

[67] N. Azizan, S. Lale, and B. Hassibi, "Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization," *arXiv preprint arXiv:1906.03830*, 2019.

[68] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014.

[69] E. Greensmith, P. L. Bartlett, and J. Baxter, "Variance reduction techniques for gradient estimates in reinforcement learning," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1471–1530, 2004.

[70] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[71] J. A. Preiss, S. M. R. Arnold, C.-Y. Wei, and M. Kloft, "Analyzing the variance of policy gradient estimators for the linear-quadratic regulator," 2019.

[72] S. M. Kakade *et al.*, "On the sample complexity of reinforcement learning," Ph.D. dissertation, University of London London, England, 2003.

[73] S. Bubeck, "Convex optimization: Algorithms and complexity," 2014.

[74] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '05, USA, 2005, p. 385–394.

APPENDIX

A. Explanation of Example 2

The intuition behind Example 2 is as follows. Let A be contractive ($\|A\| < 1$), but very close to I. As a result, with $f = 0$ and $K = 0$, any starting point x will linearly converges to 0 (with a slow rate), thus incurring finite cost. We construct a new contracting point close to x_i by adding the following expression to function f:

$$\frac{\alpha_i(x_i - x) - (A - I)x - BK_ix}{(\|x - x_i\|^2 + 1)^2}.$$

This contracting point has strength α_i and is effective when the policy K is close to K_i. With such a function f, if we start from certain states, the state will converge to this contracting point x_i, thus incurring infinite cost.

In Example 2, we construct three such contracting points, taking effects around different policy K, so that the activated policies (those incurring infinite cost) form a ring shape, leaving the center inactivated (incurring finite cost). To visualize whether a policy is activated (incurring infinite cost), we compute the limit point to which the state converges under this policy, as shown in Figure 6.

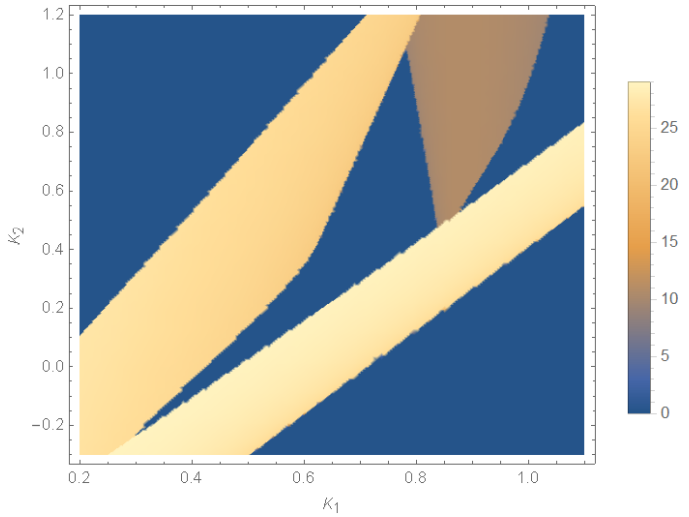


Fig. 6: The squared norm of the limit state (we use x_{100} for the plot) under different policies. For policies with infinite cost, the state converges to a non-zero contracting point. Thus, the points with non-zero value in this figure are exactly those with infinite value in Figure 2.

B. Proof of Theorem 1: Landscape Analysis of $C(K)$

The proof will be divided into three steps, corresponding to part (a), (b) and (c) of the Theorem respectively.

Step 1: We show in Lemma 1 that when $K \in \Omega$ and ℓ is bounded, then the system will be globally exponentially stable, or in other words the state trajectory will geometrically decay to the origin regardless of the initial state. This also implies boundedness of C within Ω. The proof of Lemma 1 is given in Appendix C.

Lemma 1. *When $\ell \le \frac{1-\rho_0}{4c_0}$ and when $K \in \Omega$, $\forall x_0 \in \mathbb{R}^n$, the system trajectory satisfies $\|x_t\| \le c\rho^t\|x_0\|$, with $c = 2\rho_0$ and $\rho = \frac{1+\rho_0}{2}$. As a consequence, we have $C(K)$ is finite in Ω.*

Step 2: We provide an explicit characterization of the cost function $C(K)$ and its gradient $\nabla C(K)$, and show the following Lemma 2, indicating the strong convexity and smoothness of the cost function. The proof of Lemma 2 is provided in Appendix D.

Lemma 2. *When δ, ℓ, ℓ' satisfy,*

$$\delta \le \frac{\sigma_x\sigma(1-\rho)^4}{96\Gamma^5c^7D_0^2}, \qquad \ell \le \frac{\sigma_x\sigma(1-\rho)^5}{192\Gamma^4c^9D_0^2}, \qquad \ell' \le \frac{\sigma_x\sigma(1-\rho)^5}{192\Gamma^4c^{12}D_0^3},$$

then $\Lambda(\delta) = \{K : \|K - K^*_{\text{lin}}\|_F \le \delta\} \subset \Omega$, and for all, $K, K' \in \Lambda(\delta)$,

$$C(K') - C(K) \ge \text{Tr}(K' - K)^\top \nabla C(K) + \frac{\mu}{2}\|K' - K\|_F^2,$$
$$C(K') - C(K) \le \text{Tr}(K' - K)^\top \nabla C(K) + \frac{h}{2}\|K' - K\|_F^2,$$

where $\mu = \sigma_x\sigma$, $h = 5\frac{\Gamma^4 c^4 D_0^2}{(1-\rho)^2}$. *This implies* $C(K)$ *is* μ*-strongly convex and* h*-smooth in the set* $\Lambda(\delta)$.

Step 3: We show that when K is outside of the interior of $\Lambda(\frac{\delta}{3})$, $C(K)$ is larger than $C(K^*_{\text{lin}})$. The proof of Lemma 3 is in Appendix H.

Lemma 3. *Under the conditions of Lemma 2 and if further,* ℓ, ℓ' *satisfies,*

$$\ell \le \delta \frac{\sigma\sigma_x(1-\rho)^4}{96\Gamma^4 c^8 D_0^2}, \qquad \ell' \le \delta \frac{\sigma\sigma_x(1-\rho)^4}{96\Gamma^4 c^{11} D_0^3},$$

then for all $K \in \Omega/\Lambda(\frac{\delta}{3})$, $C(K) > C(K^*_{\text{lin}})$.

The above lemma shows that $C(K)$'s minimum must be achieved in set $\Lambda(\frac{\delta}{3})$ which lies in the interior of $\Lambda(\delta)$. Since $C(K)$ is strongly convex in $\Lambda(\delta)$, $C(K)$'s minimum in Ω must be uniquely achieved at a point $K^* \in \Lambda(\frac{\delta}{3})$, which is also the unique stationary point of $C(K)$ within $\Lambda(\delta)$.

Finally, we summarize the requirements for ℓ, ℓ' and δ in the above three lemmas and provide a condition for ℓ, ℓ' and an estimate of δ below which satisfies all the conditions in Lemma 1, 2, 3,

$$\ell \le \frac{(\sigma\sigma_x)^2(1-\rho)^8}{96^2\Gamma^9 c^{15} D_0^4}, \qquad \ell' \le \frac{(\sigma\sigma_x)^2(1-\rho)^8}{96^2\Gamma^9 c^{18} D_0^5}, \qquad \delta = \frac{\sigma_x\sigma(1-\rho)^4}{96\Gamma^5 c^7 D_0^2}.$$

With this, the proof of Theorem 1 is concluded.

C. Proof of Lemma 1: Stability of the Trajectories

We in fact show a more general result in the following lemma, of which part (a) leads to Lemma 1.

Lemma 4. *Assume* $K \in \Omega$ *and* $\ell \le \frac{1-\rho_0}{4c_0}$. *Then we have the following holds.*

(a) For any $x_0 \in \mathbb{R}^n$, $\|x_t\| \le c\rho^t\|x_0\|$, *where* $c = 2c_0$ *and* $\rho = \frac{\rho_0+1}{2}$.

(b) Let $\{x_t\}$ *and* $\{x'_t\}$ *be the state tracjectories starting from* $x_0 \in \mathbb{R}^n$ *and* $x'_0 \in \mathbb{R}^n$ *respectively. Then,* $\|x_t - x'_t\| \le c\rho^t\|x_0 - x'_0\|$. *A direct consequence is that* $\|\frac{\partial x_t}{\partial x_0}\| \le c\rho^t$.

(c) Again let $\{x_t\}$ *and* $\{x'_t\}$ *be the state tracjectories starting from* $x_0 \in \mathbb{R}^n$ *and* $x'_0 \in \mathbb{R}^n$. *Then* $\|\frac{\partial x_t}{\partial x_0} - \frac{\partial x'_t}{\partial x'_0}\| \le \frac{\ell' c^3}{1-\rho}\rho^{t-1}\|x_0 - x'_0\|$.

Proof. To prove part (a), we recursively expand the system trajectory as follows,

$$x_{t+1} = (A - BK)x_t + f(x_t) = (A - BK)^{t+1}x_0 + \sum_{k=0}^{t}(A - BK)^{t-k}f(x_k).$$

Taking the norm, and using $K \in \Omega$ and the Lipschitz property of f, we have,

$$\|x_{t+1}\| \le c_0\rho_0^{t+1}\|x_0\| + \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell\|x_k\|. \tag{3}$$

We use the following simple proposition on nonnegative scalar sequences satisfying inequalities of the form in (3).

Proposition 1. *If nonnegative sequence* a_t *is such that* $a_{t+1} \le \alpha_0\lambda_1^{t+1} + \sum_{k=0}^{t}\alpha_1\lambda_2^{t-k}a_k$ *where* $\lambda_0, \lambda_1 \in (0,1)$ *and* $\alpha_0 \ge a_0$ *Then,* $a_t \le \alpha\lambda^t$ *where* α, λ *can be any positive constant satisfying* $\lambda > \lambda_2$, $\lambda \ge \lambda_1$, $\frac{\alpha_0}{\alpha} + \frac{\alpha_1}{\lambda-\lambda_2} \le 1$. *In particular, we can pick* $\alpha = 2\alpha_0$, *and* $\lambda = \max(\lambda_1, \lambda_2 + 2\alpha_1)$.

Proof. We use induction. The proposotion is clear true for $t = 0$ as $\alpha \ge \alpha_0 \ge a_0$. Assume it is true for t, then,

$$\begin{aligned}\frac{a_{t+1}}{\alpha\lambda^{t+1}} &\le \frac{\alpha_0}{\alpha}(\frac{\lambda_1}{\lambda})^{t+1} + \sum_{k=0}^{t}\alpha_1\lambda_2^{t-k}\lambda^{k-t-1} \\ &= \frac{\alpha_0}{\alpha}(\frac{\lambda_1}{\lambda})^{t+1} + \frac{\alpha_1}{\lambda}\frac{1-(\frac{\lambda_2}{\lambda})^{t+1}}{1-\frac{\lambda_2}{\lambda}} < \frac{\alpha_0}{\alpha} + \frac{\alpha_1}{\lambda-\lambda_2} \le 1.\end{aligned}$$

□

Applying Proposition 1 to (3), we have $\|x_t\| \le 2c_0\|x_0\|(\rho_0 + 2c_0\ell)^t \le c\rho^t\|x_0\|$, where we have used $\rho_0 + 2c_0\ell \le \rho_0 + 2c_0\frac{1-\rho_0}{4c_0} = \rho$.

The proof of part (b) is identical. Notice that

$$x_{t+1} - x'_{t+1} = (A - BK)(x_t - x'_t) + f(x_t) - f(x'_t)$$

$$= (A - BK)^{t+1}(x_0 - x_0') + \sum_{k=0}^{t}(A - BK)^{t-k}(f(x_k) - f(x_k')).$$

As such,

$$\|x_{t+1} - x_{t+1}'\| \leq c_0\rho_0^{t+1}\|x_0 - x_0'\| + \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell\|x_k - x_k'\|,$$

which leads to $\|x_t - x_t'\| \leq 2c_0\|x_0 - x_0'\|(\rho_0 + 2c_0\ell)^t \leq c\rho^t\|x_0 - x_0'\|$.

For part (c), we have

$$\frac{\partial x_{t+1}}{\partial x_0} = (A - BK)\frac{\partial x_t}{\partial x_0} + \frac{\partial f(x_t)}{\partial x_t}\frac{\partial x_t}{\partial x_0},$$

and therefore,

$$\begin{aligned}\frac{\partial x_{t+1}}{\partial x_0} - \frac{\partial x_{t+1}'}{\partial x_0'} &= (A - BK)(\frac{\partial x_t}{\partial x_0} - \frac{\partial x_t'}{\partial x_0'}) + \frac{\partial f(x_t)}{\partial x_t}(\frac{\partial x_t}{\partial x_0} - \frac{\partial x_t'}{\partial x_0'}) + (\frac{\partial f(x_t)}{\partial x_t} - \frac{\partial f(x_t')}{\partial x_t'})\frac{\partial x_t'}{\partial x_0'} \\ &= \sum_{k=0}^{t}(A - BK)^{t-k}\Big[\frac{\partial f(x_k)}{\partial x_k}(\frac{\partial x_k}{\partial x_0} - \frac{\partial x_k'}{\partial x_0'}) + (\frac{\partial f(x_k)}{\partial x_k} - \frac{\partial f(x_k')}{\partial x_k'})\frac{\partial x_k'}{\partial x_0'}\Big]\end{aligned}$$

Taking the norm and using the Lipschitz continuity of $\frac{\partial f(x)}{\partial x}$ in Assumption 4, we get

$$\begin{aligned}\|\frac{\partial x_{t+1}}{\partial x_0} - \frac{\partial x_{t+1}'}{\partial x_0'}\| &\leq \sum_{k=0}^{t} c_0\rho_0^{t-k}\Big[\ell\|\frac{\partial x_k}{\partial x_0} - \frac{\partial x_k'}{\partial x_0'}\| + \ell'\|x_k - x_k'\|\|\frac{\partial x_k'}{\partial x_0'}\|\Big] \\ &\leq \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell\|\frac{\partial x_k}{\partial x_0} - \frac{\partial x_k'}{\partial x_0'}\| + \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell'(c\rho^k)^2\|x_0 - x_0'\| \\ &\leq \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell\|\frac{\partial x_k}{\partial x_0} - \frac{\partial x_k'}{\partial x_0'}\| + \sum_{k=0}^{t} c_0\ell' c^2\rho^{t+k}\|x_0 - x_0'\| \\ &< \sum_{k=0}^{t} c_0\rho_0^{t-k}\ell\|\frac{\partial x_k}{\partial x_0} - \frac{\partial x_k'}{\partial x_0'}\| + c_0\ell' c^2\|x_0 - x_0'\|\frac{\rho^t}{1-\rho}.\end{aligned}$$

With this, we can invoke Proposition 1 and show that,

$$\|\frac{\partial x_t}{\partial x_0} - \frac{\partial x_t'}{\partial x_0'}\| \leq \frac{2c_0\ell' c^2}{\rho(1-\rho)}\rho^t\|x_0 - x_0'\| = \frac{\ell' c^3}{1-\rho}\rho^{t-1}\|x_0 - x_0'\|.$$

□

With the trajectory geometrically converging to zero, we also provide the following two auxiliary lemmas that will be used in the rest of the proof.

Lemma 5. *For $K \in \Omega$, define*

$$\Sigma_K = \mathbb{E}_K \sum_{t=0}^{\infty} x_t x_t^\top, \qquad \Sigma_K^{fx} = \mathbb{E}_K \sum_{t=0}^{\infty} f(x_t)x_t^\top.$$

Then, under the same conditions as in Lemma 1, we have,

$$\|\Sigma_K\| \leq C_\Sigma := \frac{c^2 D_0^2}{1-\rho}, \qquad \|\Sigma_K^{fx}\| \leq \ell C_\Sigma.$$

Proof. As a direct consequence of Lemma 1,

$$\|\Sigma_K\| \leq \mathbb{E}_K \sum_{t=0}^{\infty} \|x_t\|^2 \leq \frac{c^2}{1-\rho^2}\,\mathbb{E}\,\|x_0\|^2 \leq \frac{c^2 D_0^2}{1-\rho}.$$

Similarly, using the Lipschitz continuity of f,

$$\|\Sigma_K^{fx}\| \leq \mathbb{E}_K \sum_{t=0}^{\infty} \ell\|x_t\|^2 \leq \ell\frac{c^2 D_0^2}{1-\rho}.$$

□

Lemma 6. *For $K \in \Omega$, let P_K be the solution to the following Lyapunov equation,*

$$(A-BK)^\top P_K (A-BK) - P_K + Q + K^\top R K = 0.$$

Then, under the conditions of Lemma 1, and further when $K \in \Lambda(1)$, we have

$$\|P_K\| \le C_P := \frac{c^2}{1-\rho}\Gamma^2.$$

Proof. Note that $P_K = \sum_{t=0}^{\infty}((A-BK)^\top)^t(Q+K^\top RK)(A-BK)^t$, we have

$$\begin{aligned}\|P_K\| &\le \frac{c_0^2}{1-\rho_0^2}\|Q+K^\top RK\| \le \frac{c_0^2}{1-\rho_0}(1+\|K\|^2)\\ &\le \frac{c_0^2}{1-\rho_0}5\Gamma^2 < \frac{c^2}{1-\rho}\Gamma^2 := C_P,\end{aligned}$$

where we have used $\|K\| \le \|K-K^*_{\text{lin}}\| + \|K^*_{\text{lin}}\| \le \|K-K^*_{\text{lin}}\|_F + \|K^*_{\text{lin}}\| \le \|K^*_{\text{lin}}\| + 1 \le 2\Gamma$. □

D. Proof of Lemma 2: Strong Convexity and Smoothness

First off, note that under the conditions of Lemma 2, the conditions in Lemma 1 are satisfied, and we can use all the results in Appendix C, incluidng Lemma 4, Lemma 5 and Lemma 6. Further, it is easy to check that the conditions in this lemma also guarantees $\Lambda(\delta) = \{K : \|K-K^*_{\text{lin}}\|_F \le \delta\} \subset \Omega$ (which only requires $\delta \le \frac{1-\rho_0}{c_0\Gamma}$).

In the following, we provide a characterization of the value function, the gradient, and provide a cost differential formula. Here the value and Q function under a given controller K are defined as,

$$V_K(x) = \mathbb{E}_K\Big[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \Big| x_0 = x\Big],$$

and

$$Q_K(x,u) = \mathbb{E}_K\Big[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \Big| x_0 = x, u_0 = u\Big] = x^\top Q x + u^\top R u + V_K(Ax+Bu+f(x)).$$

The following lemma provides a characterization of the value function. The proof of Lemma 7 is given in Appendix E.

Lemma 7 (Value Function)**.** *When $K \in \Omega$, we have,*

$$V_K(x) = x^\top P_K x + g_K(x) \tag{4}$$

where P_K is the solution to the following Lyapunov equation,

$$(A-BK)^\top P_K (A-BK) - P_K + Q + K^\top R K = 0, \tag{5}$$

and function g_K is given by,

$$g_K(x) = 2\operatorname{Tr} P_K(A-BK)\sum_{t=0}^{\infty} x_t f(x_t)^\top + \operatorname{Tr} P_K \sum_{t=0}^{\infty} f(x_t)f(x_t)^\top, \tag{6}$$

where $\{x_t\}_{t=0}^{\infty}$ is the trajectory generated by controller K with initial state $x_0 = x$. Further, when $K \in \Lambda(\delta)$, and when $x, x' \in \mathbb{R}^n$ with $\|x\|, \|x'\| \le 2c^2 D_0$, we have,

$$\|\nabla g_K(x) - \nabla g_K(x')\| \le L\|x-x'\|,$$

where $L = (\ell + 2\ell' c^3 D_0)\frac{4C_P c^4}{(1-\rho)^2} = (\ell + 2\ell' c^3 D_0)\frac{4\Gamma^2 c^6}{(1-\rho)^3}$ with C_P being the upper bound on $\|P_K\|$ from Lemma 6.

Given that $C(K) = \mathbb{E}_{x\sim\mathcal{D}} V_K(x)$, the formula for $V_K(x)$ in the preceding Lemma 7 also leads to a formula for the gradient of $C(K)$, which is formally provided in the following lemma, whose proof is postponed to Appendix E.

Lemma 8 (Gradient of $C(K)$)**.** *Recall the cost function is $C(K) = \mathbb{E}_{x\sim\mathcal{D}} V_K(x)$. We have,*

$$\nabla C(K) = 2E_K\Sigma_K - 2B^\top P_K \Sigma_K^{fx} - B^\top \Sigma_K^{gx}$$

where E_K, Σ_K, Σ_K^{fx} and Σ_K^{gx} are defined as:

$$E_K = RK - B^\top P_K(A-BK) = (R+B^\top P_K B)K - B^\top P_K A, \tag{7}$$

$$\Sigma_K = \mathbb{E}_K\sum_{t=0}^{\infty} x_t x_t^\top, \quad \Sigma_K^{fx} = \mathbb{E}_K\sum_{t=0}^{\infty} f(x_t)x_t^\top, \quad \Sigma_K^{gx} = \mathbb{E}_K\sum_{t=0}^{\infty} \nabla_x g_K(x_{t+1})x_t^\top. \tag{8}$$

We also provide a formula for $C(K') - C(K)$, whose proof can be found in Appendix E.

Lemma 9 (Cost differential formula)**.** *We have for any $K, K' \in \Omega$,*

$$\begin{aligned}
&C(K') - C(K) \\
&= 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'} + \operatorname{Tr}(K'-K)^\top (R + B^\top P_K B)(K'-K)\Sigma_{K'} - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_{K'}^{fx} \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} \Big[g_K((A-BK')x_t' + f(x_t')) - g_K((A-BK)x_t' + f(x_t')) \Big]. \qquad (9)
\end{aligned}$$

With these preparations, we now proceed to prove Lemma 2, the strong convexity and smoothness of $C(K)$ within $\Lambda(\delta)$.

Proof of Lemma 2: We first focus on the strong convexity. By Lemma 9, we have for $K, K' \in \Lambda(\delta)$,

$$\begin{aligned}
&C(K') - C(K) \\
&= 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'} + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_{K'}^{fx} \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} [g_K((A-BK')x_t' + f(x_t')) - g_K((A-BK)x_t' + f(x_t'))] \\
&\overset{(a)}{\geq} 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_K + 2\operatorname{Tr}(K'-K)^\top E_K (\Sigma_{K'} - \Sigma_K) + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} \\
&\quad - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_K^{fx} + 2\operatorname{Tr}(K'-K)^\top B^\top P_K (\Sigma_K^{fx} - \Sigma_{K'}^{fx}) \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} [-\operatorname{Tr}(K'-K)^\top B^\top \nabla g_K(x_{t+1}')x_t'^\top - \frac{L}{2}\|B(K'-K)x_t'\|^2] \\
&= \operatorname{Tr}(K'-K)^\top \Big[2E_K \Sigma_K - 2B^\top P_K \Sigma_K^{fx} - \mathbb{E}_K \sum_{t=0}^{\infty} B^\top \nabla g_K(x_{t+1})x_t^\top \Big] \\
&\quad + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} \\
&\quad + 2\operatorname{Tr}(K'-K)^\top E_K (\Sigma_{K'} - \Sigma_K) + 2\operatorname{Tr}(K'-K)^\top B^\top P_K (\Sigma_K^{fx} - \Sigma_{K'}^{fx}) \\
&\quad + \operatorname{Tr}(K'-K)^\top B^\top [\mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1})x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x_{t+1}')x_t'^\top] \\
&\quad - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \frac{L}{2}\|B(K'-K)x_t'\|^2] \\
&\overset{(b)}{\geq} \operatorname{Tr}(K'-K)^\top \nabla C(K) + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} \\
&\quad - 2\|K'-K\|_F \|E_K\| \|\Sigma_{K'} - \Sigma_K\|_F - 2\|K'-K\|_F \|B\| \|P_K\| \|\Sigma_K^{fx} - \Sigma_{K'}^{fx}\|_F \\
&\quad - \|K'-K\|_F \|B\| \Big\| \mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1})x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x_{t+1}')x_t'^\top \Big\|_F \\
&\quad - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \frac{L}{2}\|B(K'-K)x_t'\|^2] \qquad (10)
\end{aligned}$$

where in step (b) we have used the gradient formula in Lemma 8, and in step (a) we have used,

$$\begin{aligned}
&g_K((A-BK)x_t' + f(x_t')) \\
&\leq g_K((A-BK')x_t' + f(x_t')) + \langle \nabla g_K((A-BK')x_t' + f(x_t')), B(K'-K)x_t' \rangle + \frac{L}{2}\|B(K'-K)x_t'\|^2 \\
&= g_K((A-BK')x_t' + f(x_t')) + \langle \nabla g_K(x_{t+1}'), B(K'-K)x_t' \rangle + \frac{L}{2}\|B(K'-K)x_t'\|^2 \\
&= g_K((A-BK')x_t' + f(x_t')) + \operatorname{Tr}(K'-K)^\top B^\top \nabla g_K(x_{t+1}')x_t'^\top + \frac{L}{2}\|B(K'-K)x_t'\|^2.
\end{aligned}$$

In the above, we have used the second part of Lemma 7 on the Lipschitz continuity of $\nabla g_K(x)$, which applies here as $K \in \Lambda(\delta)$ and since $\|(A-BK')x_t' + f(x_t')\| = \|x_{t+1}'\| \leq c\rho^{t+1}\|x_0'\| \leq cD_0$, and $\|(A-BK)x_t' + f(x_t')\| \leq \|A-BK\|\|x_t'\| + \|f(x_t')\| \leq (c+\ell)c\rho^t\|x_0'\| \leq 2c^2 D_0$ (using $\ell \leq 1 \leq c$).

Equation (10) can lead to strong convexity if we can show its first two terms dominates its last 4 terms. For this purposes, we show the following Lemma 10 and 11 to control the last 4 terms in (10). The proofs of Lemma 10 and 11 can be found in Section F and Section G respectively.

Lemma 10. *For $K \in \Lambda(\delta)$, we have,*

$$\|E_K\| \le C_E\|K - K^*_{\text{lin}}\| \le C_E\|K - K^*_{\text{lin}}\|_F \le C_E\delta,$$

where $C_E = 4\frac{\Gamma^4 c^4}{(1-\rho)^2}$.

Lemma 11. *There exists constant* $C_1 = \frac{2c^3\Gamma D_0^2}{(1-\rho)^2}, C_2 = \ell C_1, C_3 = LC_1$ *such that for all* $K, K' \in \Lambda(\delta)$,

$$\|\Sigma_{K'} - \Sigma_K\|_F \le C_1\|K' - K\|_F, \quad \|\Sigma^{fx}_{K'} - \Sigma^{fx}_K\|_F \le C_2\|K' - K\|_F,$$

$$\|\,\mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1})x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x'_{t+1})x_t'^\top\|_F \le C_3\|K - K'\|_F.$$

With the help of Lemma 10 and Lemma 11, we proceed with (10),

$$\begin{aligned}
&C(K') - C(K)\\
&\ge \operatorname{Tr}(K' - K)^\top \nabla C(K) + \operatorname{Tr}(K' - K)^\top [R + B^\top P_K B](K' - K)\Sigma_{K'}\\
&- 2C_1\|E_K\|\|K' - K\|_F^2 - 2C_2\|B\|\|P_K\|\|K' - K\|_F^2 - C_3\|B\|\|K' - K\|_F^2 - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \frac{L}{2}\|B\|^2\|x'_t\|^2\|K' - K\|^2\\
&\ge \operatorname{Tr}(K' - K)^\top \nabla C(K) + \operatorname{Tr}(K' - K)^\top [R + B^\top P_K B](K' - K)\Sigma_{K'}\\
&\quad - \left[2C_1\|E_K\| + 2C_2\|B\|\|P_K\| + C_3\|B\| + \mathbb{E}_{K'} \sum_{t=0}^{\infty} \frac{L}{2}\|B\|^2\|x'_t\|^2\right]\|K' - K\|_F^2\\
&\ge \operatorname{Tr}(K' - K)^\top \nabla C(K) + \mu\|K' - K\|_F^2\\
&\quad - \left[2C_1C_E\delta + 2C_2\Gamma C_P + C_3\Gamma + \frac{L}{2}\frac{\Gamma^2 c^2 D_0^2}{1-\rho}\right]\|K' - K\|_F^2,
\end{aligned} \tag{11}$$

where in the last inequality, $\mu = \sigma_x\sigma$, and we have used since $P_K \succeq Q$ and $\Sigma_{K'} \succeq \mathbb{E}\, x_0 x_0^\top \succeq \sigma_x I$,

$$\begin{aligned}
\operatorname{Tr}(K' - K)^\top [R + B^\top P_K B](K' - K)\Sigma_{K'} &= \operatorname{Tr}[(K' - K)\Sigma_{K'}^{1/2}]^\top [R + B^\top P_K B](K' - K)\Sigma_{K'}^{1/2}\\
&\ge \operatorname{Tr}[(K' - K)\Sigma_{K'}^{1/2}]^\top [R + B^\top Q B](K' - K)\Sigma_{K'}^{1/2}\\
&\ge \sigma \operatorname{Tr}[(K' - K)\Sigma_{K'}^{1/2}]^\top (K' - K)\Sigma_{K'}^{1/2}\\
&= \sigma \operatorname{Tr}(K' - K)\Sigma_{K'}(K' - K)^\top\\
&\ge \sigma\sigma_x\|K' - K\|_F^2.
\end{aligned}$$

From (11), it is clear that if we can show,

$$2C_1C_E\delta + 2C_2\Gamma C_P + C_3\Gamma + \frac{L}{2}\frac{\Gamma^2 c^2 D_0^2}{1-\rho} \le \frac{\mu}{2}, \tag{12}$$

then the μ-strong convexity property is proven. It remains to check our selection of δ, ℓ, ℓ' is such that (12) is true. Plug in $C_2 = \ell C_1$ and $C_3 = LC_1$, we have,

$$\begin{aligned}
2C_1C_E\delta + 2C_2\Gamma C_P + C_3\Gamma + \frac{L}{2}\frac{\Gamma^2 c^2 D_0^2}{1-\rho} &\le 2C_1C_E\delta + 2\ell\Gamma C_1 C_P + 2L\Gamma C_1\\
&\le 2C_1C_E\delta + 16\Gamma C_1[\ell + \ell' c^3 D_0]\frac{C_P c^4}{(1-\rho)^2}\\
&= 16\frac{\Gamma^5 c^7 D_0^2}{(1-\rho)^4}\delta + 32\frac{\Gamma^4 c^9 D_0^2}{(1-\rho)^5}\ell + 32\frac{\Gamma^4 c^{12} D_0^3}{(1-\rho)^5}\ell' \le \frac{\mu}{2},
\end{aligned}$$

where in the last step, we have used,

$$\begin{aligned}
\delta &\le \frac{\sigma_x\sigma}{6}\frac{(1-\rho)^4}{16\Gamma^5 c^7 D_0^2} = \frac{\sigma_x\sigma(1-\rho)^4}{96\Gamma^5 c^7 D_0^2},\\
\ell &\le \frac{\sigma_x\sigma}{6}\frac{(1-\rho)^5}{32\Gamma^4 c^9 D_0^2} = \frac{\sigma_x\sigma(1-\rho)^5}{192\Gamma^4 c^9 D_0^2},\\
\ell' &\le \frac{\sigma_x\sigma}{6}\frac{(1-\rho)^5}{32\Gamma^4 c^{12} D_0^3} = \frac{\sigma_x\sigma(1-\rho)^5}{192\Gamma^4 c^{12} D_0^3}.
\end{aligned}$$

This concludes the proof for the strong convexity. The proof for the smoothness property is similar. We follow similar steps as in (10) but reverse the direction of inequalities, getting,

$$
\begin{aligned}
&C(K') - C(K) \\
&= 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'} + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_{K'}^{fx} \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} [g_K((A-BK')x_t' + f(x_t')) - g_K((A-BK)x_t' + f(x_t'))] \\
&\le 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_K + 2\operatorname{Tr}(K'-K)^\top E_K(\Sigma_{K'} - \Sigma_K) + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} \\
&\quad - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_K^{fx} + 2\operatorname{Tr}(K'-K)^\top B^\top P_K (\Sigma_K^{fx} - \Sigma_{K'}^{fx}) \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} [-\operatorname{Tr}(K'-K)^\top B^\top \nabla g_K(x_{t+1}') x_t'^\top + \frac{L}{2}\|B(K'-K)x_t'\|^2] \\
&\le \operatorname{Tr}(K'-K)^\top \nabla C(K) + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} \\
&\quad + 2\|K'-K\|_F \|E_K\| \|\Sigma_{K'} - \Sigma_K\|_F + 2\|K'-K\|_F \|B\| \|P_K\| \|\Sigma_K^{fx} - \Sigma_{K'}^{fx}\|_F \\
&\quad + \|K'-K\|_F \|B\| \Big\| \mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1}) x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x_{t+1}') x_t'^\top \Big\|_F \\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} \frac{L}{2}\|B(K'-K)x_t'\|^2] \\
&\le \operatorname{Tr}(K'-K)^\top \nabla C(K) + \operatorname{Tr}(K'-K)^\top [R + B^\top P_K B](K'-K)\Sigma_{K'} + \frac{\mu}{2}\|K'-K\|_F^2 \\
&\le \operatorname{Tr}(K'-K)^\top \nabla C(K) + \frac{1}{2}(\mu + 2\|R + B^\top P_K B\| \|\Sigma_{K'}\|)\|K'-K\|_F^2. \qquad (13)
\end{aligned}
$$

Using the upper bound on $\|P_K\|$ and $\|\Sigma_{K'}\|$ in Lemma 6 and Lemma 5 respectively, we get

$$
\mu + 2\|R + B^\top P_K B\| \|\Sigma_{K'}\| \le \mu + 2(1 + \Gamma^2 \frac{c^2\Gamma^2}{1-\rho}) \frac{c^2 D_0^2}{1-\rho} \le 5\frac{\Gamma^4 c^4 D_0^2}{(1-\rho)^2} = h.
$$

As such, the cost function $C(K)$ is h smooth within $\Lambda(\delta)$. This concludes the proof of Lemma 2.

E. Proof of Lemma 7, 8, 9: Characterization of $C(K)$ and its Gradient.

Proof of Lemma 7. Since $K \in \Omega$, by Lemma 1, we have $V_K(x) \le \|Q + K^\top R K\| \frac{c^2}{1-\rho^2}\|x\|^2$. As such, $V_K(x)$ is finite and satisfies So $V_K(x) \to 0$ as $x \to 0$.

By Bellman equation, the value function also satisfies,

$$
V_K(x) = x^\top (Q + K^\top R K)x + V_K((A - BK)x + f(x)). \qquad (14)
$$

Define $g_K(x) = V_K(x) - x^\top P_K x$, we have

$$
x^\top P_K x + g_K(x) = x^\top (Q + K^\top R K)x + ((A-BK)x + f(x))^\top P_K((A-BK)x + f(x)) + g_K(x_1),
$$

where $x_1 = (A - BK)x + f(x)$. Since P_K satisfies (5), we have,

$$
\begin{aligned}
g_K(x) &= 2f(x)^\top P_K(A-BK)x + f(x)^\top P_K f(x) + g_K(x_1) \\
&= 2\operatorname{Tr}(P_K(A-BK)x f(x)^\top) + \operatorname{Tr} P_K f(x) f(x)^\top + g_K(x_1) \\
&= 2\operatorname{Tr} P_K(A-BK) \sum_{t=0}^{\infty} x_t f(x_t)^\top + \operatorname{Tr} P_K \sum_{t=0}^{\infty} f(x_t) f(x_t)^\top,
\end{aligned}
$$

where $\{x_t\}_{t=0}^{\infty}$ is the trajectory generated by controller K starting from $x_0 = x$. In the last step in the above equation, we have used $g_K(x_t) \to 0$ as $t \to \infty$, which is due to $g_K(x) \to 0$ as $x \to 0$ and $\|x_t\| \le c\rho^t\|x\| \to 0$ as $t \to \infty$.

Next, we show the second part of the Theorem. We first compute the gradient of $g_K(x)$ as follows,

$$
\begin{aligned}
&[\nabla g_K(x)]^\top \\
&= 2\sum_{t=0}^{\infty} \Big[f(x_t)^\top P_K(A-BK) + x_t^\top (A-BK)^\top P_K \frac{\partial f(x_t)}{\partial x_t} \Big] \frac{\partial x_t}{\partial x} + 2\sum_{t=0}^{\infty} f(x_t)^\top P_K \frac{\partial f(x_t)}{\partial x_t} \frac{\partial x_t}{\partial x} \\
&= 2\sum_{t=0}^{\infty} \Big[f(x_t)^\top P_K(A-BK) + x_{t+1}^\top P_K \frac{\partial f(x_t)}{\partial x_t} \Big] \frac{\partial x_t}{\partial x}. \qquad (15)
\end{aligned}
$$

To show that $\nabla g_K(x)$ is Lipschitz in x when $K \in \Lambda(\delta)$ and $\|x\| \le 2c^2 D_0$, we have for x, x' satsfying $\|x\|, \|x'\| \le 2c^2 D_0$,

$$\begin{aligned}
&\|\nabla g_K(x) - \nabla g_K(x')\| \\
&\le 2\sum_{t=0}^{\infty} \Big\| [f(x_t) - f(x_t')]^\top P_K (A - BK) + x_{t+1}^\top P_K \frac{\partial f(x_t)}{\partial x_t} - x_{t+1}'^\top P_K \frac{\partial f(x_t')}{\partial x_t'} \Big\| \| \frac{\partial x_t}{\partial x} \| \\
&\quad + 2\sum_{t=0}^{\infty} \Big\| f(x_t')^\top P_K (A - BK) + x_{t+1}'^\top P_K \frac{\partial f(x_t')}{\partial x_t'} \Big\| \| \frac{\partial x_t'}{\partial x'} - \frac{\partial x_t}{\partial x} \|. \qquad (16)
\end{aligned}$$

Using $\| \frac{\partial f(x)}{\partial x} - \frac{\partial f(x')}{\partial x'} \| \le \ell' \|x - x'\|$ (Assumption 4) and the fact that for any t, by Lemma 1, $\|x_t'\| \le c\|x'\| \le 2c^3 D_0 := D$, we have,

$$\begin{aligned}
&\|x_{t+1}^\top P_K \frac{\partial f(x_t)}{\partial x_t} - x_{t+1}'^\top P_K \frac{\partial f(x_t')}{\partial x_t'} \| \\
&\le \|(x_{t+1} - x_{t+1}')^\top P_K \frac{\partial f(x_t)}{\partial x_t}\| + \|x_{t+1}'^\top P_K (\frac{\partial f(x_t')}{\partial x_t'} - \frac{\partial f(x_t)}{\partial x_t})\| \\
&\le \ell C_P \|x_{t+1} - x_{t+1}'\| + D C_P \ell' \|x_t - x_t'\| \\
&\le C_P(\ell + D\ell') c \|x - x'\|, \qquad (17)
\end{aligned}$$

where in the second last inequality, we have used the bound on $\|P_K\|$ when $K \in \Lambda(\delta)$ (cf. Lemma 6), and in the last inequality, we have used Lemma 4 (b). Further, we have,

$$\Big\| (f(x_t) - f(x_t'))^\top P_K (A - BK) \Big\| \le \ell \|x_t - x_t'\| C_P \|A - BK\| \le \ell C_P c^2 \|x - x'\|, \qquad (18)$$

where we have used $\|A - BK\| \le c_0 \le c$. Also notice,

$$\Big\| f(x_t')^\top P_K (A - BK) + x_{t+1}'^\top P_K \frac{\partial f(x_t')}{\partial x_t'} \Big\| \le \ell \|x_t'\| C_P c + \|x_{t+1}'\| C_P \ell \le 2\ell D C_P c. \qquad (19)$$

Plugging in (17), (18), (19) into (16), and using $\| \frac{\partial x_t}{\partial x} \| \le c\rho^t$ (Lemma 4 (b)), $\| \frac{\partial x_t'}{\partial x'} - \frac{\partial x_t}{\partial x} \| \le \frac{\ell' c^3}{(1-\rho)} \rho^{t-1} \|x - x'\|$ (Lemma 4 (c)), we get,

$$\begin{aligned}
&\|\nabla g_K(x) - \nabla g_K(x')\| \\
&\le 2\sum_{t=0}^{\infty} \Big[\ell C_P c^2 \|x - x'\| + C_P(\ell + D\ell') c \|x - x'\| \Big] \| \frac{\partial x_t}{\partial x} \| + 2\sum_{t=1}^{\infty} 2\ell D C_P c \| \frac{\partial x_t'}{\partial x'} - \frac{\partial x_t}{\partial x} \| \\
&\le 2\Big[\ell C_P c^2 \|x - x'\| + C_P(\ell + D\ell') c \|x - x'\| \Big] \frac{c}{1-\rho} + 4\ell D C_P c \frac{\ell' c^3}{(1-\rho)^2} \|x - x'\| \\
&\le \Big[(2\ell + \ell' D) \frac{2 C_P c^3}{1-\rho} + 4\ell\ell' D \frac{C_P c^4}{(1-\rho)^2} \Big] \|x - x'\| \\
&\le (\ell + \ell' D) \frac{4 C_P c^4}{(1-\rho)^2} \|x - x'\|,
\end{aligned}$$

where in the last inequality, we have used $4\ell \le 2$. This shows $\nabla g_K(x)$ is L-Lipschitz continuous in x. □

Proof of Lemma 8. In (14), we take derivative of $V_K(x)$ w.r.t. K, and have

$$\nabla_K V_K(x) = 2RKxx^\top + \nabla_K V(x_1) + (\frac{\partial x_1}{\partial K})^\top \nabla_x V_K(x_1).$$

To proceed, note the directional derivative of x_1 w.r.t. K in the direction of Δ is $x_1'[\Delta] = -B\Delta x$. Therefore,

$$(x_1'[\Delta])^\top \nabla_x V_K(x_1) = -x^\top \Delta^\top B^\top [2P_K x_1 + \nabla_x g_K(x_1)] = \operatorname{Tr} \Delta^\top (-2B^\top P_K x_1 x^\top - B^\top \nabla g_K(x_1) x^\top)$$

This implies that

$$\begin{aligned}
\nabla_K V_K(x) &= 2RKxx^\top - 2B^\top P_K [(A - BK)x + f(x)]x^\top - B^\top \nabla_x g_K(x_1) x^\top + \nabla_K V(x_1) \\
&= (2RK - 2B^\top P_K (A - BK)) xx^\top - 2B^\top P_K f(x) x^\top - B^\top \nabla_x g_K(x_1) x^\top + \nabla_K V(x_1) \\
&= 2E_K \sum_{t=0}^{\infty} x_t x_t^\top - 2B^\top P_K \sum_{t=0}^{\infty} f(x_t) x_t^\top - B^\top \sum_{t=0}^{\infty} \nabla_x g_K(x_{t+1}) x_t^\top,
\end{aligned}$$

where $\{x_t\}$ is the trajectory starting from $x_0 = x$. Taking expectation w.r.t. x_0 and we are done. □

Proof of Lemma 9. By [7, Lemma 10], we have

$$V_{K'}(x) - V_K(x) = \sum_{t=0}^{\infty} A_K(x'_t, u'_t)$$

where $\{x'_t, u'_t\}$ is the trajectory generated by $x'_0 = x$ and $u'_t = -K'x'_t$, and $A_K(x,u) = Q_K(x,u) - V_K(x)$ is the advantage function [72].

Now, for given $u = -K'x$, we have

$$\begin{aligned}
&A_K(x,u) = Q_K(x,u) - V_K(x)\\
&= x^\top(Q + (K')^\top RK')x + V_K((A-BK')x + f(x)) - V_K(x)\\
&= x^\top(Q + (K-K+K')^\top R(K-K+K'))x + V_K((A-BK')x + f(x)) - V_K(x)\\
&= x^\top(Q + K^\top RK)x + x^\top(2(K'-K)^\top RK + (K'-K)^\top R(K'-K))x + V_K((A-BK')x + f(x)) - V_K(x)\\
&= x^\top(2(K'-K)^\top RK + (K'-K)^\top R(K'-K))x + V_K((A-BK')x + f(x)) - V_K((A-BK)x + f(x)). \qquad (20)
\end{aligned}$$

We next compute, using the formula for value function in Lemma 7,

$$\begin{aligned}
&V_K((A-BK')x + f(x)) - V_K((A-BK)x + f(x))\\
&= ((A-BK')x + f(x))^\top P_K((A-BK')x + f(x)) - ((A-BK)x + f(x))^\top P_K((A-BK)x + f(x))\\
&\quad + g_K((A-BK')x + f(x)) - g_K((A-BK)x + f(x))\\
&= 2(B(K-K')x)^\top P_K((A-BK)x + f(x)) + x^\top(K-K')^\top B^\top P_K B(K-K')x\\
&\quad + g_K((A-BK')x + f(x)) - g_K((A-BK)x + f(x))\\
&= 2x^\top(K-K')^\top B^\top P_K(A-BK)x + 2x^\top(K-K')^\top B^\top P_K f(x) + x^\top(K-K')^\top B^\top P_K B(K-K')x\\
&\quad + g_K((A-BK')x + f(x)) - g_K((A-BK)x + f(x)).
\end{aligned}$$

Plugging the above into (20), we have

$$\begin{aligned}
A_K(x,u) = {}& 2\operatorname{Tr}(K'-K)^\top[RK - B^\top P_K(A-BK)]xx^\top + \operatorname{Tr}((K'-K)^\top[R + B^\top P_K B](K'-K))xx^\top\\
&- 2\operatorname{Tr}(K'-K)^\top B^\top P_K f(x)x^\top + g_K((A-BK')x + f(x)) - g_K((A-BK)x + f(x)).
\end{aligned}$$

As a result, we have,

$$\begin{aligned}
&C(K') - C(K)\\
&= \mathbb{E}_{K'} \sum_{t=0}^{\infty} A_K(x'_t, -K'x'_t)\\
&= 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'} + \operatorname{Tr}(K'-K)^\top[R + B^\top P_K B](K'-K)\Sigma_{K'} - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma^{fx}_{K'}\\
&\quad + \mathbb{E}_{K'} \sum_{t=0}^{\infty} [g_K((A-BK')x'_t + f(x'_t)) - g_K((A-BK)x'_t + f(x'_t))].
\end{aligned}$$

□

F. Proof of Lemma 10: bounds on $\|E_K\|$

Note that $K \in \Lambda(\delta)$, $E_K = RK - B^\top P_K(A-BK)$. Further, by [7], $E_{K^*_{\text{lin}}} = 0$. Then, we have ,

$$\begin{aligned}
&\|E_K\| = \|E_K - E_{K^*_{\text{lin}}}\|\\
&\le \|R(K - K^*_{\text{lin}})\| + \|B^\top P_{K^*_{\text{lin}}} B(K - K^*_{\text{lin}})\| + \|B^\top(P_K - P_{K^*_{\text{lin}}})(A-BK)\|\\
&\le (1 + \Gamma^2 C_P)\|K - K^*_{\text{lin}}\| + \Gamma c \frac{2\Gamma^3 c^3}{(1-\rho)^2}\|K - K^*_{\text{lin}}\|\\
&\le 4\frac{\Gamma^4 c^4}{(1-\rho)^2}\|K - K^*_{\text{lin}}\|,
\end{aligned}$$

where in the second inequality, we have used Lemma 12 which is provided below. This concludes the proof of Lemma 10

Lemma 12 (Perturbation of P_K)**.** *When $K \in \Lambda(\delta)$, we have,*

$$\|P_K - P_{K^*_{\text{lin}}}\| \le \frac{2\Gamma^3 c^3}{(1-\rho)^2}\|K - K^*_{\text{lin}}\|$$

Proof. Recall that $P_K = \sum_{t=0}^{\infty}((A-BK)^\top)^t(Q+K^\top RK)(A-BK)^t$. We calculate the direction derivative of P_K w.r.t. K in the direction of Δ when $K \in \Lambda(\delta)$,

$$\begin{aligned}
&P_K'[\Delta] \\
&= \sum_{t=0}^{\infty}(((A-BK)^t)'[\Delta])^\top(Q+K^\top RK)(A-BK)^t + \sum_{t=0}^{\infty}((A-BK)^t)^\top(Q+K^\top RK)((A-BK)^t)'[\Delta] \\
&\quad + \sum_{t=0}^{\infty}((A-BK)^t)^\top(\Delta^\top RK + K^\top R\Delta)(A-BK)^t.
\end{aligned}$$

Notice that

$$((A-BK)^t)'[\Delta] = \sum_{k=1}^{t}(A-BK)^{k-1}(-B\Delta)(A-BK)^{t-k}.$$

Hence

$$\|((A-BK)^t)'[\Delta]\| \le \|B\|\|\Delta\|c_0^2 t\rho_0^{t-1} \le \|B\|c_0^2\frac{2}{1-\rho_0}(\frac{1+\rho_0}{2})^t\|\Delta\|,$$

where we have used the fact that $t\rho_0^{t-1} \le \frac{2}{1-\rho_0}(\frac{1+\rho_0}{2})^t$. As such, we have

$$\begin{aligned}
\|P_K'[\Delta]\| &\le 2\sum_{t=0}^{\infty}\|B\|\|Q+K^\top RK\|c_0\rho_0^t c_0^2\frac{2}{1-\rho_0}(\frac{1+\rho_0}{2})^t\|\Delta\| + 2\sum_{t=0}^{\infty}c_0^2\rho_0^{2t}\|K^\top R\|\|\Delta\| \\
&< 8\|B\|\|Q+K^\top RK\|\frac{c_0^3}{(1-\rho_0)^2}\|\Delta\| + 2\|K^\top R\|\frac{c_0^2}{(1-\rho_0)}\|\Delta\| \\
&\le (8\Gamma + 8\Gamma\|K\|^2 + 2\|K\|)\frac{c_0^3}{(1-\rho_0)^2}\|\Delta\|.
\end{aligned}$$

We further use that $\|K\| \le \|K_{\text{lin}}^*\| + \delta \le 2\Gamma$ (using $\delta \le 1 \le \Gamma$), then, we have,

$$\|P_K'[\Delta]\| \le 44\Gamma^3\frac{c_0^3}{(1-\rho_0)^2}\|\Delta\| \le \frac{2\Gamma^3c^3}{(1-\rho)^2}\|\Delta\|.$$

Using a simple integration argument on the line between K_{lin}^* and K, we have

$$\|P_K - P_{K_{\text{lin}}^*}\| \le \frac{2\Gamma^3c^3}{(1-\rho)^2}\|K - K_{\text{lin}}^*\|.$$

□

G. Proof of Lemma 11: Bounds on C_1, C_2, C_3

Before we start the proof, we first provide an auxiliary result on the perturbation of trajectories by a change of controller K.

Lemma 13. *For $K \in \Lambda(\delta)$, given x_0, the directional derivative of x_t w.r.t. K in the direction of Δ satisfies,*

$$\|x_t'[\Delta]\| \le \frac{c^2\Gamma}{1-\rho}\rho^t\|x_0\|\|\Delta\|.$$

As a direct consequence, for $K, K' \in \Lambda(\delta)$, let $\{x_t\}_{t=0}^{\infty}$ and $\{x_t'\}_{t=0}^{\infty}$ be two trajectories starting from the same $x_0 = x_0'$ generated by K and K' respectively. Then, we have $\|x_t - x_t'\| \le \frac{c^2\Gamma}{1-\rho}\rho^t\|x_0\|\|K'-K\|$.

Proof. The dynamical system is given by

$$x_{t+1} = (A-BK)x_t + f(x_t).$$

Taking derivative w.r.t. K in the direction of Δ, we have

$$x_{t+1}'[\Delta] = (A-BK)x_t'[\Delta] - B\Delta x_t + \frac{\partial f(x_t)}{\partial x_t}x_t'[\Delta] = \sum_{k=0}^{t}(A-BK)^{t-k}[-B\Delta x_k + \frac{\partial f(x_k)}{\partial x_k}x_k'[\Delta]].$$

Taking the norm and using the triangle inequality as well as the Lipschitz property of f, we get,

$$\|x_{t+1}'[\Delta]\| \le \sum_{k=0}^{t}c_0\rho_0^{t-k}\ell\|x_k'[\Delta]\| + \sum_{k=0}^{t}c_0\rho_0^{t-k}\|B\Delta\|\|x_k\|$$

$$\leq \sum_{k=0}^{t} c_0 \rho_0^{t-k} \ell \|x'_k[\Delta]\| + \sum_{k=0}^{t} c_0 \rho_0^{t-k} \|B\Delta\| c\rho^k \|x_0\|$$
$$\leq \sum_{k=0}^{t} c_0 \rho_0^{t-k} \ell \|x'_k[\Delta]\| + c_0 \|B\| \|\Delta\| c \|x_0\| \frac{\rho^{t+1} - \rho_0^{t+1}}{\rho - \rho_0}.$$

As such, by a simple induction argument (Proposition 1), we have

$$\|x'_t[\Delta]\| \leq 2c_0 \|B\| \|\Delta\| c \|x_0\| \frac{\rho^t}{\rho - \rho_0} = \frac{c^2 \Gamma}{1-\rho} \rho^t \|x_0\| \|\Delta\|.$$

□

We now proceed to prove Lemma 11.

Proof of Lemma 11. By definition, $\Sigma_K = \sum_{t=0}^{\infty} \mathbb{E}\, x_t x_t^\top$. For $K \in \Lambda(\delta)$, we take the directional derivative w.r.t. K in the direction of Δ, getting, we have

$$\Sigma'_K[\Delta] = \mathbb{E} \sum_{t=0}^{\infty} \left(x'_t[\Delta] x_t^\top + x_t x'_t[\Delta]^\top \right).$$

Then, using Lemma 13, we have,

$$\begin{aligned}
\|\Sigma'_K[\Delta]\|_F &\leq \mathbb{E} \sum_{t=0}^{\infty} 2 \|x'_t[\Delta]\| \|x_t\| \\
&\leq \mathbb{E} \sum_{t=0}^{\infty} 2 \frac{c^2 \Gamma}{1-\rho} \rho^t \|x_0\| \|\Delta\| c\rho^t \|x_0\| \\
&\leq \frac{2c^3 \Gamma D_0^2}{(1-\rho)^2} \|\Delta\| \\
&\leq \frac{2c^3 \Gamma D_0^2}{(1-\rho)^2} \|\Delta\|_F,
\end{aligned}$$

which, after a simple integration argument, gives a bound for C_1. Next, we consider the bound on C_2. Note that

$$\Sigma_K^{fx} = \mathbb{E} \sum_{t=0}^{\infty} f(x_t) x_t^\top.$$

Again, taking the derivative, we have,

$$(\Sigma_K^{fx})'[\Delta] = \mathbb{E} \sum_{t=0}^{\infty} [\frac{\partial f(x_t)}{\partial x_t} x'_t[\Delta] x_t^\top + f(x_t) x'_t[\Delta]^\top],$$

which leads to,

$$\|(\Sigma_K^{fx})'[\Delta]\|_F \leq \mathbb{E} \sum_{t=0}^{\infty} 2\ell \|x_t\| \|x'_t[\Delta]\| \leq \ell C_1 \|\Delta\|_F.$$

So we will get $C_2 = \ell C_1$.

Finally, we proceed to bound C_3. Recall the definition of C_3 is such that for $K, K' \in \Lambda(\delta)$,

$$\| \mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1}) x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x'_{t+1}) x_t'^\top \|_F \leq C_3 \|K - K'\|_F.$$

Fix x_0 for now with $\|x_0\| \leq D_0$, and consider the trajectories $\{x_t\}_{t=0}^{\infty}$ and $\{x'_t\}_{t=0}^{\infty}$ generated by controller K and K' starting from $x'_0 = x_0$. We have,

$$\begin{aligned}
&\|\nabla g_K(x_{t+1}) x_t^\top - \nabla g_K(x'_{t+1}) x_t'^\top\|_F \\
&\leq \|(\nabla g_K(x_{t+1}) - \nabla g_K(x'_{t+1})) x_t^\top\|_F + \|\nabla g_K(x'_{t+1})(x_t - x'_t)^\top\|_F \\
&\leq \|\nabla g_K(x_{t+1}) - \nabla g_K(x'_{t+1})\| \|x_t\| + \|\nabla g_K(x'_{t+1})\| \|x_t - x'_t\| \\
&\overset{(a)}{\leq} L \|x_{t+1} - x'_{t+1}\| \|x_t\| + L \|x'_{t+1}\| \|x_t - x'_t\| \\
&\overset{(b)}{\leq} L \frac{c^2 \Gamma}{1-\rho} \rho^{t+1} c\rho^t \|x_0\|^2 \|K' - K\| + Lc\rho^{t+1} \frac{c^2 \Gamma}{1-\rho} \rho^t \|x_0\|^2 \|K' - K\|
\end{aligned}$$

$$\leq L\frac{2c^3\Gamma D_0^2}{1-\rho}\rho^t\|K'-K\|,$$

where in inequality (a), we have used the Lipschitz continuity of $\nabla g_K(x)$ (Lemma 7), which holds here as $K \in \Lambda(\delta)$ and $\|x_{t+1}\| \leq cD_0, \|x'_{t+1}\| \leq cD_0$. In inequality (b), we have used the bound in Lemma 13. With the above bound, we can proceed to obtain C_3, getting,

$$\begin{aligned}
&\|\mathbb{E}_K \sum_{t=0}^{\infty} \nabla g_K(x_{t+1})x_t^\top - \mathbb{E}_{K'} \sum_{t=0}^{\infty} \nabla g_K(x'_{t+1})x_t'^\top\|_F \\
&\leq \sum_{t=0}^{\infty} \mathbb{E}_{K,K'} \|\nabla g_K(x_{t+1})x_t^\top - \nabla g_K(x'_{t+1})x_t'^\top\|_F \\
&\leq L\frac{2c^3\Gamma D_0^2}{(1-\rho)^2}\|K'-K\| \\
&\leq LC_1\|K'-K\|_F.
\end{aligned}$$

As a result, we can set $C_3 = LC_1$. □

H. Proof of Lemma 3: Global Optimality

By Lemma 9, we have

$$\begin{aligned}
C(K') - C(K) = {} & 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'} \\
& + \operatorname{Tr}(K'-K)^\top (R + B^\top P_K B)(K'-K)\Sigma_{K'} - 2\operatorname{Tr}(K'-K)^\top B^\top P_K \Sigma_{K'}^{fx} \\
& + \mathbb{E}_{K'} \sum_{t=0}^{\infty} \Big[g_K((A-BK')x'_t + f(x'_t)) - g_K((A-BK)x'_t + f(x'_t)) \Big].
\end{aligned}$$

Setting $K = K^*_{\text{lin}}$ in the above equation and using $E_{K^*_{\text{lin}}} = 0$ (cf. [7]), we get $\forall K \in \Omega$,

$$\begin{aligned}
C(K) - C(K^*_{\text{lin}}) = {} & \operatorname{Tr}(K-K^*_{\text{lin}})^\top (R + B^\top P_{K^*_{\text{lin}}} B)(K-K^*_{\text{lin}})\Sigma_K - 2\operatorname{Tr}(K-K^*_{\text{lin}})^\top B^\top P_{K^*_{\text{lin}}} \Sigma_K^{fx} \\
& + \mathbb{E}_K \sum_{t=0}^{\infty} \Big[g_{K^*_{\text{lin}}}((A-BK)x_t + f(x_t)) - g_{K^*_{\text{lin}}}((A-BK^*_{\text{lin}})x_t + f(x_t)) \Big] \\
\geq {} & \mu\|K-K^*_{\text{lin}}\|_F^2 - 2\|K-K^*_{\text{lin}}\|_F\|B\|\|P_{K^*_{\text{lin}}}\|\|\Sigma_K^{fx}\|_F \\
& + \mathbb{E}_K \sum_{t=0}^{\infty} \Big[g_{K^*_{\text{lin}}}((A-BK)x_t + f(x_t)) - g_{K^*_{\text{lin}}}((A-BK^*_{\text{lin}})x_t + f(x_t)) \Big],
\end{aligned} \tag{21}$$

where in the last inequality, we have used that by $R + B^\top P_{K^*_{\text{lin}}} B \succeq \sigma I$, $\Sigma_K \succeq \sigma_x I$, we have,

$$\operatorname{Tr}(K-K^*_{\text{lin}})^\top (R + B^\top P_{K^*_{\text{lin}}} B)(K-K^*_{\text{lin}})\Sigma_K \geq \sigma\sigma_x\|K-K^*_{\text{lin}}\|_F^2 = \mu\|K-K^*_{\text{lin}}\|_F^2.$$

Now we bound the last term in (21). Note that inside the expectation in the last term in (21), almost surely we have, $\|(A-BK)x_t + f(x_t)\| = \|x_{t+1}\| \leq cD_0$, and $\|(A-BK^*_{\text{lin}})x_t + f(x_t)\| \leq (c+\ell)\|x_t\| \leq 2c^2 D_0$ (using $\ell \leq 1 \leq c$). Therefore, we can invoke the second part of Lemma 7 on the smoothness of $g_{K^*_{\text{lin}}}$ and get almost surely,

$$\begin{aligned}
& g_{K^*_{\text{lin}}}((A-BK)x_t + f(x_t)) - g_{K^*_{\text{lin}}}((A-BK^*_{\text{lin}})x_t + f(x_t)) \\
& \geq -\operatorname{Tr}(B(K-K^*_{\text{lin}})x_t)^\top \nabla g_{K^*_{\text{lin}}}((A-BK)x_t + f(x_t)) - \frac{L}{2}\|B(K-K^*_{\text{lin}})x_t\|^2 \\
& \geq -\|K-K^*_{\text{lin}}\|_F\|B\|\|x_t\|L\|x_{t+1}\| - \frac{L}{2}\|B\|^2\|K-K^*_{\text{lin}}\|_F^2\|x_t\|^2 \\
& \geq -\|K-K^*_{\text{lin}}\|_F L\Gamma c^2 D_0^2 \rho^{2t} - \|K-K^*_{\text{lin}}\|_F^2 \frac{L\Gamma^2 c^2 D_0^2}{2}\rho^{2t}.
\end{aligned}$$

Plugging the above into (21) and using the easy to check fact that as $K \in \Omega$, $\|\Sigma_K^{fx}\|_F \leq \mathbb{E}\sum_{t=0}^{\infty} \ell\|x_t\|^2 \leq \ell\frac{c^2 D_0^2}{1-\rho}$, we have when $K \in \Omega/\Lambda(\frac{\delta}{3})$,

$$\begin{aligned}
C(K) - C(K^*_{\text{lin}}) &\geq \Big[\mu - \frac{1}{2}L\frac{\Gamma^2 c^2 D_0^2}{1-\rho}\Big]\|K-K^*_{\text{lin}}\|_F^2 - \Big[2\ell\frac{\Gamma^3 c^4 D_0^2}{(1-\rho)^2} + L\frac{\Gamma c^2 D_0^2}{1-\rho}\Big]\|K-K^*_{\text{lin}}\|_F \\
&> \|K-K^*_{\text{lin}}\|_F\Big[\Big(\mu - \frac{1}{2}L\frac{\Gamma^2 c^2 D_0^2}{1-\rho}\Big)\frac{\delta}{3} - 2\ell\frac{\Gamma^3 c^4 D_0^2}{(1-\rho)^2} - L\frac{\Gamma c^2 D_0^2}{1-\rho}\Big].
\end{aligned}$$

Therefore, it suffices to show that,

$$\frac{1}{2}L\frac{\Gamma^2c^2D_0^2}{1-\rho}=(\ell+2\ell'c^3D_0)\frac{2\Gamma^4c^8D_0^2}{(1-\rho)^4}\le\frac{1}{2}\mu,$$
$$2\ell\frac{\Gamma^3c^4D_0^2}{(1-\rho)^2}+L\frac{\Gamma c^2D_0^2}{1-\rho}<(\ell+\ell'c^3D_0)\frac{8\Gamma^3c^8D_0^2}{(1-\rho)^4}\le\frac{\delta\mu}{6}.$$

As such, it suffices to require

$$\ell\le\delta\frac{\sigma\sigma_x(1-\rho)^4}{96\Gamma^4c^8D_0^2},\ell'\le\delta\frac{\sigma\sigma_x(1-\rho)^4}{96\Gamma^4c^{11}D_0^3}.$$

□

I. Proof of Theorem 2 and Corollary 1: Convergence of Zeroth-Order Policy Search

We start with the following result regarding the accuracy of the gradient estimator, the proof of which is postponed to Section J.

Lemma 14. *Under the conditions of Theorem 1, when $K\in\Lambda(\frac{5}{6}\delta)$, then given e_{grad}, for any $\nu\in(0,1)$, when $r\le\min(\frac{1}{6}\delta,\frac{1}{3h}e_{grad})$,*

$$J\ge\frac{1}{e_{grad}^2}\frac{d^3}{r^2}\log\frac{4d}{\nu}\max(18(C(K^*)+2h\delta^2)^2,72C_{\max}^2),\quad T\ge\frac{2}{1-\rho_0}\log\frac{6dC_{\max}}{e_{grad}r},$$

where $d=pn$ and $C_{\max}=\frac{40\Gamma^2c_0^2}{1-\rho_0}D_0^2$, then with probability at least $1-\nu$,

$$\|\widehat{\nabla C}(K)-\nabla C(K)\|_F\le e_{grad}.$$

With the bound on the gradient estimator, we proceed to the proof of Theorem 2.

Proof of Theorem 2. Let $\mathcal{F}_m$ be the filtration generated by $\{\widehat{\nabla C}(K_{m'})\}_{m'=0}^{m-1}$. Then, we have K_m is $\mathcal{F}_m$ measurable. We define the following event,

$$\mathcal{E}_m=\{K_{m'}\in\text{Ball}(K^*,\frac{\delta}{2}),\forall m'=0,1,\dots,m\}$$
$$\cap\{\|\widehat{\nabla C}(K_{m'})-\nabla C(K_{m'})\|_F\le e_{grad},\forall m'=0,1,\dots,m-1\},$$

where $\text{Ball}(K^*,\frac{\delta}{2})=\{K:\|K-K^*\|_F\le\frac{\delta}{2}\}$, i.e. the ball centered at K^* with radius $\frac{\delta}{2}$. Clearly, $\mathcal{E}_m$ is also $\mathcal{F}_m$-measurable. We now show that conditioned on $\mathcal{E}_m$ is true, $\mathcal{E}_{m+1}$ happens with high probability, or in other words the following inequality,

$$\mathbb{E}(\mathbf{1}(\mathcal{E}_{m+1})|\mathcal{F}_m)\mathbf{1}(\mathcal{E}_m)\ge(1-\frac{\nu}{M})\mathbf{1}(\mathcal{E}_m).\tag{22}$$

To show (22), we now condition on $\mathcal{F}_m$. On event $\mathcal{E}_m$, we have by triangle inequality, $\|K_m-K_{\text{lin}}^*\|_F\le\|K^*-K_{\text{lin}}^*\|_F+\frac{\delta}{2}\le\frac{5}{6}\delta$, and hence $K_m\in\Lambda(\frac{5}{6}\delta)$. Therefore, by Lemma 14 and our selection of r,J,T, we have $\|\widehat{\nabla C}(K_m)-\nabla C(K_m)\|_F\le e_{grad}$ with probability at least $1-\frac{\nu}{M}$ (note we have replaced ν with ν/M in Lemma 14), which, as we show now, will further imply $K_{m+1}\in\text{Ball}(K^*,\frac{1}{2}\delta)$. To see this, as $K_m\in\Lambda(\frac{5}{6}\delta)$, we can use the μ-strong convexity and h-smoothness to get,

$$\begin{aligned}\|K_{m+1}-K^*\|_F&\le\|K_m-\eta\nabla C(K_m)-K^*\|_F+\eta\|\widehat{\nabla C}(K_m)-\nabla C(K_m)\|_F\\&\le(1-\eta\mu)\|K_m-K^*\|_F+\eta e_{grad}\\&\le\max(\frac{\delta}{2},\frac{1}{\mu}e_{grad})\le\frac{\delta}{2},\end{aligned}\tag{23}$$

where the second inequality is due to the contraction of gradient descent for strongly convex and smooth functions [73], and in the last step, we have used $e_{grad}\le\mu\frac{\delta}{3}$. As such, (22) is true, and taking expectation on both sides, we get,

$$\mathbb{P}(\mathcal{E}_{m+1})=\mathbb{P}(\mathcal{E}_{m+1}\cap\mathcal{E}_m)=\mathbb{E}[\mathbb{E}(\mathbf{1}(\mathcal{E}_{m+1})|\mathcal{F}_m)\mathbf{1}(\mathcal{E}_m)]\ge(1-\frac{\nu}{M})\mathbb{P}(\mathcal{E}_m).$$

As a result, we have, $\mathbb{P}(\mathcal{E}_M)\ge(1-\frac{\nu}{M})^M\mathbb{P}(\mathcal{E}_0)>1-\nu$, where we have used $\mathcal{E}_0$ is true almost surely as $K_0=K_{\text{lin}}^*\in\text{Ball}(K^*,\frac{\delta}{2})$.

Now, on the event $\mathcal{E}_M$, we have (23) is true for all $m=0,\dots,M-1$. As such, we have,

$$\|K_M-K^*\|_F\le(1-\eta\mu)^M\|K_0-K^*\|_F+\eta e_{grad}\sum_{m=0}^{M-1}(1-\eta\mu)^m$$
$$\le(1-\eta\mu)^M\frac{\delta}{3}+\frac{1}{\mu}e_{grad}$$

$$\leq \sqrt{\frac{2\varepsilon}{h}},$$

where we have used $M \geq \frac{1}{\eta\mu}\log(\delta\sqrt{\frac{h}{\varepsilon}})$, $e_{grad} \leq \frac{\mu}{2}\sqrt{\frac{\varepsilon}{h}}$. As such, by h-smoothness,

$$C(K_M) \leq C(K^*) + \frac{h}{2}\|K_M - K^*\|_F^2 \leq C(K^*) + \varepsilon,$$

which is the desired result. Note that the above is true only when conditioned on $\mathcal{E}_M$, as such the desired result is true with probability at least $1-\nu$. □

We next proceed to prove Corollary 1. Note that the only requirement on the initial point K_0 in the proof of Theorem 2 is that $K_0 \in \text{Ball}(K^*, \frac{\delta}{2})$. In the setting of Corollary 1, $K_0 = \hat{K}_{\text{lin}}$, the LQR controller based on $\hat{A}, \hat{B}$. Note that as long as $\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|_F \leq \frac{\delta}{6}$, then $\|\hat{K}_{\text{lin}} - K^*\|_F \leq \|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|_F + \|K^*_{\text{lin}} - K^*\|_F \leq \frac{\delta}{6} + \frac{\delta}{3} = \frac{\delta}{2} \Rightarrow K_0 \in \text{Ball}(K^*, \frac{\delta}{2})$. Therefore, to prove Corollary 1, we only need to show $\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|_F \leq \frac{\delta}{6}$ under the conditions of Corollary 1. This is done in the following lemma, which is a direct application of the LQR perturbation results in [21].

Lemma 15. *There exists a LQR perturbation constant* c_{per} *that depend on* A, B, Q, R *s.t. when* $\max(\|A-\hat{A}\|, \|B-\hat{B}\|) \leq \frac{\min(\delta,1)}{6\,\mathrm{c}_{\text{per}}}$, *we have* $\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|_F \leq \frac{\delta}{6}$.

Proof. Recall K^*_{lin} is the optimal LQR controller based on (A, B), and $\hat{K}_{\text{lin}}$ is the optimal LQR controller based on $(\hat{A}, \hat{B})$. We first recall the following LQR perturbation result in [21], which we rewrite using the notations in our paper.

Lemma 16. *([21, Proposition 1]) When* $\max(\|A-\hat{A}\|, \|B-\hat{B}\|) \leq \epsilon$, *then*

$$\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\| \leq \frac{7}{\sigma_{\min}(R)}(1 + \max(\|A\|, \|B\|, \|P^*\|, \|K^*_{\text{lin}}\|))^3 \max(\|\hat{P} - P^*\|, \epsilon),$$

where $\hat{P}$ *and* P^* *are the solution to the Algebraic Ricatti Equation for the system* $(\hat{A}, \hat{B})$, (A, B) *respectively, under cost matrix* (Q, R).

The above result bounds the difference $\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|$ in terms of the difference in the solution to the Algebraic Ricatti Equation $\|\hat{P} - P^*\|$. We review a separate result in [21, Proposition 3] that provides a bound on $\|\hat{P} - P^*\|$ in terms of ϵ. To state the result in [21, Proposition 3], we need to define a few constants that is based on A, B. Firstly, define ρ_A and c_A be such that $\|A^t\| \leq c_A \rho_A^t$, $\forall t \geq 0$. Further, as we have assumed the pair (A, B) is controllable in Assumption 2, there must exist[6] positive integer ℓ_{con}, and constant $\nu_{\text{con}} > 0$, such that

$$\begin{bmatrix} B, AB, \ldots, A^{\ell_{\text{con}}-1}B \end{bmatrix} \begin{bmatrix} B^\top \\ (AB)^\top \\ \vdots \\ (A^{\ell_{\text{con}}-1}B)^\top \end{bmatrix} \succeq \nu_{\text{con}}^2 I.$$

With these definitions, we state the following result.

Lemma 17. *([21, Proposition 3]) When* $\max(\|A-\hat{A}\|, \|B-\hat{B}\|) \leq \epsilon$, *then*

$$\|\hat{P} - P^*\| \leq 32\ell_{\text{con}}^{5/2}\tau_A^3(\max(1, \epsilon\tau_A + \rho_A))^{2(\ell_{\text{con}}-1)}(1 + \frac{1}{\nu_{\text{con}}})(1 + \|B\|)^2\|P^*\|\frac{\max(\|Q\|, \|R\|)}{\min(\sigma_{\min}(Q), \sigma_{\min}(R))}\epsilon,$$

as long as ϵ *is small enough s.t. the right hand side of above is upper bounded by* $\sigma_{\min}(R)$.

Define

$$\bar{\mathrm{c}}_{\text{per}} = 32\ell_{\text{con}}^{5/2}\tau_A^3(\max(1, \frac{\sigma_{\min}(R)}{\|P^*\|} + \rho_A))^{2(\ell_{\text{con}}-1)}(1 + \frac{1}{\nu_{\text{con}}})(1 + \|B\|)^2\|P^*\|\frac{\max(\|Q\|, \|R\|)}{\min(\sigma_{\min}(Q), \sigma_{\min}(R))},$$

and

$$\mathrm{c}_{\text{per}} = \frac{7}{\sigma_{\min}(R)}(1 + \max(\|A\|, \|B\|, \|P^*\|, \|K^*_{\text{lin}}\|))^3 \max(\bar{\mathrm{c}}_{\text{per}}, 1)\sqrt{n}.$$

Now we set $\epsilon = \frac{\min(\delta,1)}{6\,\mathrm{c}_{\text{per}}}$. As $\mathrm{c}_{\text{per}} > \frac{1}{\sigma_{\min}(R)}\bar{\mathrm{c}}_{\text{per}} > \frac{\|P^*\|}{\sigma_{\min}(R)}\tau_A$, we have $\epsilon\tau_A < \frac{\sigma_{\min}(R)}{\|P^*\|}$. As a result, the right hand side of the bound in Lemma 17 can be upper bounded by $\bar{\mathrm{c}}_{\text{per}}\,\epsilon < \frac{\bar{\mathrm{c}}_{\text{per}}}{\mathrm{c}_{\text{per}}} < \sigma_{\min}(R)$. Therefore, the bound in Lemma 17 holds, and we have $\|\hat{P} - P^*\| \leq \bar{\mathrm{c}}_{\text{per}}\,\epsilon$. We then combine this with Lemma 16, getting,

$$\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\|_F \leq \sqrt{n}\|\hat{K}_{\text{lin}} - K^*_{\text{lin}}\| \leq \frac{7}{\sigma_{\min}(R)}(1 + \max(\|A\|, \|B\|, \|P^*\|, \|K^*_{\text{lin}}\|))^3\sqrt{n}\max(\bar{\mathrm{c}}_{\text{per}}\,\epsilon, \epsilon) \leq \mathrm{c}_{\text{per}}\,\epsilon \leq \frac{\delta}{6},$$

[6]For example, as (A, B) is controllable, one can choose $\ell_{\text{con}} = n$, and ν_{con} to be the smallest singular value of the controllability matrix.

which concludes the proof. □

J. Proof of Lemma 14

Proof. As $r \le \frac{1}{6}\delta$ we have $K + U_j \in \Lambda(\delta)$ for all j. As such, both K and $K + U_j$ are inside $\Lambda(\delta)$, in which C is μ-strongly convex and h-smooth.

We start with a standard result in zeroth order optimization [70]. Define a "smoothed" version of the cost, $C_r(K) = \mathbb{E}_{U\sim \mathrm{Ball}(r)} C(K+U)$, where $\mathrm{Ball}(r)$ is the Ball centered at the origin with radius r (in Frobenius norm). Then by [74, Lem. 2.1],

$$\nabla C_r(K) = \frac{d}{r^2}\,\mathbb{E}_{U\sim \mathrm{Sphere}(r)}\, C(K+U)U. \tag{24}$$

Further, denote $C_j = C(K+U_j)$. With these definitions, we decompose the error in gradient estimation into three terms,

$$\begin{aligned}&\|\widehat{\nabla C}(K) - \nabla C(K)\|_F \\ &\le \underbrace{\|\nabla C_r(K) - \nabla C(K)\|_F}_{:=e_1} + \underbrace{\|\frac{1}{J}\sum_{j=1}^J \frac{d}{r^2} C_j U_j - \nabla C_r(K)\|_F}_{:=e_2} + \underbrace{\|\frac{1}{J}\sum_{j=1}^J \frac{d}{r^2}\widehat{C}_j U_j - \frac{1}{J}\sum_{j=1}^J \frac{d}{r^2} C_j U_j\|_F}_{:=e_3}.\end{aligned} \tag{25}$$

In what follows, we show that $e_1 \le \frac{1}{3}e_{grad}$ almost surely, $e_2 \le \frac{1}{3}e_{grad}$ with probability at least $1-\frac{\nu}{2}$, and $e_3 \le \frac{1}{3}e_{grad}$ with probability at least $1-\frac{\nu}{2}$. These together will lead to the desired result.

Bounding e_1. By the definition of $C_r(\cdot)$, we have $\nabla C_r(K) = \mathbb{E}_{U\sim\mathrm{Ball}(r)} \nabla C(K+U)$. As such, as $\nabla C(\cdot)$ is h-Lipschitz,

$$e_1 = \|\nabla C_r(K) - \nabla C(K)\|_F \le \mathbb{E}_{U\sim\mathrm{Ball}(r)} \|\nabla C(K+U) - \nabla C(K)\|_F \le hr \le \frac{1}{3}e_{grad},$$

where in the last step, we have used $r \le \frac{1}{3h}e_{grad}$.

Bounding e_2. For each j, $\frac{d}{r^2}C_jU_j$ is drawn i.i.d. from $\frac{d}{r^2}C(K+U)U$ with $U \sim \mathrm{Sphere}(r)$ and its expectation is $\mathbb{E}\, C_j = \nabla C_r(K)$ (cf. (24)). Further, almost surely,

$$\|\frac{d}{r^2}C_jU_j\|_F \le \frac{d}{r}C(K+U_j) \le \frac{d}{r}\big(C(K^*) + \frac{h}{2}\|K+U_j-K^*\|_F^2\big) \le \frac{d}{r}(C(K^*)+2h\delta^2).$$

As such, using Hoeffding's bound, we have with probability at least $1-\frac{\nu}{2}$,

$$e_2 = \|\frac{1}{J}\sum_{j=1}^J \frac{d}{r^2}C_jU_j - \nabla C_r(K)\|_F \le \frac{d^{1.5}}{r}(C(K^*)+2h\delta^2)\sqrt{\frac{2}{J}\log\frac{4d}{\nu}} \le \frac{1}{3}e_{grad}, \tag{26}$$

where we have used $J \ge \frac{18}{e_{grad}^2}\frac{d^3}{r^2}(C(K^*)+2h\delta^2)^2\log\frac{4d}{\nu}$.

Bounding e_3. We now condition on $\{U_j\}_{j=1}^J$ and focus on the randomness in the initial point x_0 of the trajectories generated in the gradient estimator. Let $\tilde{C}_j = \mathbb{E}_{K+U_j}\sum_{t=0}^T [x_t^\top Q x_t + u_t^\top R u_t]$, where the expectation is taken with respect to the initial state and the trajectory is generated using $K+U_j$. We further decompose e_3 into,

$$e_3 \le \underbrace{\frac{d}{r^2}\Big\|\frac{1}{J}\sum_{j=1}^J[\widehat{C}_jU_j - \tilde{C}_jU_j]\Big\|_F}_{:=e_4} + \underbrace{\frac{d}{r^2}\Big\|\frac{1}{J}\sum_{j=1}^J[\widehat{C}_jU_j - \tilde{C}_jU_j]\Big\|_F}_{:=e_5}.$$

To bound e_4, we note that the expectation of $\hat{C}_jU_j$ is $\tilde{C}_jU_j$. Further, note that by Theorem 1(a), we have $\|x_t\| \le c\rho^t\|x_0\| \le c\rho^t D_0$, where $c = 2c_0$ and $\rho = \frac{\rho_0+1}{2}$. As such,

$$\begin{aligned}|\hat{C}_j| = \sum_{t=0}^T [x_t^\top Q x_t + u_t^\top R u_t] &\le \|Q + (K+U_j)^\top R(K+U_j)\|\frac{c^2}{1-\rho^2}D_0^2 \\ &\le \frac{5\Gamma^2c^2}{1-\rho}D_0^2 := C_{\max}.\end{aligned}$$

As such, when conditioned on $\{U_j\}_{j=1}^J$, the summation in e_4 is a summation of independent random variables with zero mean and is bounded. As such, we have by Hoeffding bound, with probability at least $1-\frac{\nu}{2}$,

$$e_4 \le \frac{d^{1.5}}{r}C_{\max}\sqrt{\frac{2}{J}\log\frac{4d}{\nu}} \le \frac{1}{6}e_{grad}, \tag{27}$$

where we have used that $J \geq \frac{72d^3}{r^2 e_{grad}^2} C_{\max}^2 \log \frac{4d}{\nu}$. Finally, we have,

$$\begin{aligned}
|\tilde{C}_j - C_j| &= \left| \mathbb{E} \sum_{t=T+1}^{\infty} [x_t^\top Q x_t + u_t^\top R u_t] \right| \\
&\leq \|Q + (K + U_j)^\top R (K + U_j)\| \frac{c^2}{1-\rho^2} D_0^2 \rho^{T+1} \\
&\leq C_{\max} \rho^{T+1}.
\end{aligned}$$

As such,

$$e_5 \leq \frac{d}{r} C_{\max} \rho^{T+1} \leq \frac{1}{6} e_{grad}, \tag{28}$$

where we have used $T \geq \frac{2}{1-\rho_0} \log \frac{6dC_{\max}}{e_{grad} r}$. Combining (27) and (28), we have $e_3 \leq \frac{1}{3} e_{grad}$ with probability at least $1 - \frac{\nu}{2}$. This concludes the proof of Lemma 14. □