Policy gradient finds global optimum of nearly linear-quadratic control systems

author names withheld

Under Review for OPT 2022

Abstract

We explore reinforcement learning methods for finding the optimal policy in the nearly linear-quadratic control systems. In particular, we consider a dynamic system composed of the summation of a linear and a nonlinear component, which is governed by a policy with the same structure. Assuming that the nonlinear part is a small known kernel basis, we characterize the optimization landscape of the cost function. While the resulting landscape is generally nonconvex, we show local strong convexity and smoothness of the cost function around the global optimizer. In addition, we design a policy gradient algorithm with a carefully chosen initialization that is guaranteed to converge to the globally optimal policy with a linear rate.

1. Introduction

Reinforcement learning (RL) is one of the three basic machine learning paradigms, alongside supervised and unsupervised learning. RL is learning via trial and error, through interactions with an environment and possibly with other agents. In RL, an agent takes an action and receives a reinforcement signal in terms of a numerical reward, which encodes the outcome of the chosen action. In order to maximize the accumulated reward over time, the agent learns to select actions based on past experiences (exploitation) and/or by making new choices (exploration). In recent years, we have witnessed successful RL applications in many areas, including robotics control [16, 18], AlphaGo [27], Atari games [23], autonomous driving [17], and stock trading [7]. Despite its practical successes, theoretical understanding of RL is still limited and at its primitive stage.

In an effort to better understand the theory of RL, there has been a surge of theoretical works on the study of the Linear Quadratic Regulator (LQR) problem, which is a special class of control problems with linear dynamics and quadratic cost functions [5, 9, 13, 21, 24]. In the seminal work of [9], the authors studied an LQR problem with deterministic dynamics over an infinite horizon. They proved that the simple policy gradient method converges to the global optimal solution with a linear rate (despite nonconvexity of the objective). Their key idea is to utilize the Riccati equation (a partial-differential-equation characterization that only works for LQR problems) and show that the cost function enjoys a gradient dominant property. Later on, this result has been extended to other settings such as linear dynamics with additive or multiplicative Gaussian noise, finite-time horizon, and modifications of policy-gradient methods [5, 13, 21, 24].

Despite the desirable theoretical properties of LQR, this setting is very limited in practice due to *nonlinear nature* of many real-world dynamic systems. From a technical perspective, it is still not clear how much we can go beyond the linear setting and still maintain the desirable properties of

© .

LQR. A preliminary attempt in this direction is to study learning-based methods for linear systems perturbed by some nonlinear kernel functions of small magnitude, which is throughout denoted as *nearly linear-quadratic systems*. The motivations for considering this setting are two-fold: (1) Many nonlinear systems can be locally approximated by an LQR with a small nonlinear correction term via local expansions. (2) Analyzing the nearly linear-quadratic system provides a natural perspective to evaluate the stability of LQR systems. This could further address the question of how *robust* LQR framework is with respect to *model mis-specifications* and, more broadly speaking, how *reliable* the nearly linear-quadratic systems (including LQR problems as a special case) are.

Our Work and Contributions. In this work, we first study the optimization landscape of a special class of nonlinear control systems and propose a policy-gradient-based algorithm to find the optimal policy. Specifically, we consider nonlinear dynamics consisting of both a linear part and a nonlinear part. The nonlinear part is modeled by a linear combination of differentiable kernels with small Lipschitz coefficients. The kernel basis is known to the agent but the coefficients are not available to the agent. In addition, we allow agents to apply nonlinear control policies in the form of a sum of a linear and a nonlinear part where the nonlinear part lies in the same span of the kernel basis for the dynamics. Our analysis shows that the cost function is *locally strongly convex* in a small neighborhood containing a carefully chosen initialization point and the *globally optimal solution*. With these properties in hand, a zeroth-order policy-gradient method is proposed that is guaranteed to converge to the globally optimal solution with a linear rate.

Related Work. Our work is related to three categories of prior work:

First, our framework and analysis tools are closely related to learning-based methods for LQR problem and its variants, including the policy gradient methods in [5, 9, 12–14, 21, 24, 37] and actor-critic methods in [15, 34, 38]. All these works focus on linear systems with a class of linear policies and show the global convergence property. These works assume *linearity* in both the dynamics and control policies. In contrast, we make a step into the nonlinear world by analyzing the policy gradient method for nonlinear systems (that are "near-linear" in certain sense).

Second, our work is also related to the literature on nonlinear control systems. See [26] for a comprehensive review of this topic and [28, 29, 31, 35] for some recent developments such as feedback linearization and neural networks approximation. *While our work is largely inspired by [25], this one is different from it.* [25] considers dynamics consisting of a linear part and a small (and unknown) nonlinear part. However, the authors only consider linear policies, whereas the agent in our framework is allowed to explore nonlinear control policies, which is more general and leads to a better solution. To the best of our knowledge, this is the first theoretical work that shows global convergence for a system with both nonlinear dynamics (with continuous state and action spaces) and nonlinear control policies in the learning setting.

Finally, our work is related to the line of works on policy gradient. In addition to LQR, policy gradient methods have also been applied to learn Markov decision processes (MDPs) with finite state and action spaces. See [1, 4, 6, 8, 11, 19, 20, 30, 32, 33, 36, 38] for some recent developments that provide global convergence guarantees of the policy gradient method and its variants.

Notation. In this work, $\|\cdot\|$ is always the 2-norm for vectors and matrices, and $\|\cdot\|_F$ is the Frobenius norm. In addition, $y_1 \lesssim y_2, y_1 \asymp y_2$ and $y_1 \gtrsim y_2$ mean $y_1 \leq cy_2, y_1 = cy_2$ and $y_1 \geq cy_2$ for some absolute constant c, respectively.

2. Problem Setup

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

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

where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{n\times p}$, and a kernel basis $\phi(x) = (\phi_1, \cdots, \phi_d)^\top(x)$ with $\phi_i(x) : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, 2, \cdots, d)$ that satisfies certain Lipschitz conditions (specified in Assumption 4.1).

We focus on the class of stationary Markovian strategies which are linear combinations of the current state and kernels of the current state

$$u_t = -K_1 x_t - K_2\phi(x_t),$$

with $K_1 \in \mathbb{R}^{p\times n}$ and $K_2 \in \mathbb{R}^{p\times d}$. In addition, we consider the following domain Ω (i.e., the admissible control set) for K:

$$\Omega = \left\{K : \left\|(A - BK_1)^t\right\| \le c_1\rho_1^t, \|C - BK_2\| \le c_2\right\}, \tag{2}$$

for some $c_1 > 1, \rho_1 \in (0, 1)$, and $c_2 > 1$ (to be specified later). It is easy to show that such controller is asymptotically stable, i.e., $\|x_t\| \to 0$ as $t \to \infty$. Further, we consider the quadratic cost function $C : \mathbb{R}^{p\times(n+d)} \to \mathbb{R}$ with $K = (K_1, K_2)$:

$$C(K) = \mathbb{E}\left[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \,\middle|\, u_t = -K_1 x_t - K_2\phi(x_t)\right], \tag{3}$$

where the expectation is taken with respect to x_0 (drawn from a distribution $\mathcal{D}$). The state trajectory $\{x_t\}_{t=0}^{\infty}$ is generated via the control policy K. Here, Q and R are symmetric positive-definite matrices. The objective is to find the optimal policy K that minimizes the cost function $C(K)$.

3. Algorithm

Let us now present our policy gradient algorithm to learn the optimal control for minimizing (3). Using a zeroth-order optimization framework, Algorithm 1 provides an estimate $\widehat{\nabla C(K)}$ for the policy gradient $\nabla C(K)$. This estimate can be used in the following policy gradient update rule:

$$K^{n+1} = K^n - \eta\widehat{\nabla C(K^n)}, K^0 = K^{\text{lin}}. \tag{4}$$

where the initial policy $K^{\text{lin}} = (K_1^{\text{lin}}, K_2^{\text{lin}})$ is to be chosen by the linear approximation of the system. Specifically, K_1^{lin} is defined to be the optimal control policy for the following problem:

$$\begin{aligned} \min_{K_1} \quad & \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t\right] \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t, u_t = -K_1 x_t. \end{aligned} \tag{5}$$

It is known in optimal control literature [2, 3] that K_1^{lin} is uniquely determined when (A, B) is controllable. Let the positive definite matrix P be a solution to the Algebraic Riccati Equation (ARE),

$$P = A^\top PA + Q - A^\top PB(R + B^\top PB)^{-1}B^\top PA. \tag{6}$$

The optimal controller for the problem (5) is then given as:

$$K_1^{\text{lin}} = (R + B^\top PB)^{-1} B^\top PA. \tag{7}$$

Further, set $K_2^{\text{lin}} = (R + B^\top PB)^{-1} B^\top PC$. It is easy to see that K_2^{lin} is well defined since the matrix $R + B^\top PB$ is positive definite thus invertible. In the next section, we will show that:

- The optimal solution to the nonlinear control problem (1)-(3) can only be attained in a small neighborhood of the controller K^{lin}.
- The cost function (3) is strongly convex and smooth in a neighborhood of K^{lin}.

Utilizing these two facts, we can establish the convergence rate of Algorithm 1 to global optimality.

Algorithm 1 Policy Gradient Estimation

Input: Policy $K = (K_1, K_2)$, number of trajectories J, smoothing parameter r, and episode length T
for $j = 1, 2, \ldots, J$ **do**
  Sample a policy $\widehat{K}^j = K + U^j$, where U^j is drawn uniformly at random over matrices whose Frobenius norm is r.
  Sample $x_0 \sim \mathcal{D}$.
  for $t = 0, 1, \ldots, T$ **do**
    Set $u_t = -\widehat{K}_1^j x_t - \widehat{K}_2^j \phi(x_t)$
    Receive the cost c_t and the next state x_{t+1} from the system.
  end for
  Calculate the estimated cost $\widehat{C}_j = \sum_{t=0}^{T} c_t$
end for
return $\widehat{\nabla C(K)} = \frac{1}{J} \sum_{j=0}^{J} \frac{\widehat{D}}{r^2} \widehat{C}_j U_j$, where $\widehat{D} = p(n+d)$.

4. Main Results

In this section, we characterize the optimization landscape of the cost function and establish convergence analysis of Algorithm 1. Particularly, we first show a local strong convexity result of the cost function around the global minimum. Then, we prove that the global minimum point is close to our carefully chosen initialization. Finally, we prove the convergence of Algorithm 1. Before we state our main results, we make the following assumptions for problem (1)-(3).

Assumption 4.1 *We assume that ϕ is differentiable, $\phi(0) = 0$ and $\|\phi(x) - \phi(x')\| \leq \ell \|x - x'\|$ for some $\ell > 0$. Moreover, we assume that $\|\nabla\phi(x) - \nabla\phi(x')\| \leq \ell' \|x - x'\|$ for some $\ell' > 0$.*

Assumption 4.1 states that the kernel function ϕ is ℓ-Lipschitz and ℓ'-gradient Lipschitz. The following assumption is on the matrices Q, R and is standard in the literature [22].

Assumption 4.2 *We assume that Q, R are positive definite matrices for which $\|Q\|, \|R\| \leq 1$. Further, $R + B^\top QB \succeq \sigma I$ for some $\sigma > 0$.*

The upper bound one (on the norms of Q and R) in Assumption 4.2 is for the ease of presentation and can be generalized to any arbitrary number by rescaling the cost function. The second part of the assumption guarantees that the cost function has quadratic growth and makes the problem well-defined [25]. Our next assumption concerns the initial distribution of the state dynamics.

Assumption 4.3 *We assume that the initial distribution $\mathcal{D}$ is supported in a region with radius D_0, i.e., $\|x\| \leq D_0$ with probability one for $x \in \mathcal{D}$. Also, $\mathbb{E}\left[\psi(x_0)\psi(x_0)^\top\right] \succeq \sigma_x I$ for some $\sigma_x > 0$.*

Assumption 4.3 requires the state initial distribution to be bounded. This assumption simplifies the proof in the latter sections and can be replaced by assuming a bound on the second and the third moments [9]. Also, the covariance matrix $\mathbb{E}\left[\psi(x_0)\psi(x_0)^\top\right]$ is assumed to be bounded below from a positive constant matrix $\sigma_x I$. This "diverse covarate" assumption ensures that there is sufficient exploration (in all directions of the state space) even with a greedy algorithm. Finally, we lay out another regularity condition on the coefficient matrices (A, B) and the initializer K^{lin}.

Assumption 4.4 *The pair (A, B) is controllable. Let $K^{\text{lin}} = (K_1^{\text{lin}}, K_2^{\text{lin}})$ be defined in Section 3. Also, let $\left\|(A - BK_1^{\text{lin}})^t\right\| \leq c_1^{\text{lin}}(\rho_1^{\text{lin}})^t$, $\left\|C - BK_2^{\text{lin}}\right\| \leq c_2^{\text{lin}}$ for $\rho_1^{\text{lin}} \in (0, 1)$ and $c_1^{\text{lin}}, c_2^{\text{lin}} > 0$.*

Assumption 4.4 states that the controller K^{lin} enjoys stability property. The controllablity assumption on the pair (A, B) is standard in the literature [5].

We now state to our results. The first theorem characterizes the landscape of the cost function. It shows that the cost function is strongly convex and smooth in a neighborhood of the initialization K^{lin} when the Lipschitz constants ℓ and ℓ' are sufficiently small. Further, we prove the optimal controller K^* is inside this region. We defer the proof to Appendix A in the supplementary material.

Theorem 4.5 *Denote $\Gamma = \max\left\{\|A\|, \|B\|, \|C\|, \left\|K^{\text{lin}}\right\|, 1\right\}$. For any $c_1 \geq 2c_1^{\text{lin}}$, $\rho_1 \in \left[\frac{\rho_1^{\text{lin}}+1}{2}, 1\right)$, $c_2 \geq c_2^{\text{lin}}$, if $\ell \lesssim \frac{(1-\rho_1)^7(\sigma_x\sigma)^2}{(c_1+c_2)c_2c_1^7(1+\Gamma)^8D_0^3}$, $\ell' \lesssim \left(\frac{(1-\rho_1)^8(\sigma_x\sigma)^2}{(c_1+c_2)^2c_2^2c_1^{16}(1+\Gamma)^6D_0^4}\right)$, then*

(a) there exists a region $\Lambda(\delta) = \left\{K : \left\|K - K^{\text{lin}}\right\|_F \leq \delta\right\} \subset \Omega$ with $\delta \asymp \frac{(1-\rho_1)^4\sigma_x\sigma}{(c_1+c_2)c_1^6\Gamma^2D_0}$ such that $C(K)$ is μ-strongly convex and h-smooth in $\Lambda(\delta)$ with $\mu = \sigma_x\sigma$, $h \asymp \frac{\Gamma^4c_1^4D_0^2}{(1-\rho_1)^2}$;

(b) the global minimum of $C(K)$ is achieved at a point $K^* \in \Lambda(\delta/3)$.

Part (a) of Theorem 4.5 indicates that the cost function $C(K)$ is strongly convex and smooth within a δ-neighborhood of the initializer K^{lin}. Part (b) shows that the optimal controller K^* lies in a $\delta/3$-neighborhood of the initializer K^{lin}. Consequently, the cost function is strongly convex and smooth in a region that contains both the initialization K^{lin} and the global optimizer K^*.

Given the landscape results, if $\nabla C(K)$ is assumed to be known, starting from the initialization K^{lin}, the policy gradient descent leads to finding the global minimum of the cost function $C(K)$. Hence, it is not surprising that Algorithm 1 converges to the globally optimal solution with one-point gradient estimation. This result is formally stated in Theorem 4.6. The proof of Theorem 4.6 can be found in Appendix B in the supplementary material.

Theorem 4.6 *Assume the conditions in Theorem 4.5 hold. Let $\epsilon > 0$ and $\nu \in (0,1)$ be given. Suppose the step size $\eta < \frac{1}{h}$ and the number of gradient descent step $M \geq \frac{1}{\eta\mu} \log\left(\frac{\delta}{3}\sqrt{\frac{2h}{\epsilon}}\right)$. Further, assume the gradient estimator parameters satisfy $r \leq \min\left\{\frac{\delta}{3}, \frac{1}{3h} e_{grad}\right\}$,*

$$J \geq \frac{\widehat{D}^2}{e_{grad}^2 r^2} \log \frac{4\widehat{D}M}{\nu} \max\left\{36\left(C(K^*) + 2h\delta^2\right)^2, 144C_{\max}^2\right\}, T \geq \frac{1}{1-\rho_1} \log \frac{6\widehat{D}C_{max}}{e_{grad} r},$$

where $\widehat{D} = p(n+d)$ and $C_{\max} = \frac{24(1+\Gamma)^2 c_1^2 D_0^2}{1-\rho_1}$, and $e_{grad} = \min\left\{\frac{\delta\mu}{3}, \mu\sqrt{\frac{\epsilon}{2h}}\right\}$. Then with probability at least $1-\nu$, we have $C(K^M) - C(K^*) < \epsilon$.

This result shows that, despite the existence of nonlinear terms, finding the optimal control policy is still tractable when nonlinear terms are "sufficiently small".

5. Conclusions

We consider a nonlinear optimal control problem, characterize the local strong convexity of the cost function, and prove that the globally optimal solution is close to the carefully chosen initialization. In addition, we design a zeroth-order policy gradient algorithm and establish a convergence result under a proposed policy initialization scheme of the nonlinear control problem. We hope these results would shed light on the efficiency of policy gradient methods for nonlinear optimal control problems when the underlying models are unknown to the decision maker. The future work is to investigate the sample complexity of the algorithm and extend the analysis on quadratic cost functions to more general cost functions.

References

[1] Alekh Agarwal, Sham M. Kakade, J. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22:98:1–98:76, 2021.

[2] Brian DO Anderson and John B Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, 2007.

[3] Dimitri Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena scientific, 2012.

[4] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[5] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[6] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

[7] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2016.

[8] Dongsheng Ding, K. Zhang, Tamer Başar, and Mihailo R. Jovanović. Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*, 2020.

[9] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[10] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan 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*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

[11] Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

[12] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning robust control for LQR systems with multiplicative noise via policy gradient. *arXiv preprint arXiv:1905.13547*, 2019.

[13] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5): 3359–3391, 2021.

[14] Joao Paulo Jansch-Porto, Bin Hu, and Geir E Dullerud. Convergence guarantees of policy optimization methods for markovian jump linear systems. In *2020 American Control Conference (ACC)*, pages 2882–2887. IEEE, 2020.

[15] Zeyu Jin, Johann Michael Schmitt, and Zaiwen Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[16] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[17] Dong Li, Dongbin Zhao, Qichao Zhang, and Yaran Chen. Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. *IEEE Computational Intelligence Magazine*, 14(2):83–98, 2019.

[18] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[19] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

[20] Yanli Liu, K. Zhang, Tamer Başar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *NeurIPS*, 2020.

[21] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, and Martin J. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21:21:1–21:51, 2019.

[22] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[24] Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7474–7479. IEEE, 2019.

[25] Guannan Qu, Chenkai Yu, Steven Low, and Adam Wierman. Combining model-based and model-free methods for nonlinear control: A provably convergent policy gradient approach. *arXiv preprint arXiv:2006.07476*, 2020.

[26] Shankar Sastry. *Nonlinear Systems: Analysis, Stability, and Control*, volume 10. Springer Science & Business Media, 2013.

[27] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–503, 2016. URL http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.

[28] Jonas Umlauft and Sandra Hirche. Feedback linearization based on gaussian processes with event-triggered online learning. *IEEE Transactions on Automatic Control*, 65:4154–4169, 2020.

[29] Jonas Umlauft, Thomas Beckers, Melanie Kimmel, and Sandra Hirche. Feedback linearization using gaussian processes. *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5249–5255, 2017.

[30] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[31] Tyler Westenbroek, David Fridovich-Keil, Eric Mazumdar, Shreyas Arora, Valmik Prabhu, S Shankar Sastry, and Claire J Tomlin. Feedback linearization for uncertain systems via reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1364–1371. IEEE, 2020.

[32] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[33] Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *ICML*, 2021.

[34] Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in neural information processing systems*, 32, 2019.

[35] Guo-Qiang Zeng, Xiao-Qing Xie, Min-Rong Chen, and Jian Weng. Adaptive population extremal optimization-based pid neural network for multivariable nonlinear control systems. *Swarm and evolutionary computation*, 44:320–334, 2019.

[36] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[37] K. Zhang, Bin Hu, and Tamer Bacsar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 2021.

[38] Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Provably efficient actor-critic for risk-sensitive and robust adversarial RL: A linear-quadratic case. In *International Conference on Artificial Intelligence and Statistics*, pages 2764–2772. PMLR, 2021.

Appendix A. Proof of Theorem 4.5

Denote the value function and Q function conditioned on the initial position as

$$V_K(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \middle| x_0 = x, u_t = -K_1 x_t - K_2\phi(x_t)\right], \tag{8}$$

$$Q_K(x,u) = x^\top Q x + u^\top R u + V_K(Ax + C\phi(x) + Bu). \tag{9}$$

First, we provide a characterization of the value function, which is shown in the next lemma.

Lemma A.1 (Value Function) *The value function takes the form*

$$V_K(x) = x^\top P_{K_1} x + G_K(x), \tag{10}$$

where P_{K_1} satisfies

$$(A - BK_1)^\top P_{K_1}(A - BK_1) - P_{K_1} + Q + K_1^\top R K_1 = 0, \tag{11}$$

and $G_K(x)$ is defined as

$$\begin{aligned} G_K(x) := {} & \operatorname{Tr}\left(\left(K_2^\top R K_2 + (C - BK_2)^\top P_{K_1}(C - BK_2)\right)\sum_{t=0}^{\infty}\phi(x_t)\phi(x_t)^\top\right) \\ & +2\operatorname{Tr}\left(\left(K_1^\top R K_2 + (A - BK_1)^\top P_{K_1}(C - BK_2)\right)\sum_{t=0}^{\infty}\phi(x_t)x_t^\top\right), \end{aligned} \tag{12}$$

and $\{x_t\}_{t=0}^{\infty}$ is the trajectory generated by the policy $K = (K_1, K_2)$ starting with the initial position $x_0 = x$.

Proof By Bellman equation, the value function satisfies,

$$\begin{aligned} V_K(x) &= x^\top Q x + \Big(K_1 x + K_2\phi(x)\Big)^\top R\Big(K_1 x + K_2\phi(x)\Big) \\ &\qquad + V_K\Big(Ax - B(K_1 x + K_2\phi(x)) + C\phi(x)\Big) \\ &= x^\top(Q + K_1^\top R K_1)x + \phi(x)^\top K_2^\top R K_2\phi(x) + 2x^\top K_1^\top R K_2\phi(x) \\ &\qquad + V_K\Big((A - BK_1)x + (C - BK_2)\phi(x)\Big). \end{aligned}$$

Define $G_K(x) = V_K(x) - x^\top P_{K_1} x$, we have

$$\begin{aligned} & x^\top P_{K_1} x + G_K(x) \\ = {} & x^\top(Q + K_1^\top R K_1)x + \phi(x)^\top K_2^\top R K_2\phi(x) + 2x^\top K_1^\top R K_2\phi(x) \\ & + \Big((A - BK_1)x + (C - BK_2)\phi(x)\Big)^\top P_{K_1}\Big((A - BK_1)x + (C - BK_2)\phi(x)\Big) + G_K(x_1), \end{aligned}$$

with $x_1 = (A - BK_1)x + (C - BK_2)\phi(x)$. Since P_{K_1} satisfies (11), we have

$$G_K(x) = \phi(x)^\top\Big(K_2^\top R K_2 + (C - BK_2)^\top P_{K_1}(C - BK_2)\Big)\phi(x)$$

$$
\begin{aligned}
& +2x^\top\Big(K_1^\top RK_2 + (A-BK_1)^\top P_{K_1}(C-BK_2)\Big)\phi(x) + G_K(x_1) \\
= \;& \operatorname{Tr}\Big(\Big(K_2^\top RK_2 + (C-BK_2)^\top P_{K_1}(C-BK_2)\Big)\phi(x)\phi(x)^\top\Big) \\
& +2\operatorname{Tr}\Big(\Big(K_1^\top RK_2 + (A-BK_1)^\top P_{K_1}(C-BK_2)\Big)\phi(x)x^\top\Big) + G_K(x_1) \\
= \;& \operatorname{Tr}\Big(\Big(K_2^\top RK_2 + (C-BK_2)^\top P_{K_1}(C-BK_2)\Big)\sum_{t=0}^{\infty}\phi(x_t)\phi(x_t)^\top\Big) \\
& +2\operatorname{Tr}\Big(\Big(K_1^\top RK_2 + (A-BK_1)^\top P_{K_1}(C-BK_2)\Big)\sum_{t=0}^{\infty}\phi(x_t)x_t^\top\Big).
\end{aligned}
$$

Therefore (12) holds. ■

To proceed, define a coefficient matrix $M = (A, C)$, and a feature map $\psi(x) = (x^\top, \phi(x)^\top)^\top$. Then the dynamics (1) can be written as

$$x_{t+1} = (A - BK_1)x_t + (C - BK_2)\phi(x_t) = (M - BK)\psi(x_t). \tag{13}$$

Given Assumption 4.1, it is easy to see that $\psi(x)$ is ℓ_ψ-Lipschitz, i.e., $\|\psi(x) - \psi(x')\| \le \ell_\psi \|x - x'\|$ with $\ell_\psi := \sqrt{1+\ell^2}$. The following lemma gives us the gradient of the cost function $C(K)$.

Lemma A.2 (Gradient of $C(K)$) *The gradient of $C(K)$ is*

$$\nabla_K C(K) = 2E_K\Sigma_K^{\psi\psi} - B^\top\Sigma_K^{G\psi}, \tag{14}$$

where

$$E_K = RK - B^\top P_{K_1}(M - BK), \Sigma_K^{\psi\psi} = \mathbb{E}\sum_{t=0}^{\infty}\psi(x_t)\psi(x_t)^\top, \Sigma_K^{G\psi} = \mathbb{E}\sum_{t=0}^{\infty}\nabla G_K(x_{t+1})\psi(x_t)^\top. \tag{15}$$

Proof Recall the Bellman equation

$$V_K(x) = x^\top Qx + (K\psi(x))^\top RK\psi(x) + V_K((M - BK)\psi(x)).$$

Taking gradient in K on both sides of the Bellman equation, we have

$$\nabla_K V_K(x) = 2RK\psi(x)\psi(x)^\top + \nabla_K V(x_1) + \left(\frac{\partial x_1}{\partial K}\right)^\top \nabla_x V_K(x_1),$$

where $x_1 = (M - BK)\psi(x)$. Note the directional derivative of x_1 in K along the direction Δ is $x_1'[\Delta] = -B\Delta\psi(x)$. Then we have

$$
\begin{aligned}
x_1'[\Delta]^\top \nabla_x V_K(x_1) &= -\psi(x)^\top\Delta^\top B^\top(2P_{K_1}x_1 + \nabla G_K(x_1)) \\
&= \operatorname{Tr}\Delta^\top\left(-2B^\top P_{K_1}x_1\psi(x)^\top - B^\top\nabla G_K(x_1)\psi(x)^\top\right).
\end{aligned}
$$

This leads to

$$\nabla_K V_K(x) = 2RK\psi(x)\psi(x)^\top - 2B^\top P_{K_1}(M - BK)\psi(x)\psi(x)^\top - B^\top\nabla G_K(x_1)\psi(x)^\top + \nabla_K V(x_1)$$

$$
\begin{aligned}
&= \Big(2RK - 2B^\top P_{K_1}(M - BK)\Big)\,\psi(x)\psi(x)^\top - B^\top \nabla G_K(x_1)\psi(x)^\top + \nabla_K V(x_1) \\
&= 2E_K \sum_{t=0}^{\infty} \psi(x_t)\psi(x_t)^\top - B^\top \sum_{t=0}^{\infty} \nabla G_K(x_{t+1})\psi(x_t)^\top.
\end{aligned}
$$

Taking expectation over x_0 and then we are done. ■

With Lemma A.1 and Lemma A.2, we provide a formula for $C(K') - C(K)$, which is shown in the following cost difference lemma.

Lemma A.3 (Cost difference lemma) *For any $K = (K_1, K_2)$ and $K' = (K_1', K_2')$, we have*

$$
\begin{aligned}
C(K') - C(K) = {} & \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\Sigma_{K'}^{\psi\psi} + 2\operatorname{Tr}(K' - K)^\top E_K \Sigma_{K'}^{\psi\psi} \\
& + \mathbb{E}\sum_{t=0}^{\infty} \left[G_K((M - BK')\psi(x_t')) - G_K((M - BK)\psi(x_t'))\right]. \qquad (16)
\end{aligned}
$$

Proof By [9, Lemma 10], we have

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

where $\{x_t'\}$ is trajectory generated by $x_0' = x$ and $u_t' = -K'\psi(x_t')$, and $A_K(x, u) = Q_K(x, u) - V_K(x)$ is the advantage function.

For given $u = -K'\psi(x)$, we have

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

We next compute the last two terms

$$
\begin{aligned}
& V_K((M - BK')\psi(x)) - V_K((M - BK)\psi(x)) \\
&= \big((M - BK')\psi(x)\big)^\top P_{K_1}\big((M - BK')\psi(x)\big) - ((M - BK)\psi(x))^\top P_{K_1}((M - BK)\psi(x)) \\
&\quad + G_K\big((M - BK')\psi(x)\big) - G_K((M - BK)\psi(x)) \\
&= \psi(x)^\top (K' - K)^\top B^\top P_{K_1} B(K' - K)\psi(x) + 2\psi(x)^\top (K - K')^\top B^\top P_{K_1}(M - BK)\psi(x) \\
&\quad + G_K\big((M - BK')\psi(x)\big) - G_K((M - BK)\psi(x)).
\end{aligned}
$$

Substitution back, we have

$$
\begin{aligned}
A_K(x, u) = {} & \psi(x)^\top (K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\psi(x) \\
& + 2\psi(x)^\top (K' - K)^\top (RK - B^\top P_{K_1}(M - BK))\psi(x)
\end{aligned}
$$

$$+ G_K\left((M - BK')\psi(x)\right) - G_K\left((M - BK)\psi(x)\right).$$

In consequence, we have

$$\begin{aligned}
C(K') - C(K) &= \mathbb{E}\sum_{t=0}^{\infty} A_K(x_t', u_t') \\
&= \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\, \mathbb{E}\sum_{t=0}^{\infty} \psi(x_t')\psi(x_t')^\top \\
&\quad + 2\operatorname{Tr}(K' - K)^\top (RK - B^\top P_{K_1}(M - BK)) \sum_{t=0}^{\infty} \psi(x_t')\psi(x_t')^\top \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} \left[G_K((M - BK')\psi(x_t')) - G_K((M - BK)\psi(x_t'))\right].
\end{aligned}$$

■

Next, we show that the state trajectory has an exponential decay property regardless of the initial state. In consequence, the cost function $C(\cdot)$ is bounded.

Lemma A.4 (Stability of the trajectory $\{x_t\}$) *Assume that $K \in \Omega$ and $\ell \le \frac{1-\rho_1}{4c_1c_2}$. Then we have the following holds.*

(a) For any $x_0 \in \mathbb{R}^n$, $\|x_t\| \le c\rho^t \|x_0\|$, where $c = 2c_1$ and $\rho = \frac{\rho_1+1}{2}$.

(b) Let $\{x_t\}$ and $\{x_t'\}$ be the state trajectories starting from x_0 and x_0', respectively. Then $\|x_t - x_t'\| \le c\rho^t \|x_0 - x_0'\|$. In consequence, $\left\|\frac{\partial x_t}{\partial x_0}\right\| \le c\rho^t$.

(c) Let $\{x_t\}$ and $\{x_t'\}$ be defined as above. Then $\left\|\frac{\partial x_t}{\partial x_0} - \frac{\partial x_t'}{\partial x_0'}\right\| \le \frac{c_2\ell' c^3}{1-\rho}\rho^{t-1}\|x_0 - x_0'\|$.

Proof Let $f(x) = (C - BK_2)\phi(x)$. Then we have

$$\begin{aligned}
\left\|f(x) - f(x')\right\| &= \left\|(C - BK_2)(\phi(x) - \phi(x'))\right\| \le \|C - BK_2\| \left\|\phi(x) - \phi(x')\right\| \le c_2\ell \left\|x - x'\right\|, \\
\left\|\nabla f(x) - \nabla f(x')\right\| &= \left\|(C - BK_2)(\nabla\phi(x) - \nabla\phi(x'))\right\| = \|C - BK_2\| \left\|\nabla\phi(x) - \nabla\phi(x')\right\| \le c_2\ell' \left\|x - x'\right\|.
\end{aligned}$$

Note that we have $x_{t+1} = (A - BK_1)x_t + f(x_t)$. Applying Lemma 4 in [25] and then we are done. ■

We provide an auxiliary lemma that will be used in the rest of the proof.

Lemma A.5 *When $\left\|K - K^{\text{lin}}\right\|_F \le \delta \le 1$, we have*

$$\|P_{K_1}\| \le C_P := \frac{2c_1^2(1+\Gamma)^2}{1-\rho_1}, \tag{17}$$

where P_{K_1} is the solution of Lyapunov equation,

$$(A - BK_1)^\top P_{K_1}(A - BK_1) - P_{K_1} + Q + K_1^\top R K_1 = 0.$$

Proof Note that

$$P_{K_1} = \sum_{t=0}^{\infty} \left((A - BK_1)^\top\right)^t (Q + K_1^\top R K_1)(A - BK_1)^t.$$

Then we have

$$\|P_{K_1}\| \le \frac{c_1^2}{1-\rho_1^2}\left\|Q + K_1^\top R K_1\right\| \le \frac{c_1^2}{1-\rho_1^2}(1 + \|K\|^2) < \frac{c_1^2(1+(1+\Gamma)^2)}{1-\rho_1} < \frac{2c_1^2(1+\Gamma)^2}{1-\rho_1} =: C_P,$$

where we used the inequality $\|K\| \le \left\|K - K^{\mathrm{lin}}\right\| + \left\|K^{\mathrm{lin}}\right\| \le \delta + \Gamma \le 1 + \Gamma$. ■

We show the last result before the proof of local strong convexity property. The next lemma yields the local Lipschitzness of $G_K(x)$.

Lemma A.6 (Local Lipschitzness of $G_K(x)$) *When $\left\|K - K^{\mathrm{lin}}\right\|_F \le \delta$, and when $\|x\|, \|x'\| \le (c_1 + c_2)cD_0$, we have*

$$\left\|\nabla G_K(x) - \nabla G_K(x')\right\| \le L\left\|x - x'\right\|, \tag{18}$$

where $L = \frac{5c_2c^5(1+\Gamma)^4}{16(1-\rho)^2}\ell + \frac{3Dc_2^2c^6(1+\Gamma)^2}{16(1-\rho)^3}\ell'$ with $D = (c_1 + c_2)c^2D_0$.

Proof We first compute the gradient of $G_K(x)$ as follows

$$\begin{aligned}
[\nabla G_K(x)]^\top &= 2\sum_{t=0}^{\infty}\left[\phi(x_t)^\top (F_K^{12})^\top + x_t^\top F_K^{12}\frac{\partial \phi(x_t)}{\partial x_t}\right]\frac{\partial x_t}{\partial x} + 2\sum_{t=0}^{\infty}\phi(x_t)^\top F_K^{22}\frac{\partial \phi(x_t)}{\partial x_t}\frac{\partial x_t}{\partial x} \\
&= 2\sum_{t=0}^{\infty}\left[\phi(x_t)^\top F_K^{21} - \pi_K(x_t)^\top RK_2\frac{\partial \phi(x_t)}{\partial x_t} + x_{t+1}^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t)}{\partial x_t}\right]\frac{\partial x_t}{\partial x},
\end{aligned}$$

where we define $\pi_K(x_t) = -K_1x_t - K_2\phi(x_t)$ and

$$\begin{aligned}
F_K^{12} &= \left(F_K^{21}\right)^\top = K_1^\top RK_2 + (A - BK_1)^\top P_{K_1}(C - BK_2), \\
F_K^{22} &= K_2^\top RK_2 + (C - BK_2)^\top P_{K_1}(C - BK_2).
\end{aligned}$$

For x, x', we have

$$\begin{aligned}
&\left\|\nabla G_K(x) - \nabla G_K(x')\right\| \\
&\le 2\sum_{t=0}^{\infty}\Bigg\|\left[\phi(x_t) - \phi(x_t')\right]^\top F_K^{21} - \left[\pi_K(x_t)^\top RK_2\frac{\partial \phi(x_t)}{\partial x_t} - \pi_K(x_t')^\top RK_2\frac{\partial \phi(x_t')}{\partial x_t'}\right] \\
&\qquad + x_{t+1}^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t)}{\partial x_t} - (x_{t+1}')^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t')}{\partial x_t'}\Bigg\|\left\|\frac{\partial x_t}{\partial x}\right\| \\
&\quad + 2\sum_{t=0}^{\infty}\left\|\phi(x_t')^\top F_K^{21} - \pi_K(x_t')^\top RK_2\frac{\partial \phi(x_t')}{\partial x_t'} + (x_{t+1}')^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t')}{\partial x_t'}\right\|\left\|\frac{\partial x_t}{\partial x} - \frac{\partial x_t'}{\partial x'}\right\|.
\end{aligned}$$

We compute bounds term by term. Firstly, we have

$$\left\|\left[\phi(x_t) - \phi(x_t')\right]^\top F_K^{21}\right\| \le \ell\left\|x_t - x_t'\right\|\left\|F_K^{21}\right\| \le \ell c\left\|x - x'\right\|\left\|F_K^{21}\right\|.$$

We have the following bound on $\left\|F_K^{21}\right\|$

$$\begin{aligned}
\left\|F_K^{21}\right\| &= \left\|K_2^\top R K_1 + (C - BK_2)^\top P_{K_1}(A - BK_1)\right\| \\
&\le \|K_1\| \|R\| \|K_2\| + \|C - BK_2\| \|P_{K_1}\| \|A - BK_1\| \\
&\le \|K_1\|_F \|K_2\|_F + c_1 c_2 C_P \\
&\le \frac{1}{2}\|K\|_F^2 + c_1 c_2 C_P \\
&\le \frac{1}{2}(1+\Gamma)^2 + c_1 c_2 C_P \\
&\le \frac{5 c_2 c_1^3 (1+\Gamma)^2}{2(1-\rho_1)} =: C_F^{21},
\end{aligned}$$

where we use the assumption $\|R\| \le 1$, the bound on $\|P_{K_1}\| \le C_P$ as in Lemma 6 [25] and the bound on $\|K\|_F \le 1 + \Gamma$. It follows

$$\left\|\left[\phi(x_t) - \phi(x_t')\right]^\top F_K^{21}\right\| \le \frac{5\ell c_2 c_1^4 (1+\Gamma)^2}{1-\rho_1}\left\|x - x'\right\|.$$

Next, almost surely, we have

$$\begin{aligned}
&\left\|\pi_K^\top R K_2 \frac{\partial \phi(x_t)}{\partial x_t} - \pi_K(x_t')^\top R K_2 \frac{\partial \phi(x_t')}{\partial x_t'}\right\| \\
&\le \left\|\pi_K(x_t) - \pi_K(x_t')\right\| \|R\| \|K_2\| \left\|\frac{\partial \phi(x_t)}{\partial x_t}\right\| + \left\|\pi(x_t')\right\| \|R\| \|K_2\| \left\|\frac{\partial \phi(x_t)}{\partial x_t} - \frac{\partial \phi(x_t')}{\partial x_t'}\right\| \\
&\le \ell(\|K_1\| + \ell \|K_2\|) \|K_2\| \left\|x_t - x_t'\right\| + \ell'(\|K_1\| + \ell \|K_2\|) \|x_t\| \|K_2\| \left\|x_t - x_t'\right\| \\
&\le (\ell/2 + \ell^2) \|K\|_F^2 c \left\|x - x'\right\| + (\ell'/2 + \ell\ell') \|K\|_F^2 D c \left\|x - x'\right\| \\
&\le (\ell/2 + \ell^2 + D\ell'/2 + D\ell\ell')(1+\Gamma)^2 c \left\|x - x'\right\| \\
&= (1/2 + \ell)(\ell + D\ell')(1+\Gamma)^2 c \left\|x - x'\right\| \\
&\le 2(\ell + D\ell')(1+\Gamma)^2 c_1 \left\|x - x'\right\|,
\end{aligned}$$

where we use the bound on $\|x_t\| \le c\|x_0\| \le (c_1 + c_2)c^2 D_0 =: D$ and the fact $\ell \le 1/2$. Further, we have

$$\begin{aligned}
&\left\|x_{t+1}^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t)}{\partial x_t} - (x_{t+1}')^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t')}{\partial x_t'}\right\| \\
&\le \left\|(x_{t+1} - x_{t+1}')^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t)}{\partial x_t}\right\| + \left\|(x_{t+1}')^\top P_{K_1}(C - BK_2)\left(\frac{\partial \phi(x_t)}{\partial x_t} - \frac{\partial \phi(x_t')}{\partial x_t'}\right)\right\| \\
&\le \left\|x_{t+1} - x_{t+1}'\right\| \|P_{K_1}\| \|C - BK_2\| \left\|\frac{\partial \phi(x_t)}{\partial x_t}\right\| + \left\|x_{t+1}'\right\| \|P_{K_1}\| \|C - BK_2\| \left\|\frac{\partial \phi(x_t)}{\partial x_t} - \frac{\partial \phi(x_t')}{\partial x_t'}\right\| \\
&\le C_P c_2 \ell \left\|x_{t+1} - x_{t+1}'\right\| + D C_P c_2 \ell' \left\|x_t - x_t'\right\| \\
&\le c c_2 C_P (\ell + D\ell') \left\|x - x'\right\| \\
&\le \frac{3 c_2 c_1^3 (1+\Gamma)^2}{1-\rho_1}(\ell + D\ell') \left\|x - x'\right\|.
\end{aligned}$$

Also, notice that

$$\begin{aligned}
&\left\|\phi(x_t')^\top F_K^{21} - \pi_K(x_t')^\top R K_2 \frac{\partial \phi(x_t')}{\partial x_t'} + (x_{t+1}')^\top P_{K_1}(C - BK_2)\frac{\partial \phi(x_t')}{\partial x_t'}\right\| \\
&\le \ell \left\|x_t'\right\| C_F^{21} + \ell(\|K_1\| + \ell \|K_2\|)\left\|x_t'\right\| \|K_2\| + C_P c_2 \ell \left\|x_{t+1}'\right\| \\
&\le \ell D C_F^{21} + (\ell/2 + \ell^2)\|K\|_F^2 D + c_2 \ell C_P D \\
&\le \ell D C_F^{21} + (3/2)\ell(1+\Gamma)^2 D + c_2 \ell C_P D \\
&\le \frac{11 \ell D c_2 c_1^3 (1+\Gamma)^2}{2(1-\rho_1)}.
\end{aligned}$$

Plugging in all these results, and using $\left\|\frac{\partial x_t}{\partial x}\right\| \le c\rho^t$ and $\left\|\frac{\partial x_t}{\partial x} - \frac{\partial x_t'}{\partial x'}\right\| \le \frac{c_2 \ell' c^3}{1-\rho}\rho^{t-1}\|x - x'\|$, we have

$$\begin{aligned}
&\left\|\nabla G_K(x) - \nabla G_K(x')\right\| \\
&\le 2\sum_{t=0}^{\infty}\left[\frac{5\ell c_2 c_1^4(1+\Gamma)^2}{1-\rho_1} + 2(\ell + D\ell')(1+\Gamma)^2 c_1 + \frac{3c_2 c_1^3(1+\Gamma)^2}{1-\rho_1}(\ell + D\ell')\right]\|x - x'\|\left\|\frac{\partial x_t}{\partial x}\right\| \\
&\quad + 2\sum_{t=1}^{\infty}\frac{11\ell D c_2 c_1^3(1+\Gamma)^2}{2(1-\rho_1)}\left\|\frac{\partial x_t}{\partial x} - \frac{\partial x_t'}{\partial x'}\right\| \\
&\le \frac{2c}{1-\rho}\left(\frac{10 c_2 c_1^4(1+\Gamma)^2}{1-\rho_1}\ell + \frac{5c_2 c_1^3(1+\Gamma)^2 D}{1-\rho_1}\ell'\right)\|x - x'\| + 2\frac{11\ell D c_2 c_1^3(1+\Gamma)^2}{2(1-\rho_1)}\frac{c_2 \ell' c^3}{(1-\rho)^2}\|x - x'\| \\
&\le \left(\frac{40 c_2 c_1^5(1+\Gamma)^4}{(1-\rho_1)^2}\ell + \frac{176 D c_2^2 c_1^6(1+\Gamma)^2}{(1-\rho_1)^3}\ell'\right)\|x - x'\| \\
&= \left(\frac{5 c_2 c^5(1+\Gamma)^4}{16(1-\rho)^2}\ell + \frac{3 D c_2^2 c^6(1+\Gamma)^2}{16(1-\rho)^3}\ell'\right)\|x - x'\|,
\end{aligned}$$

which shows that $\nabla G_K(x)$ is L-Lipschitz in x. ■

With these preparations, we proceed the property of local strong convexity of the cost function $C(K)$, which is stated in the following lemma.

Lemma A.7 (Local strong convexity) *For any $c_1 \ge 2c_1^{\text{lin}}, \rho_1 \in \left[\frac{\rho_1^{\text{lin}}+1}{2}, 1\right), c_2 \ge c_2^{\text{lin}}$, if δ, ℓ, ℓ' satisfy*

$$\delta \le \frac{(1-\rho)^4 \sigma_x \sigma}{144(c_1+c_2)c^6\Gamma^2 D_0}, \ell \le \frac{(1-\rho)^4\sigma_x\sigma}{45 c_2 c^8(1+\Gamma)^6 D_0^2}, \quad \text{and} \quad \ell' \le \frac{(1-\rho)^5\sigma_x\sigma}{27(c_1+c_2)c_2^2 c^{11}(1+\Gamma)^4 D_0^3}. \tag{19}$$

Then there exists a region $\Lambda(\delta) = \left\{K : \left\|K - K^{\text{lin}}\right\|_F \le \delta\right\} \subset \Omega$ with $\mu = \sigma_x\sigma$ such that for $K, K' \in \Lambda(\delta)$, we have

$$C(K') - C(K) \ge \operatorname{Tr}\left((K' - K)^\top \nabla C(K)\right) + \frac{\mu}{2}\left\|K' - K\right\|_F^2. \tag{20}$$

Proof We first show that $\Lambda(\delta) \subset \Omega$ for any $\delta \le \min\left\{\frac{1-\rho_1}{2\Gamma c_1^{\text{lin}}}, \frac{c_2^{\text{lin}}}{\Gamma}\right\}$. Consider the following dynamics

$$x_{t+1} = (A - BK_1)x_t = (A - BK_1^{\text{lin}})x_t + B(K_1^{\text{lin}} - K_1)x_t.$$

Set $f(x) = B(K_1^{\text{lin}} - K_1)x$. Simple computation shows f has Lipschitz constant $\ell_f = \Gamma\delta$. Following the same argument in [25, Lemma 4(a)], together with the conditions, we have

$$\|x_t\| \le 2c_1^{\text{lin}}(\rho_1^{\text{lin}} + 2c_1^{\text{lin}}\ell_f)^t \|x_0\| \le c_1(\rho_1^{\text{lin}} + (1-\rho_1))^t \|x_0\| \le c_1\rho_1^t \|x_0\|.$$

Hence, we obtain

$$\left\|(A - BK_1)^t\right\| = \sup_{x_0} \frac{\left\|(A - BK_1)^t x_0\right\|}{\|x_0\|} \le c_1\rho_1^t.$$

which is desirable. Similarly, we have

$$\|C - BK_2\| \le \left\|C - BK_2^{\text{lin}}\right\| + \|B\| \left\|B_2^{\text{lin}} - K_2\right\| \le c_2^{\text{lin}} + \Gamma\delta \le 2c_2^{\text{lin}} \le c_2.$$

which follows $K \in \Omega$.

Next, we prove the local strong convexity property of $C(K)$. By Lemma A.3, we have

$$\begin{aligned}
&C(K') - C(K) \\
&= \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\Sigma_{K'}^{\psi\psi} + 2\operatorname{Tr}(K' - K)^\top E_K \Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} \left[G_K((M - BK')\psi(x_t')) - G_K((M - BK)\psi(x_t'))\right] \\
&= 2\operatorname{Tr}(K' - K)^\top E_K \Sigma_K^{\psi\psi} + 2\operatorname{Tr}(K' - K)^\top E_K(\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}) \\
&\quad + \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} \left[G_K((M - BK')\psi(x_t')) - G_K((M - BK)\psi(x_t'))\right] \\
&\ge \operatorname{Tr}(K' - K)^\top (2E_K\Sigma_K^{\psi\psi} - B^\top \Sigma_K^{G\psi}) + 2\operatorname{Tr}(K' - K)^\top E_K(\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}) \\
&\quad + \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\Sigma_{K'}^{\psi\psi} \\
&\quad + \operatorname{Tr}(K' - K)^\top B^\top \left[\mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x_{t+1})\psi(x_t)^\top - \mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x_{t+1}')\psi(x_t')^\top\right] \\
&\quad - \frac{L}{2}\mathbb{E}\sum_{t=0}^{\infty} \left\|B(K' - K)\psi(x_t')\right\|^2 \\
&\ge \operatorname{Tr}(K' - K)^\top \nabla C(K) + \operatorname{Tr}(K' - K)^\top (R + B^\top P_{K_1} B)(K' - K)\Sigma_{K'}^{\psi\psi} \\
&\quad - 2\left\|K' - K\right\|_F \|E_K\| \left\|\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}\right\|_F \\
&\quad - \left\|K' - K\right\|_F \|B\| \left\|\mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x_{t+1})\psi(x_t)^\top - \mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x_{t+1}')\psi(x_t')^\top\right\|_F \\
&\quad - \frac{L}{2}\mathbb{E}\sum_{t=0}^{\infty} \|B\|^2 \left\|K' - K\right\|_F^2 \left\|\psi(x_t')\right\|^2,
\end{aligned}$$

where we have applied the descent lemma over G_K by the Lipchitz property in Lemma A.6. To check the conditions of the lemma, since $\ell \le 1$, we note that

$$\left\|(M - BK')\psi(x_{t'})\right\| = \left\|x_{t+1}'\right\| \le cD_0,$$

$$\left\|(M-BK)\psi(x_t')\right\| \le \left\|(A-BK_1)x_t'\right\| + \left\|(C-BK_2)\phi(x_t')\right\| \le (c_1+\ell c_2)cD_0 \le (c_1+c_2)cC_0.$$

To proceed, we need the following lemma.

Lemma A.8 *For $K, K' \in \Lambda(\delta)$, there exits constants $C_E = 3(c_1+c_2)\frac{\Gamma^4 c^3}{(1-\rho)^2}$, $C_1 = \frac{4c^3\Gamma D_0^2}{(1-\rho)^2}$, $C_2 = LC_1/2$ such that*

$$\|E_K\| \le C_E \left\|K - K^{\text{lin}}\right\|, \|\Sigma_{K'} - \Sigma_K\|_F \le C_1 \left\|K'-K\right\|_F,$$
$$\left\|\mathbb{E}\sum_{t=0}^{\infty}\nabla G_K(x_{t+1})(x_t)^\top - \mathbb{E}\sum_{t=0}^{\infty}\nabla G_K(x_{t+1}')(x_t')^\top\right\|_F \le C_2\left\|K'-K\right\|_F.$$

By assumption, $\mathbb{E}\,\psi(x_0)\psi(x_0)^\top \succeq \sigma_x I$, and $R + B^\top QB \succeq \sigma I$. Following the same argument in [25], we have

$$\begin{aligned}
&\operatorname{Tr}\left((K'-K)^\top(R+B^\top P_{K_1}B)(K'-K)\Sigma_{K'}^{\psi\psi}\right)\\
&= \operatorname{Tr}\left(\left((K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)^\top(R+B^\top P_{K_1}B)(K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)\\
&\ge \operatorname{Tr}\left(\left((K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)^\top(R+B^\top QB)(K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)\\
&\ge \sigma\operatorname{Tr}\left(\left((K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)^\top(K'-K)(\Sigma_{K'}^{\psi\psi})^{1/2}\right)\\
&= \sigma\operatorname{Tr}(K'-K)^\top(\Sigma_{K'}^{\psi\psi})(K'-K)\\
&\ge \sigma_x\sigma\left\|K'-K\right\|_F^2.
\end{aligned}$$

Combining Lemma A.8 and the above lower bound, we have

$$\begin{aligned}
C(K') - C(K) \ge \operatorname{Tr}(K'-K)^\top\nabla C(K) + \mu\left\|K'-K\right\|_F^2\\
- \left[2C_1C_E + \Gamma LC_1 + \frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho}\right]\left\|K'-K\right\|_F^2,
\end{aligned}$$

where $\mu = \sigma_x\sigma$. It remains to show that

$$2C_1C_E\delta + \Gamma LC_1 + \frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho} \le \frac{\mu}{2}.$$

Write Lipschitz constant $L = \ell C_\ell + \ell' C_{\ell'}$, where $C_\ell = \frac{5c_2c^5(1+\Gamma)^4}{16(1-\rho)^2}$ and $C_{\ell'} = \frac{3Dc_2^2c^6(1+\Gamma)^2}{16(1-\rho)^3}$. Then we have

$$\begin{aligned}
2C_1C_E\delta + \Gamma LC_1 + \frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho} &\le 2C_1C_E\delta + \Gamma LC_1 + \frac{L}{2}\Gamma C_1\\
&= 2C_1C_E\delta + \frac{3}{2}\Gamma C_1(\ell C_\ell + \ell' C_{\ell'}) \le \frac{\mu}{2},
\end{aligned}$$

as long as

$$\delta \leq \frac{\mu}{12C_1C_E} = \frac{(1-\rho)^4\sigma_x\sigma}{144(c_1+c_2)c^6\Gamma^2D_0},$$
$$\ell \leq \frac{\mu}{9\Gamma C_1C_\ell} = \frac{(1-\rho)^4\sigma_x\sigma}{45c_2c^8(1+\Gamma)^6D_0^2},$$
$$\ell' \leq \frac{\mu}{9\Gamma C_1C_{\ell'}} = \frac{(1-\rho)^5\sigma_x\sigma}{27(c_1+c_2)c_2^2c^{11}(1+\Gamma)^4D_0^3}.$$

■

To prove Lemma A.8, we first need the following result, a bound on the directional derivative of the state.

Lemma A.9 *The directional derivative of x_t w.r.t. $K = (K_1, K_2)$ along the direction $\Delta = (\Delta_1, \Delta_2)$ satisfies,*

$$\left\|x_t'[\Delta]\right\| \leq \frac{2c^2\Gamma}{1-\rho}\left\|x_0\right\|\left\|\Delta\right\|. \tag{21}$$

Proof The dynamics are

$$x_{t+1} = (A - BK_1)x_t + (C - BK_2)\phi(x_t).$$

We compute the directional derivative w.r.t $K = (K_1, K_2)$ along the direction $\Delta = (\Delta_1, \Delta_2)$:

$$\begin{aligned} x_{t+1}'[\Delta] &= (A - BK_1)x_t'[\Delta] - B\Delta_1 x_t + (C - BK_2)\frac{\partial\phi(x_t)}{\partial x_t}x_t'[\Delta] - B\Delta_2\phi(x_t) \\ &= \sum_{k=0}^{t}(A - BK_1)^{t-k}\left(-B\Delta_1 x_k + (C - BK_2)\frac{\partial\phi(x_k)}{\partial x_k}x_k'[\Delta] - B\Delta\phi(x_k)\right). \end{aligned}$$

Taking the norm and applying the Lipschitz smoothness, we have

$$\begin{aligned} \left\|x_{t+1}'[\Delta]\right\| &\leq \sum_{k=0}^{t} c_1\rho_1^{t-k}\left(\|B\|\|\Delta_1\|\|x_k\| + c_2\ell\left\|x_k'[\Delta]\right\| + \|B\|\|\Delta_2\|\ell\|x_k\|\right) \\ &\leq \sum_{k=0}^{t} c_1c_2\ell\rho_1^{t-k}\left\|x_k'[\Delta]\right\| + \sum_{k=0}^{t} c_1\rho^{t-k}\|B\|\left(\|\Delta_1\| + \ell\|\Delta_2\|\right)c\rho^k\|x_0\| \\ &= \sum_{k=0}^{t} c_1c_2\ell\rho_1^{t-k}\left\|x_k'[\Delta]\right\| + c_1\rho_1\|B\|\left(\|\Delta_1\| + \ell\|\Delta_2\|\right)\|x_0\|\frac{\rho^{t+1} - \rho_1^{t+1}}{\rho - \rho_1}. \end{aligned}$$

We assume that $x_t'[\Delta] \leq \alpha\rho^t$,where $\alpha = \left(2c_1c\|x_0\|\|B\|\left(\|\Delta_1\| + \ell\|\Delta_2\|\right)\right)/(\rho - \rho_1)$. Then we have,

$$\frac{\left\|x_{t+1}'[\Delta]\right\|}{\alpha\rho^{t+1}} \leq \sum_{k=0}^{t} c_1c_2\ell\rho_1^{t-k}\frac{\alpha\rho^k}{\alpha\rho^{t+1}} + \frac{1}{2}$$

$$
\begin{aligned}
&= \frac{c_1 c_2 \ell}{\rho} \frac{(\rho_1/\rho)^{t+1} - 1}{\rho_1/\rho - 1} + \frac{1}{2} \\
&\leq \frac{c_1 c_2 \ell}{\rho - \rho_1} + \frac{1}{2} \leq 1.
\end{aligned}
$$

By induction, we conclude that

$$
\left\| x_t'[\Delta] \right\| \leq 2 c_1 c \left\| B \right\| \left\| x_0 \right\| \left(\left\| \Delta_1 \right\| + \ell \left\| \Delta_2 \right\| \right) \frac{\rho^t}{\rho - \rho_1} \leq \frac{\sqrt{2} c^2 \Gamma}{1 - \rho} \rho^t \left\| x_0 \right\| \left\| \Delta \right\|.
$$

where in the last step we use the assumption that $\ell \leq 1$ and the basic inequality $a + b \leq \sqrt{2(a^2 + b^2)}$. ■

With Lemma A.9, we prove the perturbation analysis on the covariance matrix $\Sigma_K^{\psi\psi}$.

Proof [Proof of Lemma A.8] Note that the directional derivative of $\Sigma_K^{\psi\psi}$ w.r.t. K along the direction Δ is

$$
(\Sigma_K^{\psi\psi})'[\Delta] = \mathbb{E} \sum_{t=0}^{\infty} \frac{\partial \psi(x_t)}{\partial x_t} x_t'[\Delta] \psi(x_t)^\top + \psi(x_t) \frac{\partial \psi(x_t)}{\partial x_t} x_t'[\Delta]^\top.
$$

Then we have

$$
\begin{aligned}
\left\| (\Sigma_K^{\psi\psi})'[\Delta] \right\|_F &\leq \mathbb{E} \sum_{t=0}^{\infty} 2 \ell_\psi \left\| x_t'[\Delta] \right\| \left\| \psi(x_t) \right\| \\
&\leq \mathbb{E} \sum_{t=0}^{\infty} 2 \ell_\psi \frac{\sqrt{2} c^2 \Gamma}{1 - \rho} \left\| x_0 \right\| \left\| \Delta \right\| \ell_\psi c \rho^t \left\| x_0 \right\| \\
&\leq \frac{4 c^3 \Gamma D_0^2}{(1 - \rho)^2} \left\| \Delta \right\|_F.
\end{aligned}
$$

Now, set $g(t) = \Sigma_{K + t(K' - K)}$. Since the above result holds for any K, it follows

$$
\begin{aligned}
\left\| \Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi} \right\|_F &= \left\| g(1) - g(0) \right\|_F \\
&= \left\| \int_0^1 g'(t) \mathrm{d}t \right\|_F \\
&\leq \int_0^1 \left\| g'(t) \right\|_F \mathrm{d}t \\
&\leq \frac{4 c^3 \Gamma D_0^2}{(1 - \rho)^2} \left\| K' - K \right\|_F.
\end{aligned}
$$

To prove the second inequality, we first notice that

$$
\begin{aligned}
&\left\| \nabla G_K(x_{t+1})(x_t)^\top - \nabla G_K(x_{t+1}')(x_t')^\top \right\|_F \\
&\leq \left\| \nabla G_K(x_{t+1}) - \nabla G_K(x_{t+1}') \right\| \left\| x_t \right\| + \left\| \nabla G_K(x_{t+1}') \right\| \left\| x_t - x_t' \right\| \\
&\leq L \left\| x_{t+1} - x_{t+1}' \right\| \left\| x_t \right\| + L \left\| x_{t+1}' \right\| \left\| x_t - x_t' \right\|
\end{aligned}
$$

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

which follows

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

Finally, we bound $\|E_K\|$. Note $E_{K^{\text{lin}}}=0$. Indeed, by assumption, $P_{K_1^{\text{lin}}}=P$ and it follows

$$\begin{aligned}
(R+B^\top P_{K_1^{\text{lin}}}B)K_1^{\text{lin}} &= B^\top P_{K_1^{\text{lin}}}A, \\
(R+B^\top P_{K_1^{\text{lin}}}B)K_2^{\text{lin}} &= B^\top P_{K_1^{\text{lin}}}C,
\end{aligned}$$

which gives us $E_{K^{\text{lin}}} = RK^{\text{lin}} - B^\top P_{K_1^{\text{lin}}}(M-BK^{\text{lin}})=0$. Then we have

$$\begin{aligned}
\|E_K\| &= \|E_K - E_{K^{\text{lin}}}\| \\
&\leq \left\|R(K-K^{\text{lin}})\right\| + \left\|B^\top(P_{K_1}-P_{K_1^{\text{lin}}})(M-BK)\right\| + \left\|B^\top P_{K_1^{\text{lin}}}B(K-K^{\text{lin}})\right\| \\
&\leq (1+\Gamma^2 C_P)\left\|K-K^{\text{lin}}\right\| + \Gamma\sqrt{c_1^2+c_2^2}\left\|P_{K_1}-P_{K_1^{\text{lin}}}\right\| \\
&\leq (1+\Gamma^2 C_P)\left\|K-K^{\text{lin}}\right\| + \Gamma(c_1+c_2)\frac{2\Gamma^3c^3}{(1-\rho)^2}\left\|K-K^{\text{lin}}\right\| \\
&\leq 3(c_1+c_2)\frac{\Gamma^4c^3}{(1-\rho)^2}\left\|K-K^{\text{lin}}\right\|.
\end{aligned}$$

where we used the result in [25, Lemma 12]. ■

Similarly, we show the cost function $C(K)$ is h-smooth. The formal statement is as in the following.

Lemma A.10 *Under the same conditions in Lemma A.7, with $h = 9\frac{\Gamma^4c^4D_0^2}{(1-\rho)^2}$, for any $K, K' \in \Lambda(\delta)$, we have*

$$C(K') - C(K) \leq \operatorname{Tr}\left((K'-K)^\top\nabla C(K)\right) + \frac{h}{2}\left\|K'-K\right\|_F^2. \tag{22}$$

To prove the lemma, we need the following bound on $\Sigma_K^{\psi\psi}$.

Lemma A.11 *Under the same conditions in Lemma A.4, we have*

$$\left\|\Sigma_K^{\psi\psi}\right\| \leq \frac{2c^2 D_0^2}{1-\rho}.$$

Proof By Lemma A.4, we have

$$\begin{aligned}
\left\|\Sigma_K^{\psi\psi}\right\| &\leq \mathbb{E}\sum_{t=0}^{\infty}\|\psi(x_t)\|^2 \\
&\leq \ell_\psi^2\,\mathbb{E}\sum_{t=0}^{\infty}\|x_t\|^2 \\
&\leq \frac{\ell_\psi^2 c^2}{1-\rho^2}\,\mathbb{E}\,\|x_0\|^2 \\
&\leq \frac{2c^2 D_0^2}{1-\rho}.
\end{aligned}$$

■

Proof [Proof of Lemma A.10] By Lemma A.3

$$\begin{aligned}
&C(K') - C(K) \\
&= \operatorname{Tr}(K'-K)^\top (R + B^\top P_{K_1} B)(K'-K)\Sigma_{K'}^{\psi\psi} + 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty}\left[G_K((M-BK')\psi(x_t')) - G_K((M-BK)\psi(x_t'))\right] \\
&= 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_K^{\psi\psi} + 2\operatorname{Tr}(K'-K)^\top E_K(\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}) \\
&\quad + \operatorname{Tr}(K'-K)^\top (R + B^\top P_{K_1} B)(K'-K)\Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty}\left[G_K((M-BK')\psi(x_t')) - G_K((M-BK)\psi(x_t'))\right] \\
&\leq 2\operatorname{Tr}(K'-K)^\top E_K \Sigma_K^{\psi\psi} + 2\operatorname{Tr}(K'-K)^\top E_K(\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}) \\
&\quad + \operatorname{Tr}(K'-K)^\top (R + B^\top P_{K_1} B)(K'-K)\Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} -\operatorname{Tr}\left((K'-K)^\top B^\top \nabla G_K(x_{t+1}')\psi(x_t')^\top + \frac{L}{2}\left\|B(K'-K)\psi(x_t')\right\|^2\right) \\
&= \operatorname{Tr}(K'-K)^\top \nabla C(K) + 2\operatorname{Tr}(K'-K)^\top E_K(\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}) \\
&\quad + \operatorname{Tr}(K'-K)^\top (R + B^\top P_{K_1} B)(K'-K)\Sigma_{K'}^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty}\operatorname{Tr}(K'-K)^\top B^\top\left[\nabla G_K(x_{t+1})\psi(x_t)^\top - \nabla G_K(x_{t+1}')\psi(x_t')^\top\right] \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty}\frac{L}{2}\left\|B(K'-K)\psi(x_t')\right\|^2 \\
&\leq \operatorname{Tr}(K'-K)^\top \nabla C(K) + 2\left\|K'-K\right\|_F \|E_K\| \left\|\Sigma_{K'}^{\psi\psi} - \Sigma_K^{\psi\psi}\right\|_F + \left\|K'-K\right\|_F^2 \left\|R + B^\top P_{K_1} B\right\| \left\|\Sigma_{K'}^{\psi\psi}\right\|
\end{aligned}$$

$$
\begin{aligned}
&+ \left\|K' - K\right\|_F \|B\| \left\|\mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x_{t+1})\psi(x_t)^\top - \mathbb{E}\sum_{t=0}^{\infty} \nabla G_K(x'_{t+1})\psi(x'_t)^\top\right\|_F \\
&+ \mathbb{E}\sum_{t=0}^{\infty} \frac{L}{2} \|B\|^2 \left\|K' - K\right\|_F^2 \left\|\psi(x'_t)\right\|^2 \\
\leq\ & \operatorname{Tr}(K' - K)^\top \nabla C(K) + \left(\frac{\mu}{2} + \left\|R + B^\top P_{K_1} B\right\| \left\|\Sigma_{K'}^{\psi\psi}\right\|\right) \left\|K' - K\right\|_F^2 .
\end{aligned}
$$

Using upper bound on $\|P_{K_1}\|$ and $\left\|\Sigma_{K'}^{\psi\psi}\right\|$, we get

$$
\mu + 2\left\|R + B^\top P_{K_1} B\right\| \left\|\Sigma_{K'}^{\psi\psi}\right\| \leq \mu + 2\left(1 + \Gamma^2 \frac{c^2\Gamma^2}{1-\rho}\right) \frac{2c^2 D_0^2}{1-\rho} \leq 9\frac{\Gamma^4 c^4 D_0^2}{(1-\rho)^2} =: h.
$$

■

Finally, we characterize the global optimality of $C(K)$ as in the next lemma.

Lemma A.12 (Global Optimality) *Under the conditions in Lemma A.7 and if further ℓ, ℓ' satisfy*

$$
\ell \leq \delta \frac{2(1-\rho)^3 \sigma_x \sigma}{9 c_2 c^7 \Gamma^6 D_0^2}, \quad \text{and} \quad \ell' \leq \delta \frac{2(1-\rho)^4 \sigma_x \sigma}{9(c_1 + c_2) c_2^2 c^{10} \Gamma^4 D_0^3}. \tag{23}
$$

For $K \in \Omega \setminus \Lambda(\delta/3)$, we have

$$
C(K) > C(K^{\text{lin}}).
$$

Proof Note that $E_{K^{\text{lin}}} = RK^{\text{lin}} - B^\top P_{K_1^{\text{lin}}}(M - BK^{\text{lin}}) = 0$. Then we have

$$
\begin{aligned}
C(K) - C(K^{\text{lin}}) &= 2\operatorname{Tr}(K - K^{\text{lin}})^\top E_{K^{\text{lin}}} \Sigma_K^{\psi\psi} + \operatorname{Tr}(K - K^{\text{lin}})^\top (R + B^\top P_{K_1^{\text{lin}}} B)(K - K^{\text{lin}}) \Sigma_K^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} \left[G_{K^{\text{lin}}}((M - BK))\psi(x_t) - G_{K^{\text{lin}}}((M - BK^{\text{lin}}))\psi(x_t)\right] \\
&= \operatorname{Tr}(K - K^{\text{lin}})^\top (R + B^\top P_{K_1^{\text{lin}}} B)(K - K^{\text{lin}}) \Sigma_K^{\psi\psi} \\
&\quad + \mathbb{E}\sum_{t=0}^{\infty} \left[G_{K^{\text{lin}}}((M - BK))\psi(x_t) - G_{K^{\text{lin}}}((M - BK^{\text{lin}}))\psi(x_t)\right].
\end{aligned}
$$

With the same argument as in Lemma A.7, we can show that

$$
\operatorname{Tr}(K - K^{\text{lin}})^\top (R + B^\top P_{K_1^{\text{lin}}} B)(K - K^{\text{lin}}) \Sigma_K^{\psi\psi} \geq \sigma_x \sigma \left\|K - K^{\text{lin}}\right\|_F^2 = \mu \left\|K - K^{\text{lin}}\right\|_F^2 .
$$

Also, apply the descent lemma on $-G_{K^{\text{lin}}}$, we obtain

$$
\begin{aligned}
&G_{K^{\text{lin}}}((M - BK))\psi(x_t) - G_{K^{\text{lin}}}((M - BK^{\text{lin}}))\psi(x_t) \\
&\geq -\operatorname{Tr}(B(K - K^{\text{lin}})\psi(x_t))^\top \nabla G_{K^{\text{lin}}}((M - BK)\psi(x_t)) - \frac{L}{2}\left\|B(K - K^{\text{lin}})\psi(x_t)\right\|^2 \\
&\geq -\|B\| \left\|K - K^{\text{lin}}\right\|_F \|\psi(x_t)\| L \|x_{t+1}\| - \frac{L}{2}\|B\|^2 \left\|K - K^{\text{lin}}\right\|_F^2 \|\psi(x_t)\|^2
\end{aligned}
$$

$$\geq -L\Gamma\ell_\psi c^2\rho^{2t+1}D_0^2\left\|K-K^{\text{lin}}\right\|_F - \frac{L}{2}\Gamma^2\ell_\psi^2c^2\rho^{2t}D_0^2\left\|K-K^{\text{lin}}\right\|_F^2$$
$$\geq -L\Gamma\ell_\psi c^2\rho^{2t}D_0^2\left\|K-K^{\text{lin}}\right\|_F - \frac{L}{2}\Gamma^2\ell_\psi^2c^2\rho^{2t}D_0^2\left\|K-K^{\text{lin}}\right\|_F^2.$$

Indeed, the conditions are satisfied since

$$\|(M-BK)\psi(x_t)\| = \|x_{t+1}\| \leq c\,\|x_0\| \leq cD_0,$$
$$\left\|(M-BK^{\text{lin}})\psi(x_t)\right\| = \left\|(A-BK_1^{\text{lin}})x_t + (C-BK_2^{\text{lin}})\phi(x_t)\right\| \leq (c_1+\ell c_2)cD_0 \leq (c_1+c_2)cD_0.$$

Then we have

$$C(K)-C(K^{\text{lin}}) \geq \left[\mu - \frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho}\right]\left\|K-K^{\text{lin}}\right\|_F^2 - \frac{L\Gamma\ell_\psi c^2D_0^2}{1-\rho}\left\|K-K^{\text{lin}}\right\|_F.$$

Since $\left\|K-K^{\text{lin}}\right\|_F > \delta/3$, it suffices to show that

$$\left[\mu - \frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho}\right]\left\|K-K^{\text{lin}}\right\|_F - \frac{L\Gamma\ell_\psi c^2D_0^2}{1-\rho} \geq 0.$$

This condition is indeed satisfied since

$$\frac{L}{2}\frac{\Gamma^2\ell_\psi^2c^2D_0^2}{1-\rho} \leq \frac{\mu}{2},$$
$$\frac{L\Gamma\ell_\psi c^2D_0^2}{1-\rho} \leq \frac{\mu\delta}{6},$$

as long as

$$\ell \leq \delta\frac{2(1-\rho)^3\sigma_x\sigma}{9c_2c^7(1+\Gamma)^6D_0^2},$$
$$\ell' \leq \delta\frac{2(1-\rho)^4\sigma_x\sigma}{9(c_1+c_2)c_2^2c^{10}(1+\Gamma)^4D_0^3}.$$

■

Combining Lemma A.7 and A.10, we prove part (a) of Theorem 4.5. Further, by substituting δ chosen in Lemma A.7 into Lemma A.12, we finish the proof of part (b) of Theorem 4.5.

Appendix B. Proof of Theorem 4.6

We first characterize the gradient estimation as in the following lemma.

Lemma B.1 *Under conditions in Theorem 4.5, when $K \in \Lambda(2\delta/3)$, then given $e_{grad} > 0$, for any $\nu \in (0,1)$, when $r \leq \min\left\{\frac{\delta}{3}, \frac{1}{3h}e_{grad}\right\}$,*

$$J \geq \frac{\widehat{D}^2}{e_{grad}^2r^2}\log\frac{4\widehat{D}}{\nu}\max\left\{36\left(C(K^*)+2h\delta^2\right)^2, 144C_{\max}^2\right\}, T \geq \frac{1}{1-\rho_1}\log\frac{6\widehat{D}C_{max}}{e_{grad}r},$$

where $\widehat{D} = p(n+d)$ and $C_{\max} = \frac{24(1+\Gamma)^2c_1^2D_0^2}{1-\rho_1}$, then with probability at least $1-\nu$,

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

Proof Denote $C_r(K) = \mathbb{E}_{U\sim\text{Ball}(r)}\, C(K+U)$, where Ball$(r)$ is the ball with radius r in Frobenius norm centered at the origin. Then by [10, Lemma 1], we have

$$\nabla C_r(K) = \frac{\widehat{D}}{r^2}\,\mathbb{E}_{U\sim\text{Sphere}(r)}\, C(K+U)U.$$

Further, define $C_j = C(K+U_j)$, where $U_j \sim$ Sphere(r). Then, the error in gradient estimation can be decomposed into three parts,

$$\begin{aligned}&\left\|\widehat{\nabla C}(K) - C(K)\right\|_F \\ &\le \underbrace{\|\nabla C_r(K) - \nabla C(K)\|_F}_{:=e_1} + \underbrace{\left\|\frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}C_jU_j - \nabla C_r(K)\right\|_F}_{:=e_2} + \underbrace{\left\|\frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}\widehat{C}_jU_j - \frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}C_jU_j\right\|_F}_{:=e_3}.\end{aligned}$$

Firstly, by definition, $\nabla C_r(K) = \mathbb{E}_{U\sim\text{Ball}(r)}\nabla C(K+U)$. Since $r \le \frac{\delta}{3}$, $K+U \in \Lambda(\delta)$, in which the cost function $C(\cdot)$ is μ-strongly convex and h-smooth. Then we have

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

where we have used $r \le \frac{1}{3h}e_{\text{grad}}$.

Next, notice that $\left\{\frac{\widehat{D}}{r^2}C_jU_j\right\}_{j=1}^{J}$ are i.i.d. with expectation $\nabla C_r(K)$. Again, by h-smoothness,

$$\left\|\frac{\widehat{D}}{r^2}C_jU_j\right\|_F \le \frac{\widehat{D}}{r}C_j \le \frac{\widehat{D}}{r}\left(C(K^*) + \frac{h}{2}\|K+U_j-K^*\|_F^2\right) \le \frac{\widehat{D}}{r}\left(C(K^*) + 2h\delta^2\right).$$

Then, by matrix Bernstein inequality, we have

$$\mathbb{P}\left(e_2 \le \frac{e_{\text{grad}}}{3}\right) \ge 1 - 2\widehat{D}\exp\left(-\frac{(e_{\text{grad}}J/3)^2}{4J\left((\widehat{D}/r)(C(K^*)+2h\delta^2)\right)^2}\right) \ge 1-\nu/2,$$

where we have used $J \ge \frac{36\widehat{D}^2}{e_{\text{grad}}^2 r^2}\left(C(K^*)+2h\delta^2\right)^2\log\frac{4\widehat{D}}{\nu}$.

Finally, to bound e_3, we further decompose it into two parts. Define $\tilde{C}_j = \mathbb{E}\sum_{t=0}^{T}\left[x_t^\top Qx_t + u_t^\top Ru_t\right]$, where $u_t = -(K+U_j)\psi(x_t)$. Then, we have

$$e_3 \le \underbrace{\left\|\frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}\widehat{C}_jU_j - \frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}\tilde{C}_jU_j\right\|_F}_{:=e_4} + \underbrace{\left\|\frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}\tilde{C}_jU_j - \frac{1}{J}\sum_{j=1}^{J}\frac{\widehat{D}}{r^2}C_jU_j\right\|_F}_{:=e_5}.$$

Note that

$$\left|\hat{C}_j\right| = \sum_{t=0}^{\infty}\left[x_t^\top Qx_t + u_t^\top Ru_t\right]$$

$$
\begin{aligned}
&\leq \sum_{t=0}^{\infty} \|x_t\|^2 \|Q\| + \|\psi(x_t)\|^2 \left\|(K+U_j)^\top R(K+U_j)\right\| \\
&\leq \left(\|Q\| + \ell_\psi^2 \left\|(K+U_j)^\top R(K+U_j)\right\|\right) \sum_{t=0}^{\infty} \|x_t\|^2 \\
&\leq (1+2(2\Gamma)^2) \sum_{t=0}^{\infty} c^2 \rho^{2t} D_0^2 \\
&\leq \frac{3(1+\Gamma)^2 c^2 D_0^2}{1-\rho} =: C_{\max},
\end{aligned}
$$

where we have used the fact $K+U_j \in \Lambda(\delta)$ and thus $\|K+U_j\| \leq 1+\Gamma$. Since, $\mathbb{E}\left[\hat{C}_j U_j - \tilde{C}_j U_j \middle| U_j\right] = 0$, by matrix Bernstein inequality, with probability at least $1-\nu/2$,

$$
\mathbb{P}\left(e_4 \leq \frac{e_{\text{grad}}}{6}\right) \geq 1 - 2\widehat{D} \exp\left(-\frac{(e_{\text{grad}} J/6)^2}{4JC_{\max}^2}\right) \geq 1 - \frac{\nu}{2},
$$

where we have used $J \geq \frac{144\widehat{D}^2 C_{\max}^2}{e_{\text{grad}}^2 r^2} \log \frac{4\widehat{D}}{\nu}$. Also, we have

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

As such, we have

$$
e_5 \leq \frac{\widehat{D}}{r} C_{\max} \rho^{2(T+1)} \leq \frac{1}{6} e_{\text{grad}},
$$

where we have used $T \geq \frac{1}{1-\rho_1} \log \frac{6\widehat{D}C_{\max}}{e_{\text{grad}} r}$. Hence, $e_3 \leq \frac{1}{3} e_{\text{grad}}$ with probability at least $1 - \frac{\nu}{2}$, which completes the proof. ■

With Lemma B.1, we prove convergence rate of Algorithm 1.

Proof [Proof of Theorem 4.6] Let $\mathcal{F}_m$ be the filtration generated by $\left\{\widehat{\nabla C}(K^{m'})\right\}_{m'=0}^{m-1}$. Define the following event:

$$
\begin{aligned}
\mathcal{E}_m = &\left\{K^{m'} \in \text{Ball}(K^*, \delta/3), m' = 0, \ldots, m\right\} \\
&\cap \left\{\left\|\widehat{\nabla C}(K^{m'}) - \nabla C(K^{m'})\right\|_F \leq e_{\text{grad}}, m' = 0, \ldots, m-1\right\},
\end{aligned}
$$

where $\text{Ball}(K^*, \delta/3) = \{K : \|K - K^*\|_F \leq \delta/3\}$. It is easy to see that both K^m and the event $\mathcal{E}_m$ are $\mathcal{F}_m$-measurable. We want to show the following inequality:

$$
\mathbb{E}\left[1(\mathcal{E}_{m+1}) | \mathcal{F}_m\right] 1(\mathcal{E}_m) \geq \left(1 - \frac{\nu}{M}\right) 1(\mathcal{E}_m), \tag{24}
$$

i.e., if event $\mathcal{E}_m$ is true, conditioned on $\mathcal{F}_m$, the event $\mathcal{E}_{m+1}$ happens with high probability. Note that on event $\mathcal{E}^m$, we have $\left\|K^m - K^{\text{lin}}\right\|_F \leq \|K^m - K^*\|_F + \left\|K^{\text{lin}} - K^*\right\|_F \leq 2\delta/3$, which follows that $K^m \in \Lambda(\delta)$. Under our selection of parameters, with probability at least $1 - \nu/M$, we have $\left\|\widehat{\nabla C}(K^m) - \nabla C(K^m)\right\|_F \leq e_{\text{grad}}$. It follows that $K^{m+1} \in \text{Ball}(K^*, \delta/3)$. Indeed, by μ-strong convexity and h-smoothness,

$$\begin{aligned}\left\|K^{m+1} - K^*\right\|_F &\leq \|K^m - \eta\nabla C(K^m) - K^*\|_F + \eta\left\|\widehat{\nabla C}(K^m) - \nabla C(K^m)\right\|_F \\ &\leq (1-\eta\mu)\|K^m - K^*\|_F + \eta e_{\text{grad}} \\ &\leq (1-\eta\mu)\frac{\delta}{3} + \eta e_{\text{grad}} \leq \frac{\delta}{3},\end{aligned}$$

where we have used $e_{\text{grad}} \leq \frac{\delta\mu}{3}$ in the last inequality. As such, taking the expectation of (24) on both sides, we have

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

Unrolling this recursive relation, we obtain $\mathbb{P}(\mathcal{E}_m) \geq \left(1 - \frac{\nu}{M}\right)^M \mathbb{P}(\mathcal{E}_0) = \left(1 - \frac{\nu}{M}\right)^M \geq 1 - \nu$. Now, on event $\mathcal{E}_m$, we also have

$$\begin{aligned}\left\|K^M - K^*\right\|_F &\leq (1-\eta\mu)^M \left\|K^0 - K^*\right\|_F + \eta e_{\text{grad}} \sum_{m=0}^{M-1}(1-\eta\mu)^m \\ &\leq (1-\eta\mu)^M \frac{\delta}{3} + \frac{e_{\text{grad}}}{\mu} \\ &\leq \sqrt{\frac{2\epsilon}{h}},\end{aligned}$$

where we have used $M \geq \frac{1}{\eta\mu}\log\left(\frac{\delta}{3}\sqrt{\frac{2h}{\epsilon}}\right)$ and $e_{\text{grad}} \leq \mu\sqrt{\frac{\epsilon}{2h}}$. Finally, by h-smoothness, we have

$$C(K^M) \leq C(K^*) + \frac{h}{2}\left\|K^M - K^*\right\|_F^2 \leq C(K^*) + \epsilon,$$

which is desirable. ■