Sample Complexity of the Linear Quadratic Regulator: A Reinforcement Learning Lens

Amirreza Neshaei Moghaddam   AMIRNESHA@UCLA.EDU
Department of Electrical and Computer Engineering
University of California, Los Angeles

Alex Olshevsky   ALEXOLS@BU.EDU
Department of Electrical and Computer Engineering
Boston University

Bahman Gharesifard   GHARESIFARD@UCLA.EDU
Department of Electrical and Computer Engineering
University of California, Los Angeles

Abstract

We provide the first known algorithm that provably achieves ε-optimality within $\widetilde{\mathcal{O}}(1/\varepsilon)$ function evaluations for the discounted discrete-time LQR problem with unknown parameters, without relying on two-point gradient estimates. These estimates are known to be unrealistic in many settings, as they depend on using the exact same initialization, which is to be selected randomly, for two different policies. Our results substantially improve upon the existing literature outside the realm of two-point gradient estimates, which either leads to $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ rates or heavily relies on stability assumptions.

Keywords: Linear Quadratic Regulator, Model-free Control, Policy Gradient

1. Introduction

The Linear-Quadratic Regulator (LQR) has been used as a benchmark in optimal control theory since the sixties, see Lee and Markus (1967). The key distinguishing property of LQR problems is that the optimal controller is linear and can be fully characterized by the celebrated Riccati equation (Bertsekas, 1995). Naturally, with the recent increase in interest in model-free and data-driven methods, the study of LQR problems has resurfaced in the literature in scenarios where the model parameters are unknown and either need to be estimated, or model-free strategies need to be used. Even though such settings fall within the realm of adaptive control, the majority of classical studies addressing this issue have concentrated on system identification or examining asymptotic outcomes (Lai and Wei, 1982; Chen and Guo, 1987; Chen and Zhang, 1990; Campi and Kumar, 1998; Bittanti and Campi, 2006).

Recently, the problem has been examined from a machine learning standpoint in both online and offline contexts. In online settings, least-square estimators have been demonstrated to achieve sublinear regret. This area has seen extensive research focusing on the details of these estimations (Abbasi-Yadkori and Szepesvári, 2011; Cohen et al., 2019; Mania et al., 2019; Akbari et al., 2022; Simchowitz and Foster, 2020). This paper focuses on the offline

setting and builds on a sequence of breakthrough results through a reinforcement learning lens, starting with Fazel et al. (2018). By establishing a gradient domination/Polyak-Lojasiewicz property, the results of Fazel et al. (2018) first demonstrate that exact gradient descent, in the model-based case, converges to the global optimal solution, despite the non-convex landscape of the LQR problem under study. Using this and in the model-free settings, gradient estimations are derived from samples of the cost function value, leading to policy gradient methods. For the undiscounted discrete-time LQR under the random initialization setting, global convergence guarantees are provided using so-called one-point gradient estimates. As also explicitly pointed out in later work by Malik et al. (2020), the convergence rate for obtaining an ε-optimal policy established in Fazel et al. (2018) is only of the order $\widetilde{\mathcal{O}}(1/\varepsilon^4)$ in zero-order evaluations. Note that by zero-order methods, we mean a setup where gradients are not available and can only be approximated using samples of the function value. The two most common such methods in the LQR problem are the one-point and two-point estimates where the former is obtained from a single function evaluation and the latter from two different such evaluations[1].

The next significant development related to our work is presented in Malik et al. (2020), which considers the discounted discrete-time LQR and employs zero-order methods for gradient estimation. For the essential case of one-point gradient estimation, an enhanced analysis is proposed. This analysis does not rely on stability assumptions (i.e., it does not assume a priori that the policies remain stable throughout the algorithm), yet improves the convergence rate reported in Fazel et al. (2018) from $\widetilde{\mathcal{O}}(1/\varepsilon^4)$ to $\widetilde{\mathcal{O}}(1/\varepsilon^2)$. Remarkably, with a two-point gradient estimate, ε-optimality can be achieved using only $\widetilde{\mathcal{O}}(1/\varepsilon)$ function evaluations. Similar findings are reported in Mohammadi et al. (2022), which are somewhat restrictive in terms of scaling of probability bounds with respect to dimensions. The substantial improvement in Malik et al. (2020) stems from the application of sharp probabilistic estimates on stability regions using martingale techniques, a method we also heavily rely on. It should be noted that in both mentioned works, a constant learning rate is employed for the policy update. Interestingly, it is not difficult to observe that there is no advantage in using time-varying learning rates when the technique developed in Malik et al. (2020) is applied directly.

It is worth pointing out the literature related to the discrete-time LQR problem with time-average cost. For instance, Yang et al. (2019) employs an actor-critic approach to achieve a sample complexity of $\widetilde{\mathcal{O}}(1/\varepsilon^5)$. Similarly, using actor-critic methods, Zhou and Lu (2023) demonstrates that a sample complexity of $\widetilde{\mathcal{O}}(1/\varepsilon)$ is achievable, assuming almost sure stability and boundedness of the policy size throughout the algorithm. However, assuming boundedness is often unrealistic, and assuming stability is even less so, given the system's inherently noisy dynamics. For example, this issue is echoed in the recent work by Ju et al. (2023), which presupposes the boundedness of policies at every iteration.

As part of our contributions, and somewhat inspired by REINFORCE (Williams, 1992; Sutton et al., 1999), we propose a different gradient estimate scheme. Our approach relies on a new take on using policy gradient for gradient estimation based on appropriate sampling of deterministic policies, and only requires a single noisy cost evaluation, unlike two-point methods that require two evaluations under an identical noise realization (Malik et al., 2020).

---

1. We give the formal definitions of these estimates in equations (7) and (8).

We are able to achieve high-probability upper bounds on our gradient estimations using moment concentration inequalities. Coupled with the adoption of time-varying learning rates, our methodology enables us to reach a $\widetilde{\mathcal{O}}(1/\varepsilon)$ convergence rate, circumventing the need for two-point gradient estimations.

Similar to Malik et al. (2020), our gradient estimate relies on an oracle that returns noisy zero-order evaluations of the cost function. Moreover, we assume access to a single state observation drawn randomly from the discounted state distribution. We consider this assumption milder than that of Dean et al. (2020), which requires access to an entire state trajectory, or Malik et al. (2020), whose two-point method implicitly assumes the ability to both observe and *select* a specific random initial state for a second policy rollout-something that is rarely feasible in realistic systems.

2. Problem Statement

We start with a few mathematical notations that will be used throughout. For arbitrary matrix $M \in \mathbb{R}^{m \times n}$, we use $\|M\|$, $\|M\|_F$, and $\sigma_{\min}(M)$ to denote the 2-norm, Frobenius norm, and the minimum singular value of M respectively. In addition, for a square matrix $\tilde{M} \in \mathbb{R}^{n \times n}$, $\rho(\tilde{M})$ denotes the spectral radius of $\tilde{M}$, $\mathrm{tr}(\tilde{M})$ the trace of $\tilde{M}$, and $\mathcal{K}(\tilde{M})$ the Kreiss constant of $\tilde{M}$:

$$\mathcal{K}(\tilde{M}) := \sup_{|z|>1, z \in \mathbb{C}} (|z|-1)\|(zI - \tilde{M})^{-1}\|.$$

We also use $\langle M_1, M_2 \rangle := \mathrm{tr}(M_1^\top M_2)$ to denote the inner product of the matrices $M_1, M_2 \in \mathbb{R}^{m \times n}$.

Let us now define the problem under study. We consider the discrete-time infinite-horizon discounted LQR problem

$$\min \mathbb{E}\left[\sum_{t \geq 0} \gamma^t c_t\right] \quad \text{s.t.} \quad x_{t+1} = Ax_t + Bu_t + z_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the system state at time t, initialized (deterministically or randomly) at x_0; $u_t \in \mathbb{R}^m$ is the control input at time t; and $z_t \in \mathbb{R}^n$ is the additive noise of the system at time t. The stage cost is defined as

$$c_t := x_t^\top Q x_t + u_t^\top R u_t,$$

where $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times m}$ are positive-definite matrices that parameterize the quadratic costs. The system matrices are $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. In most of what follows, we assume that the pair (A, B) is controllable.

As noted above, randomness is introduced in two different ways in the above problem formulation: through the initialization or as an added disturbance to the dynamics. This has led to two separate scenarios considered in the literature:

- **Random initialization:** where it is assumed that the additive noise z_t is zero for all $t \geq 0$, and that the initial state x_0 is randomly chosen from an initial distribution $\mathcal{D}_0$. Given the initial state x_0, we let $\mathcal{C}_{\mathrm{init}, \gamma}(K; x_0)$ be the random variable representing

the cost of implementing the linear policy $K \in \mathbb{R}^{m \times n}$, i.e., choosing $u_t = -Kx_t$ for $t \geq 0$, from the initial state x_0:

$$\mathcal{C}_{\text{init}}(K; x_0) := \sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t), \tag{2}$$

where $0 < \gamma \leq 1$ is the discount factor, and the dynamics is given by (1) with $z_t = 0$. That is, in this case the trajectories satisfy the dynamics

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ u_t &= -Kx_t. \end{aligned} \tag{3}$$

Naturally, the objective is to minimize the population cost defined as

$$\mathcal{C}_{\text{init}}(K) := \mathbb{E}_{x_0 \sim \mathcal{D}_0}[\mathcal{C}_{\text{init}}(K; x_0)]$$

over choices of the policy K.

- **Noisy dynamics:** where it is assumed z_t is drawn i.i.d. for each t from a distribution $\mathcal{D}_{\text{add}}$, and that the initial state x_0 is set deterministically to zero. Given a sequence of random variables $\mathcal{Z} = \{z_t\}_{t \geq 0}$, we let $\mathcal{C}_{\text{dyn}}(K; \mathcal{Z})$ be the random variable representing the cost of implementing the linear policy K on a system where the additive noise is drawn from $\mathcal{Z}$, i.e.,

$$\mathcal{C}_{\text{dyn}}(K; \mathcal{Z}) := \sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t),$$

where we have set $x_0 = 0$, the dynamics is given by (1) with $u_t = -Kx_t$ for each $t \geq 0$, and $0 < \gamma < 1$ is the discount factor. In contrast to the random initialization setting, the discount factor in this setting obeys $\gamma < 1$ to prevent the cost from diverging to infinity for all K due to the accumulation of noise over time. Once again, the objective is to minimize the population cost

$$\mathcal{C}_{\text{dyn}}(K) := \mathbb{E}_{\mathcal{Z} \sim \mathcal{D}_{\text{add}}^{\mathbb{N}}}[\mathcal{C}_{\text{dyn}}(K; \mathcal{Z})].$$

By classical results in optimal control theory (Kalman et al., 1960; Lee and Markus, 1967), the optimal controller in both cases is linear and can be expressed as $u_t = -K^* x_t$ where $t \geq 0$ and $K^* \in \mathbb{R}^{m \times n}$ is the controller gain, and can be explicitly computed. When the system matrices are known, which is not the case in this paper, the policy K^* can be derived as follows

$$K^* = \gamma(R + \gamma B^\top PB)^{-1} B^\top PA,$$

where P denotes the unique positive definite solution to the discounted discrete-time algebraic Riccati equation (Bertsekas, 1995):

$$P = \gamma A^\top PA - \gamma^2 A^\top PB(R + \gamma B^\top PB)^{-1} B^\top PA + Q.$$

Throughout this paper, we closely follow the notation and terminology that is introduced in the seminal work of Malik et al. (2020). To start, for a random variable $v \sim \mathcal{D}$ where $\mathcal{D} \in \{\mathcal{D}_0, \mathcal{D}_{\text{add}}\}$, we assume that

$$\mathbb{E}[v] = 0, \quad \mathbb{E}[vv^\top] = I, \text{ and } \|v\|^2 \leq C_m \text{ a.s.} \tag{4}$$

where as per usual, "a.s." refers to almost surely. The assumption on the covariance being identity is without loss of generality, see Malik et al. (2020). Moreover, it is noteworthy to mention that using the definition (2) with the trajectories following (3), the cost for the random initialization setting can be rewritten as

$$\mathcal{C}_{\text{init}}(K; x_0) = x_0^\top P_K x_0, \tag{5}$$

where P_K is the symmetric positive semi-definite solution to the fixed point equation:

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

Consequently, it also holds that

$$\begin{aligned}
\mathcal{C}_{\text{init}}(K) &= \mathbb{E}_{x_0 \sim \mathcal{D}_0}[\mathcal{C}_{\text{init}}(K; x_0)] \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}_0}[x_0^\top P_K x_0] \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}_0}[\text{tr}(P_K x_0 x_0^\top)] \\
&= \text{tr}(P_K \mathbb{E}_{x_0 \sim \mathcal{D}_0}[x_0 x_0^\top]) \\
&\overset{(i)}{=} \text{tr}(P_K),
\end{aligned} \tag{6}$$

where (i) follows from assumption (4) on the randomness. Although this formulation is stated for the cost under the random initialization setting, it turns out that the two costs are essentially equivalent when the respective systems are driven by noise with the same first two moments, in the sense that is shown in Lemma 4 to follow. For this reason, we focus on the random initialization scenario henceforth.

Let us now state the problem that we consider throughout this paper. We recall here that we assume that the pair (A, B) is controllable, however, *unknown*. A policy K is said to stabilize the system (A, B) if we have $\rho(A - BK) < 1$. Note that by the controllability assumption, there exists some policy K satisfying the condition $\rho(A - BK) < 1$. Furthermore, we assume access to *some* stable policy K_0; this is a mild assumption that can be satisfied in a variety of ways; we refer the reader to Fazel et al. (2018); Dean et al. (2018). We use K_0 to initialize our algorithms, which we shortly introduce.

With this in mind, the main objective of this paper is to find an ε-optimal policy $\hat{K}$, i.e., one satisfying

$$\mathcal{C}_{\text{init}}(\hat{K}) - \mathcal{C}_{\text{init}}(K^*) \leq \varepsilon,$$

where K^* is an optimal policy. The proposed scheme in the literature crucially involves forming an estimation of the gradient of the cost function (2), which is then used for a gradient update with an appropriate learning rate.

To make our later comparisons precise and to clarify the discussions emphasized earlier, we now recall the standard forms of the one-point and two-point estimates. The one-point estimate at a policy $K \in \mathbb{R}^{m \times n}$ is computed as

$$g_r^1(K) := \mathcal{C}_{\text{init}}(K + rU; x_0) \cdot \frac{mn}{r} U, \tag{7}$$

for a smoothing radius $r \in \mathbb{R}$ and a random matrix $U \in \mathbb{R}^{m \times n}$ drawn uniformly over matrices with unit Frobenius norm. The two-point estimate instead uses

$$g_r^2(K) := [\mathcal{C}_{\text{init}}(K + rU; x_0) - \mathcal{C}_{\text{init}}(K - rU; x_0)] \cdot \frac{mn}{2r} U, \tag{8}$$

which requires cost evaluations under two different policies, $K + rU$ and $K - rU$, with respect to the *same* initial condition x_0. This is often unrealistic in practice, since x_0 is typically random and not something the algorithm can choose or reproduce across rollouts. The estimator we propose later avoids this assumption and instead works by just using a single noisy cost evaluation along one perturbed trajectory.

In accordance with this, we present an algorithm here, displayed as Algorithm 1, where we use an estimate inspired by the REINFORCE method (Williams, 1992; Sutton et al., 1999) with a time-varying learning rate to achieve ε-optimality. Below, we present a brief roadmap of the key contributions and supporting arguments developed in this paper.
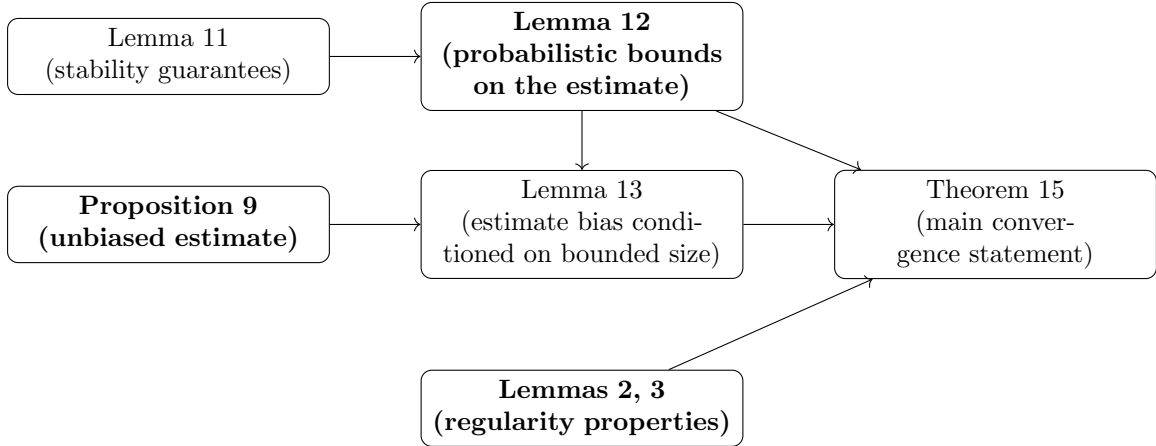

Figure 1: Roadmap of the main technical results.

In this diagram, we omit most intermediate steps and highlight (in bold) the main components that the convergence theorem ultimately depends on. Among these, **Lemma 12** and **Proposition 9** are the core novel contributions of this paper. The regularity properties (Lemmas 2, 3), which we will discuss in detail in the next section, are adapted from prior work Malik et al. (2020) and included here for completeness.

6

2.1 Regularity Properties

We introduce some notations related to the regularity properties of the cost functions; these will play a crucial role in some of our bounds; the next few results are borrowed from Malik et al. (2020).

Lemma 1 (LQR cost is locally Lipschitz) *(Malik et al., 2020, Lemma 4) Given any linear policy K with finite cost, there exist positive scalars $(\lambda_K, \widetilde{\lambda}_K, \zeta_K)$, depending on the function value $\mathcal{C}_{\mathrm{init}}(K)$, such that for all policies K' satisfying $\|K' - K\|_F \leq \zeta_K$, and for all initial states x_0, we have*

$$|\mathcal{C}_{\mathrm{init}}(K') - \mathcal{C}_{\mathrm{init}}(K)| \leq \lambda_K \|K' - K\|_F, \ \ and$$
$$|\mathcal{C}_{\mathrm{init}}(K'; x_0) - \mathcal{C}_{\mathrm{init}}(K; x_0)| \leq \widetilde{\lambda}_K \|K' - K\|_F.$$

Lemma 2 (LQR cost has locally Lipschitz gradients) *(Malik et al., 2020, Lemma 5) Given any linear policy K with finite cost, there exist positive scalars (β_K, ϕ_K), depending on the function value $\mathcal{C}_{\mathrm{init}}(K)$, such that for all policies K' satisfying $\|K' - K\|_F \leq \beta_K$, we have*

$$\|\nabla \mathcal{C}_{\mathrm{init}}(K') - \nabla \mathcal{C}_{\mathrm{init}}(K)\|_F \leq \phi_K \|K' - K\|_F.$$

Lemma 3 (LQR satisfies PL) *(Malik et al., 2020, Lemma 6) There exists a universal constant $\mu_{lqr} > 0$ such that for all stable policies K, we have*

$$\|\nabla \mathcal{C}_{\mathrm{init}}(K)\|_F^2 \geq \mu_{lqr} \left(\mathcal{C}_{\mathrm{init}}(K) - \mathcal{C}_{\mathrm{init}}(K^*) \right), \tag{9}$$

where K^* is a global minimizer of the cost function $\mathcal{C}_{\mathrm{init}}$.

For the sake of exposition, these properties are stated here without specifying the various smoothness and PL constants. The explicit expressions for $\{\lambda_K, \widetilde{\lambda}_K, \phi_K, \beta_K, \zeta_K, \mu_{\mathrm{lqr}}\}$ in terms of the parameters of the LQR problem are provided in Malik et al. (2020, Appendix A). Remark 5 to follow will provide further elaboration on these parameters as well.

Lemma 4 (Equivalence of population costs up to scaling) *(Malik et al., 2020, Lemma 7) For all policies K, we have*

$$\mathcal{C}_{\mathrm{dyn}}(K) = \frac{\gamma}{1 - \gamma} \mathcal{C}_{\mathrm{init}}(K).$$

This result shows that the noisy dynamics and random initialization population costs behave identically when their respective sources of randomness have the same first two moments. Therefore, we focus on the random initialization cost from now on and remind the reader that $\mathcal{C}(K) := \mathcal{C}_{\mathrm{init}}(K)$ for ease of notation.

We define the set

$$\mathcal{G}^{\mathrm{lqr}} := \{K \mid \mathcal{C}(K) - \mathcal{C}(K^*) \leq 10\mathcal{C}(K_0)\}. \tag{10}$$

Since $\mathcal{C}$ is (ζ_K, λ_K) locally Lipschitz and (β_K, ϕ_K) locally smooth, both properties hold simultaneously within a Frobenius norm radius $\omega_K := \min\{\beta_K, \zeta_K\}$ of a point $K \in \mathcal{G}^{\mathrm{lqr}}$. We define the quantities

$$\phi_{\mathrm{lqr}} := \sup_{K \in \mathcal{G}^{\mathrm{lqr}}} \phi_K, \qquad \lambda_{\mathrm{lqr}} := \sup_{K \in \mathcal{G}^{\mathrm{lqr}}} \lambda_K, \quad \text{and} \quad \omega_{\mathrm{lqr}} := \inf_{K \in \mathcal{G}^{\mathrm{lqr}}} \omega_K.$$

It is noteworthy to mention that these values are non-zero and finite, and their explicit formulation is provided in Malik et al. (2020, Appendix A), see Remark 5 to follow for further clarification.

Observe that by the definition of these quantities, one can immediately show that for any $K \in \mathcal{G}^{\mathrm{lqr}}$ and $K' \in \mathbb{R}^{m \times n}$ such that $\|K' - K\|_F \leq \omega_{\mathrm{lqr}}$, we have that

$$|\mathcal{C}(K') - \mathcal{C}(K)| \leq \lambda_{\mathrm{lqr}} \|K' - K\|_F, \text{ and}$$
$$\|\nabla \mathcal{C}(K') - \nabla \mathcal{C}(K)\|_F \leq \phi_{\mathrm{lqr}} \|K' - K\|_F.$$

Remark 5 We now describe how to specify the set of parameters $\{\lambda_K, \widetilde{\lambda_K}, \phi_K, \beta_K, \zeta_K, \mu_{\mathrm{lqr}}\}$ in our setting. We start by recalling that a set of parameters $\{c_{K_0}, c_{K_1}, \ldots, c_{K_9}\}$ is defined in Malik et al. (2020, Appendix A), which notably depend on $\mathcal{C}(K)$. Subsequently, by replacing said $\mathcal{C}(K)$ with $\sup_{K \in \mathcal{G}^{\mathrm{lqr}}} \mathcal{C}(K)$, they obtain a set of constants $\{\widetilde{c_{K_0}}, \widetilde{c_{K_1}}, \ldots, \widetilde{c_{K_9}}\}$ which are independent of K. For ease of access for the reader, we point out that

$$\omega_{\mathrm{lqr}} = \widetilde{c_{K_9}}, \quad \phi_{\mathrm{lqr}} = \widetilde{c_{K_7}}, \quad \text{and} \quad \lambda_{\mathrm{lqr}} = \widetilde{c_{K_8}}.$$

Moreover, it holds that $\max\{\|K\|, \|\nabla \mathcal{C}(K)\|_F\} \leq \widetilde{c_{K_1}}$ for any $K \in \mathcal{G}^{\mathrm{lqr}}$, see Malik et al. (2020, Appendix A) and Fazel et al. (2018, Lemma 22). Note that the only required modification in the values of $\widetilde{c_{K_0}}, \widetilde{c_{K_1}}, \ldots, \widetilde{c_{K_9}}$ for our case is having $10\mathcal{C}(K_0) + \mathcal{C}(K^*)$ as $\sup_{K \in \mathcal{G}^{\mathrm{lqr}}} \mathcal{C}(K)$ instead of Malik et al. (2020)'s $10\mathcal{C}(K_0) - 9\mathcal{C}(K^*)$, due to the difference in our definition of $\mathcal{G}^{\mathrm{lqr}}$ in (10). $\diamond$

We now provide an informal statement of our main result, which shows that our proposed algorithm obtains an ε-optimal policy after $\widetilde{\mathcal{O}}(1/\varepsilon)$ iterations. As we outline precisely later, this algorithm forms an estimate $\widehat{\nabla \mathcal{C}}(K_t)$ of the gradient at a given time t and updates the policy K_t with a time-varying learning rate α_t.

Theorem 6 (Informal Statement of Our Main Result): *If the step-size is chosen as $\alpha_t = C \frac{1}{t+N}$ with N "large enough", i.e., $N \sim \mathcal{O}\left((\log \frac{1}{\delta})^{3/2}\right)$ for any chosen δ, and C being a known constant, then after $T = \mathcal{O}\left(\frac{1}{\varepsilon}(\log \frac{1}{\delta})^{3/2}\right)$ iterations, provided the discount factor exceeds a constant threshold strictly less than 1, we have that*

$$\mathcal{C}(K_T) - \mathcal{C}(K^*) \leq \varepsilon$$

with a probability of at least $4/5 - \delta T$. In particular, choosing δ proportional to $1/T$, we attain $\mathcal{C}(K_T) - \mathcal{C}(K^*)$ with a constant probability with a sample complexity of $\widetilde{\mathcal{O}}(1/\varepsilon)$.

A precise version of this result is given later in Theorem 15, with the corresponding algorithm formally stated in Algorithm 1.

8

Let us first point out that this result substantially improves the ones in the literature by achieving a $\widetilde{\mathcal{O}}(1/\varepsilon)$ rate without any additional assumptions. The best previous result achieves a convergence rate of $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ (Malik et al., 2020) in this setting. Indeed, $\widetilde{\mathcal{O}}(1/\varepsilon)$ rates were only available using so-called two-point estimates which re-use randomness (e.g., require being able to initialize the system at a given x_0). Note that the limitations of this assumption become especially evident in the noisy dynamics setting, where access to cost evaluations of two different policies is required under the exact same infinite sequence of additive noise. This is significantly more restrictive than in the random initialization setting, which only requires matching a single random variable—namely, the initial condition. In both cases, however, this coupling is difficult to realize in practice, as one must have perfect control over a simulator to use such estimates; one cannot implement them for black-box systems with unknown dynamics which need to learn in the real world, for example. In contrast, our result only uses gradient estimates with a single zero-order evaluation at each step.

We now begin the process of collecting the essentials needed to articulate our theorem precisely and to prove this result, beginning with a fresh examination of the policy gradient that we employ for gradient estimation.

3. Policy Gradient

Most formulations of the policy gradient require probabilistic policies; in contrast, as can be seen in (3), we have used a deterministic policy given by $u_t = -Kx_t$. To remedy, we use a control input $u_{\hat{t}}$, to be defined shortly, where $\hat{t}$ is sampled at random from the distribution $\mu_\gamma(t) := (1-\gamma)\gamma^t$, where $t \in \{0,1,2,\cdots\}$. Keeping this in mind, we now compute

$$\widehat{\nabla\mathcal{C}}(K) := \frac{1}{1-\gamma}Q^K(x_{\hat{t}}, u_{\hat{t}})\nabla_K \log \pi_K(u_{\hat{t}}|x_{\hat{t}}), \tag{11}$$

where the control input $u_{\hat{t}}$ is drawn from the Gaussian distribution $\mathcal{N}(-Kx_{\hat{t}}, \sigma^2 I_m)$ for some $\sigma > 0$ only for the selected iteration $\hat{t}$, and $x_{\hat{t}} = (A - BK)^{\hat{t}}x_0$ with $x_0 \sim \mathcal{D}$ as before. Note that

$$\mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\widehat{\nabla\mathcal{C}}(K)\right] = \sum_{t=0}^{\infty}\gamma^t Q^K(x_t, u_t)\nabla_K \log \pi_K(u_t|x_t),$$

where

$$\pi_K(u_t|x_t) = \frac{1}{\sqrt{(2\pi)^m(\sigma^2)^m}}e^{-\frac{(u_t+Kx_t)^\top(u_t+Kx_t)}{2\sigma^2}}, \tag{12}$$

and

$$\begin{aligned}
Q^K(x_t, u_t) &:= x_t^\top Q x_t + u_t^\top R u_t + \gamma\mathcal{C}_{\text{init}}(K; x_{t+1}) \\
&= x_t^\top Q x_t + u_t^\top R u_t + \gamma\mathcal{C}_{\text{init}}(K; Ax_t + Bu_t) \\
&\stackrel{(i)}{=} x_t^\top Q x_t + u_t^\top R u_t + \gamma(Ax_t + Bu_t)^\top P_K(Ax_t + Bu_t),
\end{aligned} \tag{13}$$

where (i) is on account of (5). Note that we can also rewrite $u_{\hat{t}} \sim \mathcal{N}(-Kx_{\hat{t}}, \sigma^2 I_m)$ as

$$u_{\hat{t}} = -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}, \tag{14}$$

9

where $\eta_{\hat{t}} \sim \mathcal{N}(0, I_m)$. Moreover, we have the following lemma to provide an alternative way of representing (11).

Lemma 7 *The gradient estimate in* (11) *can be modified to get*

$$\widehat{\nabla \mathcal{C}}(K) = -\frac{1}{\sigma(1-\gamma)} Q^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}})\eta_{\hat{t}}x_{\hat{t}}^\top. \tag{15}$$

Proof Following (11), we have that

$$\begin{aligned}
\widehat{\nabla \mathcal{C}}(K) =& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\nabla_K \log \pi_K(u_{\hat{t}}|x_{\hat{t}}) \\
\overset{(i)}{=}& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\nabla_K \left(-\frac{(u_{\hat{t}} + Kx_{\hat{t}})^\top (u_{\hat{t}} + Kx_{\hat{t}})}{2\sigma^2} \right) \\
=& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\nabla_K \left(-\frac{u_{\hat{t}}^\top u_{\hat{t}} + 2u_{\hat{t}}^\top Kx_{\hat{t}} + x_{\hat{t}}^\top K^\top Kx_{\hat{t}}}{2\sigma^2} \right) \\
=& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\nabla_K \left(-\frac{\operatorname{tr}\left(2x_{\hat{t}}u_{\hat{t}}^\top K\right) + \operatorname{tr}\left(x_{\hat{t}}x_{\hat{t}}^\top K^\top K\right)}{2\sigma^2} \right),
\end{aligned} \tag{16}$$

where (i) follows from (12). Now note that

$$\nabla_K \operatorname{tr}\left(2x_{\hat{t}}u_{\hat{t}}^\top K\right) = \nabla_K \operatorname{tr}\left(\left(2u_{\hat{t}}x_{\hat{t}}^\top\right)^\top K\right) = \nabla_K \left\langle 2u_{\hat{t}}x_{\hat{t}}^\top, K \right\rangle = 2u_{\hat{t}}x_{\hat{t}}^\top, \tag{17}$$

and

$$\begin{aligned}
\nabla_K \operatorname{tr}\left(x_{\hat{t}}x_{\hat{t}}^\top K^\top K\right) =& \nabla_{K_1} \operatorname{tr}\left(x_{\hat{t}}x_{\hat{t}}^\top K^\top K_1\right) + \nabla_{K_2} \operatorname{tr}\left(x_{\hat{t}}x_{\hat{t}}^\top K_2^\top K\right) \\
=& \nabla_{K_1} \operatorname{tr}\left(\left(Kx_{\hat{t}}x_{\hat{t}}^\top\right)^\top K_1\right) + \nabla_{K_2} \operatorname{tr}\left(K_2^\top \left(Kx_{\hat{t}}x_{\hat{t}}^\top\right)\right) \\
=& \nabla_{K_1} \left\langle Kx_{\hat{t}}x_{\hat{t}}^\top, K_1 \right\rangle + \nabla_{K_2} \left\langle Kx_{\hat{t}}x_{\hat{t}}^\top, K_2 \right\rangle \\
=& 2Kx_{\hat{t}}x_{\hat{t}}^\top.
\end{aligned} \tag{18}$$

As a result, combining (17) and (18) with (16) yields

$$\begin{aligned}
\widehat{\nabla \mathcal{C}}(K) =& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\left(-\frac{1}{2\sigma^2}\left(2(Kx_{\hat{t}}x_{\hat{t}}^\top + u_{\hat{t}}x_{\hat{t}}^\top)\right)\right) \\
=& \frac{1}{1-\gamma} Q^K(x_{\hat{t}}, u_{\hat{t}})\left(-\frac{(u_{\hat{t}} + Kx_{\hat{t}})}{\sigma^2}x_{\hat{t}}^\top\right) \\
\overset{(i)}{=}& -\frac{1}{\sigma(1-\gamma)} Q^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}})\eta_{\hat{t}}x_{\hat{t}}^\top,
\end{aligned}$$

where (i) follows from (14). This finishes the proof. ■

We now provide the following remark on the computation of $Q^K(x_{\hat{t}}, u_{\hat{t}})$.

Remark 8 *The Q-function in* (15) *represents the cost-to-go from time step $\hat{t}$. Using the quadratic stage cost $c_t := x_t^\top Q x_t + u_t^\top R u_t$, we can write*

$$Q^K(x_{\hat{t}}, u_{\hat{t}}) = \sum_{t=\hat{t}}^{\infty} \gamma^{t-\hat{t}} c_t,$$

where the dynamics follow (3) *with control*

$$u_t = \begin{cases} -Kx_t + \sigma\eta_t, & \text{if } t = \hat{t}, \\ -Kx_t, & \text{otherwise,} \end{cases}$$

and $x_0 \sim \mathcal{D}$. This is analogous to the zero-order oracle in Malik et al. (2020), which computes

$$\mathcal{C}(K; x_0) := \sum_{t=0}^{\infty} \gamma^t c_t \quad \text{with } u_t = -Kx_t.$$

Accordingly, we also assume access to an oracle that returns a single noisy evaluation of such costs under the given policy.

Taking the alternative formulation of our gradient estimate provided in Lemma 7 into consideration, we introduce the algorithm

---
Algorithm 1 LQR with Policy Gradient

1: Given iteration number $T \geq 1$, initial policy $K_0 \in \mathbb{R}^{m \times n}$, noise parameter σ, and step size $\alpha_t > 0$
2: **for** $t \in \{0, 1, \ldots, T-1\}$ **do**
3:      Sample $x_0 \sim \mathcal{D}$, $\hat{t} \sim \mu_\gamma$, and $\eta_{\hat{t}} \sim \mathcal{N}(0, I_m)$
4:      Simulate K_t for $\hat{t}$ steps starting from x_0 and observe $x_{\hat{t}}$.
5:      $u_{\hat{t}} \leftarrow -K_t x_{\hat{t}} + \sigma\eta_{\hat{t}}$
6:      $\widehat{\nabla\mathcal{C}}(K_t) \leftarrow -\frac{1}{\sigma(1-\gamma)}\eta_{\hat{t}} x_{\hat{t}}^\top Q^{K_t}(x_{\hat{t}}, u_{\hat{t}})$
7:      $K_{t+1} \leftarrow K_t - \alpha_t \widehat{\nabla\mathcal{C}}(K_t)$
    return K_T

Before we state the next result, note that one can compute

$$\nabla\mathcal{C}(K) = 2((R + \gamma B^\top P_K B)K - \gamma B^\top P_K A)\mathbb{E}_{x_0 \sim \mathcal{D}}\left[\sum_{t=0}^{\infty} \gamma^t x_t x_t^\top\right]; \qquad (19)$$

a proof can be found in Fazel et al. (2018) for the undiscounted case, where $\gamma = 1$, and in Malik et al. (2020) for the discounted case. The following proposition plays a key role in our constructions.

Proposition 9 *Suppose $u_{\hat{t}} \sim \mathcal{N}(-Kx_{\hat{t}}, \sigma^2 I_m)$ as before. Then for any given K,*

$$\mathbb{E}[\widehat{\nabla\mathcal{C}}(K)] = \nabla\mathcal{C}(K).$$

Proof Following (15),

$$\mathbb{E}[\widehat{\nabla\mathcal{C}}(K)]$$

$$= \mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\mathbb{E}_{x_0\sim\mathcal{D}}\left[\mathbb{E}_{\eta_{\hat{t}}\sim\mathcal{N}(0,I_m)}\left[\widehat{\nabla\mathcal{C}}(K)\big|\hat{t},x_0\right]\Big|\hat{t}\right]\right]$$

$$\overset{(i)}{=} \mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\mathbb{E}_{x_0\sim\mathcal{D}}\left[-\frac{1}{\sigma^2(1-\gamma)}\mathbb{E}_{\eta_{\hat{t}}\sim\mathcal{N}(0,I_m)}\left[Q(x_{\hat{t}},-Kx_{\hat{t}}+\sigma\eta_{\hat{t}})(\sigma\eta_{\hat{t}})\big|\hat{t},x_0\right]x_{\hat{t}}^\top\Big|\hat{t}\right]\right]$$

$$\overset{(ii)}{=} \frac{1}{1-\gamma}\mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\mathbb{E}_{x_0\sim\mathcal{D}}\left[\mathbb{E}_{\eta_{\hat{t}}\sim\mathcal{N}(0,I_m)}\left[-\nabla_u Q^K(x_{\hat{t}},u)\Big|_{u=-Kx_{\hat{t}}+\sigma\eta_{\hat{t}}}\big|\hat{t},x_0\right]x_{\hat{t}}^\top\Big|\hat{t}\right]\right], \quad (20)$$

where (i) follows from $x_{\hat{t}}$ being determined when given x_0 and $\hat{t}$, and (ii) from Stein's lemma (Stein, 1981). Using (13), we compute

$$\nabla_u Q^K(x_{\hat{t}},u) = \nabla_u\left(x_{\hat{t}}^\top Q x_{\hat{t}} + u^\top R u + \gamma(Ax_{\hat{t}}+Bu)^\top P_K(Ax_{\hat{t}}+Bu)\right)$$

$$= 2Ru + 2\gamma B^\top P_K B u + 2\gamma B^\top P_K A x_{\hat{t}},$$

which evaluated at $u = -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}$ yields

$$\nabla_u Q^K(x_{\hat{t}},u)\Big|_{u=-Kx_{\hat{t}}+\sigma\eta_{\hat{t}}} = 2\left((R+\gamma B^\top P_K B)(-Kx_{\hat{t}}+\sigma\eta_{\hat{t}}) + \gamma B^\top P_K A x_{\hat{t}}\right).$$

Substituting in (20), we obtain

$$\mathbb{E}[\widehat{\nabla\mathcal{C}}(K)]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\mathbb{E}_{x_0\sim\mathcal{D}}\left[2\left((R+\gamma B^\top P_K B)K - \gamma B^\top P_K A\right)x_{\hat{t}}x_{\hat{t}}^\top\big|\hat{t}\right]\right]$$

$$= \frac{2}{1-\gamma}\mathbb{E}_{\hat{t}\sim\mu_\gamma}\left[\left((R+\gamma B^\top P_K B)K - \gamma B^\top P_K A\right)(A-BK)^{\hat{t}}\mathbb{E}_{x_0\sim\mathcal{D}}[x_0 x_0^\top]\left((A-BK)^{\hat{t}}\right)^\top\right]$$

$$= 2\left((R+\gamma B^\top P_K B)K - \gamma B^\top P_K A\right)\sum_{t=0}^\infty \gamma^t(A-BK)^t\mathbb{E}_{x_0\sim\mathcal{D}}[x_0 x_0^\top]\left((A-BK)^t\right)^\top$$

$$\overset{(i)}{=} 2\left((R+\gamma B^\top P_K B)K - \gamma B^\top P_K A\right)\mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^\infty \gamma^t(A-BK)^t x_0 x_0^\top\left((A-BK)^t\right)^\top\right]$$

$$\overset{(ii)}{=} 2\left((R+\gamma B^\top P_K B)K - \gamma B^\top P_K A\right)\mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^\infty \gamma^t x_t x_t^\top\right]$$

$$\overset{(iii)}{=} \nabla\mathcal{C}(K),$$

where (i) is done by employing the linearity of expectation along with replacing $\hat{t}$ by t as it is just a sum variable from that equation forward, (ii) is due to $x_t = (A-BK)^t x_0$, and (iii) follows from (19). ∎

Remark 10 (Extension beyond LQR) [2] *The construction in* (15) *is not automatically restricted to linear-quadratic control, but instead relies on the following assumption on the Q-values which is satisfied in the LQR setting. Suppose the action-value function satisfies*

$$Q^\mu(s, a) = a^\top H(s) a + b(s)^\top a + c(s) \tag{21}$$

with $H(s) = H(s)^\top \in \mathbb{R}^{m \times m}$. *Then* $\nabla_a Q^\mu(s, a) = 2H(s)a + b(s)$ *is affine in* a. *Let* $\eta \sim \mathcal{N}(0, I_m)$, *independent of* s, *and write* $a_\theta(s) = \mu_\theta(s)$. *For* $f(\eta) := Q^\mu(s, a_\theta(s) + \sigma\eta)$ *we have* $\nabla_\eta f(\eta) = \sigma \nabla_a Q^\mu(s, a_\theta(s) + \sigma\eta)$. *Stein's lemma (Stein, 1981) yields*

$$\mathbb{E}_\eta[\eta f(\eta)] = \mathbb{E}_\eta[\nabla_\eta f(\eta)] = \sigma \mathbb{E}_\eta[\nabla_a Q^\mu(s, a_\theta(s) + \sigma\eta)] = \sigma \nabla_a Q^\mu(s, a_\theta(s)),$$

where the last equality uses linearity of the integrand in a. *Hence*

$$\mathbb{E}_\eta[\sigma^{-1} Q^\mu(s, a_\theta(s) + \sigma\eta) \eta] = \nabla_a Q^\mu(s, a_\theta(s)).$$

Combining this with the deterministic policy gradient of Silver et al. (2014, Theorem 1),

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim \rho^\mu}[\nabla_\theta \mu_\theta(s)^\top \nabla_a Q^\mu(s, a)]_{a = a_\theta(s)},$$

gives the unbiased estimator

$$\widehat{\nabla J}(\theta) = \nabla_\theta \mu_\theta(s)^\top [\sigma^{-1} Q^\mu(s, a_\theta(s) + \sigma\eta) \eta], \qquad s \sim \rho^\mu, \ \eta \sim \mathcal{N}(0, I_m).$$

Linear actor: If $a_\theta(s) = \Theta s$, *then* $\nabla_\theta a_\theta(s) = I_m \otimes s^\top$, *so that*

$$\nabla_\theta a_\theta(s)^\top \nabla_a Q^\mu(s, a) = (I_m \otimes s) \nabla_a Q^\mu(s, a) = \text{vec}[\nabla_a Q^\mu(s, a) s^\top],$$

using the identity $(I_m \otimes s)u = \text{vec}(us^\top)$. *Unvectorising recovers the familiar matrix form* $\nabla_\Theta J(\Theta) = \mathbb{E}_{s \sim \rho^\mu}[\nabla_a Q^\mu(s, a) s^\top]$, *and the estimator in* (15) *follows by substituting the Stein-based replacement for* $\nabla_a Q^\mu$.

It may therefore be possible to extend the gradient estimators discussed here beyond the LQR setting by establishing that equation (21) *holds (perhaps approximately) for various classes of nonlinear systems.*

Before moving on to the next result, we define the undiscounted cost

$$\mathcal{C}_{\text{und}}(K) = \mathbb{E}_{x_0 \sim \mathcal{D}}\left[\sum_{t=0}^\infty (x_t^\top Q x_t + u_t^\top R u_t)\right],$$

subject to (3).

---

2. Notation in this remark follows Silver et al. (2014) rather than the LQR-specific symbols used elsewhere in the paper: $s \in \mathcal{S}$ is the state, $a \in \mathbb{R}^m$ the action, μ_θ the (deterministic) policy, ρ^μ the (improper) discounted state distribution, Q^μ the action–value function, $J(\theta) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t]$ the performance objective, and $\eta \sim \mathcal{N}(0, I_m)$ the Gaussian exploration noise.

Lemma 11 *Suppose K_0 is stable and suppose that*

$$\gamma \in \left(1 - \frac{\sigma_{\min}(Q)}{11 \mathcal{C}_{\mathrm{und}}(K_0)}, 1 \right).$$

Then

$$\sup_{K \in \mathcal{G}^{lqr}} \rho(A - BK) \leq \frac{1}{\sqrt{\gamma}} \sqrt{1 - \frac{\sigma_{\min}(Q)}{10 \mathcal{C}(K_0) + \mathcal{C}(K^*)}}; \tag{22}$$

in particular, the set $\mathcal{G}^{lqr}$ in (10) only contains stable policies.

This result shows that this assumption on γ ensures stability of the policies in the $\mathcal{G}^{lqr}$ set. When γ is small, the cost becomes heavily concentrated on early time steps and places less emphasis on the asymptotic behavior, which can lead to the optimal policy being unstable (Postoyan et al., 2017, Example 1). The assumption on γ serves to exclude such degenerate behavior by making instability more costly. Moreover, this condition on γ is tied to the particular definition of $\mathcal{G}^{lqr}$, and can be relaxed by tightening the required upper bound on the optimality gap in its definition—provided the resulting set still allows the analysis to achieve a sufficiently high confidence level. A more detailed discussion is given in Remark 21 in Appendix A.

Before we provide the proof, we point out that the condition on stability of K_0 readily implies that $\mathcal{C}_{\mathrm{und}}(K_0)$ is finite.

Proof Suppose $\tilde{K}$ satifies $\rho(A - B\tilde{K}) \geq 1$. Then we have

$$\begin{aligned}
\mathcal{C}(\tilde{K}) &= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t) \right] \\
&\geq \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \mathbb{E} \| (A - B\tilde{K})^t x_0 \|^2 \\
&= \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \mathbb{E}[\mathrm{tr}(((A - B\tilde{K})^t)^\top (A - B\tilde{K})^t x_0 x_0^\top)] \\
&\overset{(i)}{=} \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \| (A - B\tilde{K})^t \|_F^2 \\
&\geq \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \rho((A - B\tilde{K})^t)^2 \\
&\overset{(ii)}{\geq} \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \\
&= \frac{\sigma_{\min}(Q)}{1 - \gamma}, \tag{23}
\end{aligned}$$

where (i) comes from the linearity of expectation along with the assumption on the noise from (4), and (ii) follows from the instability of $\tilde{K}$ and that $\rho(A^t) = (\rho(A))^t$ which holds for any square matrix A. Now as a result of this, if we also show $\sup_{K \in \mathcal{G}^{lqr}} \mathcal{C}(K) < \frac{\sigma_{\min}(Q)}{1-\gamma}$,

14

we have proved stability of every K in the set $\mathcal{G}^{\text{lqr}}$. We do so as follows:

$$\frac{\sigma_{\min}(Q)}{1-\gamma} \overset{(i)}{>} 11\mathcal{C}_{\text{und}}(K_0) \overset{(ii)}{\geq} 11\mathcal{C}(K_0) \geq 10\mathcal{C}(K_0) + \mathcal{C}(K^*) \overset{(iii)}{\geq} \sup_{K \in \mathcal{G}^{\text{lqr}}} \mathcal{C}(K),$$

where (i) comes from the assumption on γ, (ii) from the fact that for a given policy, the undiscounted cost is not less than the discounted cost, and (iii) from the definition of the set $\mathcal{G}^{\text{lqr}}$ from (10). This proves the second claim.

For the first part, since for any $K \in \mathcal{G}^{\text{lqr}}$ we have that $\rho(A - BK) < 1$, we conclude that

$$\mathcal{C}(K) = \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t) \right]$$

$$\overset{(i)}{\geq} \sum_{t=0}^{\infty} \gamma^t \sigma_{\min}(Q) \rho((A - BK)^t)^2$$

$$= \sigma_{\min}(Q) \sum_{t=0}^{\infty} (\gamma(\rho(A - BK))^2)^t$$

$$\overset{(ii)}{=} \frac{\sigma_{\min}(Q)}{1 - \gamma(\rho(A - BK))^2},$$

where (i) is done the same way as (23) and (ii) follows from $\gamma(\rho(A - BK))^2 < 1$ for $K \in \mathcal{G}^{\text{lqr}}$. As a result, for $K \in \mathcal{G}^{\text{lqr}}$, we have that

$$1 - \gamma(\rho(A - BK))^2 \geq \frac{\sigma_{\min}(Q)}{\mathcal{C}(K)} \Rightarrow$$

$$\gamma(\rho(A - BK))^2 \leq 1 - \frac{\sigma_{\min}(Q)}{\mathcal{C}(K)} \Rightarrow$$

$$\rho(A - BK) \leq \frac{1}{\sqrt{\gamma}} \sqrt{1 - \frac{\sigma_{\min}(Q)}{\mathcal{C}(K)}},$$

which after taking a supremum gives

$$\sup_{K \in \mathcal{G}^{\text{lqr}}} \rho(A - BK) \leq \frac{1}{\sqrt{\gamma}} \sup_{K \in \mathcal{G}^{\text{lqr}}} \sqrt{1 - \frac{\sigma_{\min}(Q)}{\mathcal{C}(K)}} = \frac{1}{\sqrt{\gamma}} \sqrt{1 - \frac{\sigma_{\min}(Q)}{10\mathcal{C}(K_0) + \mathcal{C}(K^*)}},$$

concluding the proof. ■

We next introduce a high probability upper bound on our gradient estimate on any $K \in \mathcal{G}^{\text{lqr}}$.

Lemma 12 *Suppose $\delta \in (0, \frac{1}{e}]$, and γ is chosen as in Lemma 11. Then for any $K \in \mathcal{G}^{lqr}$, we have that*

$$\|\widehat{\nabla \mathcal{C}}(K)\|_F \leq \frac{\xi_3}{1 - \gamma} \left(\log \frac{1}{\delta} \right)^{3/2}$$

with probability at least $1 - \delta$, where $\xi_1, \xi_2, \xi_3 \in \mathbb{R}$ are given by

$$\xi_1 := \left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\gamma(10\mathcal{C}(K_0) + \mathcal{C}(K^*)) \right) e^3 n^3 \bar{\mathcal{K}}^3 C_m^{3/2} \tag{24}$$

$$\xi_2 := \left(2\|R\| + 2\gamma\|B\|^2(10\mathcal{C}(K_0) + \mathcal{C}(K^*)) \right) en\bar{\mathcal{K}} C_m^{1/2} \tag{25}$$

$$\xi_3 := \frac{1}{\sigma} \left(\xi_1 5^{1/2} m^{1/2} \right) + \sigma \left(\xi_2 5^{3/2} m^{3/2} \right), \tag{26}$$

where $\bar{\mathcal{K}}$ is a positive constant. Moreover,

$$\mathbb{E}\|\widehat{\nabla\mathcal{C}}(K)\|_F^2 \leq \frac{\xi_4}{(1-\gamma)^2}, \tag{27}$$

where

$$\xi_4 := \frac{1}{\sigma^2}\xi_1^2 m + 2\xi_1\xi_2 m(m+2) + \sigma^2\xi_2^2 m(m+2)(m+4). \tag{28}$$

Proof Using the formulation of $\widehat{\nabla\mathcal{C}}(K)$ derived in (15), we have

$$\|\widehat{\nabla\mathcal{C}}(K)\|_F = \left\|\frac{1}{\sigma(1-\gamma)}\eta_{\hat{t}}x_{\hat{t}}^\top Q^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}})\right\|_F$$
$$\leq \frac{1}{\sigma(1-\gamma)}\|\eta_{\hat{t}}\|\|x_{\hat{t}}\|Q^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}). \tag{29}$$

First, note that

$$\|x_{\hat{t}}\| = \|(A-BK)^{\hat{t}}x_0\| \leq \|(A-BK)^{\hat{t}}\|\|x_0\| \overset{(i)}{\leq} \sup_{t\geq 0}\|(A-BK)^t\|C_m^{1/2}, \tag{30}$$

where (i) follows from the assumption on the initial state noise mentioned in (4).

Sublemma 3.1 *We have that*

$$\sup_{K\in\mathcal{G}^{lqr}}\sup_{t\geq 0}\|(A-BK)^t\| \tag{31}$$

is finite.

Proof of Sublemma 3.1. We start by arguing that $\mathcal{G}^{\mathrm{lqr}}$ is a compact set. First, note that since $\|K\| \leq \widetilde{c_{K_1}}$ (see Remark 5) for any $K \in \mathcal{G}^{\mathrm{lqr}}$, the set $\mathcal{G}^{\mathrm{lqr}}$ is bounded. Secondly, since $\mathcal{C}(K)$ is locally Lipschitz in $\mathcal{G}^{\mathrm{lqr}}$, it is also continuous, and hence, by the definition of $\mathcal{G}^{\mathrm{lqr}}$ in (10), we have that $\mathcal{G}^{\mathrm{lqr}}$ is the pre-image of the closed interval $[0, 10\mathcal{C}(K_0) + \mathcal{C}(K^*)]$ under a continuous map $\mathcal{C} : \mathcal{G}^{\mathrm{lqr}} \to \mathbb{R}$, implying $\mathcal{G}^{\mathrm{lqr}}$ is closed as well. As a result of this, we have that $\mathcal{G}^{\mathrm{lqr}}$ is compact. Now we move on to show why (31) is finite.

First, let us define

$$S(x_0; K) := \sum_{t=0}^{\infty}\|x_t\|^2,$$

where $x_{t+1} = (A - BK)x_t$. Moreover, we let

$$
\begin{aligned}
S(K) &:= \mathbb{E}_{x_0 \sim \mathcal{D}} S(x_0; K) \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \|x_t\|^2 \right] \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \|(A - BK)^t x_0\|^2 \right] \\
&= \sum_{t=0}^{\infty} \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\operatorname{tr} \left(\left((A - BK)^t \right)^\top (A - BK)^t x_0 x_0^\top \right) \right] \\
&= \sum_{t=0}^{\infty} \|(A - BK)^t\|_F^2 \\
&\geq \sum_{t=0}^{\infty} \|(A - BK)^t\|^2,
\end{aligned}
$$

which after taking the square root of both sides gives

$$
\begin{aligned}
\sqrt{S(K)} &\geq \sqrt{\sum_{t=0}^{\infty} \|(A - BK)^t\|^2} \\
&\geq \sup_{t \geq 0} \|(A - BK)^t\|.
\end{aligned}
$$

As a result, we have that

$$
\sup_{t \geq 0} \|(A - BK)^t\| \leq \sqrt{S(K)},
$$

which after taking a supremum over $\mathcal{G}^{\text{lqr}}$ yields

$$
\sup_{K \in \mathcal{G}^{\text{lqr}}} \sup_{t \geq 0} \|(A - BK)^t\| \leq \sup_{K \in \mathcal{G}^{\text{lqr}}} \sqrt{S(K)}.
$$

Now it suffices to show $\sup_{K \in \mathcal{G}^{\text{lqr}}} \sqrt{S(K)}$ is finite, which we prove by contradiction. Suppose that this is not the case. Therefore, there exists a sequence $\{K_j\}_{j=1}^{\infty}$ such that $\sqrt{S(K_j)} \xrightarrow{j \to \infty} \infty$. By compactness, we can pick a convergent subsequence whose limit we denote by $\bar{K}$. We will abuse notation and henceforth use K_j to refer to the subsequence; observe that K_j should also satisfy $\sqrt{S(K_j)} \xrightarrow{j \to \infty} \infty$.

Now since $\bar{K} \in \mathcal{G}^{\text{lqr}}$, we have from Lemma 11 that $A - B\bar{K}$ is strictly stable, and thus, there exists a Lyapunov function $V(x) = x^\top \bar{P} x$ where $\bar{P}$ is a positive definite matrix that satisfies

$$
(A - B\bar{K})^\top \bar{P}(A - B\bar{K}) - \bar{P} = -I.
$$

Therefore, for j large enough,

$$
(A - BK_j)^\top \bar{P}(A - BK_j) - \bar{P} \preceq -(1/2)I.
$$

Then

$$V((A - BK_j)x) - V(x) = x^T(A - BK_j)\bar{P}(A - BK_j)x - x^T\bar{P}x$$
$$\leq -(1/2)\|x\|^2$$
$$= -\frac{1}{2\lambda_{\max}(\bar{P})}\left(\lambda_{\max}(\bar{P})\|x\|^2\right)$$
$$\overset{(i)}{\leq} -\frac{1}{2\lambda_{\max}(\bar{P})}V(x),$$

where (i) is due to the fact that $V(x) \leq \lambda_{\max}(\bar{P})\|x\|^2$. Thus,

$$V((A - BK_j)x) \leq \left(1 - \frac{1}{2\lambda_{\max}(\bar{P})}\right)V(x). \tag{32}$$

As a result, we have that

$$S(x_0; K_j) = \sum_{t=0}^{\infty} \|x_t\|^2$$
$$\overset{(i)}{\leq} \frac{1}{\lambda_{\min}(\bar{P})} \sum_{t=0}^{\infty} V((A - BK_j)^t x_0)$$
$$\overset{(ii)}{\leq} \frac{1}{\lambda_{\min}(\bar{P})} \sum_{t=0}^{\infty} \left(1 - \frac{1}{2\lambda_{\max}(\bar{P})}\right)^t V(x_0)$$
$$\leq \frac{2\lambda_{\max}(\bar{P})}{\lambda_{\min}(\bar{P})}V(x_0)$$
$$\leq \frac{2\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}\|x_0\|^2,$$

where (i) follows from $V(x) \geq \lambda_{\min}(\bar{P})\|x\|^2$ and (ii) from (32). Now taking an expectation over $x_0 \sim \mathcal{D}$ yields

$$S(K_j) \leq \frac{2\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}\mathbb{E}_{x_0\sim\mathcal{D}}\|x_0\|^2$$
$$= \frac{2\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}\mathbb{E}_{x_0\sim\mathcal{D}}\operatorname{tr}(x_0 x_0^\top)$$
$$= \frac{2\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}\operatorname{tr}\left(\mathbb{E}_{x_0\sim\mathcal{D}}[x_0 x_0^\top]\right)$$
$$= \frac{2\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}\operatorname{tr}(I_n)$$
$$= \frac{2n\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})},$$

and hence,

$$\sqrt{S(K_j)} \leq \sqrt{\frac{2n\lambda_{\max}^2(\bar{P})}{\lambda_{\min}(\bar{P})}},$$

18

which is finite, resulting in a contradiction, concluding the proof of Sublemma 3.1. ◇

We now continue with the proof of Lemma 12. Let us first make a remark. By the Kreiss matrix theorem (Leveque and Trefethen, 1984; Spijker, 1991), we have that

$$\mathcal{K}(A - BK) \leq \sup_{t \geq 0} \|(A - BK)^t\| \leq e \, n \, \mathcal{K}(A - BK). \tag{33}$$

Consequently, we can define the following constant

$$\bar{\mathcal{K}} := \sup_{K \in \mathcal{G}^{\mathrm{lqr}}} \mathcal{K}(A - BK), \tag{34}$$

which is finite as a result of (33) and Sublemma 3.1. Combining (33) and (34) with (30) gives

$$\|x_{\hat{t}}\| \leq e \, n \, C_m^{1/2} \, \mathcal{K}(A - BK) \leq e \, n \, C_m^{1/2} \, \bar{\mathcal{K}}, \tag{35}$$

for any $\hat{t} \geq 0$. Moreover,

$$
\begin{aligned}
Q^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}) =& x_{\hat{t}}^\top Q x_{\hat{t}} + (-Kx_{\hat{t}} + \sigma\eta_{\hat{t}})^\top R(-Kx_{\hat{t}} + \sigma\eta_{\hat{t}}) \\
& + \gamma((A - BK)x_{\hat{t}} + \sigma B\eta_{\hat{t}})^\top P_K((A - BK)x_{\hat{t}} + \sigma B\eta_{\hat{t}}) \\
\overset{(i)}{\leq}& \|Q\|e^2 n^2 \bar{\mathcal{K}}^2 C_m + \|R\|\|-Kx_{\hat{t}} + \sigma\eta_{\hat{t}}\|^2 + \gamma\|P_K\|\|x_{\hat{t}+1} + \sigma B\eta_{\hat{t}}\|^2 \\
\overset{(ii)}{\leq}& \|Q\|e^2 n^2 \bar{\mathcal{K}}^2 C_m + 2\|R\| \left(\|K\|^2\|x_{\hat{t}}\|^2 + \sigma^2\|\eta_{\hat{t}}\|^2\right) \\
& + 2\gamma\mathcal{C}(K) \left(\|x_{\hat{t}+1}\|^2 + \sigma^2\|B\|^2\|\eta_{\hat{t}}\|^2\right) \\
\overset{(iii)}{\leq}& \left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\gamma(10\mathcal{C}(K_0) + \mathcal{C}(K^*))\right) e^2 n^2 \bar{\mathcal{K}}^2 C_m \\
& + \left(2\sigma^2\|R\| + 2\gamma\sigma^2\|B\|^2(10\mathcal{C}(K_0) + \mathcal{C}(K^*))\right) \|\eta_{\hat{t}}\|^2, \tag{36}
\end{aligned}
$$

where (i) follows from (35), (ii) from $\|P_K\| \leq \mathrm{tr}(P_K)$ along with $\mathrm{tr}(P_K) = \mathcal{C}(K)$ as shown in (6), and (iii) from the fact that $\|K\| \leq \widetilde{c_{K_1}}$ for any $K \in \mathcal{G}^{\mathrm{lqr}}$ (see Remark 5) along with reapplying (35) and using the upper bound obtained on $\mathcal{C}(K)$ by the definition of the set $\mathcal{G}^{\mathrm{lqr}}$. Now applying the derived bounds (35) and (36) on (29), we conclude that

$$
\begin{aligned}
\|\widehat{\nabla\mathcal{C}}(K)\|_F \leq& \frac{\left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\gamma(10\mathcal{C}(K_0) + \mathcal{C}(K^*))\right) e^3 n^3 \bar{\mathcal{K}}^3 C_m^{3/2}}{\sigma(1 - \gamma)} \|\eta_{\hat{t}}\| \\
& + \frac{\sigma^2 \left(2\|R\| + 2\gamma\|B\|^2(10\mathcal{C}(K_0) + \mathcal{C}(K^*))\right) en\bar{\mathcal{K}}C_m^{1/2}}{\sigma(1 - \gamma)} \|\eta_{\hat{t}}\|^3 \\
=& \frac{1}{1 - \gamma} \left(\frac{1}{\sigma}\xi_1\|\eta_{\hat{t}}\| + \sigma\xi_2\|\eta_{\hat{t}}\|^3\right). \tag{37}
\end{aligned}
$$

Furthermore, since $\eta_{\hat{t}} \sim \mathcal{N}(0, I_m)$ for any $\hat{t}$, $\|\eta_{\hat{t}}\|^2$ is distributed according to the chi-squared distribution with m degrees of freedom ($\|\eta_{\hat{t}}\|^2 \sim \chi^2(m)$ for any $\hat{t}$). Therefore, the standard Laurent and Massart (2000) bounds suggest that for arbitrary $y > 0$, we have that

$$\mathbb{P}\{\|\eta_{\hat{t}}\|^2 \geq m + 2\sqrt{my} + 2y\} \leq e^{-y}.$$

19

Now since by our assumption $0 < \delta \leq 1/e$, it holds that $y = m \log \frac{1}{\delta} \geq m$ and thus

$$\mathbb{P}\{\|\eta_{\hat{t}}\|^2 \geq 5y\} \leq \mathbb{P}\{\|\eta_{\hat{t}}\|^2 \geq m + 2\sqrt{my} + 2y\} \leq e^{-y},$$

which after substituting y with its value $m \log \frac{1}{\delta}$ gives

$$\mathbb{P}\{\|\eta_{\hat{t}}\|^2 \geq 5m \log \frac{1}{\delta}\} \leq e^{-m \log \frac{1}{\delta}} = \delta^m \leq \delta.$$

As a result, we have $\|\eta_{\hat{t}}\| \leq 5^{1/2} m^{1/2} (\log \frac{1}{\delta})^{1/2}$ and consequently

$$\|\eta_{\hat{t}}\|^3 \leq 5^{3/2} m^{3/2} (\log \frac{1}{\delta})^{3/2}$$

with probability at least $1 - \delta$, which after applying on (37) yields

$$
\begin{aligned}
\|\widehat{\nabla \mathcal{C}}(K)\|_F &\leq \frac{1}{1-\gamma} \left(\frac{1}{\sigma} \xi_1 5^{1/2} m^{1/2} \left(\log \frac{1}{\delta} \right)^{1/2} + \sigma \xi_2 5^{3/2} m^{3/2} \left(\log \frac{1}{\delta} \right)^{3/2} \right) \\
&\leq \frac{1}{1-\gamma} \left(\frac{1}{\sigma} \xi_1 5^{1/2} m^{1/2} + \sigma \xi_2 5^{3/2} m^{3/2} \right) \left(\log \frac{1}{\delta} \right)^{3/2} \\
&= \frac{\xi_3}{1-\gamma} \left(\log \frac{1}{\delta} \right)^{3/2},
\end{aligned}
$$

proving the first claim.

As for the second claim, note that using (37), we have

$$\|\widehat{\nabla \mathcal{C}}(K)\|_F^2 \leq \frac{1}{(1-\gamma)^2} \left(\frac{1}{\sigma^2} \xi_1^2 \|\eta_{\hat{t}}\|^2 + 2\xi_1\xi_2 \|\eta_{\hat{t}}\|^4 + \sigma^2 \xi_2^2 \|\eta_{\hat{t}}\|^6 \right). \tag{38}$$

Now since $\|\eta_{\hat{t}}\| \sim \chi(m)$ whose moments are known, taking an expectation on both sides of (38) results in

$$
\begin{aligned}
\mathbb{E}\|\widehat{\nabla \mathcal{C}}(K)\|_F^2 &\leq \frac{1}{(1-\gamma)^2} \left(\frac{1}{\sigma^2} \xi_1^2 \mathbb{E}\|\eta_{\hat{t}}\|^2 + 2\xi_1\xi_2 \mathbb{E}\|\eta_{\hat{t}}\|^4 + \sigma^2 \xi_2^2 \mathbb{E}\|\eta_{\hat{t}}\|^6 \right) \\
&= \frac{1}{(1-\gamma)^2} \left(\frac{1}{\sigma^2} \xi_1^2 m + 2\xi_1\xi_2 m(m+2) + \sigma^2 \xi_2^2 m(m+2)(m+4) \right) \\
&= \frac{\xi_4}{(1-\gamma)^2},
\end{aligned}
$$

concluding the proof. ∎

Following Lemma 12, we now define the following event for each iteration t of Algorithm 1:

$$\mathcal{A}_t = \left\{ \|\widehat{\nabla \mathcal{C}}(K_t)\|_F \leq \frac{\xi_3}{1-\gamma} \left(\log \frac{1}{\delta} \right)^{3/2} \right\}. \tag{39}$$

Having this, we introduce the following lemma:

Lemma 13 *Suppose $\delta \in (0, e^{-3/2}]$, and γ is chosen as in Lemma 11. Then for any given $K_t \in \mathcal{G}^{lqr}$, we have that*

$$\|\mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t}] - \nabla \mathcal{C}(K_t)\|_F \le \frac{3\xi_3}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}.$$

Proof Following Proposition 9, we have that

$$\begin{aligned}
\nabla \mathcal{C}(K_t) &= \mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)] \\
&= \mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t}] + \mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t^c}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t}] - \nabla \mathcal{C}(K_t)\|_F \\
&= \|\mathbb{E}[\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t^c}]\|_F \\
&\overset{(i)}{\le} \mathbb{E}\left[\|\widehat{\nabla \mathcal{C}}(K_t)1_{\mathcal{A}_t^c}\|_F\right] \\
&= \mathbb{E}\left[\|\widehat{\nabla \mathcal{C}}(K_t)\|_F 1_{\mathcal{A}_t^c}\right] \\
&\overset{(ii)}{\le} \mathbb{E}\left[\|\widehat{\nabla \mathcal{C}}(K_t)\|_F 1_{\left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right\}}\right] \\
&= \mathbb{P}\left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right\} \mathbb{E}\left[\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \,\Big|\, \|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right], \quad (40)
\end{aligned}$$

where (i) follows from Jensen's inequality and (ii) from the fact that

$$\mathcal{A}_t^c = \left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F > \frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}\right\} \subseteq \left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}\right\}.$$

Moreover, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \,\Big|\, \|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right] \\
&= \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma} + \frac{\int_{\frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}}^{\infty} \mathbb{P}\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge z\}\, dz}{\mathbb{P}\left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right\}}. \quad (41)
\end{aligned}$$

Now recall from Lemma 12 that

$$\mathbb{P}\left\{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F \ge \frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}\right\} \le \delta \quad (42)$$

for arbitrary δ, which implies

$$\mathbb{P}\left\{\|\widehat{\nabla\mathcal{C}}(K_t)\|_F \geq z\right\} \leq e^{-\left(\frac{z(1-\gamma)}{\xi_3}\right)^{2/3}}. \tag{43}$$

Now combining (43), (41), and (40) yields

$$\|\mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}] - \nabla\mathcal{C}(K_t)\|_F$$

$$\leq \mathbb{P}\left\{\|\widehat{\nabla\mathcal{C}}(K_t)\|_F \geq \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma}\right\} \frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{1-\gamma} + \int_{\frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}}^{\infty} e^{-\left(\frac{z(1-\gamma)}{\xi_3}\right)^{2/3}} dz$$

$$\overset{(i)}{\leq} \frac{\xi_3}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2} + \frac{\xi_3}{1-\gamma}\int_{\left(\log\frac{1}{\delta}\right)^{3/2}}^{\infty} e^{-u^{2/3}} du$$

$$= \frac{\xi_3}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2} + \frac{\xi_3}{1-\gamma}\left(\frac{3}{2}\delta\left(\log\frac{1}{\delta}\right)^{1/2} + \frac{3}{4}\sqrt{\pi}\,\mathrm{erfc}\left(\sqrt{\log\frac{1}{\delta}}\right)\right)$$

$$\overset{(ii)}{\leq} \frac{\xi_3}{1-\gamma}\left(\delta\left(\log\frac{1}{\delta}\right)^{3/2} + \frac{3}{2}\delta\left(\log\frac{1}{\delta}\right)^{1/2} + \frac{3}{4}\sqrt{\pi}\delta\right)$$

$$\overset{(iii)}{\leq} \frac{3\xi_3}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2},$$

where (i) follows from (42) along with a change of variables $u = \left(\frac{1-\gamma}{\xi_3}\right)z$ in the integral, (ii) from the fact that $\mathrm{erfc}\left(\sqrt{\log\frac{1}{\delta}}\right) \leq \delta$, and (iii) from $\delta \leq e^{-3/2}$. This concludes the proof. ∎

Before introducing the next lemma, let us denote the optimality gap of iterate t of the algorithm by

$$\Delta_t := \mathcal{C}(K_t) - \mathcal{C}(K^*). \tag{44}$$

Moreover, let $\mathcal{F}_t$ denote the σ-algebra containing the randomness up to iteration t of Algorithm 1 (including K_t but not $\widehat{\nabla\mathcal{C}}(K_t)$). We then define

$$\tau_1 := \min\left\{t \mid \Delta_t > 10\mathcal{C}(K_0)\right\}, \tag{45}$$

which is a stopping time with respect to $\mathcal{F}_t$.

Lemma 14 *Suppose $\delta \in (0, e^{-3/2}]$, γ is as suggested in Lemma 11, and the update rule follows*

$$K_{t+1} = K_t - \alpha_t\widehat{\nabla\mathcal{C}}(K_t)$$

with a step-size α_t such that for all $t \in \{0, 1, 2, \dots\}$,

$$\alpha_t \leq \frac{\omega_{lqr}}{\frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}}.$$

22

Then for any $t \in \{0, 1, 2, \dots\}$, we have

$$\mathbb{E}[\Delta_{t+1}1_{\mathcal{A}_t}|\mathcal{F}_t]1_{\tau_1>t} \leq \left((1 - \mu_{lqr}\alpha_t)\Delta_t + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{lqr}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2} \right)1_{\tau_1>t},$$

where Δ_t and $\mathcal{A}_t$ are defined in (44) and (39) respectively.

Proof First, note that by the definition of τ_1 in (45), $\tau_1 > t$ implies $K_t \in \mathcal{G}^{lqr}$. In addition, since $\alpha_t \leq \frac{\omega_{lqr}}{\frac{\xi_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}}$, the event $\mathcal{A}_t$ implies that

$$\|K_{t+1} - K_t\|_F = \|\alpha_t\widehat{\nabla\mathcal{C}}(K_t)\|_F \leq \omega_{lqr}.$$

Thus, by local smoothness of $\mathcal{C}(K_t)$, see Lemma 2, it holds that

$$(\Delta_{t+1} - \Delta_t)1_{\tau_1>t}1_{\mathcal{A}_t} = (\mathcal{C}(K_{t+1} - \mathcal{C}(K_t))1_{\tau_1>t}1_{\mathcal{A}_t}$$
$$\leq \left(\langle\nabla\mathcal{C}(K_t), K_{t+1} - K_t\rangle + \frac{\phi_{lqr}}{2}\|K_{t+1} - K_t\|_F^2 \right)1_{\tau_1>t}1_{\mathcal{A}_t}$$
$$= \left(-\alpha_t\left\langle\nabla\mathcal{C}(K_t), \widehat{\nabla\mathcal{C}}(K_t)\right\rangle + \frac{\phi_{lqr}\alpha_t^2}{2}\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2 \right)1_{\tau_1>t}1_{\mathcal{A}_t},$$

which after taking an expectation conditioned on $\mathcal{F}_t$ gives

$$\mathbb{E}[\Delta_{t+1}1_{\tau_1>t}1_{\mathcal{A}_t}|\mathcal{F}_t] - \mathbb{E}[\Delta_t 1_{\tau_1>t}1_{\mathcal{A}_t}|\mathcal{F}_t]$$
$$\leq -\alpha_t\left\langle\nabla\mathcal{C}(K_t), \mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\tau_1>t}1_{\mathcal{A}_t}|\mathcal{F}_t]\right\rangle + \frac{\phi_{lqr}}{2}\alpha_t^2\mathbb{E}[\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2 1_{\tau_1>t}1_{\mathcal{A}_t}|\mathcal{F}_t].$$

Since Δ_t and $1_{\tau_1>t}$ are determined by $\mathcal{F}_t$,

$$\mathbb{E}[\Delta_{t+1}1_{\mathcal{A}_t}|\mathcal{F}_t]1_{\tau_1>t}$$
$$\leq \left(\Delta_t\mathbb{E}[1_{\mathcal{A}_t}|\mathcal{F}_t] - \alpha_t\left\langle\nabla\mathcal{C}(K_t), \mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}|\mathcal{F}_t]\right\rangle + \frac{\phi_{lqr}}{2}\alpha_t^2\mathbb{E}[\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2 1_{\mathcal{A}_t}|\mathcal{F}_t] \right)1_{\tau_1>t}$$
$$\overset{(i)}{\leq} \left(\Delta_t - \alpha_t\left\langle\nabla\mathcal{C}(K_t), \mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}|\mathcal{F}_t]\right\rangle + \frac{\phi_{lqr}}{2}\alpha_t^2\mathbb{E}[\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2|\mathcal{F}_t] \right)1_{\tau_1>t}$$
$$= \Delta_t1_{\tau_1>t} - \alpha_t\left\langle\nabla\mathcal{C}(K_t), \nabla\mathcal{C}(K_t) + \mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}|\mathcal{F}_t] - \nabla\mathcal{C}(K_t)\right\rangle1_{\tau_1>t}$$
$$\quad + \frac{\phi_{lqr}}{2}\alpha_t^2\mathbb{E}[\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2|\mathcal{F}_t]1_{\tau_1>t}$$
$$= \Delta_t1_{\tau_1>t} - \alpha_t\langle\nabla\mathcal{C}(K_t), \nabla\mathcal{C}(K_t)\rangle1_{\tau_1>t}$$
$$\quad - \alpha_t\left\langle\nabla\mathcal{C}(K_t), \mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}|\mathcal{F}_t] - \nabla\mathcal{C}(K_t)\right\rangle1_{\tau_1>t} + \frac{\phi_{lqr}}{2}\alpha_t^2\mathbb{E}[\|\widehat{\nabla\mathcal{C}}(K_t)\|_F^2|\mathcal{F}_t]1_{\tau_1>t}$$
$$\overset{(ii)}{\leq} \Delta_t1_{\tau_1>t} - \alpha_t\|\nabla\mathcal{C}(K_t)\|_F^2 1_{\tau_1>t}$$
$$\quad + \alpha_t\|\nabla\mathcal{C}(K_t)\|_F\|\mathbb{E}[\widehat{\nabla\mathcal{C}}(K_t)1_{\mathcal{A}_t}|\mathcal{F}_t] - \nabla\mathcal{C}(K_t)\|_F1_{\tau_1>t} + \frac{\phi_{lqr}}{2}\alpha_t^2\frac{\xi_4}{(1-\gamma)^2}1_{\tau_1>t}$$
$$\overset{(iii)}{\leq} \Delta_t1_{\tau_1>t} - \alpha_t\mu_{lqr}\Delta_t1_{\tau_1>t} + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t1_{\tau_1>t} + \frac{\phi_{lqr}}{2}\alpha_t^2\frac{\xi_4}{(1-\gamma)^2}1_{\tau_1>t}$$
$$= \left((1 - \mu_{lqr}\alpha_t)\Delta_t + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{lqr}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2} \right)1_{\tau_1>t},$$

where (i) follows from $1_{\mathcal{A}_t} \le 1$, (ii) from Lemma 12, and (iii) from applying the PL inequality (9), the fact that $\|\nabla \mathcal{C}(K_t)\|_F \le \widetilde{c_{K_1}}$ for any $K_t \in \mathcal{G}^{\text{lqr}}$ (see Remark 5), and Lemma 13. This finishes the proof of Lemma 14. ∎

We are now in a position to state a precise version of our main result.

Theorem 15 *Suppose K_0 is stable and γ is as suggested in Lemma 11. If the step-size α_t is chosen as*

$$\alpha_t = \frac{2}{\mu_{lqr}} \frac{1}{t+N} \quad for \quad N = \max\left\{N_1, \frac{2}{\mu_{lqr}} \frac{\xi_3 \left(\log \frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\omega_{lqr}}\right\}, \tag{46}$$

where

$$N_1 = \max\left\{2, \frac{4\phi_{lqr}\xi_4}{\mu_{lqr}^2(1-\gamma)^2} \frac{2}{\mathcal{C}(K_0)}\right\}, \tag{47}$$

then for a given error tolerance ε such that $\mathcal{C}(K_0) \ge \frac{\varepsilon}{20}$, and δ chosen arbitrarily to satisfy

$$\delta \le \min\left\{2 \times 10^{-5}, \left(\frac{\phi_{lqr}\xi_4\omega_{lqr}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\right)^3 \varepsilon^3, \right.$$

$$\left.\left(\frac{\phi_{lqr}\xi_4}{480(1-\gamma)\mu_{lqr}\xi_3\widetilde{c_{K_1}}N_1\mathcal{C}(K_0)}\right)^3 \varepsilon^3, \left(\frac{\mu_{lqr}(1-\gamma)}{240\xi_3\widetilde{c_{K_1}}}\right)^3 \varepsilon^3\right\}, \tag{48}$$

the iterate K_T of Algorithm 1 after

$$T = \frac{40}{\varepsilon} N\mathcal{C}(K_0) \tag{49}$$

steps satisfies

$$\mathcal{C}(K_T) - \mathcal{C}(K^*) \le \varepsilon$$

with a probability of at least $4/5 - \delta T$.

It is essential to re-emphasize that, as also evident from the statement of Theorem 15, there is no reliance on an assumption that the policy remains stable throughout the algorithm; rather, the result is proven to hold with a certain probability. In particular, the instances of the algorithm that lead to instability at any iteration before T are factored into the failure probability $1/5 + \delta T$.

In Appendix A, we show that the success probability in Theorem 15 can be improved from $4/5 - \delta T$ to $1 - \delta T$ by averaging a batch of gradient estimates to reduce variance and obtain an estimate that is close to the true gradient with high probability. Moreover, in Appendix B, we show how our gradient estimation method and convergence analysis, developed for the random initialization setting, can be naturally extended to the noisy dynamics setting.

The proof of Theorem 15 relies on an intermediate result, namely Proposition 18, which we establish next. Before doing so, we provide some observations regarding the statement of the theorem. First, we have the following remark for δ:

Remark 16 (Selection of δ for the probability of failure) The δT term in the probability of failure stated in Theorem 15 can be adjusted arbitrarily; however, since T depends on N which depends on δ itself, we add some further discussion here. If we want the δT term to be less than some arbitrary small δ', it needs to hold that

$$\delta T = \delta \frac{40}{\varepsilon} \max \left\{ N_1 \mathcal{C}(K_0), \frac{2\xi_3 \mathcal{C}(K_0)}{\mu_{\mathrm{lqr}}\omega_{\mathrm{lqr}}(1-\gamma)} \left(\log \frac{1}{\delta}\right)^{3/2} \right\} \leq \delta'.$$

Therefore, δ first needs to satisfy

$$\frac{40}{\varepsilon} N_1 \mathcal{C}(K_0)\delta \leq \delta' \Rightarrow \delta \leq \frac{\delta'\varepsilon}{40 N_1 \mathcal{C}(K_0)}, \tag{50}$$

and secondly,

$$\frac{80\xi_3 \mathcal{C}(K_0)}{\mu_{\mathrm{lqr}}\omega_{\mathrm{lqr}}(1-\gamma)} \frac{1}{\varepsilon}\delta \left(\log \frac{1}{\delta}\right)^{3/2} \leq \delta' \Rightarrow \delta \left(\log \frac{1}{\delta}\right)^{3/2} \leq \frac{\mu_{\mathrm{lqr}}\omega_{\mathrm{lqr}}(1-\gamma)}{80\xi_3 \mathcal{C}(K_0)}\delta'\varepsilon. \tag{51}$$

Now since $a^3 \left(\log \frac{1}{a^3}\right)^{3/2} \leq a$ for any $a \in (0,1)$, for (51) to hold, it would suffice to have

$$\delta \leq \left(\frac{\mu_{\mathrm{lqr}}\omega_{\mathrm{lqr}}(1-\gamma)}{80\xi_3 \mathcal{C}(K_0)}\right)^3 (\delta'\varepsilon)^3. \tag{52}$$

Note that (52) is only a loose sufficient bound on δ that can be improved (for instance, the exponents in (52) can be reduced from 3 to 2 considering the other requirements on δ in (48)); however, since the dependence of T on δ is logarithmic, the looser requirement only adds a constant and does not change the order.

As a result, adding (50) and (52) to the existing requirements on δ in (48), we will have

$$\delta \leq \min \left\{ 2 \times 10^{-5}, \left(\frac{\phi_{\mathrm{lqr}}\xi_4\omega_{\mathrm{lqr}}}{960\xi_3^2 \widetilde{c_{K_1}} \mathcal{C}(K_0)}\right)^3 \varepsilon^3, \left(\frac{\phi_{\mathrm{lqr}}\xi_4}{480(1-\gamma)\mu_{\mathrm{lqr}}\xi_3 \widetilde{c_{K_1}} N_1 \mathcal{C}(K_0)}\right)^3 \varepsilon^3, \right.$$

$$\left. \left(\frac{\mu_{\mathrm{lqr}}(1-\gamma)}{240\xi_3 \widetilde{c_{K_1}}}\right)^3 \varepsilon^3, \frac{\delta'\varepsilon}{40 N_1 \mathcal{C}(K_0)}, \left(\frac{\mu_{\mathrm{lqr}}\omega_{\mathrm{lqr}}(1-\gamma)}{80\xi_3 \mathcal{C}(K_0)}\right)^3 (\delta'\varepsilon)^3 \right\},$$

which will lead to the result of Theorem 15 holding with probability $4/5 - \delta'$ after

$$T \sim \frac{N}{\varepsilon} \sim \mathcal{O}\left(\frac{1}{\varepsilon}\left(\log \frac{1}{(\delta'\varepsilon)^3}\right)^{3/2}\right) = \mathcal{O}\left(\frac{1}{\varepsilon}\left(\log \frac{1}{\delta'} + \log \frac{1}{\varepsilon}\right)^{3/2}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon}\right)$$

iterations of Algorithm 1. $\diamond$

Secondly, we find it worthwile to provide the following observation on the choice of σ:

Remark 17 (Selection of σ and its impact on T) Note that the value of σ in (12) is at our discretion, so one natural question would be regarding the asymptotic analysis of σ and its impact on our rate T. Observe that the only effect of σ on T is through ξ_3 and ξ_4 defined in (26) and (28) respectively. Taking everything else as constants, following the choice of T and N suggested in Theorem 15, we have that $T \geq \mathcal{O}(\max\{\xi_3, \xi_4\})$. Now since both ξ_3 and ξ_4 will grow unbounded as σ approaches either zero or infinity, so does T. Therefore, we choose a non-zero value for σ instead. An optimal value can be derived, but given that this only affects the constants in the rate, we opt for $\sigma = 1$. $\diamond$

Thirdly, note that for any $K_t \in \mathcal{G}^{\mathrm{lqr}}$, by our choice of α_t and N in Theorem 15, we have

$$
\begin{aligned}
\|K_{t+1} - K_t\|_F &= \|\alpha_t \widehat{\nabla \mathcal{C}}(K_t)\|_F \\
&= \frac{2}{\mu_{\mathrm{lqr}}} \frac{1}{t+N} \|\widehat{\nabla \mathcal{C}}(K_t)\|_F \\
&\leq \frac{2}{\mu_{\mathrm{lqr}}} \frac{1}{N} \|\widehat{\nabla \mathcal{C}}(K_t)\|_F \\
&\overset{\text{(i)}}{\leq} \omega_{\mathrm{lqr}} \frac{\|\widehat{\nabla \mathcal{C}}(K_t)\|_F}{\frac{\xi_3}{1-\gamma} \left(\log \frac{1}{\delta}\right)^{3/2}},
\end{aligned}
\tag{53}
$$

where (i) follows from (46). Now applying Lemma 12 on (53) yields

$$
\|K_{t+1} - K_t\|_F \leq \omega_{\mathrm{lqr}} = \inf_{K \in \mathcal{G}^{\mathrm{lqr}}} \omega_K
$$

with probability at least $1 - \delta$, where $\omega_K = \min\{\beta_K, \zeta_K\}$. This implies that the local Lipschitzness and local smoothness properties of the cost hold for the update at iteration t with probability at least $1 - \delta$.

Fourthly, to help unravel the logical reasoning elucidated in the proof, we introduce the following stopping times:

$$
\tau_2 := \min\left\{ t \geq 1 \;\Big|\; \|\widehat{\nabla \mathcal{C}}(K_{t-1})\|_F > \frac{\xi_3}{1-\gamma} \left(\log \frac{1}{\delta}\right)^{3/2} \right\}
$$
$$
\tau := \min\{\tau_1, \tau_2\},
\tag{54}
$$

with τ_1 previously defined in (45). Essentially, one can observe that as long as $t < \tau_1$ and $t+1 < \tau_2$, it holds that $K_t \in \mathcal{G}^{\mathrm{lqr}}$ and $\|K_{t+1} - K_t\|_F \leq \omega_{\mathrm{lqr}}$, implying that local Lipschitzness and local smoothness properties of the cost hold until that iteration. By the definition of τ in (54), we have that

$$
1_{\tau > t} = 1_{\tau_1 > t} 1_{\tau_2 > t}.
\tag{55}
$$

Moreover, following the definition of $\mathcal{A}_t$ in (39), it also holds that

$$
1_{\tau_2 > t+1} = 1_{\tau_2 > t} 1_{\mathcal{A}_t}.
\tag{56}
$$

Finally, we note that the idea of introducing a stopping time (54), which helps identify the failure of the algorithm and is also used to define a stopped process later on, is inspired by Malik et al. (2020). However, despite the similarity of our forthcoming statements to those in the proof of Malik et al. (2020, Theorem 8), the paths we take to prove said statements are considerably different due to the differences in how we defined our stopping time (and subsequently the stopped process to be defined later on), our gradient estimation method, the time-varying learning rate, etc.

Having covered all of the above, we are now ready to present the following proposition:

Proposition 18 *Under the parameter settings of Theorem 15, we have*

$$
\mathbb{E}[\Delta_T 1_{\tau > T}] \leq \frac{\varepsilon}{20}.
$$

Furthermore, the event $\{\tau > T\}$ happens with a probability of at least $\frac{17}{20} - \delta T$.

Proof The following provides us with a stepping stone for proving the first claim:

Sublemma 3.2 *Under the parameter settings of Theorem 15, we have that*

$$\mathbb{E}[\Delta_t 1_{\tau>t}] \leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{t+N},$$

for all $t \in [T]$.

Proof of Sublemma 3.2. We prove this result by induction on t as follows:
  Base case ($t = 0$):

$$\Delta_0 1_{\tau>0} \leq \Delta_0 \leq \mathcal{C}(K_0) = \frac{N\mathcal{C}(K_0)}{0+N} \leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{0+N},$$

which after taking expectation proves the claim for $t = 0$.
  Inductive step: Let $k \in [T-1]$ be fixed and assume that

$$\mathbb{E}[\Delta_k 1_{\tau>k}] \leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{k+N} \tag{57}$$

holds (the inductive hypothesis). Observe that

$$
\begin{aligned}
\mathbb{E}[\Delta_{k+1} 1_{\tau>k+1}] &\overset{\text{(i)}}{=} \mathbb{E}[\Delta_{k+1} 1_{\tau_1>k+1} 1_{\tau_2>k+1}] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}[\Delta_{k+1} 1_{\tau_1>k} 1_{\tau_2>k} 1_{\mathcal{A}_k}] \\
&= \mathbb{E}[\mathbb{E}[\Delta_{k+1} 1_{\tau_1>k} 1_{\tau_2>k} 1_{\mathcal{A}_k} | \mathcal{F}_k]] \\
&\overset{\text{(iii)}}{=} \mathbb{E}[\mathbb{E}[\Delta_{k+1} 1_{\mathcal{A}_k} | \mathcal{F}_k] 1_{\tau_1>k} 1_{\tau_2>k}],
\end{aligned}
\tag{58}
$$

where (i) follows from (55), (ii) from equation (56) along with the fact that $1_{\tau_1>k+1} \leq 1_{\tau_1>k}$, and (iii) is due to $1_{\tau_2>k}$ and $1_{\tau_1>k}$ being determined by $\mathcal{F}_k$. By Lemma 14, we have that

$$
\begin{aligned}
&\left(\mathbb{E}[\Delta_{k+1} 1_{\mathcal{A}_k} | \mathcal{F}_k] 1_{\tau_1>k}\right) 1_{\tau_2>k} \\
&\leq \left(\left(\left(1 - \mu_{\text{lqr}}\alpha_k\right)\Delta_k + \frac{3\xi_3 \widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_k + \frac{\phi_{\text{lqr}}\alpha_k^2}{2}\frac{\xi_4}{(1-\gamma)^2}\right) 1_{\tau_1>k}\right) 1_{\tau_2>k} \\
&\overset{\text{(i)}}{=} \left(\left(1 - \frac{2}{k+N}\right)\Delta_k + \frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{\mu_{\text{lqr}}(1-\gamma)}\frac{1}{k+N} + \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\frac{1}{(k+N)^2}\right) 1_{\tau>k}, \tag{59}
\end{aligned}
$$

where (i) follows from (55) along with replacing α_k with its value in (46). Now due to the choice of δ in (48), we have that

$$\delta \leq \left(\frac{\mu_{\text{lqr}}(1-\gamma)}{240\xi_3\widetilde{c_{K_1}}}\right)^3 \varepsilon^3,$$

which after noting that $a^3\left(\log\frac{1}{a^3}\right)^{3/2} \leq a$ for any $a \in (0,1)$ implies

$$\delta\left(\log\frac{1}{\delta}\right)^{3/2} \leq \frac{\mu_{\text{lqr}}(1-\gamma)}{240\xi_3\widetilde{c_{K_1}}}\varepsilon \Rightarrow \frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{\mu_{\text{lqr}}(1-\gamma)} \leq \frac{\varepsilon}{40}. \tag{60}$$

27

Applying (60) on (59) yields

$$\mathbb{E}[\Delta_{k+1}1_{\mathcal{A}_k}|\mathcal{F}_k]1_{\tau_1>k}1_{\tau_2>k}$$
$$\leq \left(\left(1-\frac{2}{k+N}\right)\Delta_k + \frac{\varepsilon}{40}\frac{1}{k+N} + \frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}\frac{1}{(k+N)^2}\right)1_{\tau>k}$$
$$\leq \left(1-\frac{2}{k+N}\right)\Delta_k1_{\tau>k} + \frac{\varepsilon}{40}\frac{1}{k+N} + \frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}\frac{1}{(k+N)^2},$$

which after taking expectation results in

$$\mathbb{E}[\mathbb{E}[\Delta_{k+1}1_{\mathcal{A}_k}|\mathcal{F}_k]1_{\tau_1>k}1_{\tau_2>k}]$$
$$\leq \left(1-\frac{2}{k+N}\right)\mathbb{E}[\Delta_k1_{\tau>k}] + \frac{\varepsilon}{40}\frac{1}{k+N} + \frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}\frac{1}{(k+N)^2}. \qquad (61)$$

Combining the hypothesis (57) and inequality (58) with (61), we obtain

$$\mathbb{E}[\Delta_{k+1}1_{\tau>k+1}]$$
$$\leq \left(1-\frac{2}{k+N}\right)\left(\frac{\varepsilon}{40}+\frac{N\mathcal{C}(K_0)}{k+N}\right) + \frac{\varepsilon}{40}\frac{1}{k+N} + \frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}\frac{1}{(k+N)^2}$$
$$\leq \frac{\varepsilon}{40} + \left(1-\frac{1}{k+N}\right)\frac{N\mathcal{C}(K_0)}{k+N} - \frac{1}{(k+N)^2}\left(N\mathcal{C}(K_0)-\frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}\right)$$
$$\overset{(i)}{\leq} \frac{\varepsilon}{40} + \left(\frac{k+N-1}{(k+N)^2}\right)N\mathcal{C}(K_0)$$
$$\leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{k+N+1},$$

where (i) follows from the fact that

$$N\mathcal{C}(K_0) \geq N_1\mathcal{C}(K_0) \geq \left(\frac{4\phi_{\mathrm{lqr}}\xi_4}{\mu_{\mathrm{lqr}}^2(1-\gamma)^2}\frac{2}{\mathcal{C}(K_0)}\right)\mathcal{C}(K_0) = \frac{8\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2} \geq \frac{2\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2\mu_{\mathrm{lqr}}^2}.$$

This proves the claim for $k+1$, completing the inductive step. $\qquad \diamond$

Now due to Sublemma 3.2 and the choice of T from (49) in Theorem 15,

$$\mathbb{E}[\Delta_T1_{\tau>T}] \leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{T+N} \leq \frac{\varepsilon}{40} + \frac{N\mathcal{C}(K_0)}{T} = \frac{\varepsilon}{20},$$

which finishes the proof of the first claim of Proposition 18. Now before moving on to the second claim, we introduce the following sublemma:

Sublemma 3.3 *Under the parameter setup of Theorem 15, we have that for all $t \in [T]$,*

$$\frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{lqr}\xi_4}{2(1-\gamma)^2}\alpha_t^2 + \frac{4\phi_{lqr}\xi_4}{(1-\gamma)^2\mu_{lqr}^2}\frac{1}{t+N+1} \leq \frac{4\phi_{lqr}\xi_4}{(1-\gamma)^2\mu_{lqr}^2}\frac{1}{t+N}. \qquad (62)$$

Proof of Sublemma 3.3. First, substituting α_t with its value in (46), inequality (62) becomes

$$\frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\mu_{\text{lqr}}}\frac{1}{t+N} + \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\left(\frac{1}{(t+N)^2} + \frac{2}{t+N+1}\right) \leq \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\left(\frac{2}{t+N}\right)$$

$$\iff \frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\mu_{\text{lqr}}}\frac{1}{t+N} \leq \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\left(\frac{2}{t+N} - \frac{2}{t+N+1} - \frac{1}{(t+N)^2}\right)$$

$$\iff \frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\mu_{\text{lqr}}}\frac{1}{t+N} \leq \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\left(\frac{2}{(t+N)(t+N+1)} - \frac{1}{(t+N)^2}\right)$$

$$\iff \frac{6\xi_3\widetilde{c_{K_1}}\delta\left(\log\frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\mu_{\text{lqr}}}\frac{1}{t+N} \leq \frac{2\phi_{\text{lqr}}\xi_4}{(1-\gamma)^2\mu_{\text{lqr}}^2}\left(\frac{t+N-1}{(t+N)^2(t+N+1)}\right)$$

$$\iff \delta\left(\log\frac{1}{\delta}\right)^{3/2} \leq \frac{\phi_{\text{lqr}}\xi_4}{3\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{t+N-1}{(t+N)(t+N+1)}\right). \tag{63}$$

Note that for the right-hand side of (63), we have for all $t \in [T]$ that

$$\frac{\phi_{\text{lqr}}\xi_4}{3\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{t+N-1}{t+N}\frac{1}{t+N+1}\right) \overset{(i)}{\geq} \frac{\phi_{\text{lqr}}\xi_4}{6\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{1}{t+N+1}\right)$$

$$\geq \frac{\phi_{\text{lqr}}\xi_4}{6\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{1}{T+N+1}\right)$$

$$\overset{(ii)}{\geq} \frac{\phi_{\text{lqr}}\xi_4}{12\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{1}{T}\right), \tag{64}$$

where (i) follows from the fact that $\frac{t+N-1}{t+N} \geq \frac{1}{2}$ which is due to $N \geq 2$, and (ii) from $\mathcal{C}(K_0) \geq \frac{\varepsilon}{20}$ under the settings of Theorem 15, which results in

$$T = \frac{40}{\varepsilon}N\mathcal{C}(K_0) \geq 2N \geq N+1 \Rightarrow \frac{1}{T+N+1} \geq \frac{1}{2T}.$$

As a result of (63) and (64), in order to conclude the proof Sublemma 3.3, it would suffice to show that

$$\delta\left(\log\frac{1}{\delta}\right)^{3/2} \leq \frac{\phi_{\text{lqr}}\xi_4}{12\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\left(\frac{1}{T}\right)$$

$$= \frac{\phi_{\text{lqr}}\xi_4}{12\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\frac{\varepsilon}{40}\frac{1}{N\mathcal{C}(K_0)}$$

$$= \frac{\phi_{\text{lqr}}\xi_4}{12\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\frac{\varepsilon}{40}\frac{1}{\max\left\{N_1\mathcal{C}(K_0), \frac{2\mathcal{C}(K_0)}{\mu_{\text{lqr}}}\frac{\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\omega_{\text{lqr}}}\right\}}$$

$$= \frac{\phi_{\text{lqr}}\xi_4}{12\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}}\frac{\varepsilon}{40}\min\left\{\frac{1}{N_1\mathcal{C}(K_0)}, \frac{\mu_{\text{lqr}}\omega_{\text{lqr}}(1-\gamma)}{2\mathcal{C}(K_0)\xi_3\left(\log\frac{1}{\delta}\right)^{3/2}}\right\}. \tag{65}$$

29

For (65) to hold, we need two inequalities to hold as a result of the $\min\{.,.\}$ operator. First, we require

$$\delta\left(\log\frac{1}{\delta}\right)^{3/2} \leq \frac{\phi_{\text{lqr}}\xi_4}{480\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}N_1\mathcal{C}(K_0)}\varepsilon. \tag{66}$$

Now since $a^3\left(\log\frac{1}{a^3}\right)^{3/2} \leq a$ for all $a \in (0,1)$ and the choice of δ in (48), i.e.,

$$\delta \leq \left(\frac{\phi_{\text{lqr}}\xi_4}{480\xi_3\widetilde{c_{K_1}}(1-\gamma)\mu_{\text{lqr}}N_1\mathcal{C}(K_0)}\right)^3\varepsilon^3,$$

we conclude that (66) holds for the parameter setup of Theorem 15.

Secondly, it needs to hold that

$$\delta\left(\log\frac{1}{\delta}\right)^{3/2} \leq \frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)\left(\log\frac{1}{\delta}\right)^{3/2}}\varepsilon$$

$$\iff \delta\left(\log\frac{1}{\delta}\right)^3 \leq \frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\varepsilon. \tag{67}$$

Now if

$$\frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\varepsilon \leq 0.028,$$

for any $\delta \leq \left(\frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\right)^3\varepsilon^3$, we have that

$$\delta\left(\log\frac{1}{\delta}\right)^3 \leq \frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\varepsilon,$$

and if

$$\frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\varepsilon > 0.028,$$

it would suffice to have that

$$\delta\left(\log\frac{1}{\delta}\right)^3 \leq 0.028,$$

which would hold for any $\delta \leq 2 \times 10^{-5}$. As a result, due to the choice of δ in (48), i.e.,

$$\delta \leq \min\left\{2 \times 10^{-5}, \left(\frac{\phi_{\text{lqr}}\xi_4\omega_{\text{lqr}}}{960\xi_3^2\widetilde{c_{K_1}}\mathcal{C}(K_0)}\right)^3\varepsilon^3\right\},$$

we have that (67) will also hold under the parameter setup of Theorem 15. Finally, since both (66) and (67) hold for δ as chosen in (48), inequality (65) is satisfied, finishing the proof. $\diamond$

We now prove the second claim. Even though our proof strategy mimics the one in Malik et al. (2020), the structure of the stopping times in (45) and (54) makes the arguments more involved. Note that this difference in the definition of the stopping time (and subsequently the stopped process) can be attributed to the fact that in contrast to Malik et al. (2020)'s one scenario (leaving the stable region) which may lead their algorithm to fail, there are

two possible scenarios that may cause the failure of our algorithm. We start by introducing the stopped process

$$Y_t := \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} + \frac{4\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N} \quad \text{for each } t \in [T].$$

We next show that this process is a supermartingale. First, we have that

$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t]$$

$$=\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1}|\mathcal{F}_t] + \frac{4\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N+1}$$

$$=\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1} \left(1_{\tau_1 \leq t} + 1_{\tau_1 > t}\right)|\mathcal{F}_t] + \frac{4\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N+1}$$

$$=\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1} 1_{\tau_1 \leq t}|\mathcal{F}_t] + \mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1} 1_{\tau_1 > t}|\mathcal{F}_t] + \frac{4\phi_{\mathrm{lqr}}\xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N+1}. \quad (68)$$

Then for the first term on the right-hand side of (68), it holds that

$$\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1} 1_{\tau_1 \leq t}|\mathcal{F}_t] \leq \mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t} 1_{\tau_1 \leq t}|\mathcal{F}_t]$$
$$=1_{\tau_2 > t} \mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_1 \leq t}|\mathcal{F}_t]$$
$$=1_{\tau_2 > t} \mathbb{E}[\Delta_{\tau_1 \wedge t} 1_{\tau_1 \leq t}|\mathcal{F}_t]$$
$$=\Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} 1_{\tau_1 \leq t}. \quad (69)$$

As for the second term, we have

$$\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_2 > t+1} 1_{\tau_1 > t}|\mathcal{F}_t]$$

$$\overset{(i)}{=}\mathbb{E}[\Delta_{\tau_1 \wedge t+1} 1_{\tau_1 > t} 1_{\tau_2 > t} 1_{\mathcal{A}_t}|\mathcal{F}_t]$$

$$=\mathbb{E}[\Delta_{t+1} 1_{\tau_1 > t} 1_{\tau_2 > t} 1_{\mathcal{A}_t}|\mathcal{F}_t]$$

$$=\mathbb{E}[\Delta_{t+1} 1_{\mathcal{A}_t}|\mathcal{F}_t] 1_{\tau_1 > t} 1_{\tau_2 > t}$$

$$\overset{(ii)}{\leq} \left((1-\mu_{\mathrm{lqr}}\alpha_t)\Delta_t + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{\mathrm{lqr}}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2}\right)1_{\tau_1 > t} 1_{\tau_2 > t}$$

$$= \left(\left(1-\frac{2}{t+N}\right)\Delta_t + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{\mathrm{lqr}}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2}\right)1_{\tau_1 > t} 1_{\tau_2 > t}$$

$$\overset{(iii)}{\leq} \Delta_t 1_{\tau_1 > t} 1_{\tau_2 > t} + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{\mathrm{lqr}}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2}$$

$$\overset{(iv)}{=} \Delta_{\tau_1 \wedge t} 1_{\tau_1 > t} 1_{\tau_2 > t} + \frac{3\xi_3\widetilde{c_{K_1}}}{1-\gamma}\delta\left(\log\frac{1}{\delta}\right)^{3/2}\alpha_t + \frac{\phi_{\mathrm{lqr}}\alpha_t^2}{2}\frac{\xi_4}{(1-\gamma)^2}, \quad (70)$$

where (i) follows from (56), (ii) from Lemma 14, (iii) from $1_{\tau_1 > t} 1_{\tau_2 > t} \leq 1$ along with the fact that $\frac{2}{t+N} \leq 1$ for all $t \in [T]$, and (iv) from $\Delta_t 1_{\tau_1 > t} = \Delta_{\tau_1 \wedge t} 1_{\tau_1 > t}$.

31

Combining (68), (69), and (70), we obtain that for all $t \in [T]$,

$$
\begin{aligned}
\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq & \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} 1_{\tau_1 \leq t} + \Delta_{\tau_1 \wedge t} 1_{\tau_1 > t} 1_{\tau_2 > t} \\
& + \frac{3\xi_3 \widetilde{c_{K_1}}}{1-\gamma} \delta \left(\log \frac{1}{\delta}\right)^{3/2} \alpha_t + \frac{\phi_{\mathrm{lqr}} \alpha_t^2}{2} \frac{\xi_4}{(1-\gamma)^2} + \frac{4\phi_{\mathrm{lqr}} \xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N+1} \\
\overset{(i)}{\leq} & \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} (1_{\tau_1 \leq t} + 1_{\tau_1 > t}) + \frac{4\phi_{\mathrm{lqr}} \xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N} \\
= & \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} + \frac{4\phi_{\mathrm{lqr}} \xi_4}{(1-\gamma)^2 \mu_{\mathrm{lqr}}^2} \frac{1}{t+N} \\
= & Y_t,
\end{aligned}
$$

where (i) follows from Sublemma 3.3. This proves the claim that Y_t is a supermartingale. Moreover, define the following events:

$$
\begin{aligned}
\mathcal{E}_1 & := \{\tau_2 \leq \tau_1 \text{ and } \tau_2 \in [T]\} \\
\mathcal{E}_2 & := \{\tau_1 < \tau_2 \text{ and } \tau_1 \in [T]\} \\
\mathcal{E}_3 & := \left\{\max_{t \in [T]} \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} \geq 10\mathcal{C}(K_0)\right\},
\end{aligned}
$$

and hence, we have $\mathbb{P}\{\tau \leq T\} = \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2)$. Now since $\tau_2 \leq \tau_1$ in $\mathcal{E}_1$ suggests that $\|\widehat{\nabla \mathcal{C}}(K_{\tau_2-1})\|_F > \frac{\xi_3}{1-\gamma} \left(\log \frac{1}{\delta}\right)^{3/2}$ despite $\Delta_{\tau_2-1} \leq 10\mathcal{C}(K_0)$ (which implies $\mathcal{K}_{\tau_2-1} \in \mathcal{G}^{\mathrm{lqr}}$), after applying union bound on the result of Lemma 12, we have

$$
\mathbb{P}(\mathcal{E}_1) \leq \delta T. \tag{71}
$$

Furthermore, note that $\tau_1 < \tau_2$ in $\mathcal{E}_2$ implies that $\Delta_{\tau_1 \wedge \tau_1} 1_{\tau_2 > \tau_1} = \Delta_{\tau_1}$ and since $\tau_1 \in [T]$, it holds that

$$
\max_{t \in [T]} \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} \geq \Delta_{\tau_1 \wedge \tau_1} 1_{\tau_2 > \tau_1} = \Delta_{\tau_1} \overset{(i)}{>} 10\mathcal{C}(K_0),
$$

where (i) follows the definition of τ_1. As a result of this, we have that $\mathcal{E}_2$ implies $\mathcal{E}_3$, and consequently, $\mathbb{P}(\mathcal{E}_2) \leq \mathbb{P}(\mathcal{E}_3)$. Finally, since $Y_t \geq \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t}$ for all $t \in [T]$, we have that

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_2) &\leq \mathbb{P}(\mathcal{E}_3) \\
&= \mathbb{P}\left\{ \max_{t \in [T]} \Delta_{\tau_1 \wedge t} 1_{\tau_2 > t} \geq 10 \mathcal{C}(K_0) \right\} \\
&\leq \mathbb{P}\left\{ \max_{t \in [T]} Y_t \geq 10 \mathcal{C}(K_0) \right\} \\
&\overset{(i)}{\leq} \frac{\mathbb{E}[Y_0]}{10 \mathcal{C}(K_0)} \\
&= \frac{\Delta_{\tau_1 \wedge 0} 1_{\tau_2 > 0} + \frac{4 \phi_{\text{lqr}} \xi_4}{(1-\gamma)^2 \mu_{\text{lqr}}^2} \frac{1}{N}}{10 \mathcal{C}(K_0)} \\
&\overset{(ii)}{\leq} \frac{\Delta_0 + \mathcal{C}(K_0)/2}{10 \mathcal{C}(K_0)} \\
&\leq \frac{\mathcal{C}(K_0) + \mathcal{C}(K_0)/2}{10 \mathcal{C}(K_0)} \\
&= \frac{3}{20},
\end{aligned}
\tag{72}
$$

where (i) follows from applying Doob/Ville's inequality for supermartingales, and (ii) from the condition on the choice of N in Theorem 15. Employing the acquired probability bounds (71) and (72), we observe that

$$
\begin{aligned}
\mathbb{P}\{\tau \leq T\} &= \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) \\
&\leq \delta T + \frac{3}{20},
\end{aligned}
$$

which verifies the second claim of Proposition 18, concluding the proof. ∎

The proof of our main result is a straightforward corollary:

Proof [Proof of Theorem 15] We now show how Proposition 18 can be employed to validate the claims of Theorem 15. Note that

$$
\begin{aligned}
\mathbb{P}\{\Delta_T \geq \varepsilon\} &\leq \mathbb{P}\{\Delta_T 1_{\tau > T} \geq \varepsilon\} + \mathbb{P}\{1_{\tau \leq T} = 1\} \\
&\overset{(i)}{\leq} \frac{1}{\varepsilon} \mathbb{E}[\Delta_T 1_{\tau > T}] + \mathbb{P}\{\tau \leq T\} \\
&\overset{(ii)}{\leq} \frac{1}{20} + \frac{3}{20} + \delta T = \frac{1}{5} + \delta T,
\end{aligned}
$$

where (i) follows from Markov's inequality and (ii) follows from Proposition 18. ∎

In the next section, we present a brief simulation study using two representative examples from Malik et al. (2020) to empirically validate our theoretical guarantees and compare convergence rates.

4. Simulation Studies

We now revisit several examples introduced from the previous literature (specifically from Malik et al., 2020) and show empirically that our performance does indeed match $\tilde{O}(\epsilon^{-1})$ guaranteed by our theoretical results.

We begin with the following LQR problem:

$$A = \begin{bmatrix} 1 & 0 & -10 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \ Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \ R = \begin{bmatrix} 5 & -3 & 0 \\ -3 & 5 & -2 \\ 0 & -2 & 5 \end{bmatrix},$$

under the random initialization setting, where the initial state x_0 is sampled uniformly from the set of signed canonical basis vectors, yielding a mean-zero distribution. The discount factor is set to $\gamma = 0.9$. We initialize with a policy K_0 satisfying $\mathcal{C}_{\text{init}}(K_0) - \mathcal{C}_{\text{init}}(K^*) = 11.716$, use a constant step-size of 10^{-4}, and set the batch size to $N_s = 10^3$. This example was previously considered in Malik et al. (2020) under a two-point gradient estimation scheme, where an empirical sample complexity of approximately $\mathcal{O}(\epsilon^{-1})$ was observed (see Malik et al., 2020, Figure 2 (b)). As shown in Figure 2(a), our method achieves a fitted rate of approximately $\mathcal{O}(\epsilon^{-1.03})$, in line with our theoretical guarantees of $\widetilde{\mathcal{O}}(\epsilon^{-1})$, with the small discrepancy likely due to logarithmic factors.

We next consider a second example from Malik et al. (2020), this time under the noisy dynamics setting:

$$A = 0.1 I_3, \quad B = 0.01 I_3, \quad Q = 100 I_3, \quad R = 100 I_3,$$

with the discount factor again set to $\gamma = 0.9$. The system is subject to additive Gaussian noise with zero mean and covariance $\frac{1}{25} I_3$. We initialize with a policy K_0 such that $\mathcal{C}_{\text{dyn}}(K_0) = \mathcal{C}_{\text{dyn}}(K^*) + 3.12$, and apply a time-varying step-size given by

$$\alpha_t = \max\left(\frac{1}{60t + 2000}, 2 \cdot 10^{-5} \right),$$

along with a batch size of $N_s = 3000$. This choice allows us to apply the time-varying step-size scheme from Theorem 15 (or, equivalently, from Corollary 25 in Appendix B for the noisy dynamics setting), although we note that a constant step-size performs similarly well in practice. The same problem was studied in Malik et al. (2020) under a one-point estimation scheme, where a sample complexity of approximately $\mathcal{O}(\epsilon^{-2})$ was observed (see Malik et al., 2020, Figure 2 (c)). As can be seen from Figure 2(b), our empirical rate is approximately $\mathcal{O}(\epsilon^{-0.87})$, satisfying our theoretical guarantee of at most $\widetilde{\mathcal{O}}(\epsilon^{-1})$.

5. Summary and Discussion

We have provided an algorithm with ε-optimality guarantees with a provable convergence rate of $\widetilde{\mathcal{O}}(1/\varepsilon)$ for the discounted discrete-time LQR problem in the model-free setting. This was made possible by employing a gradient estimation technique inspired by REINFORCE, combined with a time-varying step-size. Our results contrast from the ones obtained by two-point methods—which make the stronger assumption of access to cost for two different
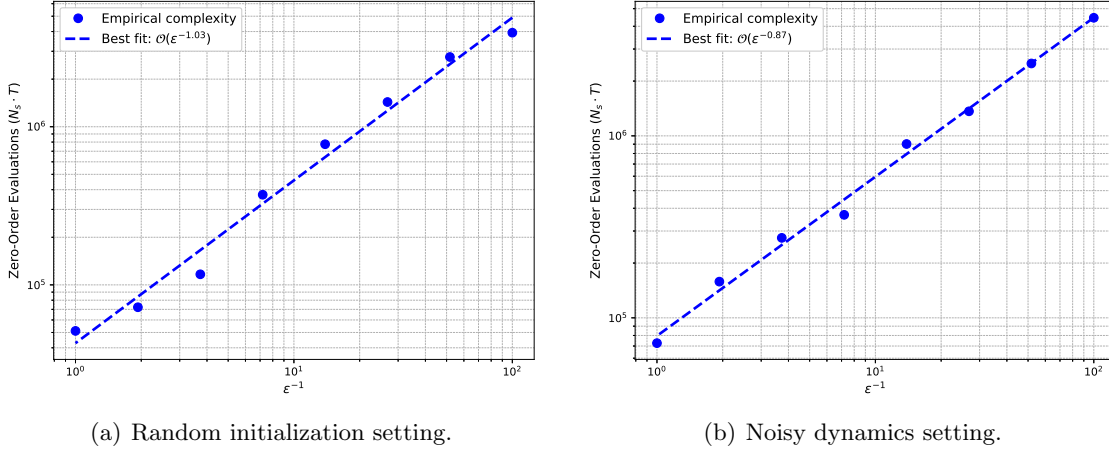
(a) Random initialization setting. (b) Noisy dynamics setting.

Figure 2: Empirical zero-order evaluations required by the policy gradient method to achieve ϵ-optimality. Dashed lines indicate the best-fit lines in the log-log scale. The plots were generated by averaging 20 runs of Algorithm 1.

policies with the same realization of all system randomness—as well as results that assume stability of the obtained policies throughout the algorithm.

An interesting future direction would be to investigate an actor-critic approach that could maintain the rate without requiring further assumptions. Moreover, one could consider an extension of the presented results for the undiscounted case; in particular, the current analysis of gradient estimation with one zero-order evaluation per iteration heavily relies on sampling from a distribution whose definition relies on the discount factor be strictly less that one.

Appendix A. Improved Probability Guarantee via Minibatching

Following the useful comments of an anonymous reviewer, we dedicate this section to addressing our constant probability guarantees in Theorem 15. To that end, and inspired by the approach in Malik et al. (2020, Appendix E), we propose a minibatched gradient estimation method, in which we average a sufficiently large number of i.i.d. copies of our original gradient estimate to obtain a more accurate approximation of the true gradient with high probability. Consider the minibatch gradient estimate

$$\overline{\nabla \mathcal{C}}_{N_s}(K) := \frac{1}{N_s} \sum_{i=1}^{N_s} \widehat{\nabla \mathcal{C}}_i(K), \tag{73}$$

where each $\widehat{\nabla \mathcal{C}}_i(K)$ is an i.i.d. copy of $\widehat{\nabla \mathcal{C}}(K)$ in (15). We provide the following lemma regarding the concentration of this averaged estimate around its expectation, which is equal to the actual gradient as shown in Propositon 9.

Lemma 19 *Suppose $K \in \mathcal{G}^{lqr}$, γ is chosen as in Lemma 11, and $\delta > 0$ chosen to satisfy*

$$\delta \leq \min\left\{e^{-3/2}, \frac{1-\gamma}{3\xi_3}\sqrt{\frac{\mu_{lqr}\varepsilon}{8}}\right\}. \tag{74}$$

If N_s is selected such that

$$N_s \geq \left\lceil \max\left\{5000, 8\left(\log\frac{2}{\delta}\right)^3, \frac{2048\xi_3^2}{9(1-\gamma)^2\mu_{lqr}}\frac{1}{\varepsilon}\left(\log\frac{2(mn+1)}{\delta}\right)^2, \right.\right.$$
$$\left.\left. \frac{128\xi_4}{\mu_{lqr}(1-\gamma)^2}\frac{1}{\varepsilon}\log\frac{2(mn+1)}{\delta}\right\}\right\rceil = \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon}\right), \tag{75}$$

then the minibatch averaged estimate (73) *satisfies*

$$\|\overline{\nabla\mathcal{C}}_{N_s}(K) - \nabla\mathcal{C}(K)\|_F \leq \sqrt{\frac{\mu_{lqr}\varepsilon}{8}},$$

with probability at least $1 - \delta$.

Proof Let us define the following event

$$\mathcal{B}_i = \left\{\|\widehat{\nabla\mathcal{C}}_i(K)\|_F \leq \frac{\xi_3}{1-\gamma}\left(\log\frac{2N_s}{\delta}\right)^{3/2}\right\},$$

which holds with probability at least $1 - \frac{\delta}{2N_s}$ for each i. As a result, following Lemma 13, we have for all $i \in \{1, 2, \ldots, N_s\}$ that

$$\|\mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}] - \nabla\mathcal{C}(K)\|_F \leq \frac{3\xi_3}{1-\gamma}\frac{\delta}{2N_s}\left(\log\frac{2N_s}{\delta}\right)^{3/2}, \tag{76}$$

where $\mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}]$ holds the same value for all i. Moreover, note that

$$\overline{\nabla\mathcal{C}}_{N_s}(K) - \nabla\mathcal{C}(K) = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i^c} + \widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i} - \nabla\mathcal{C}(K)\right)$$

$$= \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i^c} + \widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i} - \mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}]\right)$$

$$+ \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}] - \nabla\mathcal{C}(K)$$

$$= \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i^c} + \widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i} - \mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}]\right)$$

$$+ \mathbb{E}[\widehat{\nabla\mathcal{C}}_1(K)1_{\mathcal{B}_1}] - \nabla\mathcal{C}(K). \tag{77}$$

Let us now define

$$S_i := \widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i} - \mathbb{E}[\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i}],$$

36

so we can use (77) to write

$$\|\overline{\nabla\mathcal{C}}_{N_s}(K) - \nabla\mathcal{C}(K)\|_F$$

$$\leq \frac{1}{N_s}\sum_{i=1}^{N_s}\|\widehat{\nabla\mathcal{C}}_i(K)1_{\mathcal{B}_i^c}\|_F + \|\frac{1}{N_s}\sum_{i=1}^{N_s}S_i\|_F + \|\mathbb{E}[\widehat{\nabla\mathcal{C}}_1(K)1_{\mathcal{B}_1}] - \nabla\mathcal{C}(K)\|_F$$

$$\stackrel{(i)}{\leq} \frac{1}{N_s}\sum_{i=1}^{N_s}\|\widehat{\nabla\mathcal{C}}_i(K)\|_F 1_{\mathcal{B}_i^c} + \|\frac{1}{N_s}\sum_{i=1}^{N_s}S_i\|_F + \frac{3\xi_3}{1-\gamma}\frac{\delta}{2N_s}\left(\log\frac{2N_s}{\delta}\right)^{3/2}, \qquad (78)$$

where (i) follows from (76). For the first term in (78), we have

$$\mathbb{P}\left\{\frac{1}{N_s}\sum_{i=1}^{N_s}\|\widehat{\nabla\mathcal{C}}_i(K)\|_F 1_{\mathcal{B}_i^c} = 0\right\} \geq \mathbb{P}\left\{\bigcap_{i=1}^{N_s}\mathcal{B}_i\right\} \geq 1 - \sum_{i=1}^{N_s}\mathbb{P}\{\mathcal{B}_i^c\} \geq 1 - \frac{\delta}{2}. \qquad (79)$$

Additionally, we can use the matrix Bernstein theorem to bound the second term in (78) with high probability (Tropp, 2015, Theorem 1.6.2). In order to do so, first observe that S_i's are i.i.d. random matrices and satisfy

$$\mathbb{E}[S_i] = 0, \text{ and } \|S_i\|_F \leq \frac{2\xi_3}{1-\gamma}\left(\log\frac{2N_s}{\delta}\right)^{3/2}$$

for all $i \in \{1, 2, \ldots, N_s\}$. Now let

$$Z := \sum_{i=1}^{N_s}S_i.$$

We have

$$\mathbb{E}[\|Z\|_F^2] = \mathbb{E}[\text{tr}(Z^\top Z)]$$
$$= \sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\mathbb{E}[\text{tr}(S_i^\top S_j)]$$
$$= \sum_{i=1}^{N_s}\mathbb{E}[\|S_i\|_F^2]$$
$$= N_s\mathbb{E}[\|S_1\|_F^2]$$
$$\leq N_s\mathbb{E}[\|\widehat{\nabla\mathcal{C}}_1(K)1_{\mathcal{B}_1}\|_F^2]$$
$$\stackrel{(i)}{\leq} N_s\frac{\xi_4}{(1-\gamma)^2},$$

where (i) follows from (27) in Lemma 12. As a result, following Tropp (2015, Theorem 1.6.2), along with an additional vectorization of the matrices (to transfer the results from

2-norm to Frobenius norm), we have

$$\mathbb{P}\{\|\frac{1}{N_s}\sum_{i=1}^{N_s} S_i\|_F \geq t\} = \mathbb{P}\{\|Z\|_F \geq N_s t\}$$

$$\leq (mn+1)\exp\left(-\frac{N_s^2 t^2}{2N_s\frac{\xi_4}{(1-\gamma)^2} + \frac{4\xi_3}{3(1-\gamma)}\left(\log\frac{2N_s}{\delta}\right)^{3/2} N_s t}\right)$$

$$= (mn+1)\exp\left(-\frac{N_s t^2}{2\frac{\xi_4}{(1-\gamma)^2} + \frac{4\xi_3}{3(1-\gamma)}\left(\log\frac{2N_s}{\delta}\right)^{3/2} t}\right). \tag{80}$$

Now letting $t = \frac{1}{2}\sqrt{\frac{\mu_{\text{lqr}}\varepsilon}{8}}$ and selecting N_s as suggested in (75) lets us write (80) as

$$\mathbb{P}\left\{\|\frac{1}{N_s}\sum_{i=1}^{N_s} S_i\|_F \geq \frac{1}{2}\sqrt{\frac{\mu_{\text{lqr}}\varepsilon}{8}}\right\} \leq \frac{\delta}{2}. \tag{81}$$

For the third term in (78), note that due to the choice of N_s in (75), we have

$$\frac{3\xi_3}{1-\gamma}\frac{\delta}{2}\frac{\left(\log\frac{2N_s}{\delta}\right)^{3/2}}{N_s} \overset{(i)}{\leq} \frac{3\xi_3}{1-\gamma}\frac{\delta}{2}\frac{1}{\sqrt{N_s}}$$

$$\leq \frac{3\xi_3}{1-\gamma}\frac{\delta}{2}$$

$$\overset{(ii)}{\leq} \frac{1}{2}\sqrt{\frac{\mu_{\text{lqr}}\varepsilon}{8}}, \tag{82}$$

where (i) follows from the choice of N_s in (75), and (ii) from (74). Finally, applying (79), (81), and (82), along with union bound, on (78) concludes the proof. ∎

We are now in a position to present the following result:

Theorem 20 *Suppose K_0 is stable, γ is as suggested in Lemma 11, and the update rule follows*

$$K_{t+1} = K_t - \alpha\overline{\nabla\mathcal{C}}_{N_s}(K_t) \tag{83}$$

with a constant step-size α satisfying

$$\alpha \leq \min\left\{\frac{\omega_{lqr}}{\widetilde{c_{K_1}} + \sqrt{\frac{\mu_{lqr}\mathcal{C}(K_0)}{8}}}, \frac{1}{4\phi_{lqr}}, \frac{4}{\mu_{lqr}}\right\}. \tag{84}$$

Then for a given error tolerance $\varepsilon \in (0, \mathcal{C}(K_0)]$, and for any δ satisfying (74), the update rule (83), with $N_s \sim \widetilde{\mathcal{O}}(1/\varepsilon)$ chosen according to (75), guarantees that after

$$T = \frac{4}{\alpha\mu_{lqr}}\log\left(\frac{2\mathcal{C}(K_0)}{\varepsilon}\right) \tag{85}$$

38

iterations, we have

$$\mathcal{C}(K_T) - \mathcal{C}(K^*) \le \varepsilon,$$

with a probability of at least $1 - \delta T$.

Proof First, assume that $K_t \in \mathcal{G}^{\text{lqr}}$, then since N_s is chosen as in (75), we have from Lemma 19 that

$$\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F \le \sqrt{\frac{\mu_{\text{lqr}}\varepsilon}{8}},$$

with probability at least $1 - \delta$. Hence, conditioned on this event, we have the following bound

$$
\begin{aligned}
\|\alpha\overline{\nabla\mathcal{C}}_{N_s}(K_t)\|_F &\le \alpha\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t) + \nabla\mathcal{C}(K_t)\|_F \\
&\le \alpha\left(\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F + \|\nabla\mathcal{C}(K_t)\|_F\right) \\
&\le \alpha\left(\sqrt{\frac{\mu_{\text{lqr}}\varepsilon}{8}} + \widetilde{c_{K_1}}\right) \\
&\overset{(i)}{\le} \alpha\left(\sqrt{\frac{\mu_{\text{lqr}}\mathcal{C}(K_0)}{8}} + \widetilde{c_{K_1}}\right) \\
&\overset{(ii)}{\le} \omega_{\text{lqr}},
\end{aligned}
\tag{86}
$$

where (i) follows from $\varepsilon \le \mathcal{C}(K_0)$ and (ii) from the choice of α in (84). Note that (86) ensures that our step-size is small enough for Lipschitz and smoothness properties to hold. Consequently, we can write

$$
\begin{aligned}
\Delta_{t+1} - \Delta_t &= \mathcal{C}(K_{t+1}) - \mathcal{C}(K_t) \\
&\le -\left\langle \nabla\mathcal{C}(K_t), \alpha\overline{\nabla\mathcal{C}}_{N_s}(K_t)\right\rangle + \frac{\phi_{\text{lqr}}}{2}\alpha^2\|\overline{\nabla\mathcal{C}}_{N_s}(K_t)\|_F^2 \\
&= -\alpha\left\langle \nabla\mathcal{C}(K_t), \nabla\mathcal{C}(K_t) + \overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\right\rangle \\
&\quad + \frac{\phi_{\text{lqr}}}{2}\alpha^2\left(\|\nabla\mathcal{C}(K_t) + (\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t))\|_F^2\right) \\
&\le -\alpha\|\nabla\mathcal{C}(K_t)\|_F^2 + \alpha\|\nabla\mathcal{C}(K_t)\|_F\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F \\
&\quad + \phi_{\text{lqr}}\alpha^2\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2 + \phi_{\text{lqr}}\alpha^2\|\nabla\mathcal{C}(K_t)\|_F^2 \\
&\le -\alpha\|\nabla\mathcal{C}(K_t)\|_F^2 + \frac{\alpha}{2}\left(\|\nabla\mathcal{C}(K_t)\|_F^2 + \|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2\right) \\
&\quad + \phi_{\text{lqr}}\alpha^2\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2 + \phi_{\text{lqr}}\alpha^2\|\nabla\mathcal{C}(K_t)\|_F^2 \\
&= -\frac{\alpha}{2}\|\nabla\mathcal{C}(K_t)\|_F^2 + \phi_{\text{lqr}}\alpha^2\|\nabla\mathcal{C}(K_t)\|_F^2 \\
&\quad + \left(\frac{\alpha}{2} + \phi_{\text{lqr}}\alpha^2\right)\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2 \\
&\overset{(i)}{\le} -\frac{\alpha}{2}\|\nabla\mathcal{C}(K_t)\|_F^2 + \frac{\alpha}{4}\|\nabla\mathcal{C}(K_t)\|_F^2 + \left(\frac{\alpha}{2} + \frac{\alpha}{4}\right)\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2 \\
&\le -\frac{\alpha}{4}\|\nabla\mathcal{C}(K_t)\|_F^2 + \alpha\|\overline{\nabla\mathcal{C}}_{N_s}(K_t) - \nabla\mathcal{C}(K_t)\|_F^2 \\
&\overset{(ii)}{\le} -\frac{\alpha\mu_{\text{lqr}}}{4}\Delta_t + \alpha\frac{\mu_{\text{lqr}}\varepsilon}{8},
\end{aligned}
\tag{87}
$$

where (i) follows from the fact that $\alpha\phi_{\mathrm{lqr}} \leq 1/4$ due to the choice of α in (84), and (ii) from the PL inequality (9). Rearranging (87) yields

$$\Delta_{t+1} \leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)\Delta_t + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8}. \tag{88}$$

With this in place, we use strong induction to finalize the proof. For each time $i \in \{1, 2, \ldots, T\}$, let $\mathscr{E}_i$ denote the event that $\Delta_i \leq 10\mathcal{C}(K_0)$ (implying $K_i \in \mathcal{G}^{\mathrm{lqr}}$) and $\Delta_i \leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)\Delta_{i-1} + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8}$. We claim that for each $t \in \mathbb{N}$, it holds that

$$\mathbb{P}\left\{\cap_{i=1}^t \mathscr{E}_i\right\} \geq 1 - \delta t.$$

We demonstrate this by induction as follows:

Base case ($t = 0$): Since $K_0 \in \mathcal{G}^{\mathrm{lqr}}$, we have by Lemma 19 and inequality (88) that

$$\Delta_1 \leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)\Delta_0 + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8}.$$

Moreover, since $\alpha \leq \frac{4}{\mu_{\mathrm{lqr}}}$ and $\varepsilon \leq \mathcal{C}(K_0)$, we have that $\Delta_1 \leq \Delta_0 + \frac{1}{2}\mathcal{C}(K_0) \leq 10\mathcal{C}(K_0)$. Thus, we have shown that $\mathscr{E}_1$ holds with probability at least $1 - \delta$, establishing the base case.

Inductive step: By induction hypothesis, we have that the event $\cap_{i=1}^t \mathscr{E}_i$ holds with probability at least $1 - \delta t$. Conditioned on this event, we have by Lemma 19 and inequality (88) that with probability at least $1 - \delta$, the following holds

$$\begin{aligned}
\Delta_{t+1} &\leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)\Delta_t + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8} \\
&\leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^{t+1}\Delta_0 + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8}\sum_{i=0}^t\left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^i \\
&\leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^{t+1}\Delta_0 + \alpha\frac{\mu_{\mathrm{lqr}}\varepsilon}{8}\sum_{i=0}^\infty\left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^i \\
&= \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^{t+1}\Delta_0 + \frac{\varepsilon}{2},
\end{aligned} \tag{89}$$

and since $\varepsilon \leq \mathcal{C}(K_0)$, we also have $\Delta_{t+1} \leq 10\mathcal{C}(K_0)$. Now combining this with a union bound shows that $\cap_{i=1}^{t+1}\mathscr{E}_i$ holds with a probability of at least $1 - (\delta t + \delta) = 1 - \delta(t+1)$, completing the inductive step.

Finally, conditioned on $\cap_{i=1}^T\mathscr{E}_i$, similar to (89), we obtain

$$\begin{aligned}
\Delta_T &\leq \left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^T\Delta_0 + \frac{\varepsilon}{2} \\
&\overset{(i)}{\leq} \left[\left(1 - \frac{\alpha\mu_{\mathrm{lqr}}}{4}\right)^{\frac{4}{\alpha\mu_{\mathrm{lqr}}}}\right]^{\log\left(\frac{2\mathcal{C}(K_0)}{\varepsilon}\right)}\Delta_0 + \frac{\varepsilon}{2} \\
&\leq \left(e^{-1}\right)^{\log\left(\frac{2\mathcal{C}(K_0)}{\varepsilon}\right)}\Delta_0 + \frac{\varepsilon}{2} \\
&= \frac{\varepsilon}{2\mathcal{C}(K_0)}\Delta_0 + \frac{\varepsilon}{2} \\
&\leq \varepsilon,
\end{aligned}$$

where (i) follows from the choice of T in (85). This, along with recalling $\mathbb{P}\left\{\cap_{i=1}^{T}\mathscr{E}_i\right\} \geq 1-\delta T$, concludes the proof. ∎

Remark 21 *As discussed after Lemma 11, the condition on γ depends only on the cost bound used to define the set $\mathcal{G}^{lqr}$. In particular, from the induction step in the proof of Theorem 20, one can deduce that this bound can be tightened from $10\mathcal{C}(K_0) + \mathcal{C}(K^*)$ to $\mathcal{C}(K_0)+\mathcal{C}(K^*)$, while still preserving the convergence guarantees (i.e., achieving ε-optimality with probability exceeding $1-\delta T$ for small enough ε). This effectively enlarges the allowable range of γ; for example, the alternative set*

$$\mathcal{G}'^{lqr} = \{K \mid \mathcal{C}(K) - \mathcal{C}(K^*) \leq \mathcal{C}(K_0)\}$$

admits any γ in the interval $\left(1 - \frac{\sigma_{\min}(Q)}{2\mathcal{C}_{und}(K_0)}, 1\right)$, which is more permissive than the condition stated in Lemma 11.

Appendix B. Extension to Noisy Dynamics Setting

In this section, we show how everything from the Random Initialization setting trasnfers into the noisy dynamics setup. We begin by establishing an exponential decay bound on $\|(A - BK)^t\|$, which serves as a key technical tool for the results that follow.

B.1 Exponential Decay of $\|(A - BK)^t\|$

Before we introduce the next result, let us define

$$M := \sqrt{\frac{10\mathcal{C}_{\text{init}}(K_0) + \mathcal{C}_{\text{init}}(K^*)}{\lambda_{\min}(Q)}}, \quad \text{and}$$

$$r := \sqrt{1 - \frac{0.5\lambda_{\min}(Q)}{10\mathcal{C}_{\text{init}}(K_0) + \mathcal{C}_{\text{init}}(K^*) - 0.5\lambda_{\min}(Q)}} \ \in (0, 1). \tag{90}$$

Lemma 22 *Suppose $\gamma \in \left(1 - \frac{0.5\sigma_{\min}(Q)}{11\mathcal{C}_{und}(K_0)}, 1\right)$. Then for any $K \in \mathcal{G}^{lqr}$, it holds that*

$$\|(A - BK)^t\|_2 \leq Mr^t.$$

Proof Let P_K denote the unique positive-definite solution of the discrete algebraic Riccati equation

$$P_K = Q + K^\top RK + \gamma(A - BK)^\top P_K(A - BK).$$

Re-arranging gives the Lyapunov inequality

$$\gamma(A - BK)^\top P_K(A - BK) = P_K - \left(Q + K^\top RK\right) \preceq (1 - a_K)P_K,$$

where

$$a_K = \frac{\lambda_{\min}\left(Q + K^\top RK\right)}{\lambda_{\max}(P_K)} \in (0, 1].$$

Define $b_K := \sqrt{1 - a_K} \in (0, 1)$; then

$$\gamma(A - BK)^\top P_K(A - BK) \preceq b_K^2 P_K,$$

and hence,

$$\left[(A - BK)^t\right]^\top P_K(A - BK)^t \preceq \frac{b_K^2}{\gamma} \left[(A - BK)^{t-1}\right]^\top P_K(A - BK)^{t-1}$$

$$\preceq \cdots \preceq \left(\frac{b_K^2}{\gamma}\right)^t P_K. \qquad (91)$$

Equip $\mathbb{R}^n$ with the quadratic norm $\|x\|_{P_K} := \sqrt{x^\top P_K x}$. From (91), we obtain

$$\|(A - BK)^t x\|_{P_K} \leq \left(\frac{b_K}{\sqrt{\gamma}}\right)^t \|x\|_{P_K} \quad \forall x \in \mathbb{R}^n,$$

hence for every integer $t \geq 0$

$$\|(A - BK)^t\|_{P_K} \leq \left(\frac{b_K}{\sqrt{\gamma}}\right)^t,$$

where $\|.\|_{P_K}$ is the P_K-induced matrix norm. Because all norms on a finite-dimensional space are equivalent,

$$\|x\|_2^2 \leq \lambda_{\min}^{-1}\|x\|_{P_K}^2, \qquad \|x\|_{P_K}^2 \leq \lambda_{\max}(P_K)\|x\|_2^2,$$

so the operator norm induced by $\|\cdot\|_2$ satisfies

$$\|(A - BK)^t\|_2 \leq \sqrt{\frac{\lambda_{\max}(P_K)}{\lambda_{\min}(P_K)}}\|(A - BK)^t\|_{P_K} \leq \sqrt{\frac{\lambda_{\max}(P_K)}{\lambda_{\min}(P_K)}}\left(\frac{b_K}{\sqrt{\gamma}}\right)^t.$$

Now note that we have that for $K \in \mathcal{G}^{\mathrm{lqr}}$,

$$10\mathcal{C}_{\mathrm{init}}(K_0) + \mathcal{C}_{\mathrm{init}}(K^*) \geq \mathcal{C}_{\mathrm{init}}(K) = \mathrm{tr}(P_K) \geq \lambda_{\max}(P_K),$$

and hence,

$$\lambda_{\max}(P_K) \leq 10\mathcal{C}_{\mathrm{init}}(K_0) + \mathcal{C}_{\mathrm{init}}(K^*) =: \lambda_1.$$

As a result of this, all the previously used values for bounding $\|(A - BK)^t\|$ can be bounded by constants independent of K:

$$\lambda_{\min}(P_K) \geq \lambda_{\min}(Q) =: \lambda_2$$

$$a_K \geq \frac{\lambda_{\min}(Q)}{\lambda_{\max}(P_K)} \geq \frac{\lambda_2}{\lambda_1}$$

$$b_K = \sqrt{1 - a_K} \leq \sqrt{1 - \frac{\lambda_2}{\lambda_1}}$$

$$\sqrt{\frac{\lambda_{\max}(P_K)}{\lambda_{\min}(P_K)}} \leq \sqrt{\frac{\lambda_1}{\lambda_2}}.$$

Now since by assumption,

$$\gamma \geq 1 - \frac{0.5\sigma_{\min}(Q)}{11\mathcal{C}_{\mathrm{und}}(K_0)} \geq 1 - \frac{0.5\lambda_2}{\lambda_1},$$

we also conclude that

$$\left(\frac{b_K}{\sqrt{\gamma}}\right) \leq \sqrt{\frac{\lambda_1 - \lambda_2}{\lambda_1 - 0.5\lambda_2}} = \sqrt{1 - \frac{0.5\lambda_2}{\lambda_1 - 0.5\lambda_2}};$$

therefore,

$$\|(A - BK)^t\|_2 \leq \sqrt{\frac{\lambda_1}{\lambda_2}} \left(1 - \frac{0.5\lambda_2}{\lambda_1 - 0.5\lambda_2}\right)^{t/2}, \tag{92}$$

which is independent of K as long as we are withing the $\mathcal{G}^{\mathrm{lqr}}$ set. Substituting the values of λ_1 and λ_2 finishes the proof. ∎

Finally, note that for the noisy dynamics setting, if we let

$$\mathcal{G}^{\mathrm{lqr}}_{\mathrm{dyn}} = \{K \mid \mathcal{C}_{\mathrm{dyn}}(K) - \mathcal{C}_{\mathrm{dyn}}(K^*) \leq 10\mathcal{C}_{\mathrm{dyn}}(K_0)\}, \tag{93}$$

since $\mathcal{C}_{\mathrm{dyn}}(K) = \frac{\gamma}{1-\gamma}\mathcal{C}_{\mathrm{init}}(K)$ due to Lemma 4, this set is the exact same as (10) in the random initialization setting. Therefore, all the bounds leading to (92) hold with exactly the same values for the noisy dynamics case as well.

The exponential decay bound established in Lemma 22 plays a crucial role in bounding the gradient estimate under the noisy dynamics setup. We now turn to this estimate, show that it remains unbiased and admits similar concentration bounds in this setting, and re-establish the main convergence guarantees for both standard and minibatched policy updates.

B.2 Gradient Estimation and Convergence Results

Suppose $K \in \mathcal{G}^{\mathrm{lqr}}_{\mathrm{dyn}}$, with $\mathcal{G}^{\mathrm{lqr}}_{\mathrm{dyn}}$ defined in (93). Now let us define $Q^K_{\mathrm{dyn}}(x_{\hat{t}}, u_{\hat{t}})$ as

$$Q^K_{\mathrm{dyn}}(x_{\hat{t}}, u_{\hat{t}}) := x_{\hat{t}}^\top Q x_{\hat{t}} + u_{\hat{t}}^\top R u_{\hat{t}} + \sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}} x_t^\top (Q + K^\top R K) x_t,$$

where

$$x_{\hat{t}+1} = A x_{\hat{t}} + B u_{\hat{t}} + z_{\hat{t}},$$

and

$$x_{t+1} = (A - BK)x_t + z_t,$$

for all $t \neq \hat{t}$, with $x_0 = 0$ and i.i.d. additive noise sequence $z_t \sim \mathcal{D}$ for all t. As a result, for every $t \geq \hat{t} + 1$,

$$x_t = (A - BK)^{t-\hat{t}-1}\big(Ax_{\hat{t}} + Bu_{\hat{t}}\big) + \sum_{i=0}^{t-\hat{t}-1} (A - BK)^{t-\hat{t}-1-i} z_{\hat{t}+i},$$

43

which is affine in $u_{\hat{t}}$. Combining this with the fact that each stage cost $x_t^\top (Q + K^\top RK)x_t$ is quadratic in x_t yields a quadratic function of $u_{\hat{t}}$. Therefore,

$$Q_{\mathrm{dyn}}^K(x_{\hat{t}}, u_{\hat{t}}) = \underbrace{x_{\hat{t}}^\top Q x_{\hat{t}}}_{\text{independent of } u_{\hat{t}}} + \underbrace{u_{\hat{t}}^\top R u_{\hat{t}}}_{\text{quadratic in } u_{\hat{t}}} + \underbrace{\sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}} x_t^\top (Q + K^\top RK)x_t}_{\text{quadratic in } u_{\hat{t}}}$$

is quadratic in $u_{\hat{t}}$, satisfying the condtion in Remark 10. Following this, we have that the gradient estimate

$$\widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}(K) := -\frac{1}{\sigma(1-\gamma)} Q_{\mathrm{dyn}}^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}})\eta_{\hat{t}} x_{\hat{t}}^\top \tag{94}$$

satisfies

Corollary 23 *Suppose $\hat{t} \sim \mu_\gamma$ and $\eta_{\hat{t}} \sim \mathcal{N}(0, I_m)$ as before. Then for any given K,*

$$\mathbb{E}[\widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}(K)] = \nabla \mathcal{C}_{\mathrm{dyn}}(K).$$

The proof of this is a direct consequence of Remark 10. We now introduce a result similar to 12 where we provide some bounds on this gradient estimate in the noisy dynamics setting.

Lemma 24 *Suppose $\delta \in (0, \frac{1}{e}]$, and γ is chosen as in Lemma 22. Then for any $K \in \mathcal{G}^{lqr}$, we have that*

$$\|\widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}(K)\|_F \le \frac{\tilde{\xi}_3}{1-\gamma}\left(\log\frac{1}{\delta}\right)^{3/2}$$

with probability at least $1 - \delta$, where $\tilde{\xi}_1, \tilde{\xi}_2, \tilde{\xi}_3 \in \mathbb{R}$ are given by

$$\tilde{\xi}_1 := \frac{M^3 C_m^{3/2}}{(1-r)^3}\left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\gamma\left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2\right)\frac{(M^2 r + 2)^2}{1-\gamma}\right)$$

$$\tilde{\xi}_2 := \frac{2M C_m^{1/2}}{1-r}\left(\|R\| + \gamma\left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2\right)\frac{M^2\|B\|^2}{1-\gamma}\right)$$

$$\tilde{\xi}_3 := \frac{1}{\sigma}\left(\tilde{\xi}_1 5^{1/2} m^{1/2}\right) + \sigma\left(\tilde{\xi}_2 5^{3/2} m^{3/2}\right),$$

where M and r are defined in (90). Moreover,

$$\mathbb{E}\|\widehat{\nabla \mathcal{C}}(K)\|_F^2 \le \frac{\tilde{\xi}_4}{(1-\gamma)^2},$$

where

$$\tilde{\xi}_4 := \frac{1}{\sigma^2}\tilde{\xi}_1^2 m + 2\tilde{\xi}_1\tilde{\xi}_2 m(m+2) + \sigma^2\tilde{\xi}_2^2 m(m+2)(m+4).$$

Proof First, note that since $x_0 = 0$ in this setting, it holds that

$$x_{\hat{t}} = \sum_{i=0}^{\hat{t}-1}(A - BK)^i z_{\hat{t}-1-i},$$

and hence,

$$\|x_{\hat{t}}\| \leq \sum_{i=0}^{\hat{t}-1} \|(A-BK)^i\|\|z_{\hat{t}-1-i}\| \overset{(i)}{\leq} \sum_{i=0}^{\hat{t}-1}(Mr^i)C_m^{1/2} \leq MC_m^{1/2}\sum_{i=0}^{\infty}r^i = \frac{MC_m^{1/2}}{1-r}, \qquad (95)$$

where (i) follows from Lemma 22 and assumption (4) on the additive noise. Moreover, we have

$$x_{\hat{t}+1} = (A-BK)x_{\hat{t}} + \sigma B\eta_{\hat{t}} + z_{\hat{t}},$$

and thus,

$$\|x_{\hat{t}+1}\| \leq \|A-BK\|\|x_{\hat{t}}\| + \sigma\|B\|\|\eta_{\hat{t}}\| + C_m^{1/2} \leq (Mr)\frac{MC_m^{1/2}}{1-r} + \sigma\|B\|\|\eta_{\hat{t}}\| + C_m^{1/2}. \qquad (96)$$

Additionally, for all $t \geq \hat{t}+1$, we can write

$$x_t = (A-BK)^{t-\hat{t}-1}x_{\hat{t}+1} + \sum_{i=0}^{t-\hat{t}-2}(A-BK)^i z_{t-1-i},$$

and hence,

$$
\begin{aligned}
\|x_t\| &\leq M\|x_{\hat{t}+1}\| + \sum_{i=0}^{t-\hat{t}-2}(Mr^i)C_m^{1/2} \\
&\overset{(i)}{\leq} M\left(\frac{M^2 r C_m^{1/2}}{1-r} + \sigma\|B\|\|\eta_{\hat{t}}\| + C_m^{1/2}\right) + \frac{MC_m^{1/2}}{1-r} \\
&\leq \frac{MC_m^{1/2}(M^2 r+2)}{1-r} + \sigma M\|B\|\|\eta_{\hat{t}}\|,
\end{aligned}
\qquad (97)
$$

45

where (i) follows from (96). We are now in a position to show the following upper bound:

$$Q_{\text{dyn}}^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}})$$

$$= x_{\hat{t}}^\top Q x_{\hat{t}} + (-Kx_{\hat{t}} + \sigma\eta_{\hat{t}})^\top R(-Kx_{\hat{t}} + \sigma\eta_{\hat{t}}) + \sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}} x_t^\top (Q + K^\top RK)x_t$$

$$\leq \|Q\| \|x_{\hat{t}}\|^2 + \|R\| \| -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}\|^2 + \sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}} \left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) \|x_t\|^2$$

$$\overset{(i)}{\leq} \|Q\| \|x_{\hat{t}}\|^2 + 2\|R\| \left(\|K\|^2 \|x_{\hat{t}}\|^2 + \sigma^2 \|\eta_{\hat{t}}\|^2 \right)$$

$$+ \sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}} \left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) \left(\frac{MC_m^{1/2}(M^2r + 2)}{1-r} + \sigma M\|B\|\|\eta_{\hat{t}}\| \right)^2$$

$$\overset{(ii)}{\leq} \|Q\|\frac{M^2C_m}{(1-r)^2} + 2\|R\| \left(\widetilde{c_{K_1}}^2 \frac{M^2C_m}{(1-r)^2} + \sigma^2 \|\eta_{\hat{t}}\|^2 \right)$$

$$+ 2\left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) \left(\frac{M^2C_m(M^2r + 2)^2}{(1-r)^2} + \sigma^2 M^2 \|B\|^2 \|\eta_{\hat{t}}\|^2 \right) \sum_{t=\hat{t}+1}^{\infty} \gamma^{t-\hat{t}}$$

$$= \frac{M^2C_m}{(1-r)^2} \left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) (M^2r + 2)^2 \frac{\gamma}{1-\gamma} \right)$$

$$+ 2\sigma^2 \left(\|R\| + \left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) M^2 \|B\|^2 \frac{\gamma}{1-\gamma} \right) \|\eta_{\hat{t}}\|^2, \tag{98}$$

where (i) follows from (97) and (ii) from (95). Combining (95) and (98), we have

$$\|\widehat{\nabla\mathcal{C}_{\text{dyn}}}(K)\|_F$$

$$\leq \frac{1}{\sigma(1-\gamma)} Q_{\text{dyn}}^K(x_{\hat{t}}, -Kx_{\hat{t}} + \sigma\eta_{\hat{t}}) \|x_{\hat{t}}\| \|\eta_{\hat{t}}\|$$

$$\overset{(i)}{\leq} \frac{1}{\sigma(1-\gamma)} \frac{M^3 C_m^{3/2}}{(1-r)^3} \left(\|Q\| + 2\|R\|\widetilde{c_{K_1}}^2 + 2\gamma \left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) \frac{(M^2r + 2)^2}{1-\gamma} \right) \|\eta_{\hat{t}}\|$$

$$+ \frac{\sigma}{1-\gamma} \frac{2MC_m^{1/2}}{1-r} \left(\|R\| + \gamma \left(\|Q\| + \|R\|\widetilde{c_{K_1}}^2 \right) \frac{M^2 \|B\|^2}{1-\gamma} \right) \|\eta_{\hat{t}}\|^3$$

$$= \frac{1}{1-\gamma} \left(\frac{1}{\sigma} \tilde{\xi}_1 \|\eta_{\hat{t}}\| + \sigma\tilde{\xi}_2 \|\eta_{\hat{t}}\|^3 \right),$$

which resembles the expression of the bound (37) shown for the random initialization setting. Therefore, the rest of the proof follows exactly like that of Lemma 12, after substituting ξ_1, ξ_2 with $\tilde{\xi}_1, \tilde{\xi}_2$ respectively. ∎

Now note that as a consequence of Lemma 4, and as also pointed out in Malik et al. (2020), $\mathcal{C}_{\text{dyn}}(K)$ is also $(\frac{\gamma}{1-\gamma}\phi_K, \beta_K)$ locally smooth, $(\frac{\gamma}{1-\gamma}\lambda_K, \zeta_K)$ locally Lipschitz, and globally

$\frac{\gamma}{1-\gamma}\mu_{\text{lqr}}$-PL. Now similar to before, we recall $\omega_K = \min\{\beta_K, \zeta_K\}$ and define the quantities

$$\phi_{\text{dyn}} := \sup_{K \in \mathcal{G}_{\text{dyn}}^{\text{lqr}}} \frac{\gamma}{1-\gamma}\phi_K = \frac{\gamma}{1-\gamma} \sup_{K \in \mathcal{G}^{\text{lqr}}} \phi_K = \frac{\gamma}{1-\gamma}\phi_{\text{lqr}},$$

$$\lambda_{\text{dyn}} := \sup_{K \in \mathcal{G}_{\text{dyn}}^{\text{lqr}}} \frac{\gamma}{1-\gamma}\lambda_K = \frac{\gamma}{1-\gamma} \sup_{K \in \mathcal{G}^{\text{lqr}}} \lambda_K = \frac{\gamma}{1-\gamma}\lambda_{\text{lqr}},$$

$$\omega_{\text{dyn}} := \inf_{K \in \mathcal{G}_{\text{dyn}}^{\text{lqr}}} \omega_K = \inf_{K \in \mathcal{G}^{\text{lqr}}} \omega_K = \omega_{\text{lqr}},$$

$$\mu_{\text{dyn}} := \frac{\gamma}{1-\gamma}\mu_{\text{lqr}},$$

where the equalities in the first three lines follow from $\mathcal{G}_{\text{dyn}}^{\text{lqr}} = \mathcal{G}^{\text{lqr}}$, which holds due to the cost equivalence shown in Lemma 4. Building on this, along with employing Corollary 23 and Lemma 24, we have all the necessary tools to provide the equivalent convergence result of Theorem 15 for the noisy dynamics setting.

Corollary 25 *Suppose K_0 is stable and γ is as suggested in Lemma 22, and the update rule follows*

$$K_{t+1} = K_t - \alpha_t \widehat{\nabla \mathcal{C}_{\text{dyn}}}(K_t). \tag{99}$$

If the step-size α_t is chosen as

$$\alpha_t = \frac{2}{\mu_{dyn}}\frac{1}{t+N} \quad for \quad N = \max\left\{N_1, \frac{2}{\mu_{dyn}}\frac{\tilde{\xi}_3 \left(\log \frac{1}{\delta}\right)^{3/2}}{(1-\gamma)\omega_{dyn}}\right\},$$

where

$$N_1 = \max\left\{2, \frac{4\phi_{dyn}\tilde{\xi}_4}{\mu_{dyn}^2(1-\gamma)^2}\frac{2}{\mathcal{C}_{\text{dyn}}(K_0)}\right\},$$

then for a given error tolerance ε such that $\mathcal{C}_{\text{dyn}}(K_0) \geq \frac{\varepsilon}{20}$, and δ chosen arbitrarily to satisfy

$$\delta \leq \min\left\{2 \times 10^{-5}, \left(\frac{\phi_{dyn}\tilde{\xi}_4\omega_{dyn}}{960\tilde{\xi}_3^2\widetilde{c_{K_1}}\mathcal{C}_{\text{dyn}}(K_0)}\right)^3 \varepsilon^3, \right.$$

$$\left. \left(\frac{\phi_{dyn}\tilde{\xi}_4}{480(1-\gamma)\mu_{dyn}\tilde{\xi}_3\widetilde{c_{K_1}}N_1\mathcal{C}_{\text{dyn}}(K_0)}\right)^3 \varepsilon^3, \left(\frac{\mu_{dyn}(1-\gamma)}{240\tilde{\xi}_3\widetilde{c_{K_1}}}\right)^3 \varepsilon^3\right\},$$

the iterate K_T of (99) after

$$T = \frac{40}{\varepsilon}N\mathcal{C}_{\text{dyn}}(K_0)$$

steps satisfies

$$\mathcal{C}(K_T) - \mathcal{C}(K^*) \leq \varepsilon$$

with a probability of at least $4/5 - \delta T$.

Furthermore, we can also extend the minibatched gradient estimation argument to this setting. Let

$$\overline{\nabla \mathcal{C}_{\mathrm{dyn}}}_{N_s}(K) := \frac{1}{N_s} \sum_{i=1}^{N_s} \widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}_i(K),$$

where each $\widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}_i(K)$ is an i.i.d. copy of $\widehat{\nabla \mathcal{C}_{\mathrm{dyn}}}(K)$ in (94). We are now in a position to provide a convergence result similar to Theorem 20 for the noisy dynamics setting.

Corollary 26 *Suppose K_0 is stable, γ is as suggested in Lemma 22, and the update rule follows*

$$K_{t+1} = K_t - \alpha \overline{\nabla \mathcal{C}_{\mathrm{dyn}}}_{N_s}(K_t) \tag{100}$$

with a constant step-size α satisfying

$$\alpha \le \min \left\{ \frac{\omega_{dyn}}{\widetilde{c_{K_1}} + \sqrt{\frac{\mu_{dyn} \mathcal{C}_{\mathrm{dyn}}(K_0)}{8}}}, \frac{1}{4\phi_{dyn}}, \frac{4}{\mu_{dyn}} \right\}.$$

Then for a given error tolerance $\varepsilon \in (0, \mathcal{C}_{\mathrm{dyn}}(K_0)]$, and for any $\delta \le \min \left\{ e^{-3/2}, \frac{1-\gamma}{3\tilde{\xi}_3} \sqrt{\frac{\mu_{dyn}\varepsilon}{8}} \right\}$, the update rule (100), with $N_s \sim \widetilde{\mathcal{O}}(1/\varepsilon)$ chosen according to

$$N_s \ge \left\lceil \max \left\{ 5000, 8 \left(\log \frac{2}{\delta} \right)^3, \frac{2048 \tilde{\xi}_3^2}{9(1-\gamma)^2 \mu_{dyn}} \frac{1}{\varepsilon} \left(\log \frac{2(mn+1)}{\delta} \right)^2, \right. \right.$$
$$\left. \left. \frac{128 \tilde{\xi}_4}{\mu_{dyn}(1-\gamma)^2} \frac{1}{\varepsilon} \log \frac{2(mn+1)}{\delta} \right\} \right\rceil = \widetilde{\mathcal{O}} \left(\frac{1}{\varepsilon} \right),$$

guarantees that after

$$T = \frac{4}{\alpha \mu_{dyn}} \log \left(\frac{2\mathcal{C}_{\mathrm{dyn}}(K_0)}{\varepsilon} \right)$$

iterations, we have

$$\mathcal{C}_{\mathrm{dyn}}(K_T) - \mathcal{C}_{\mathrm{dyn}}(K^*) \le \varepsilon,$$

with a probability of at least $1 - \delta T$.

References

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*, pages 1–26, 2011.

M. Akbari, B. Gharesifard, and T. Linder. Logarithmic regret in adaptive control of noisy linear quadratic regulator systems using hints. *Journal of Machine Learning Research*, 2022. submitted, arXiv preprint arXiv:2210.16303.

D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995. ISBN 1886529124.

S. Bittanti and M. C. Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

M. C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

H. Chen and L. Guo. Optimal adaptive control and consistent parameter estimates for ARMAX model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4): 845–867, 1987.

H. Chen and J. Zhang. Identification and adaptive control for systems with unknown orders, delay, and coefficients. *IEEE Transactions on Automatic Control*, 35(8):866–877, 1990.

A. Cohen, T. Koren, and Y. Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. Proceedings of Machine Learning Research, 2019.

S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 633–679, 2020.

M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

C. Ju, G. Kotsalis, and G. Lan. A model-free first-order method for linear quadratic regulator with $\widetilde{O}(1/\varepsilon)$ sampling complexity, 2023.

R. E. Kalman et al. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, 5(2):102–119, 1960.

T. Lai and C. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

E. B. Lee and L. Markus. *Foundations of optimal control theory*. Wiley, New York, 1967.

J. R. Leveque and L. N. Trefethen. On the resolvent condition in the Kreiss matrix theorem. *BIT Numerical Mathematics*, 24:584–591, 1984.

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. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

H. Mania, S. Tu, and B. Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, 2022.

R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz. Stability analysis of discrete-time infinite-horizon optimal control with discounted cost. *IEEE Transactions on Automatic Control*, 62(6):2736–2749, 2017. doi: 10.1109/TAC.2016.2616644.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 387–395. PMLR, 22–24 Jun 2014.

M. Simchowitz and D. Foster. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pages 8937–8948. Proceedings of Machine Learning Research, 2020.

M. N. Spijker. On a conjecture by le Veque and Trefethen related to the Kreiss matrix theorem. *BIT Numerical Mathematics*, 9:551–555, 1991.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. ISSN 00905364.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and trends in machine learning*, 8(1-2), 2015. ISSN 1935-8237.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Z. Yang, Y. Chen, M. Hong, and Z. Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Neural Information Processing Systems*, 2019.

M. Zhou and J. Lu. Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24(222): 1–34, 2023.