Simple Policies for Long Horizons: Policy Gradient in Finite-Horizon LQ Control

Zhijie Chen* Anran Hu†

Abstract

Finite-horizon linear-quadratic (LQ) control generally requires time-varying policies, whose complexity grows with the horizon. We study a simpler alternative that restricts learning to a single stationary linear feedback policy. We show that this restriction is asymptotically cost-free for long horizons: stationary policies achieve an $O(H^{-1})$ performance gap relative to the finite-horizon optimum. By establishing finite-horizon-to-average-cost approximations for both costs and policy gradients, we connect the finite-horizon learning problem to the infinite-horizon average-cost LQ problem and obtain convergence guarantees in both known-model and model-free settings. The stationary parameterization has a horizon-independent policy dimension and more favorable horizon dependence in the theoretical convergence bound, with an $O(H^{-1})$ approximation error as the trade-off. Numerical experiments on synthetic LQR and optimal liquidation demonstrate the computational advantages of learning stationary policies in long-horizon problems.

1 Introduction

The linear-quadratic regulator (LQR) is a classical model in optimal control and reinforcement learning. Its linear dynamics and quadratic cost structure make it analytically tractable while retaining many of the fundamental challenges of sequential decision making. LQ control can be formulated in both discrete and continuous time, and over finite or infinite horizons. When the system dynamics are known, optimal feedback policies are characterized by Riccati equations [3, 5]. LQ models also arise in applications such as optimal trade execution [2].

When the dynamics are unknown, LQR has become an important benchmark for learning-based control. In discrete time, a substantial literature studies model-based and adaptive-control approaches, where the learner estimates the system dynamics or updates the controller online, with performance characterized through regret or sample-complexity guarantees [1, 10, 18, 8]. Related learning problems have also been studied in continuous time. In particular, logarithmic-regret guarantees have been established for episodic continuous-time finite-horizon LQ reinforcement learning with unknown coefficients [4].

A complementary line of work optimizes the controller directly rather than first identifying the underlying dynamics. Policy-gradient methods optimize over a parameterized policy class, and global convergence guarantees have been established for infinite-horizon discrete-time LQR despite the nonconvexity of the policy-optimization problem [11, 20]. Related policy-optimization methods have also been studied for continuous-time LQ control [7]. When the model is unknown, zeroth-order and derivative-free methods further allow descent directions to be estimated directly from rollout costs [17].

*Columbia University, Department of Industrial Engineering and Operations Research; zc2807@columbia.edu.
†Columbia University, Department of Industrial Engineering and Operations Research; ah4277@columbia.edu.

For finite-horizon noisy LQR, policy-gradient convergence has been established in both known- and unknown-model settings [16]; related finite-horizon exploratory LQ problems have also been studied in continuous time [13]. A key difference from the infinite-horizon setting, however, is that the finite-horizon optimal control is generally time-varying. A direct policy-gradient method therefore optimizes over a sequence of feedback matrices $\{K_h\}_{h=0}^{H-1}$, whose number of parameters grows linearly with the horizon.

This growth is particularly relevant from a learning perspective. A stationary linear policy requires learning only a single feedback matrix, whereas a general time-varying linear policy contains H feedback matrices. Thus, the number of policy parameters grows linearly with H for the time-varying parameterization, while remaining fixed for the stationary one. In the model-free setting, the simpler parameterization also reduces the cost of gradient estimation. A stationary policy has only one feedback matrix whose gradient needs to be estimated, whereas a time-varying policy has H feedback matrices and requires estimating a separate gradient component for each time step. Thus, the sampling burden of the time-varying parameterization grows with the horizon, while that of the stationary parameterization does not. Hence, for long horizons, the additional flexibility of a time-varying policy comes with an increasingly larger optimization and sampling burden.

This motivates the central question of this paper: *when the horizon is long, can a much simpler stationary policy still achieve near-optimal finite-horizon performance?* We study stationary linear feedback policies $u_h = -Kx_h$, which we call *simple policies*. As discussed above, this restriction offers substantial advantages in optimization and gradient estimation for long horizons. The central issue is therefore to understand the trade-off between this reduction in learning complexity and the loss of flexibility relative to the full time-varying policy class.

The stationary restriction is not directly handled by standard finite-horizon dynamic programming, which naturally produces a time-varying optimal policy. By contrast, the infinite-horizon average-cost LQR admits a stationary optimal control and has a well-understood policy-optimization geometry [11, 20]. This suggests using the average-cost problem as an analytical surrogate for the long-horizon finite-horizon problem.

This viewpoint is inspired in part by the fictitious-discount framework in [14], which studies stationary-policy approximations and infinite-horizon surrogate formulations for finite-horizon episodic reinforcement learning of finite-state-action Markov Decision Processes. In a similar spirit, we optimize over a single stationary feedback control for the finite-horizon LQR and analyze it through the infinite-horizon average-cost problem. The central question is whether the loss from this stationary restriction becomes negligible as the horizon grows.

Our analysis develops a finite-horizon-to-average-cost bridge that makes this idea precise. We show that the average-cost optimal stationary policy is $O(H^{-1})$-optimal for the finite-horizon time-average problem, and that finite-horizon costs and policy gradients of stabilizing stationary policies approach their average-cost counterparts at the same order. This allows the favorable policy-optimization structure of the average-cost LQR to be transferred to finite-horizon simple policy gradient. We establish convergence guarantees in both known-model and model-free settings, with the latter using a one-point zeroth-order estimator based only on rollout costs.

An important consequence is that the convergence rate of the simple-policy method does not deteriorate as the horizon increases in the same way as the direct time-varying parameterization. Our sufficient contraction factor for simple policy gradient has no explicit dependence on H. In contrast, under uniformly bounded and nondegenerate state second moments, the corresponding contraction factor in the time-varying analysis scales as $1-O(H^{-1})$ for step sizes of the same order. Thus, the theoretical contraction bound favors the simple-policy method by a factor of order H. The trade-off is approximation: the stationary policy class does not generally contain the finite-horizon optimum, and our convergence guarantee therefore has an $O(H^{-1})$ residual error. Importantly, this

approximation error decreases as the horizon grows, precisely in the regime where the optimization and learning advantages of the stationary parameterization become most pronounced.

We complement the theory with numerical experiments comparing simple policy gradient with a time-varying policy-gradient benchmark. The experiments include a synthetic LQR and an optimal liquidation application. They show that the stationary parameterization becomes increasingly attractive as the horizon grows, achieving competitive finite-horizon performance while substantially reducing the optimization and computational burden.

Organization. The rest of the paper is organized as follows. Section 2 introduces the LQR problem. Section 3 establishes the finite-to-average bridge. Section 4 presents the convergence analysis of simple policy gradient in the known-model and unknown-model settings. Section 5 reports the numerical experiments, and Section 6 concludes.

Notation. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices, while $\|\cdot\|_F$ denotes the Frobenius norm.

2 Problem set-up

In this section, we introduce the finite-horizon and infinite-horizon average-cost LQR problems considered in this paper. We consider the discrete-time, time-homogeneous stochastic linear system

$$x_{h+1} = Ax_h + Bu_h + w_h, \qquad h \geq 0, \tag{2.1}$$

where $x_h \in \mathbb{R}^d$ is the state, $u_h \in \mathbb{R}^k$ is the control, $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ are system matrices, and $w_h \in \mathbb{R}^d$ is the process noise.

For a finite horizon $H \geq 1$, a finite-horizon policy is a sequence of Markov feedback maps [1] $\pi = \{\pi_h\}_{h=0}^{H-1}$, where $\pi_h : \mathbb{R}^d \to \mathbb{R}^k$ and $u_h = \pi_h(x_h)$. We denote the policy class by Π_H. The finite-horizon time-average LQR objective under a policy $\pi \in \Pi_H$ is

$$J_H(\pi) := \frac{1}{H}\mathbb{E}^\pi\left[\sum_{h=0}^{H-1}\left(x_h^\top Q x_h + u_h^\top R u_h\right)\right]. \tag{2.2}$$

We denote the optimal finite-horizon cost by $C_H^\star := \inf_\pi J_H(\pi)$.

Our main interest is in stationary linear feedback policies of the form

$$u_h = -Kx_h, \qquad K \in \mathbb{R}^{k\times d},$$

which we refer to as *simple policies*. For such a policy, we write

$$C_H(K) := \frac{1}{H}\mathbb{E}\left[\sum_{h=0}^{H-1} x_h^\top\left(Q + K^\top R K\right)x_h\right].$$

For a stationary linear policy K, whenever the limit exists, define the corresponding infinite-horizon average cost by

$$C(K) := \lim_{H\to\infty} C_H(K) = \lim_{H\to\infty}\frac{1}{H}\mathbb{E}\left[\sum_{h=0}^{H-1} x_h^\top\left(Q + K^\top R K\right)x_h\right]. \tag{2.3}$$

[1] Restricting to Markov policies is without loss of optimality for the finite-horizon LQR; the same optimal value is obtained if general adapted controls are allowed.

We next state the basic assumptions for the finite-horizon and average-cost LQR problems. Additional positive-definiteness conditions will be imposed later when needed for the policy-gradient analysis.

Assumption 2.1. *The process noises $\{w_h\}_{h\geq 0}$ are i.i.d., have zero mean and finite second moment, and are independent of the initial state x_0. Specifically,*

$$\mathbb{E}[w_h] = 0, \qquad \mathbb{E}[w_h w_h^\top] = W \succeq 0, \qquad \mathbb{E}[x_0 x_0^\top] = \Sigma_0 \succeq 0.$$

The cost matrices satisfy $Q \succeq 0$ and $R \succ 0$.

For a square matrix M, let

$$\rho(M) := \max\{|\lambda| : \lambda \in \sigma(M)\}$$

denote its spectral radius, where $\sigma(M)$ denotes the spectrum of M. We call M *Schur stable* if $\rho(M) < 1$.

Assumption 2.2. *The pair (A, B) is stabilizable; that is, there exists $K \in \mathbb{R}^{k\times d}$ such that $A - BK$ is Schur stable: $\rho(A - BK) < 1$. We call any such feedback control K* stabilizing.

We denote the set of stabilizing feedback controls by $\mathcal{K} := \left\{K \in \mathbb{R}^{k\times d} : \rho(A - BK) < 1\right\}$. The optimal average cost over stabilizing stationary policies is denoted by $C^\star := \inf_{K\in\mathcal{K}} C(K)$.

Assumption 2.3. *The pair $(A, Q^{1/2})$ is detectable; equivalently, the pair $(A^\top, Q^{1/2})$ is stabilizable.*

Under Assumption 2.1, the finite-horizon LQR problem admits an optimal linear feedback policy, which is generally time-varying. Under Assumptions 2.1, 2.2 and 2.3, the infinite-horizon average-cost LQR admits an optimal stabilizing stationary linear policy. These are classical LQR results; see, e.g., [3, 5, 6]. We recall their Riccati characterizations below.

Lemma 2.1. *Suppose Assumption 2.1 holds. Then the finite-horizon LQR admits an optimal linear feedback policy*

$$u_h^\star = -K_h^\star x_h, \qquad h = 0, \dots, H-1,$$

where

$$K_h^\star = \left(R + B^\top P_{h+1}^\star B\right)^{-1} B^\top P_{h+1}^\star A.$$

The matrices $\{P_h^\star\}_{h=0}^H$ form the unique positive-semidefinite solution of the Riccati difference equation

$$P_h^\star = Q + A^\top P_{h+1}^\star A - A^\top P_{h+1}^\star B \left(R + B^\top P_{h+1}^\star B\right)^{-1} B^\top P_{h+1}^\star A,$$

with terminal condition $P_H^\star = 0$.

Lemma 2.2. *Under Assumptions 2.1, 2.2 and 2.3,, the infinite-horizon average-cost LQR admits an optimal stabilizing stationary linear policy*

$$u_h^\star = -K^\star x_h,$$

where

$$K^\star = \left(R + B^\top P^\star B\right)^{-1} B^\top P^\star A.$$

Here $P^\star \succeq 0$ is the unique stabilizing solution of the discrete algebraic Riccati equation

$$P^\star = Q + A^\top P^\star A - A^\top P^\star B \left(R + B^\top P^\star B\right)^{-1} B^\top P^\star A. \tag{2.4}$$

Equivalently, $P^\star$ satisfies

$$P^\star = Q + (K^\star)^\top R K^\star + (A - BK^\star)^\top P^\star (A - BK^\star). \tag{2.5}$$

3 Bridging finite-horizon LQR and average-cost LQR

In this section, we establish the connection between the finite-horizon time-average LQR and the infinite-horizon average-cost LQR. We first show that the average-cost optimal stationary policy is $O(H^{-1})$-optimal for the finite-horizon problem. As a consequence, restricting the finite-horizon problem to stationary linear policies is asymptotically cost-free as $H \to \infty$. We then show that, for any stabilizing stationary policy, both the finite-horizon cost and its policy gradient converge to their average-cost counterparts at rate $O(H^{-1})$. These results form the finite-to-average bridge used in the policy-gradient analysis of Section 4.

3.1 Near-optimality of stationary policies for long horizons

We first quantify the finite-horizon performance loss incurred by using the optimal stationary policy of the infinite-horizon average-cost LQR. Let $K^\star$ denote the average-cost optimal feedback controller introduced in Lemma 2.2. We show that its finite-horizon cost differs from the unrestricted finite-horizon optimum by at most $O(H^{-1})$.

Lemma 3.1. *Under Assumptions 2.1, 2.2 and 2.3, there exists a constant $L_P > 0$, independent of H, such that*

$$0 \le C_H(K^\star) - C_H^\star \le \frac{L_{\mathrm{gap}}}{H}, \tag{3.1}$$

where $L_{\mathrm{gap}} := L_P\big(\mathrm{Tr}(\Sigma_0) + \mathrm{Tr}(W)\big)$.

Proof. For the unrestricted finite-horizon optimum, the standard stochastic LQR representation gives

$$C_H^\star = \frac{1}{H}\left[\mathrm{Tr}(\Sigma_0 P_0^\star) + \sum_{h=0}^{H-1} \mathrm{Tr}(W P_{h+1}^\star)\right]. \tag{3.2}$$

By the closed-loop form of the algebraic Riccati equation,

$$P^\star = Q + (K^\star)^\top R K^\star + (A - BK^\star)^\top P^\star (A - BK^\star).$$

Using $x_{h+1} = (A - BK^\star)x_h + w_h$ and the independence and zero-mean assumptions on w_h, we obtain

$$\mathbb{E}[x_{h+1}^\top P^\star x_{h+1}] = \mathbb{E}\left[x_h^\top (A - BK^\star)^\top P^\star (A - BK^\star) x_h\right] + \mathrm{Tr}(WP^\star).$$

Hence,

$$\mathbb{E}\left[x_h^\top \big(Q + (K^\star)^\top R K^\star\big) x_h\right] = \mathbb{E}[x_h^\top P^\star x_h] - \mathbb{E}[x_{h+1}^\top P^\star x_{h+1}] + \mathrm{Tr}(WP^\star).$$

Summing over $h = 0, \dots, H-1$ yields

$$C_H(K^\star) = \frac{1}{H}\Big[\mathrm{Tr}(\Sigma_0 P^\star) - \mathbb{E}[x_H^\top P^\star x_H] + H\,\mathrm{Tr}(WP^\star)\Big]. \tag{3.3}$$

Subtracting (3.2) from (3.3), and using $P^\star \succeq 0$, gives

$$C_H(K^\star) - C_H^\star \le \frac{1}{H}\left[\mathrm{Tr}(\Sigma_0)\|P^\star - P_0^\star\| + \mathrm{Tr}(W)\sum_{h=0}^{H-1}\|P^\star - P_{h+1}^\star\|\right].$$

By the geometric convergence of the finite-horizon Riccati recursion to the stabilizing solution of the algebraic Riccati equation [15, Corollary 2], there exist constants $c > 0$ and $\varrho \in (0,1)$, independent of H, such that

$$\|P_h^\star - P^\star\| \le c\,\varrho^{\,H-h}, \qquad h = 0, \dots, H.$$

Indeed, by time homogeneity, $P_h^\star$ is the initial Riccati iterate of a finite-horizon problem with remaining horizon $H-h$ and terminal condition zero. Consequently,

$$\sum_{h=0}^{H} \|P_h^\star - P^\star\| \le c \sum_{j=0}^{H} \varrho^j \le \frac{c}{1-\varrho} =: L_P,$$

where L_P is independent of H. Substituting this bound into the preceding inequality gives

$$C_H(K^\star) - C_H^\star \le \frac{L_P}{H}\big(\mathrm{Tr}(\Sigma_0) + \mathrm{Tr}(W)\big),$$

which proves the result. □

In particular, since $K^\star$ is itself a stationary policy, Lemma 3.1 shows that there exists a stationary linear policy whose finite-horizon performance gap is $O(H^{-1})$. Thus, restricting attention to stationary policies is asymptotically cost-free as H goes to ∞.

3.2 Finite-to-average approximation of costs and gradients

Lemma 3.1 establishes that stationary policies are sufficient for near-optimal finite-horizon performance when H is large. To analyze policy-gradient optimization over this stationary policy class, we now compare the finite-horizon stationary-policy objective $C_H(K)$ with the infinite-horizon average-cost objective $C(K)$.

The key observation is that, for a stabilizing controller K, the effect of the initial condition contributes only a transient $O(1)$ term to the total cost. After normalization by the horizon, this yields an $O(H^{-1})$ discrepancy in both the objective and its gradient.

For a stationary policy K, let $\Sigma_h^K := \mathbb{E}[x_h x_h^\top]$ denote the second moment of the state at time h. If K is stabilizing, let Σ_K denote the unique positive-semidefinite solution of the Lyapunov equation

$$\Sigma_K = (A-BK)\Sigma_K(A-BK)^\top + W,$$

and let P_K denote the unique positive-semidefinite solution of

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

Lemma 3.2. *Under Assumptions 2.1, 2.2 and 2.3, let K be stabilizing, and suppose $Q \succ 0$ and $W \succ 0$. Then*

$$|C_H(K) - C(K)| \le \frac{L_{\mathrm{cost}}(K)}{H}, \tag{3.4}$$

where

$$L_{\mathrm{cost}}(K) := \frac{C(K)}{\sigma_{\min}(W)}\left(\mathrm{Tr}(\Sigma_0) + \frac{C(K)}{\sigma_{\min}(Q)}\right).$$

Proof. For notational convenience, write $A_K := A - BK$ and $M_K := Q + K^\top R K$. Under the stationary policy $u_h = -Kx_h$, the state second moments satisfy $\Sigma_{h+1}^K = A_K \Sigma_h^K A_K^\top + W$. Subtracting the Lyapunov equation $\Sigma_K = A_K \Sigma_K A_K^\top + W$ and iterating gives

$$\Sigma_h^K - \Sigma_K = A_K^h(\Sigma_0 - \Sigma_K)(A_K^\top)^h. \tag{3.5}$$

Since K is stabilizing,

$$C_H(K) = \frac{1}{H}\sum_{h=0}^{H-1} \mathrm{Tr}(M_K \Sigma_h^K), \qquad C(K) = \mathrm{Tr}(M_K \Sigma_K).$$

Hence, by (3.5),

$$C_H(K) - C(K) = \frac{1}{H} \sum_{h=0}^{H-1} \mathrm{Tr}\Big(M_K A_K^h (\Sigma_0 - \Sigma_K)(A_K^\top)^h \Big) = \frac{1}{H} \mathrm{Tr}(S_{K,H}(\Sigma_0 - \Sigma_K)) \,,$$

where $S_{K,H} := \sum_{h=0}^{H-1} (A_K^\top)^h M_K A_K^h$.

Since K is stabilizing, the Lyapunov equation for P_K admits the convergent series representation

$$P_K = \sum_{h=0}^{\infty} (A_K^\top)^h M_K A_K^h,$$

and therefore $0 \preceq S_{K,H} \preceq P_K$. It follows that

$$|C_H(K) - C(K)| \le \frac{\|S_{K,H}\|}{H} \big\|\Sigma_0 - \Sigma_K\big\|_* \le \frac{\|P_K\|}{H} \left(\mathrm{Tr}(\Sigma_0) + \mathrm{Tr}(\Sigma_K)\right),$$

where $\|\cdot\|_*$ denotes the nuclear norm and we used $\Sigma_0, \Sigma_K \succeq 0$.

By Lemma A.1,

$$\|P_K\| \le \mathrm{Tr}(P_K) \le \frac{C(K)}{\sigma_{\min}(W)}, \qquad \mathrm{Tr}(\Sigma_K) \le \frac{C(K)}{\sigma_{\min}(Q)}.$$

Combining the above inequalities yields

$$|C_H(K) - C(K)| \le \frac{1}{H} \frac{C(K)}{\sigma_{\min}(W)} \left(\mathrm{Tr}(\Sigma_0) + \frac{C(K)}{\sigma_{\min}(Q)} \right),$$

which proves the result. □

We next establish the corresponding approximation for the policy gradients.

Lemma 3.3. *Under Assumptions 2.1, 2.2 and 2.3, let K be a stabilizing policy and suppose $Q \succ 0$ and $W \succ 0$. Then there exists a constant $L_{\mathrm{grad}}(K) > 0$, independent of H, such that*

$$\|\nabla C_H(K) - \nabla C(K)\|_F \le \frac{L_{\mathrm{grad}}(K)}{H}. \tag{3.6}$$

An explicit choice of $L_{\mathrm{grad}}(K)$ is given in (A.9) *in Appendix A.*

Proof. For a fixed horizon H, let $\{P_{K,h}^{(H)}\}_{h=0}^{H}$ be defined by

$$P_{K,h}^{(H)} = Q + K^\top R K + (A - BK)^\top P_{K,h+1}^{(H)} (A - BK), \qquad P_{K,H}^{(H)} = 0.$$

Thus, $P_{K,h}^{(H)}$ is the finite-horizon cost-to-go matrix associated with using the fixed stationary policy K from time h onward. Define

$$E_{K,h}^{(H)} := \Big(R + B^\top P_{K,h+1}^{(H)} B \Big) K - B^\top P_{K,h+1}^{(H)} A,$$

and

$$E_K := \Big(R + B^\top P_K B \Big) K - B^\top P_K A.$$

Following [16, Lem. 3.5] and [11, Lem. 1], the finite- and infinite-horizon policy gradients are

$$\nabla C_H(K) = \frac{2}{H}\sum_{h=0}^{H-1} E_{K,h}^{(H)}\Sigma_h^K, \qquad \nabla C(K) = 2E_K\Sigma_K.$$

Hence

$$\nabla C_H(K) - \nabla C(K) = \frac{2}{H}\sum_{h=0}^{H-1}\Big[\big(E_{K,h}^{(H)} - E_K\big)\Sigma_h^K + E_K(\Sigma_h^K - \Sigma_K)\Big]. \tag{3.7}$$

The two terms in (3.7) correspond to two different boundary effects. First, subtracting the finite- and infinite-horizon Lyapunov equations gives

$$P_K - P_{K,h+1}^{(H)} = (A_K^\top)^{H-h-1} P_K A_K^{H-h-1},$$

and hence

$$E_{K,h}^{(H)} - E_K = B^\top (A_K^\top)^{H-h-1} P_K A_K^{H-h}. \tag{3.8}$$

Thus, this term captures the terminal-horizon effect.

Second, the state second moments satisfy

$$\Sigma_h^K - \Sigma_K = A_K^h(\Sigma_0 - \Sigma_K)(A_K^\top)^h, \tag{3.9}$$

which captures the transient effect of the initial state.

Since K is stabilizing, the closed-loop powers decay geometrically. More precisely, Lemma A.3 in Appendix A shows that

$$\sum_{h=0}^{\infty} \|A_K^h\|^2 < \infty,$$

with an explicit bound depending only on $C(K)$ and the system parameters. Consequently, the two transient sums in (3.7) are uniformly bounded in H. Therefore, there exists a constant $L_{\mathrm{grad}}(K)$, independent of H, such that

$$\|\nabla C_H(K) - \nabla C(K)\|_F \le \frac{L_{\mathrm{grad}}(K)}{H}.$$

The detailed estimates for the two transient terms and an explicit choice of $L_{\mathrm{grad}}(K)$ are provided in Appendix A; see (A.9).

□

4 Simple Policy Gradient for Finite-Horizon LQR

A direct convergence analysis of simple policy gradient for the finite-horizon LQR problem is challenging because the finite-horizon objective is inherently time dependent, whereas the policy class considered here is restricted to stationary stabilizing feedback controls. We therefore adopt an indirect approach and use the infinite-horizon average-cost LQR problem as an analytical surrogate.

Let $K^\star$ denote the optimal stabilizing controller for the infinite-horizon problem. Our analysis relies on two finite-to-average approximation results. First, Lemma 3.1 shows that $C_H(K^\star) - C_H^\star = O(\frac{1}{H})$, and hence $K^\star$ is near-optimal for the finite-horizon problem when H is sufficiently large. Second, Lemma 3.3 establishes that $\|\nabla C_H(K) - \nabla C(K)\|_F = O(\frac{1}{H})$ over the relevant set of stabilizing policies. Consequently, an update based on the finite-horizon gradient $\nabla C_H(K)$ can be viewed as an inexact policy-gradient update for the infinite-horizon objective $C(K)$, with an approximation error that vanishes as H increases.

This constitutes the main proof strategy for establishing the convergence of simple policy gradient in the finite-horizon setting.

4.1 Known-model case

We first give the policy updating rule and the main convergence theorem.

Algorithm 1 Policy updating with known model

1: Input: K_0, H, η, N
2: **for** $n = 0$ to $N - 1$ **do**
3: Compute $\nabla C_H(K_n)$
4: $K_{n+1} = K_n - \eta \nabla C_H(K_n)$
5: **end for**
6: Return K_N

Let K_0 be a stabilizing policy and set $C_0 := C(K_0)$. Define the average-cost sublevel set

$$\mathcal{K}_0 := \{K \in \mathcal{K} : C(K) \le 2C_0\}.$$

We use the uniform constants

$$\overline{L}_{\mathrm{grad}} := \sup_{K \in \mathcal{K}_0} L_{\mathrm{grad}}(K),$$

and

$$\overline{L}_{\mathrm{cost}} := \frac{2C_0}{\sigma_{\min}(W)} \left(\mathrm{Tr}(\Sigma_0) + \frac{2C_0}{\sigma_{\min}(Q)} \right).$$

By Lemma 3.2 and (A.9), both constants are finite and depend polynomially on C_0 and the problem parameters.

Theorem 4.1. *Under Assumptions 2.1, 2.2 and 2.3, suppose $Q \succ 0$ and $W \succ 0$. Let K_0 be a stabilizing policy and set $C_0 := C(K_0)$. Consider the iterates generated by Algorithm 1. There exist constants $\eta_0, H_0, L_1, L_2 > 0$, independent of H and N, such that for any*

$$0 < \eta \le \eta_0, \qquad H \ge H_0,$$

we have, for every $N \ge 0$,

$$C_H(K_N) - C_H^\star \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + \frac{L_1}{H^2} + \frac{L_2}{H},$$

where $\gamma := \eta \frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}$. Moreover, H_0, L_1, L_2 and η_0^{-1} can be chosen to depend polynomially on C_0 and the system parameters. The dependence of these constants is detailed in the proof and Appendix B.

Proof. Choose η_0 and H_0 as specified in Appendix B. For $0 < \eta \le \eta_0$ and $H \ge H_0$, Lemma D.2 implies that

$$K_n \in \mathcal{K}_0, \qquad n \ge 0.$$

In particular,

$$\|\nabla C_H(K_n) - \nabla C(K_n)\|_F \le \frac{\overline{L}_{\mathrm{grad}}}{H}.$$

Applying Lemma B.3 and using $C(K_n) \le 2C_0$, we obtain

$$C(K_{n+1}) - C(K^\star) \le (1-\gamma)\big(C(K_n) - C(K^\star)\big) + \eta\Theta_0 \frac{\overline{L}_{\mathrm{grad}}^2}{H^2},$$

where $\Theta_0 := \frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}$. Iterating the above inequality gives

$$C(K_N) - C(K^\star) \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + \eta\Theta_0 \frac{\overline{L}_{\text{grad}}^2}{H^2} \sum_{j=0}^{N-1} (1-\gamma)^j.$$

By the choice of η_0, we have $0 < \gamma < 1$, and hence $\sum_{j=0}^{N-1}(1-\gamma)^j \le \frac{1}{\gamma}$. Since $\gamma = \eta \frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}$, we conclude that

$$C(K_N) - C(K^\star) \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + \frac{L_1}{H^2},$$

where one may take

$$L_1 := \Theta_0 \overline{L}_{\text{grad}}^2 \frac{\|\Sigma_{K^\star}\|}{\sigma_{\min}^2(W)\sigma_{\min}(R)}.$$

It remains to transfer the bound to the finite-horizon objective. We write

$$\begin{aligned} C_H(K_N) - C_H^\star = C(K_N) - C(K^\star) &+ \big(C_H(K_N) - C(K_N)\big) \\ &+ \big(C(K^\star) - C_H(K^\star)\big) + \big(C_H(K^\star) - C_H^\star\big). \end{aligned}$$

Since $C(K_N) \le 2C_0$ and $C(K^\star) \le C_0$, Lemma 3.2 gives

$$|C_H(K_N) - C(K_N)| + |C_H(K^\star) - C(K^\star)| \le \frac{2\overline{L}_{\text{cost}}}{H}.$$

Moreover, Lemma 3.1 gives

$$C_H(K^\star) - C_H^\star \le \frac{L_{\text{gap}}}{H}.$$

Therefore,

$$C_H(K_N) - C_H^\star \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + \frac{L_1}{H^2} + \frac{L_2}{H},$$

where

$$L_2 := 2\overline{L}_{\text{cost}} + L_{\text{gap}}.$$

The claimed polynomial dependence follows from the explicit choices of η_0 and H_0 in Appendix B, Lemma 3.2, and the bounds in Appendix A.

□

Remark 4.2 (Comparison with time-varying policy gradient)**.** *For the simple-policy method,*

$$\gamma_{\text{simple}} = \eta_{\text{simple}} \frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|} = \Theta(\eta_{\text{simple}}),$$

with no explicit dependence on H. In contrast, the contraction factor in the time-varying analysis of [16] has the form

$$\gamma_{\text{direct}} = 2\eta_{\text{direct}} \frac{\underline{\sigma}_H^2 \sigma_{\min}(R)}{\left\|\sum_{h=0}^{H-1} \Sigma_h^\star\right\|}, \qquad \underline{\sigma}_H := \min_h \sigma_{\min}(\Sigma_h^\star).$$

If the state second moments are uniformly bounded and nondegenerate, then $\gamma_{\text{direct}} = \Theta(\eta_{\text{direct}}/H)$. Thus, for comparable step sizes, the sufficient contraction bound for the simple-policy method is better by a factor of order H. Its sufficient step-size bound can also be chosen independently of H for sufficiently large H.

The trade-off is the restriction to a stationary policy class: our guarantee contains an approximation error

$$\varepsilon_H = \frac{L_1}{H^2} + \frac{L_2}{H} = O(H^{-1}),$$

whereas the time-varying parameterization directly optimizes the finite-horizon objective. These are sufficient theoretical bounds and need not be tight.

4.2 Unknown-model case

We next consider the model-free setting, in which the system parameters are unknown. Rather than first estimating the system dynamics, we optimize the policy directly from observed rollout costs. Since $\nabla C_H(K)$ is unavailable, we use a zeroth-order gradient estimator based on randomly perturbed policies [9, 19]. We show below that its sampling and smoothing errors can be controlled, allowing the resulting update to be analyzed as an inexact policy-gradient step.

Algorithm 2 Policy update with unknown model

1: Input: K_0, H, η, N, m, r
2: **for** $n = 0, \ldots, N-1$ **do**
3: Compute $\widehat{\nabla}_n$ using Algorithm 3
4: $K_{n+1} = K_n - \eta \widehat{\nabla}_n$
5: **end for**
6: Return K_N

For a single rollout under policy K, define the empirical finite-horizon cost

$$\widehat{C}_H(K) := \frac{1}{H} \sum_{h=0}^{H-1} \left(x_h^\top Q x_h + u_h^\top R u_h \right).$$

Thus, $\widehat{C}_H(K)$ is a single-sample realization of $C_H(K)$. The zeroth-order estimator used in our analysis is given in Algorithm 3. It is adapted from [16, Alg. 1], with the key difference that we perturb a single stationary feedback matrix K, rather than the time-dependent feedback matrices of a non-stationary policy.

Algorithm 3 Policy Gradient Estimation

1: Input: K, H, m, r, $D := kd$
2: **for** $i = 1$ to m **do**
3: Sample a perturbation $U_i \sim \mathrm{Unif}(\mathbb{S}_r)$ ($\mathbb{S}_r := \{U \in \mathbb{R}^{k \times d} : \|U\|_F = r\}$)
4: Roll out one trajectory $\{x_h^{(i)}\}_{h=0}^{H-1}$ under the perturbed policy $K + U_i$
5: Compute

$$\widehat{C}_H(K + U_i) := \frac{1}{H} \sum_{h=0}^{H-1} \left[(x_h^{(i)})^\top Q x_h^{(i)} + (u_h^{(i)})^\top R u_h^{(i)} \right], \qquad u_h^{(i)} = -(K + U_i) x_h^{(i)}.$$

6: **end for**
7: Return

$$\widehat{\nabla} := \frac{1}{m} \sum_{i=1}^{m} \frac{D}{r^2} \widehat{C}_H(K + U_i) U_i \tag{4.1}$$

Algorithm 3 is a one-point zeroth-order estimator. To describe its target, define the ball-smoothed finite-horizon cost function for $r > 0$ as

$$C_H^r(K) := \mathbb{E}_{V\sim \mathrm{Unif}(\mathbb{B}_r)}\left[C_H(K+V)\right], \quad \mathbb{B}_r := \{V \in \mathbb{R}^{k\times d} : \|V\|_F \le r\}.$$

By the standard spherical smoothing identity [12, Lemma 2.1],

$$\nabla C_H^r(K) = \frac{D}{r^2}\mathbb{E}_{U\sim \mathrm{Unif}(S_r)}[C_H(K+U)U], \tag{4.2}$$

where $D := kd$. Thus, Algorithm 3 estimates the gradient of the smoothed objective C_H^r. Since

$$\mathbb{E}[\widehat{C}_H(K+U) \mid U] = C_H(K+U),$$

the estimator in Algorithm 3 satisfies $\mathbb{E}[\widehat{\nabla}] = \nabla C_H^r(K)$. Thus, Algorithm 3 is an unbiased estimator of the gradient of the smoothed objective C_H^r. The difference between $\nabla C_H^r(K)$ and $\nabla C_H(K)$ is controlled by the local perturbation bounds in Appendix C, while concentration of the rollout estimator requires the following assumption.

Assumption 4.1. *For each rollout, the initial state and process noises are independent of the policy perturbation. The initial states, process noises, and policy perturbations used in different rollouts and different iterations are mutually independent. Moreover:*

1. *Initial state: $x_0 = \mu_0 + \widetilde{W}_0 z_0$, where $z_0 \in \mathbb{R}^d$ has independent, mean-zero, sub-Gaussian components with sub-Gaussian parameter σ_0^2, and $\widetilde{W}_0 \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*
2. *Noise process: $w_h = \widetilde{W} v_h$, where $\{v_h\}_{h\ge 0}$ are i.i.d. and independent of x_0. Each $v_h \in \mathbb{R}^d$ has independent, mean-zero, sub-Gaussian components with sub-Gaussian parameter σ_w^2, and $\widetilde{W} \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*

Assumption 4.1 is a mild extension of [16, Assumption 4.3] that allows the initial state to have a nonzero mean μ_0. The additional mean term only introduces a linear sub-Gaussian term in the rollout-cost concentration argument.

We are now ready to state the main model-free convergence result. Unlike the fixed high-probability formulations in [11, 16], we keep the confidence parameter explicit: the guarantee holds for any prescribed $\delta \in (0,1)$, with logarithmic dependence of the sample size on $1/\delta$.

Theorem 4.3. *Under Assumptions 2.1, 2.2, 2.3, and 4.1, suppose $Q \succ 0$ and $W \succ 0$. Let K_0 be stabilizing and set $C_0 := C(K_0)$. There exist constants $e_0, \eta_0^{\mathrm{mf}}, L_{\mathrm{mf}} > 0$, independent of H and N, such that the following holds. For any $N \ge 1$ and any*

$$0 < \zeta \le \frac{e_0}{2}, \qquad 0 < \epsilon \le L_{\mathrm{mf}} e_0^2, \qquad 0 < \delta < 1,$$

there exist thresholds $r_0(\epsilon)$, $H_0(\epsilon)$, and $m_0(\zeta,\delta,N,r)$ such that, for any

$$0 < \eta \le \eta_0^{\mathrm{mf}}, \qquad 0 < r \le r_0(\epsilon), \qquad H \ge H_0(\epsilon), \qquad m \ge m_0(\zeta,\delta,N,r),$$

we have, with probability at least $1-\delta$,

$$C_H(K_N) - C_H^\star \le (1-\gamma)^N\big(C_0 - C(K^\star)\big) + 2L_{\mathrm{mf}}\zeta^2 + \epsilon,$$

where $\gamma := \eta \frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}$. Moreover, e_0, $(\eta_0^{\mathrm{mf}})^{-1}$, and L_{mf} are polynomially bounded in C_0 and the system and noise parameters. The quantities $r_0(\epsilon)^{-1}$, $H_0(\epsilon)$, and $m_0(\zeta,\delta,N,r)$ have polynomial dependence on the corresponding accuracy parameters, while m_0 depends logarithmically on N/δ. Explicit choices are given in Appendix D.

Proof. By Lemma D.3, with probability at least $1-\delta$,

$$K_n \in \mathcal{K}_0, \qquad n = 0, \dots, N.$$

We work on this event. Since $h^f_{\mathrm{grad}}(K_n) \le \overline{h}^f_{\mathrm{grad}}$ and $L_{\mathrm{grad}}(K_n) \le \overline{L}_{\mathrm{grad}}$, Lemma D.2 gives, for $n = 0, \dots, N-1$,

$$C(K_{n+1}) - C(K^\star) \le (1-\gamma)\big(C(K_n) - C(K^\star)\big) + \eta\Theta_0 \left(\zeta + \overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H}\right)^2.$$

Iterating this inequality and using $\sum_{j=0}^{N-1}(1-\gamma)^j \le \frac{1}{\gamma}$ yields

$$C(K_N) - C(K^\star) \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + L_{\mathrm{mf}} \left(\zeta + \overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H}\right)^2.$$

As in the proof of Theorem 4.1, Lemmas 3.1 and 3.2 imply

$$C_H(K_N) - C_H^\star \le C(K_N) - C(K^\star) + \frac{L_2}{H}.$$

By the definitions of $r_0(\epsilon)$ and $H_0(\epsilon)$,

$$\overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H} \le \frac{1}{2}\sqrt{\frac{\epsilon}{L_{\mathrm{mf}}}}, \qquad \frac{L_2}{H} \le \frac{\epsilon}{2}.$$

Hence,

$$\begin{aligned} L_{\mathrm{mf}} \left(\zeta + \overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H}\right)^2 &\le 2L_{\mathrm{mf}}\zeta^2 + 2L_{\mathrm{mf}} \left(\overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H}\right)^2 \\ &\le 2L_{\mathrm{mf}}\zeta^2 + \frac{\epsilon}{2}. \end{aligned}$$

Combining the preceding bounds gives

$$C_H(K_N) - C_H^\star \le (1-\gamma)^N \big(C_0 - C(K^\star)\big) + 2L_{\mathrm{mf}}\zeta^2 + \epsilon,$$

as claimed. □

Remark 4.4. *The stationary parameterization also reduces the sampling cost of the zeroth-order gradient estimator. To make this comparison concrete, consider the coordinate-wise estimator analyzed in [16, Lem. 4.11] for a time-varying policy $\boldsymbol{K} = (K_0, \dots, K_{H-1})$. There, each component $\nabla_h C_H(\boldsymbol{K})$ is estimated using m independently perturbed trajectories, for $h = 0, \dots, H-1$. Thus, estimating the full policy gradient requires mH trajectories per update.*

In contrast, under our stationary parameterization, the same perturbation of K is used at every time step, and the entire gradient $\nabla C_H(K)$ is estimated from m trajectories. Hence, for these two zeroth-order estimators, the number of trajectories per policy-gradient update is reduced by a factor of H. This difference becomes particularly significant for large horizons.

5 Numerical experiments

We compare simple policy gradient with the time-varying policy-gradient benchmark in two examples. The first is a synthetic LQR satisfying the assumptions of our analysis. The second is an optimal liquidation problem, which serves as an application and a stress test outside some of the standing assumptions.

5.1 Synthetic LQR

Set-up. We use the four-dimensional LQR instance from [16, Section 5], with

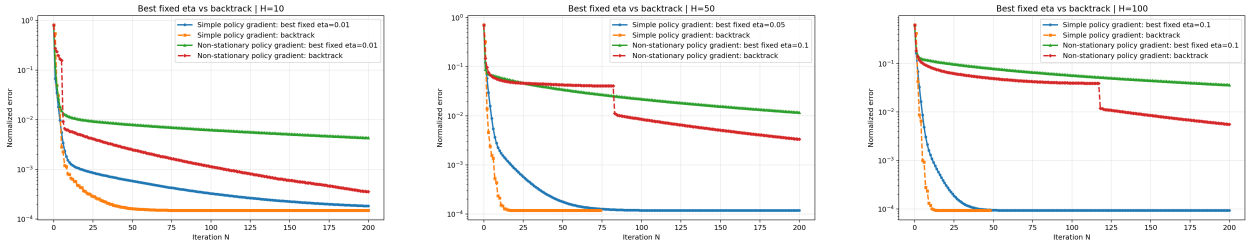
$$A = \begin{bmatrix} 0.5 & 0.05 & 0.1 & 0.2 \\ 0.0 & 0.2 & 0.3 & 0.1 \\ 0.06 & 0.1 & 0.2 & 0.4 \\ 0.05 & 0.2 & 0.15 & 0.1 \end{bmatrix}, B = \begin{bmatrix} -0.05 & -0.01 \\ -0.005 & -0.01 \\ -1.0 & -0.01 \\ -0.01 & -0.9 \end{bmatrix}, Q = \begin{bmatrix} 1.0 & 0.2 & -0.005 & 0.015 \\ 0.2 & 1.1 & 0.15 & 0.0 \\ -0.005 & 0.15 & 0.9 & -0.08 \\ 0.015 & 0.0 & -0.08 & 0.88 \end{bmatrix}, R = \begin{bmatrix} 0.4 & -0.25 \\ -0.25 & 0.7 \end{bmatrix}.$$

We take $\mu_0 = [5.0, 2.0, 8.0, 5.0]^\top$, $\mathrm{cov}(x_0) = \mathrm{diag}(0.1, 0.3, 1.0, 0.5)$, and $W = \mathrm{diag}(0.1, 0.5, 0.2, 0.3)$. The initial policy is chosen as $\{K_0\}_{ij} = 0.05$ for all i, j.

5.1.1 Known-model comparison

Figure 1 compares fixed step sizes with backtracking line search for both policy parameterizations and for $H \in \{10, 50, 100\}$ when the model is known. Backtracking line search gives more reliable descent across all three horizons and is substantially less sensitive to the choice of a single global step size. This becomes particularly useful as H increases, since the optimization geometry of the time-varying parameterization changes with the horizon. We therefore always use backtracking line search in the remaining experiments unless otherwise specified, so that the comparison is not driven by an unfavorable fixed step-size choice for either method.

The figure also already suggests a horizon effect in the comparison between the two policy classes: the relative performance of the simple policy improves as H grows. We investigate this effect more directly in the model-free experiments below.



(a) $H = 10$ (b) $H = 50$ (c) $H = 100$

Figure 1: Comparison of descent methods in both settings and all horizons.

5.1.2 Model-free comparison

We use $m = 200$ rollouts per gradient estimate and tune the smoothing radius over

$$r \in \{0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 20\}$$

for each method and horizon. The selected values are $r = 0.5$ for the simple policy for all three horizons, and $r = 5, 10, 10$ for the time-varying benchmark at $H = 10, 50, 100$, respectively. We report 50 independent runs for $H = 10, 50$ and 25 runs for $H = 100$; solid curves show the mean and shaded regions the 5%–95% quantiles. For both examples, we report the normalized suboptimality

$$\mathrm{Err}(K_n) := \frac{C_H(K_n) - C_H^\star}{C_H(K_0) - C_H^\star}.$$

The benchmark value $C_H^\star$ is computed from the corresponding finite-horizon Riccati recursion. For each method and horizon, the smoothing radius is selected from the above grid using separate pilot

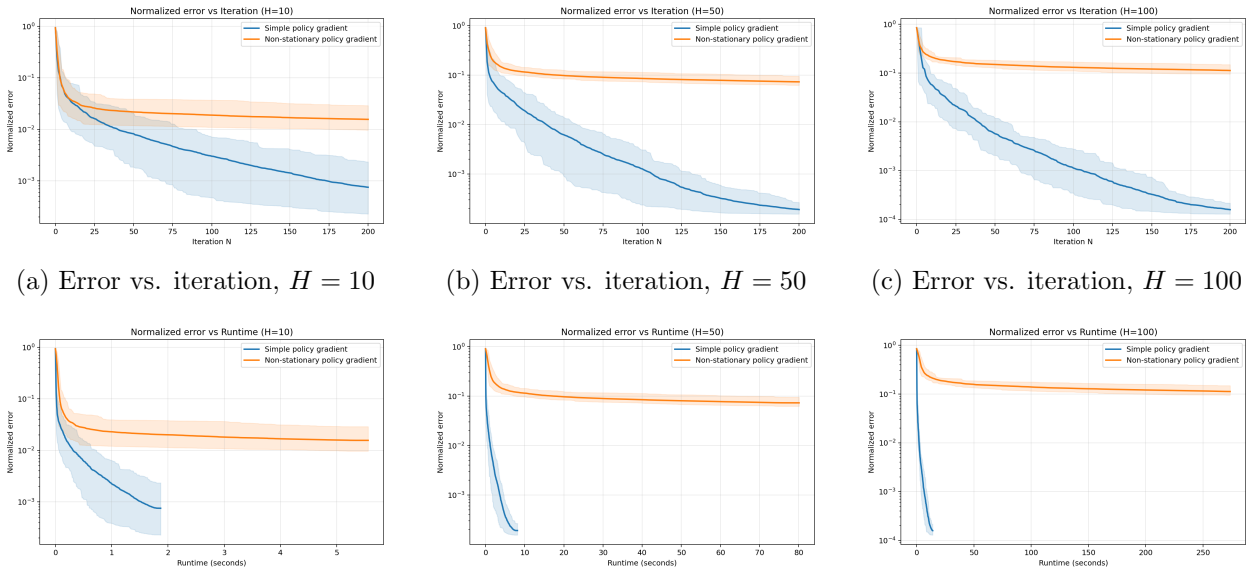
runs, based on the smallest final normalized error median. For the stationary estimator, $m = 200$ denotes the total number of rollouts per update. For the coordinate-wise time-varying estimator, we use $m = 200$ rollouts for each time component, resulting in $200H$ rollouts per full policy-gradient update.

Results. Figure 2 compares the two methods in terms of both optimization iterations and wall-clock runtime. Two features are particularly clear.

First, the advantage of the simple-policy parameterization becomes more pronounced as the horizon increases. For a time-varying policy, the number of feedback matrices to be learned grows linearly with H, whereas the simple policy always optimizes a single matrix K. Thus, increasing the horizon makes the direct policy class progressively higher-dimensional, while leaving the dimension of the simple-policy optimization problem unchanged. The iteration plots show that this difference is not merely a per-iteration implementation issue: for larger H, the simple policy also makes more effective progress in terms of the number of policy updates.

The runtime plots amplify this difference. In the zeroth-order setting, the coordinate-wise time-varying estimator must estimate the gradient of each time-dependent feedback matrix, whereas a single batch of rollouts estimates the gradient of the stationary policy. Consequently, the computational cost of one update grows much more rapidly with H for the time-varying benchmark. This explains why the separation between the methods is substantially more visible when performance is plotted against runtime rather than iteration count.

These results illustrate the tradeoff underlying our analysis. Restricting to a stationary policy reduces the expressiveness of the policy class and can therefore introduce a finite-horizon approximation error. On the other hand, Sections 3 and 4 show that this approximation error becomes small for long horizons, while the optimization problem remains horizon-independent in dimension. The experiments exhibit precisely this regime: as H grows, the cost of restricting the policy class becomes less important, while the computational benefit of optimizing over a single feedback matrix becomes increasingly significant.



(a) Error vs. iteration, $H = 10$ (b) Error vs. iteration, $H = 50$ (c) Error vs. iteration, $H = 100$

(d) Error vs. runtime, $H = 10$ (e) Error vs. runtime, $H = 50$ (f) Error vs. runtime, $H = 100$

Figure 2: Comparison of two methods in different horizons.

5.2 Optimal liquidation

We next consider the discrete-time Almgren–Chriss liquidation example used in [16]. Let S_h denote the asset price, q_h the remaining inventory, and u_h the number of shares sold at time h. The model can be written directly in LQ form as

$$x_{h+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} x_h + \begin{pmatrix} -\gamma \\ -1 \end{pmatrix} u_h + \begin{pmatrix} \sigma Z_{h+1} \\ 0 \end{pmatrix}, \qquad x_h = \begin{pmatrix} S_h \\ q_h \end{pmatrix},$$

where γ is the permanent price-impact coefficient and $Z_h \sim N(0,1)$.

Following the standard quadratic mean–variance formulation [2, 16], the control-dependent part of the objective is

$$\sum_{h=0}^{H-1} \left(\delta u_h^2 + \phi\sigma^2 q_h^2\right) + (\delta + \phi\sigma^2) q_H^2, \qquad \delta := \beta - \frac{\gamma}{2},$$

up to terms independent of the policy. Thus,

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & \phi\sigma^2 \end{pmatrix}, \qquad R = \delta, \qquad Q_H = \begin{pmatrix} 0 & 0 \\ 0 & \delta + \phi\sigma^2 \end{pmatrix}.$$

Setup. We take

$$\beta = 1.03 \times 10^{-5}, \qquad \gamma = 7.27 \times 10^{-6}, \qquad \sigma = 0.107, \qquad \phi = 5 \times 10^{-6},$$

with

$$S_0 \sim N(200, 1), \qquad q_0 \sim N(500, 1), \qquad (K_0)_{ij} = -0.2.$$

The parameters are calibrated from the AAPL data used in [16]. We only study the model-free setup for this problem. We again use $m = 200$ and tune r separately for each method and horizon over

$$\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10\}.$$

The selected radii for the simple policy are $0.05, 0.01, 0.01$ for $H = 10, 50, 100$, respectively, while those for the time-varying benchmark are $0.5, 0.5, 1$.

Relation to the theory. This example is intentionally outside some of the assumptions of our main analysis. In particular, the running state-cost and noise matrices are only positive semidefinite, the terminal cost differs from the running cost, and the liquidation dynamics do not admit a Schur-stabilizing stationary feedback control. We therefore use this example as an empirical stress test rather than as an instance covered directly by the theoretical guarantees.

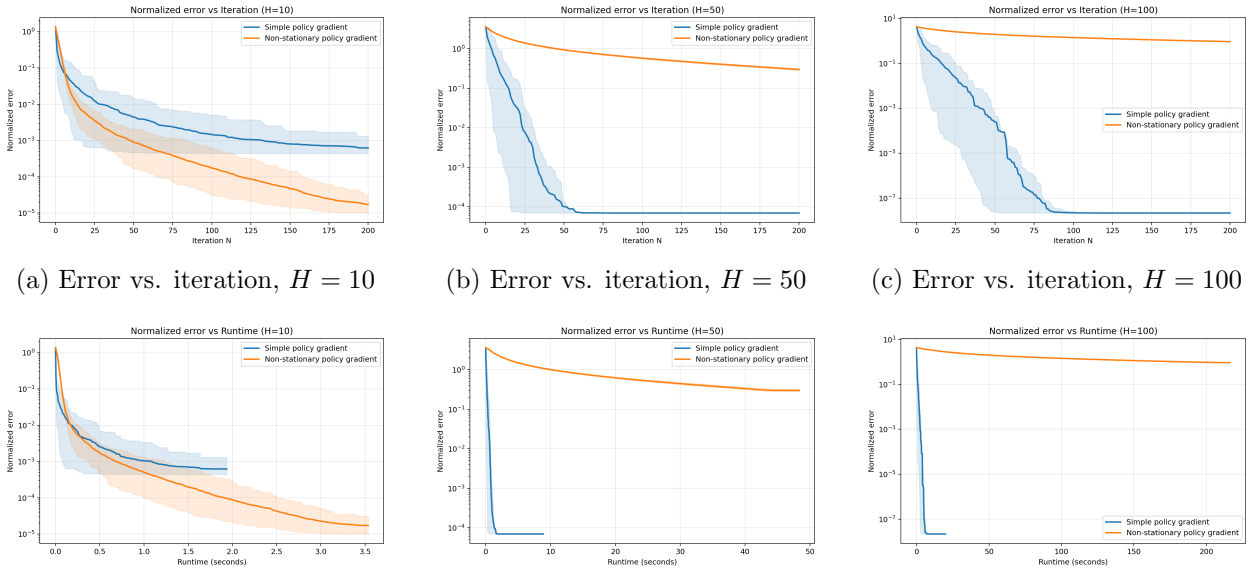
Results. We use 50 independent runs for $H = 10, 50$ and 25 runs for $H = 100$, with the same plotting convention as in the synthetic experiment. Figure 3 shows a more nuanced comparison than the synthetic example. At the shorter horizon, the time-varying policy can be competitive with, and in some cases outperform, the stationary policy. This is natural: when the liquidation horizon is short, boundary and terminal effects are important, and a time-dependent feedback rule has additional flexibility to adapt its behavior as the terminal time approaches. A stationary policy cannot reproduce this time dependence exactly.

The comparison changes as the horizon grows. For $H = 50$ and especially $H = 100$, the benefit of optimizing the much smaller stationary policy class becomes increasingly visible. The time-varying

method must learn a separate feedback control at every time step, while the stationary method continues to optimize only one control. At the same time, the relative importance of the terminal boundary becomes smaller over a long horizon. These two effects work in the same direction: the additional flexibility of the time-varying policy becomes less valuable, while its optimization and sampling burden continues to grow with H.

The distinction between the iteration and runtime plots is also informative. A comparison against iterations isolates, to some extent, the optimization behavior of the two parameterizations, whereas the runtime plots additionally account for the substantially different cost of constructing a policy-gradient estimate. The larger separation in runtime for long horizons therefore reflects both the easier optimization problem and the lower per-update computational cost associated with the stationary parameterization.

Despite these departures from the theoretical setting, the same qualitative horizon effect persists. This suggests that the computational motivation for stationary policies extends beyond the regime covered directly by our guarantees.



(a) Error vs. iteration, $H = 10$ (b) Error vs. iteration, $H = 50$ (c) Error vs. iteration, $H = 100$

(d) Error vs. runtime, $H = 10$ (e) Error vs. runtime, $H = 50$ (f) Error vs. runtime, $H = 100$

Figure 3: Comparison of two methods in different horizons in liquidation application.

6 Conclusion

We studied a simple stationary-policy approach to finite-horizon linear-quadratic control when the horizon is long. Although the finite-horizon optimal policy is generally time-varying, we showed that optimizing over a single stabilizing feedback control incurs only an $O(1/H)$ approximation error. At the same time, the resulting optimization problem has a dimension independent of the horizon, and its policy-gradient convergence rate does not deteriorate with H.

We further developed a model-free zeroth-order policy-gradient method for this stationary parameterization and established high-probability convergence guarantees. The numerical experiments support the same qualitative picture: the additional flexibility of a time-varying policy can be useful for short horizons, while the computational and sampling advantages of the stationary parameterization become increasingly pronounced as the horizon grows.

Several extensions are natural. In particular, it would be interesting to understand when similar long-horizon simplifications hold for time-varying dynamics and costs, more general terminal objectives, and control problems beyond the stabilizing LQ setting.

References

[1] 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*, volume 19 of *Proceedings of Machine Learning Research*, pages 1–26. PMLR, 2011.

[2] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3(2):5–39, 2001.

[3] B. D. O. Anderson and J. B. Moore. *Optimal Control: Linear Quadratic Methods*. Dover Publications, 2007.

[4] M. Basei, X. Guo, A. Hu, and Y. Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 4 edition, 2012.

[6] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 4 edition, 2017.

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

[8] A. Cohen, T. Koren, and Y. Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1300–1309. PMLR, 2019.

[9] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009.

[10] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[11] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 2018.

[12] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[13] M. Giegrich, C. Reisinger, and Y. Zhang. Convergence of policy gradient methods for finite-horizon exploratory linear-quadratic control problems. *SIAM Journal on Control and Optimization*, 62(2):1060–1090, 2024.

[14] X. Guo, A. Hu, and J. Zhang. Theoretical guarantees of fictitious discount algorithms for episodic reinforcement learning and global convergence of policy gradient methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6774–6782, 2022.

[15] W. W. Hager and L. L. Horowitz. Convergence and stability properties of the discrete riccati operator equation and the associated optimal control and filtering problems. *SIAM Journal on Control and Optimization*, 14(2):295–312, 1976.

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

[17] 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.

[18] 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.

[19] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[20] 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 *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Important bounds derivations

In this appendix, we establish several auxiliary bounds for stabilizing stationary policies. These bounds are used in the proof of Lemmas 3.2, 3.3 and in the subsequent policy-gradient analysis.

Lemma A.1. *Suppose K is stabilizing and $Q \succ 0$, $W \succ 0$. Then*

$$\mathrm{Tr}(P_K) \le \frac{C(K)}{\sigma_{\min}(W)}, \qquad \mathrm{Tr}(\Sigma_K) \le \frac{C(K)}{\sigma_{\min}(Q)}.$$

Consequently, $\|P_K\|$, $\|P_K\|_F \le \frac{C(K)}{\sigma_{\min}(W)}$, and $\|\Sigma_K\|$, $\|\Sigma_K\|_F \le \frac{C(K)}{\sigma_{\min}(Q)}$.

Proof. Let $M_K := Q + K^\top R K$. Since K is stabilizing, $C(K) = \mathrm{Tr}(M_K \Sigma_K)$. Because $M_K \succeq Q \succeq \sigma_{\min}(Q) I$, we have

$$C(K) \ge \sigma_{\min}(Q)\,\mathrm{Tr}(\Sigma_K),$$

which yields $\mathrm{Tr}(\Sigma_K) \le \frac{C(K)}{\sigma_{\min}(Q)}$.

For the bound on P_K, using

$$M_K = P_K - (A - BK)^\top P_K (A - BK),$$

we obtain

$$\begin{aligned} C(K) &= \mathrm{Tr}(M_K \Sigma_K) \\ &= \mathrm{Tr}(P_K \Sigma_K) - \mathrm{Tr}\Big((A-BK)^\top P_K (A-BK)\Sigma_K\Big) \\ &= \mathrm{Tr}\Big(P_K\big(\Sigma_K - (A-BK)\Sigma_K(A-BK)^\top\big)\Big) \\ &= \mathrm{Tr}(P_K W), \end{aligned}$$

where the last equality follows from the Lyapunov equation for Σ_K. Since $W \succeq \sigma_{\min}(W) I$ and $P_K \succeq 0$,

$$C(K) = \mathrm{Tr}(P_K W) \geq \sigma_{\min}(W)\,\mathrm{Tr}(P_K),$$

and hence $\mathrm{Tr}(P_K) \leq \frac{C(K)}{\sigma_{\min}(W)}$.

Finally, since P_K and Σ_K are positive semidefinite,

$$\|P_K\| \leq \|P_K\|_F \leq \mathrm{Tr}(P_K), \qquad \|\Sigma_K\| \leq \|\Sigma_K\|_F \leq \mathrm{Tr}(\Sigma_K),$$

which proves the remaining bounds. □

We next introduce several operators that will also be used in the perturbation analysis. For a symmetric matrix X, define

$$\mathcal{F}_K(X) := (A-BK)X(A-BK)^\top, \quad \mathcal{F}_K^h(X) = (A-BK)^h X[(A-BK)^\top]^h, \quad h = 0,1,2,\ldots. \tag{A.1}$$

Then define the corresponding infinite sum operator:

$$\mathcal{T}_K(X) := \sum_{h=0}^{\infty} (A-BK)^h X[(A-BK)^\top]^h. \tag{A.2}$$

We also define the corresponding transpose operators

$$\begin{aligned} &\mathcal{F}_K^\top(X) := (A-BK)^\top X(A-BK), \quad \mathcal{F}_K^{\top,h}(X) = [(A-BK)^\top]^h X(A-BK)^h, \\ &\mathcal{T}_K^\top(X) := \sum_{h=0}^{\infty} [(A-BK)^\top]^h X(A-BK)^h. \end{aligned} \tag{A.3}$$

Finally define the induced norm of $\|\mathcal{T}_K\|$ and $\|\mathcal{T}_K^\top\|$ as

$$\|\mathcal{T}_K\| := \sup_{X \neq 0} \frac{\|\mathcal{T}_K(X)\|}{\|X\|}, \qquad \|\mathcal{T}_K^\top\| := \sup_{X \neq 0} \frac{\|\mathcal{T}_K^\top(X)\|}{\|X\|}.$$

Lemma A.2. *For any stabilizing policy K,*

$$\mathcal{T}_K = (I - \mathcal{F}_K)^{-1}, \qquad \mathcal{T}_K^\top = (I - \mathcal{F}_K^\top)^{-1}.$$

If, in addition, $Q \succ 0$ and $W \succ 0$, then

$$\|\mathcal{T}_K\| \leq \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}, \qquad \|\mathcal{T}_K^\top\| \leq \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}.$$

Proof. The first two identities follows from the same resolvent argument used in [11, Lemma 18].

We next establish the norm bounds. By the Lyapunov equation for Σ_K, $\Sigma_K = \mathcal{T}_K(W)$. For any symmetric matrix X,

$$-\|X\| I \preceq X \preceq \|X\| I.$$

Since $\mathcal{T}_K$ is a positive linear operator,

$$-\|X\| \mathcal{T}_K(I) \preceq \mathcal{T}_K(X) \preceq \|X\| \mathcal{T}_K(I),$$

and therefore $\|\mathcal{T}_K(X)\| \le \|X\| \|\mathcal{T}_K(I)\|$. Since $W \succeq \sigma_{\min}(W) I$,

$$\mathcal{T}_K(I) \preceq \frac{1}{\sigma_{\min}(W)} \mathcal{T}_K(W) = \frac{1}{\sigma_{\min}(W)} \Sigma_K.$$

Using Lemma A.1,

$$\|\mathcal{T}_K(X)\| \le \|X\| \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}.$$

Thus $\|\mathcal{T}_K\| \le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}$.

Similarly, writing $M_K := Q + K^\top R K$, the Lyapunov equation for P_K gives $P_K = \mathcal{T}_K^\top(M_K)$. Since $M_K \succeq Q \succeq \sigma_{\min}(Q) I$,

$$\mathcal{T}_K^\top(I) \preceq \frac{1}{\sigma_{\min}(Q)} \mathcal{T}_K^\top(Q) \preceq \frac{1}{\sigma_{\min}(Q)} P_K.$$

Hence, again by Lemma A.1,

$$\|\mathcal{T}_K^\top\| \le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}.$$

□

Lemma A.3. *Suppose K is stabilizing and $Q \succ 0$, $W \succ 0$. Let $A_K := A - BK$, and define $\alpha_K := \frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)}$. Then $0 < \alpha_K \le 1$, and*

$$(A_K^\top)^h P_K A_K^h \preceq (1 - \alpha_K)^h P_K, \qquad h \ge 0. \tag{A.4}$$

Consequently,

$$\|A_K^h\|^2 \le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)} (1 - \alpha_K)^h, \qquad h \ge 0, \tag{A.5}$$

and

$$\sum_{h=0}^{\infty} \|A_K^h\|^2 \le \frac{C(K)^2}{\sigma_{\min}^2(Q)\sigma_{\min}^2(W)}. \tag{A.6}$$

Proof. Since $\Sigma_K \succeq W$,

$$C(K) = \mathrm{Tr}\Big((Q + K^\top R K)\Sigma_K\Big) \ge \mathrm{Tr}(Q\Sigma_K) \ge \sigma_{\min}(Q)\sigma_{\min}(W).$$

Hence $0 < \alpha_K \le 1$.

Recall that $P_K = Q + K^\top R K + A_K^\top P_K A_K$. Hence $A_K^\top P_K A_K \preceq P_K - Q$. By Lemma A.1,

$$P_K \preceq \frac{C(K)}{\sigma_{\min}(W)} I,$$

and therefore

$$Q \succeq \sigma_{\min}(Q) I \succeq \frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)} P_K = \alpha_K P_K.$$

It follows that $A_K^\top P_K A_K \preceq (1-\alpha_K) P_K$. Iterating this inequality gives

$$(A_K^\top)^h P_K A_K^h \preceq (1-\alpha_K)^h P_K.$$

Moreover, since $P_K \succeq Q \succeq \sigma_{\min}(Q) I$, for every $x \in \mathbb{R}^d$,

$$\begin{aligned}
\sigma_{\min}(Q)\|A_K^h x\|^2 &\le x^\top (A_K^\top)^h P_K A_K^h x \\
&\le (1-\alpha_K)^h x^\top P_K x \\
&\le (1-\alpha_K)^h \frac{C(K)}{\sigma_{\min}(W)} \|x\|^2.
\end{aligned}$$

Taking the supremum over $\|x\| = 1$ yields

$$\|A_K^h\|^2 \le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)} (1-\alpha_K)^h.$$

Finally,

$$\begin{aligned}
\sum_{h=0}^{\infty} \|A_K^h\|^2 &\le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)} \sum_{h=0}^{\infty} (1-\alpha_K)^h \\
&= \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)} \frac{1}{\alpha_K} = \frac{C(K)^2}{\sigma_{\min}^2(Q)\sigma_{\min}^2(W)}.
\end{aligned}$$

□

Details for the proof of Lemma 3.3

We provide explicit bounds for the two transient terms appearing in (3.7). Recall that $\Gamma_K := \sum_{h=0}^{\infty} \|A_K^h\|^2$, which, by Lemma A.3, satisfies

$$\Gamma_K \le \frac{C(K)^2}{\sigma_{\min}^2(Q)\sigma_{\min}^2(W)}. \tag{A.7}$$

First, the state second moments satisfy

$$\Sigma_h^K = A_K^h \Sigma_0 (A_K^\top)^h + \sum_{t=0}^{h-1} A_K^t W (A_K^\top)^t.$$

Hence,

$$\sup_{h \ge 0} \|\Sigma_h^K\| \le \Gamma_K \big(\|\Sigma_0\| + \|W\|\big). \tag{A.8}$$

We first consider the terminal-horizon term. By (3.8),

$$\|E_{K,h}^{(H)} - E_K\|_F \le \|B\| \|P_K\|_F \|A_K^{H-h-1}\| \|A_K^{H-h}\|.$$

Therefore, using (A.8) and the Cauchy–Schwarz inequality,

$$\sum_{h=0}^{H-1} \left\| (E_{K,h}^{(H)} - E_K)\Sigma_h^K \right\|_F \le \|B\|\|P_K\|_F \sup_{h\ge 0} \|\Sigma_h^K\| \sum_{j=0}^{H-1} \|A_K^j\|\|A_K^{j+1}\| \\ \le \|B\|\|P_K\|_F \Gamma_K^2 \big(\|\Sigma_0\| + \|W\|\big).$$

For the initial-state transient, by (3.9),

$$\sum_{h=0}^{H-1} \left\| E_K(\Sigma_h^K - \Sigma_K) \right\|_F \le \|E_K\|_F \sum_{h=0}^{H-1} \|A_K^h\|^2 \|\Sigma_0 - \Sigma_K\|_F \\ \le \|E_K\|_F \Gamma_K \left(\|\Sigma_0\|_F + \|\Sigma_K\|_F\right) \\ \le \|E_K\|_F \Gamma_K \left(\|\Sigma_0\|_F + \frac{C(K)}{\sigma_{\min}(Q)}\right),$$

Combining the preceding two estimates with (3.7), we obtain

$$\|\nabla C_H(K) - \nabla C(K)\|_F \le \frac{L_{\mathrm{grad}}(K)}{H},$$

where one may take

$$L_{\mathrm{grad}}(K) := 2\|B\|\|P_K\|_F \Gamma_K^2 \big(\|\Sigma_0\| + \|W\|\big) + 2\|E_K\|_F \Gamma_K \left(\|\Sigma_0\|_F + \frac{C(K)}{\sigma_{\min}(Q)}\right). \tag{A.9}$$

Moreover,

$$C(K) = \mathrm{Tr}\Big((Q + K^\top R K)\Sigma_K\Big) \ge \mathrm{Tr}(K^\top R K\, \Sigma_K) \ge \sigma_{\min}(W)\sigma_{\min}(R)\|K\|_F^2,$$

and hence

$$\|K\|_F \le \sqrt{\frac{C(K)}{\sigma_{\min}(W)\sigma_{\min}(R)}}. \tag{A.10}$$

Finally,

$$E_K = RK - B^\top P_K A_K,$$

and therefore

$$\|E_K\|_F \le \|R\|\sqrt{\frac{C(K)}{\sigma_{\min}(W)\sigma_{\min}(R)}} + \|B\|\frac{C(K)}{\sigma_{\min}(W)} \left(\|A\| + \|B\|\sqrt{\frac{C(K)}{\sigma_{\min}(W)\sigma_{\min}(R)}}\right). \tag{A.11}$$

Together with Lemma A.1 and (A.7), these estimates show that $L_{\mathrm{grad}}(K)$ is bounded polynomially in $C(K)$ and the system parameters, and is independent of H.

B Known-model policy-gradient analysis

In this appendix, we establish the auxiliary results used in the proof of Theorem 4.1. We first recall the gradient-domination and almost-smoothness properties of the average-cost LQR objective, then derive a one-step descent inequality for the biased gradient $\nabla C_H(K)$, and finally show that the iterates remain in a fixed average-cost sublevel set.

Lemma B.1 (Gradient domination)**.** *Under Assumptions 2.1, 2.2 and 2.3, and suppose $W \succ 0$, for any stabilizing policy K,*

$$C(K) - C(K^\star) \le \frac{\|\Sigma_{K^\star}\|}{4\,\sigma_{\min}^2(W)\sigma_{\min}(R)}\,\|\nabla C(K)\|_F^2, \tag{B.1}$$

where $K^\star$ is the optimal policy for the average-cost LQR.

The proof follows the standard policy-improvement argument used in [11, Lemma 11]. In our average-cost setting, Σ_K is the stationary second-moment matrix and satisfies

$$\Sigma_K = (A - BK)\Sigma_K(A - BK)^\top + W \succeq W.$$

Hence $\sigma_{\min}(\Sigma_K) \ge \sigma_{\min}(W)$, which replaces the lower bound on the state covariance used in the infinite-sum formulation of [11].

Lemma B.2 (Almost smoothness)**.** *Under Assumptions 2.1, 2.2 and 2.3, let K and K' be two stabilizing policies, we have*

$$C(K') - C(K) = 2\,\mathrm{Tr}\left(\Sigma_{K'}(K' - K)^\top E_K\right) + \mathrm{Tr}\left(\Sigma_{K'}(K' - K)^\top\big(R + B^\top P_K B\big)(K' - K)\right). \tag{B.2}$$

This is the average-cost policy-improvement identity corresponding to [11, Lemma 12].

We next show the one-step analysis. This follows the similar proof route as [11], but instead of direct policy gradient, we incorporate a biased gradient here.

Lemma B.3 (One-step analysis in known model case)**.** *Under Assumptions 2.1, 2.2 and 2.3, suppose $Q \succ 0$ and $W \succ 0$, for any stabilizing policy K, following the update rule $K' := K - \eta\nabla C_H(K)$, if*

$$\eta \le \min\left\{\frac{1}{16}\left(\frac{\sigma_{\min}(Q)\,\sigma_{\min}(W)}{C(K)}\right)^2 \frac{1}{\|B\|\,\|\nabla C_H(K)\|_F(1 + \|A - BK\|)},\ \frac{3\sigma_{\min}(Q)}{32\,C(K)\|R + B^\top P_K B\|}\right\}, \tag{B.3}$$

then K' is stabilizing and

$$\begin{aligned} C(K') - C(K^*) \le{}& \left(1 - \eta\,\frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^*}\|}\right)(C(K) - C(K^*)) \\ &+ \eta\left(\frac{16C(K)^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\|\nabla C_H(K) - \nabla C(K)\|_F^2. \end{aligned}$$

Proof. We first verify that K' is stabilizing. Let $\Delta := K' - K$, $A_K := A - BK$. Since

$$P_K - A_K^\top P_K A_K = Q + K^\top R K \succeq \sigma_{\min}(Q) I,$$

we have

$$P_K - A_{K'}^\top P_K A_{K'} \succeq \Big(\sigma_{\min}(Q) - 2\|A_K\|\|P_K\|\|B\|\|\Delta\|_F - \|P_K\|\|B\|^2\|\Delta\|_F^2\Big) I.$$

By the first step-size condition,

$$\|\Delta\|_F \le \frac{1}{16}\left(\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)}\right)^2 \frac{1}{\|B\|(1 + \|A_K\|)}.$$

Together with $\|P_K\| \le \frac{C(K)}{\sigma_{\min}(W)}$ and $\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)} \le 1$, this implies $P_K - A_{K'}^\top P_K A_{K'} \succ 0$. By the discrete-time Lyapunov criterion, $A_{K'}$ is Schur stable.

Define $e_H(K) := \nabla C_H(K) - \nabla C(K)$ and $S_K := R + B^\top P_K B$. by Lemma B.2 and $E_K = \frac{1}{2}\nabla C(K)\Sigma_K^{-1}$, we have

$$\begin{aligned}
C(K')-C(K) =& -\eta \mathrm{Tr}\big(\Sigma_{K'}(\nabla C(K)+e_H(K))^\top \nabla C(K)\Sigma_K^{-1}\big)+\eta^2 \mathrm{Tr}\big(\Sigma_{K'}(\nabla C(K)+e_H(K))^\top S_K(\nabla C(K)+e_H(K))\big)\\
=& -\eta \mathrm{Tr}\big(\Sigma_{K'}\nabla C(K)^\top \nabla C(K)\Sigma_K^{-1}\big) - \eta \mathrm{Tr}\big(\Sigma_{K'} e_H(K)^\top \nabla C(K)\Sigma_K^{-1}\big)\\
&+\eta^2 \mathrm{Tr}\big(\Sigma_{K'}\nabla C(K)^\top S_K \nabla C(K)\big)+2\eta^2 \mathrm{Tr}\big(\Sigma_{K'} e_H(K)^\top S_K \nabla C(K)\big)+\eta^2 \mathrm{Tr}\big(\Sigma_{K'} e_H(K)^\top S_K e_H(K)\big)\\
\le& -\eta\left(1 - \frac{\|\Sigma_{K'}-\Sigma_K\|}{\sigma_{\min}(W)} - \eta\|\Sigma_{K'}\|\|S_K\|\right)\mathrm{Tr}(\nabla C(K)^\top \nabla C(K))\\
&+\Big(\frac{\eta}{4}+\eta^2\|\Sigma_{K'}\|\|S_K\|\Big)\mathrm{Tr}(\nabla C(K)^\top \nabla C(K))+\Big(\eta\frac{\|\Sigma_{K'}\|^2}{\sigma_{\min}^2(W)}+2\eta^2\|\Sigma_{K'}\|\|S_K\|\Big)\mathrm{Tr}(e_H(K)^\top e_H(K))\\
=& -\eta\left(\frac{3}{4} - \frac{\|\Sigma_{K'}-\Sigma_K\|}{\sigma_{\min}(W)} - 2\eta\|\Sigma_{K'}\|\|S_K\|\right)\mathrm{Tr}(\nabla C(K)^\top \nabla C(K))\\
&+\Big(\eta\frac{\|\Sigma_{K'}\|^2}{\sigma_{\min}^2(W)}+2\eta^2\|\Sigma_{K'}\|\|S_K\|\Big)\mathrm{Tr}(e_H(K)^\top e_H(K)),
\end{aligned} \tag{B.4}$$

here the inequality derivation uses $\Sigma_{K'} \succeq \Sigma_K - \|\Sigma_{K'}-\Sigma_K\| I$, $\Sigma_K \succeq \sigma_{\min}(W) I$, and Young's inequality.

Moreover, the same covariance perturbation estimate as in [11, Lemma 16], using $\Sigma_K \succeq W$, gives

$$\|\Sigma_{K'}-\Sigma_K\| \le 4\left(\frac{C(K)}{\sigma_{\min}(Q)}\right)^2 \frac{\|B\|(1+\|A-BK\|)}{\sigma_{\min}(W)}\|K'-K\|_F.$$

Together with the first step-size condition in (B.3), this yields

$$\frac{\|\Sigma_{K'}-\Sigma_K\|}{\sigma_{\min}(W)} \le \frac{1}{4}. \tag{B.5}$$

Since $\Sigma_{K'} \succeq W$, we have $\|\Sigma_{K'}\| \ge \sigma_{\min}(W)$, and therefore

$$\|\Sigma_{K'}\| \le \|\Sigma_{K'}-\Sigma_K\| + \|\Sigma_K\| \le \frac{\sigma_{\min}(W)}{4} + \frac{C(K)}{\sigma_{\min}(Q)} \le \frac{\|\Sigma_{K'}\|}{4} + \frac{C(K)}{\sigma_{\min}(Q)},$$

and therefore, $\|\Sigma_{K'}\| \le \frac{4C(K)}{3\sigma_{\min}(Q)}$. When η satisfies the second term in (B.3), we have

$$\eta\,\|\Sigma_{K'}\|\,\|S_K\| \le \frac{1}{8}, \tag{B.6}$$

then

$$\frac{3}{4} - \frac{\|\Sigma_{K'}-\Sigma_K\|}{\sigma_{\min}(W)} - 2\eta\|\Sigma_{K'}\|\,\|S_K\| \ge \frac{3}{4} - \frac{1}{4} - \frac{1}{4} = \frac{1}{4},$$

which ensure the descent property in gradient.

Inserting (B.5) and (B.6) and upper-bound of $\|\Sigma_{K'}\|$ into (B.4), also combining the (B.1), we get the final inequality claimed in lemma. □

We now give the condition on H such that all $C(K_n)$ are bounded. This is important since many H-independent upper-bounds rely on $C(K)$.

We first specify a step size that satisfies the conditions of Lemma B.3 uniformly over the sublevel set

$$\mathcal{K}_0 := \{K \in \mathcal{K} : C(K) \le 2C_0\}.$$

For any $K \in \mathcal{K}_0$, Lemma A.1 and (A.10) give

$$\|K\|_F \le \sqrt{\frac{2C_0}{\sigma_{\min}(W)\sigma_{\min}(R)}}.$$

Hence

$$\|A - BK\| \le A_0 := \|A\| + \|B\|\sqrt{\frac{2C_0}{\sigma_{\min}(W)\sigma_{\min}(R)}}.$$

Moreover, $\|P_K\| \le \frac{2C_0}{\sigma_{\min}(W)}$, and therefore

$$\|R + B^\top P_K B\| \le S_0 := \|R\| + \frac{2C_0\|B\|^2}{\sigma_{\min}(W)}.$$

Also, since $P_{K,h}^{(H)} \preceq P_K$, we have uniformly in H and h,

$$\|E_{K,h}^{(H)}\|_F \le E_0 := \|R\|\sqrt{\frac{2C_0}{\sigma_{\min}(W)\sigma_{\min}(R)}} + \|B\|\frac{2C_0}{\sigma_{\min}(W)}\left(\|A\| + \|B\|\sqrt{\frac{2C_0}{\sigma_{\min}(W)\sigma_{\min}(R)}}\right).$$

Furthermore, by Lemma A.2,

$$\sup_{h\ge 0}\|\Sigma_h^K\| \le \frac{2C_0}{\sigma_{\min}(Q)\sigma_{\min}(W)}\big(\|\Sigma_0\| + \|W\|\big).$$

Consequently,

$$\sup_{H\ge 1}\|\nabla C_H(K)\|_F \le G_0 := 2E_0\frac{2C_0}{\sigma_{\min}(Q)\sigma_{\min}(W)}\big(\|\Sigma_0\| + \|W\|\big).$$

We may therefore choose

$$\eta_0 := \min\left\{\frac{1}{16}\left(\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{2C_0}\right)^2\frac{1}{\|B\|G_0(1+A_0)}, \frac{3\sigma_{\min}(Q)}{64C_0S_0}, \frac{\|\Sigma_{K^\star}\|}{2\sigma_{\min}^2(W)\sigma_{\min}(R)}\right\}. \tag{B.7}$$

Thus, for every $K \in \mathcal{K}_0$, every $H \ge 1$, and every $0 < \eta \le \eta_0$, the conditions of Lemma B.3 hold uniformly. Moreover, the last term in (B.7) ensures that $0 < \gamma \le 1/2$.

Lemma B.4. *Under Assumptions 2.1, 2.2 and 2.3, suppose $Q \succ 0$ and $W \succ 0$. Let $0 < \eta \le \eta_0$. If*

$$H \ge H_0 := \overline{L}_{\mathrm{grad}}\left[\left(\frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\frac{\|\Sigma_{K^\star}\|}{\sigma_{\min}(Q)\sigma_{\min}^3(W)\sigma_{\min}(R)}\right]^{1/2}, \tag{B.8}$$

then the iterates of Algorithm 1 satisfy

$$K_n \in \mathcal{K}, \qquad C(K_n) \le 2C_0, \qquad n \ge 0.$$

Proof. We argue by induction. Clearly $K_0 \in \mathcal{K}_0$. Suppose $K_n \in \mathcal{K}_0$. Since $\eta \le \eta_0$, Lemma B.3 applies to K_n. In particular, K_{n+1} is stabilizing and

$$C(K_{n+1}) - C(K^\star) \le (1-\gamma)\big(C(K_n) - C(K^\star)\big) + \eta\left(\frac{16C(K_n)^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\|\nabla C_H(K_n) - \nabla C(K_n)\|_F^2,$$

where $\gamma := \eta\frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}$.

By the induction hypothesis and Lemma 3.3, $\|\nabla C_H(K_n) - \nabla C(K_n)\|_F \le \frac{\overline{L}_{\mathrm{grad}}}{H}$, and hence

$$C(K_{n+1}) - C(K^\star) \le (1-\gamma)\big(C(K_n) - C(K^\star)\big) + \eta\left(\frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\frac{\overline{L}_{\mathrm{grad}}^2}{H^2}. \tag{B.9}$$

By the condition on H,

$$\left(\frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\frac{\overline{L}_{\mathrm{grad}}^2}{H^2} \le \frac{\sigma_{\min}(Q)\sigma_{\min}^3(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}.$$

Therefore,

$$\eta\left(\frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\frac{\overline{L}_{\mathrm{grad}}^2}{H^2} \le \eta\frac{\sigma_{\min}(Q)\sigma_{\min}^3(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|} = \gamma\,\sigma_{\min}(Q)\sigma_{\min}(W).$$

Since $C_0 \ge \sigma_{\min}(Q)\sigma_{\min}(W)$, we obtain

$$\eta\left(\frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\frac{\overline{L}_{\mathrm{grad}}^2}{H^2} \le \gamma C_0.$$

Hence (B.9) gives

$$C(K_{n+1}) - C(K^\star) \le (1-\gamma)\big(C(K_n) - C(K^\star)\big) + \gamma C_0.$$

By the induction hypothesis $C(K_n) \le 2C_0$, and $0 < \gamma \le 1/2$ by the definition of η_0,

$$\begin{aligned} C(K_{n+1}) - C(K^\star) &\le (1-\gamma)(2C_0 - C(K^\star)) + \gamma C_0 \\ &= 2C_0 - C(K^\star) - \gamma\big(C_0 - C(K^\star)\big) \le 2C_0 - C(K^\star), \end{aligned}$$

where the last inequality follows from $C_0 \ge C(K^\star)$. Thus $C(K_{n+1}) \le 2C_0$. Together with Lemma B.3, which ensures that K_{n+1} is stabilizing, this completes the induction. □

C Perturbation analysis

This appendix collects the local perturbation bounds needed in the model-free analysis. Standard perturbation arguments for stationary infinite-horizon LQR are given in [11, Lemmas 16, 19–20, 24–25], while finite-horizon perturbation bounds for time-varying policies are established in [16, Lemmas 4.7–4.8].

The main point needed here is slightly different: since we optimize over a stationary stabilizing policy, the local Lipschitz constants for the finite-horizon cost and gradient can be chosen independently of the horizon H. We record this property below.

Lemma C.1. *Under Assumptions 2.1, 2.2 and 2.3, suppose $Q \succ 0$ and $W \succ 0$. Let K be stabilizing and suppose*

$$\|K' - K\|_F \le \delta_K,$$

where

$$\delta_K := \min\left\{ \frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{4C(K)\|B\|(1+\|A-BK\|)}, 1 \right\}.$$

Then K' is stabilizing, and there exist constants $h^f_{\mathrm{cost}}(K)$, $h^f_{\mathrm{grad}}(K)$, and $h^a_{\mathrm{cost}}(K)$, independent of H, such that

$$|C_H(K') - C_H(K)| \le h^f_{\mathrm{cost}}(K)\|K' - K\|_F,$$
$$\|\nabla C_H(K') - \nabla C_H(K)\|_F \le h^f_{\mathrm{grad}}(K)\|K' - K\|_F,$$

and

$$|C(K') - C(K)| \le h^a_{\mathrm{cost}}(K)\|K' - K\|_F.$$

Moreover, these constants are polynomially bounded in $C(K)$ and the system parameters.

Proof. Let

$$\Delta := K' - K, \qquad A_K := A - BK, \qquad M_K := Q + K^\top R K.$$

The standard operator perturbation argument of [11, Lemmas 19–20], together with Lemma A.2, gives

$$\|\mathcal{F}_{K'} - \mathcal{F}_K\| \le c_F(K)\|\Delta\|_F,$$

and

$$\|\mathcal{T}_{K'}\| \le 2\|\mathcal{T}_K\|, \qquad \|\mathcal{T}^\top_{K'}\| \le 2\|\mathcal{T}^\top_K\|,$$

where $c_F(K)$ is polynomially bounded in $C(K)$ and the system parameters. The same perturbation condition also implies that K' is stabilizing; this follows, for example, from the Lyapunov perturbation argument used in the proof of Lemma B.3.

Since

$$\Sigma^K_h = \mathcal{F}^h_K(\Sigma_0) + \sum_{t=0}^{h-1} \mathcal{F}^t_K(W),$$

and $\Sigma_0, W \succeq 0$, positivity of $\mathcal{F}_K$ gives $0 \preceq \mathcal{F}^h_K(\Sigma_0) \preceq \mathcal{T}_K(\Sigma_0)$ and $0 \preceq \sum_{t=0}^{h-1} \mathcal{F}^t_K(W) \preceq \mathcal{T}_K(W)$. Hence, by Lemma A.2,

$$\sup_{h\ge 0} \|\Sigma^K_h\| \le \|\mathcal{T}_K\|\big(\|\Sigma_0\| + \|W\|\big) \le \frac{C(K)}{\sigma_{\min}(Q)\sigma_{\min}(W)}\big(\|\Sigma_0\| + \|W\|\big). \tag{C.1}$$

Moreover,

$$\Sigma^{K'}_h - \Sigma^K_h = \sum_{t=0}^{h-1} \mathcal{F}^{h-1-t}_{K'}\Big[(\mathcal{F}_{K'} - \mathcal{F}_K)(\Sigma^K_t)\Big].$$

Using positivity of $\mathcal{F}_{K'}$, (C.1), and $\sum_{j=0}^{h-1} \mathcal{F}^j_{K'}(I) \preceq \mathcal{T}_{K'}(I)$, we obtain

$$\sup_{h\ge 0} \|\Sigma^{K'}_h - \Sigma^K_h\|_F \le L_\Sigma(K)\|\Delta\|_F, \tag{C.2}$$

where $L_\Sigma(K)$ is independent of H and polynomially bounded in $C(K)$ and the system parameters.

For the gradient perturbation, recall from the proof of Lemma 3.3 that

$$P^{(H)}_{K,h} = M_K + A_K^\top P^{(H)}_{K,h+1} A_K, \qquad P^{(H)}_{K,H} = 0.$$

We also define

$$E_{K,h}^{(H)} := \big(R + B^\top P_{K,h+1}^{(H)} B\big)K - B^\top P_{K,h+1}^{(H)} A,$$

so that

$$\nabla C_H(K) = \frac{2}{H} \sum_{h=0}^{H-1} E_{K,h}^{(H)} \Sigma_h^K.$$

The quantities $P_{K',h}^{(H)}$ and $E_{K',h}^{(H)}$ are defined analogously with K replaced by K'.

Since $P_{K,h}^{(H)} \preceq P_K$, subtracting the recursions for K' and K and using the same resolvent argument gives

$$\sup_{\substack{H\geq 1 \\ 0\leq h\leq H}} \|P_{K',h}^{(H)} - P_{K,h}^{(H)}\|_F \leq L_P(K)\|\Delta\|_F, \tag{C.3}$$

where $L_P(K)$ is independent of H and polynomially bounded in $C(K)$ and the system parameters. Indeed, the difference satisfies

$$P_{K',h}^{(H)} - P_{K,h}^{(H)} = \mathcal{F}_{K'}^\top \big(P_{K',h+1}^{(H)} - P_{K,h+1}^{(H)}\big) + (M_{K'} - M_K) + (\mathcal{F}_{K'}^\top - \mathcal{F}_K^\top)(P_{K,h+1}^{(H)}),$$

and all forcing terms are $O(\|\Delta\|_F)$ uniformly in H.

Next,

$$E_{K',h}^{(H)} - E_{K,h}^{(H)} = \big(R + B^\top P_{K',h+1}^{(H)} B\big)\Delta + B^\top \big(P_{K',h+1}^{(H)} - P_{K,h+1}^{(H)}\big)(BK - A).$$

Therefore,

$$\|E_{K',h}^{(H)} - E_{K,h}^{(H)}\|_F \leq \|R + B^\top P_{K',h+1}^{(H)} B\|\|\Delta\|_F + \|B\|\|P_{K',h+1}^{(H)} - P_{K,h+1}^{(H)}\|_F \|BK - A\|.$$

Since $P_{K,h}^{(H)} \preceq P_K$, (C.3), and Lemma A.1 imply

$$\sup_{H,h} \|P_{K',h}^{(H)}\| \leq \frac{C(K)}{\sigma_{\min}(W)} + L_P(K),$$

where we used $\|\Delta\|_F \leq 1$. Together with

$$\|BK - A\| \leq \|A\| + \|B\| \sqrt{\frac{C(K)}{\sigma_{\min}(W)\sigma_{\min}(R)}},$$

this yields

$$\sup_{\substack{H\geq 1 \\ 0\leq h\leq H-1}} \|E_{K',h}^{(H)} - E_{K,h}^{(H)}\|_F \leq L_E(K)\|\Delta\|_F, \tag{C.4}$$

where $L_E(K)$ is independent of H and polynomially bounded in $C(K)$ and the system parameters.

We now obtain the finite-horizon cost bound. Writing $M_{K'} := Q + (K')^\top R K'$, we have

$$C_H(K) = \frac{1}{H} \sum_{h=0}^{H-1} \mathrm{Tr}(M_K \Sigma_h^K).$$

Using (C.1)–(C.2), together with

$$\|M_{K'} - M_K\|_F = O(\|\Delta\|_F),$$

gives

$$|C_H(K') - C_H(K)| \le h^f_{\mathrm{cost}}(K)\|\Delta\|_F,$$

with $h^f_{\mathrm{cost}}(K)$ independent of H.

Similarly,

$$\nabla C_H(K') - \nabla C_H(K) = \frac{2}{H}\sum_{h=0}^{H-1}\Big[(E^{(H)}_{K',h} - E^{(H)}_{K,h})\Sigma^{K'}_h + E^{(H)}_{K,h}(\Sigma^{K'}_h - \Sigma^K_h)\Big].$$

By (C.1) and (C.2), $\Sigma^{K'}_h$ is uniformly bounded. Moreover, $P^{(H)}_{K,h} \preceq P_K$ and Lemma A.1 imply that $E^{(H)}_{K,h}$ is uniformly bounded in H and h. Combining these bounds with (C.4) gives

$$\|\nabla C_H(K') - \nabla C_H(K)\|_F \le h^f_{\mathrm{grad}}(K)\|\Delta\|_F,$$

again with a constant independent of H.

Finally, the average-cost perturbation bound follows from the same resolvent argument as [11, Lemma 24]. In our additive-noise average-cost setting,

$$C(K) = \mathrm{Tr}(P_K W), \qquad P_K = \mathcal{T}_K^\top(M_K),$$

so the same argument, with W playing the role of the initial second-moment matrix, gives

$$|C(K') - C(K)| \le h^a_{\mathrm{cost}}(K)\|\Delta\|_F.$$

The bounds in Lemmas A.1 and A.2, together with the bounds displayed above, show that all constants are polynomially bounded in $C(K)$ and the system parameters. □

D Model-free policy-gradient analysis

In this appendix, we establish the auxiliary results used in the proof of Theorem 4.3. We first control the error of the zeroth-order gradient estimator, then derive a one-step descent inequality, and finally choose the algorithmic parameters uniformly over the sublevel set $\mathcal{K}_0$ and show that all iterates remain in this set with high probability.

Lemma D.1 (Gradient-estimation error)**.** *Let K be stabilizing, and fix $\zeta > 0$ and $\delta \in (0,1)$. Suppose*

$$0 < r \le r_K := \min\left\{\delta_K,\ \frac{C_H(K)}{4h^f_{\mathrm{cost}}(K)},\ \frac{C(K)}{h^a_{\mathrm{cost}}(K)}\right\}.$$

Then there exist constants $\nu_K, \alpha_K > 0$, polynomially bounded in $C(K)$ and the system and noise parameters, such that if

$$m \ge c\left(D\log 5 + \log\frac{1}{\delta}\right)\max\left\{\frac{D^2 C_H(K)^2}{r^2\zeta^2},\ \frac{D^2\nu_K^2}{r^2\zeta^2},\ \frac{D\alpha_K}{r\zeta}\right\},$$

for a universal constant $c > 0$, then

$$\mathbb{P}\left(\|\widehat{\nabla}(K) - \nabla C^r_H(K)\|_F \le \zeta\right) \ge 1 - \delta.$$

Proof. Define the intermediate estimator

$$\widetilde{\nabla}(K) := \frac{1}{m}\sum_{i=1}^{m} \frac{D}{r^2} C_H(K+U_i)U_i.$$

$\mathbb{E}[\widetilde{\nabla}(K)] = \nabla C_H^r(K)$. We decompose

$$\|\widehat{\nabla}(K) - \nabla C_H^r(K)\|_F \le \|\widehat{\nabla}(K) - \widetilde{\nabla}(K)\|_F + \|\widetilde{\nabla}(K) - \nabla C_H^r(K)\|_F. \tag{D.1}$$

Since $r \le \delta_K$, Lemma C.1 applies to every $K+U$ with $\|U\|_F = r$. In particular,

$$|C_H(K+U) - C_H(K)| \le h^f_{\mathrm{cost}}(K) r,$$

and

$$|C(K+U) - C(K)| \le h^a_{\mathrm{cost}}(K) r.$$

Since $r \le C_H(K)/(4h^f_{\mathrm{cost}}(K))$ and $r \le C(K)/h^a_{\mathrm{cost}}(K)$, we obtain, uniformly over $\|U\|_F = r$,

$$C_H(K+U) \le \frac{5}{4} C_H(K) \le 2C_H(K), \tag{D.2}$$

and

$$C(K+U) \le 2C(K). \tag{D.3}$$

We first control the random-direction error $\widetilde{\nabla}(K) - \nabla C_H^r(K)$. Define

$$X_i := \frac{D}{r^2}\left(C_H(K+U_i)\,\mathrm{vec}(U_i) - \mathbb{E}_U[C_H(K+U)\,\mathrm{vec}(U)]\right).$$

Then $\mathbb{E}[X_i] = 0$ and

$$\mathrm{vec}\big(\widetilde{\nabla}(K) - \nabla C_H^r(K)\big) = \frac{1}{m}\sum_{i=1}^{m} X_i.$$

By (D.2) and $\|U_i\|_F = r$,

$$\left\| \frac{D}{r^2} C_H(K+U_i)\,\mathrm{vec}(U_i) \right\| \le \frac{2DC_H(K)}{r},$$

and hence, after centering,

$$\|X_i\| \le \frac{4DC_H(K)}{r}. \tag{D.4}$$

Let $\mathcal{N}$ be a 1/2-net of the unit sphere S^{D-1} with $|\mathcal{N}| \le 5^D$. For each fixed $v \in \mathcal{N}$, the random variables $\langle v, X_i\rangle$ are independent, mean zero, and uniformly bounded by the right-hand side of (D.4). Thus, by a standard scalar Bernstein (or Hoeffding) inequality,

$$\mathbb{P}\left(\left| \frac{1}{m}\sum_{i=1}^{m} \langle v, X_i\rangle \right| > \frac{\zeta}{4} \right) \le 2\exp\left(-c_1 m \frac{r^2\zeta^2}{D^2 C_H(K)^2} \right),$$

for a universal constant $c_1 > 0$. Using the standard 1/2-net inequality

$$\|x\| \le 2 \sup_{v\in\mathcal{N}} |\langle v, x\rangle|$$

and taking a union bound over $\mathcal{N}$ gives

$$\mathbb{P}\left(\|\widetilde{\nabla}(K) - \nabla C_H^r(K)\|_F > \frac{\zeta}{2}\right) \le 2 \cdot 5^D \exp\left(-c_1 m \frac{r^2\zeta^2}{D^2 C_H(K)^2}\right). \tag{D.5}$$

We next control the rollout error $\widehat{\nabla}(K) - \widetilde{\nabla}(K)$. We first establish a uniform concentration bound for the rollout costs. Fix U with $\|U\|_F = r$, and write

$$K_U := K + U, \qquad A_U := A - BK_U, \qquad M_U := Q + K_U^\top R K_U.$$

Under Assumption 4.1, the state process can be decomposed as

$$x_h = \bar{x}_h + \widetilde{x}_h, \qquad \bar{x}_h := A_U^h \mu_0,$$

where $\widetilde{x}_h$ is a centered linear function of the independent sub-Gaussian variables appearing in the initial state and process noises. Consequently, $\widehat{C}_H(K_U) - C_H(K_U)$ is the sum of a centered quadratic form in these sub-Gaussian variables and a centered linear form.

The quadratic term is controlled by the same Hanson–Wright argument as in [16, Lemma 4.10]. The nonzero initial mean μ_0 produces only the additional linear term. Since K_U is stabilizing, the closed-loop powers A_U^h are geometrically summable, and hence this linear term is sub-Gaussian, with parameters controlled by $C(K_U)$ and the system and noise parameters. By (D.3), $C(K_U) \le 2C(K)$ uniformly over $\|U\|_F = r$. Therefore, there exist constants $\nu_K, \alpha_K > 0$, independent of U and H and polynomially bounded in $C(K)$ and the system and noise parameters, such that

$$\widehat{C}_H(K+U) - C_H(K+U) \in \mathrm{SE}(\nu_K^2, \alpha_K) \tag{D.6}$$

uniformly over $\|U\|_F = r$. Now define

$$Y_i := \frac{D}{r^2}\big(\widehat{C}_H(K+U_i) - C_H(K+U_i)\big)\,\mathrm{vec}(U_i).$$

Conditional on U_i, the rollout cost is unbiased, and hence $\mathbb{E}[Y_i] = 0$. Moreover,

$$\mathrm{vec}\big(\widehat{\nabla}(K) - \widetilde{\nabla}(K)\big) = \frac{1}{m}\sum_{i=1}^m Y_i.$$

For any $v \in S^{D-1}$, $|\langle v, \mathrm{vec}(U_i)\rangle| \le r$. Therefore, by (D.6), $\langle v, Y_i\rangle$ is sub-exponential with parameters $\left(\frac{D^2\nu_K^2}{r^2}, \frac{D\alpha_K}{r}\right)$. Bernstein's inequality for independent mean-zero sub-exponential random variables then gives

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m \langle v, Y_i\rangle\right| > \frac{\zeta}{4}\right) \le 2\exp\left[-c_2 m \min\left\{\frac{r^2\zeta^2}{D^2\nu_K^2}, \frac{r\zeta}{D\alpha_K}\right\}\right],$$

for a universal constant $c_2 > 0$. Using the same 1/2-net $\mathcal{N}$ and taking a union bound,

$$\mathbb{P}\left(\|\widehat{\nabla}(K) - \widetilde{\nabla}(K)\|_F > \frac{\zeta}{2}\right) \le 2 \cdot 5^D \exp\left[-c_2 m \min\left\{\frac{r^2\zeta^2}{D^2\nu_K^2}, \frac{r\zeta}{D\alpha_K}\right\}\right]. \tag{D.7}$$

Taking the universal constant c in the statement sufficiently large, the stated lower bound on m makes the right-hand sides of (D.5) and (D.7) at most $\delta/2$ each. Therefore, by a union bound and (D.1),

$$\|\widehat{\nabla}(K) - \nabla C_H^r(K)\|_F \le \zeta$$

with probability at least $1 - \delta$.

□

Lemma D.2 (One-step analysis in the model-free case)**.** *Under Assumptions 2.1, 2.2, 2.3, and 4.1, suppose $Q \succ 0$ and $W \succ 0$. Let K be stabilizing, and choose r, m, ζ, δ so that the conditions of Lemma D.1 hold. Define*

$$\mathcal{E}_K := \zeta + h^f_{\mathrm{grad}}(K) r + \frac{L_{\mathrm{grad}}(K)}{H}.$$

Consider the update

$$K' := K - \eta \widehat{\nabla}(K).$$

If

$$\eta \le \min\left\{ \frac{1}{16}\left(\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)}\right)^2 \frac{1}{\|B\|\big(\|\nabla C(K)\|_F + \mathcal{E}_K\big)(1+\|A-BK\|)},\ \frac{3\sigma_{\min}(Q)}{32 C(K)\|R + B^\top P_K B\|} \right\}, \tag{D.8}$$

then, with probability at least $1-\delta$, K' is stabilizing and

$$\begin{aligned} C(K') - C(K^\star) \le\ & \left(1 - \eta \frac{\sigma^2_{\min}(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}\right)\big(C(K) - C(K^\star)\big) \\ & + \eta\left(\frac{16 C(K)^2}{9\sigma^2_{\min}(Q)\sigma^2_{\min}(W)} + \frac{1}{4}\right)\mathcal{E}_K^2. \end{aligned} \tag{D.9}$$

Proof. By Lemma D.1, with probability at least $1-\delta$,

$$\|\widehat{\nabla}(K) - \nabla C^r_H(K)\|_F \le \zeta. \tag{D.10}$$

We work on this event.

We first relate the gradient estimator to the average-cost gradient. Since $r \le \delta_K$, Lemma C.1 applies to every $K+V$ with $\|V\|_F \le r$. By the definition of the smoothed objective and differentiation under the expectation,

$$\nabla C^r_H(K) = \mathbb{E}_{V \sim \mathrm{Unif}(B_r)}[\nabla C_H(K+V)].$$

Therefore,

$$\begin{aligned} \|\nabla C^r_H(K) - \nabla C_H(K)\|_F &\le \mathbb{E}_V\big[\|\nabla C_H(K+V) - \nabla C_H(K)\|_F\big] \\ &\le h^f_{\mathrm{grad}}(K) r. \end{aligned}$$

Combining this bound with (D.10) and Lemma 3.3 gives

$$\begin{aligned} \|\widehat{\nabla}(K) - \nabla C(K)\|_F &\le \|\widehat{\nabla}(K) - \nabla C^r_H(K)\|_F \\ &\quad + \|\nabla C^r_H(K) - \nabla C_H(K)\|_F + \|\nabla C_H(K) - \nabla C(K)\|_F \\ &\le \zeta + h^f_{\mathrm{grad}}(K) r + \frac{L_{\mathrm{grad}}(K)}{H} = \mathcal{E}_K. \end{aligned} \tag{D.11}$$

In particular,

$$\|\widehat{\nabla}(K)\|_F \le \|\nabla C(K)\|_F + \mathcal{E}_K. \tag{D.12}$$

We next verify stability of the updated policy. Let

$$\Delta := K' - K = -\eta \widehat{\nabla}(K), \qquad A_K := A - BK.$$

By (D.8) and (D.12),

$$\|\Delta\|_F \le \frac{1}{16}\left(\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{C(K)}\right)^2 \frac{1}{\|B\|(1+\|A_K\|)}.$$

This is exactly the perturbation bound used in the proof of Lemma B.3. Hence the same Lyapunov argument yields

$$P_K - (A - BK')^\top P_K (A - BK') \succ 0,$$

and therefore K' is stabilizing.

Now define the gradient error $e(K) := \widehat{\nabla}(K) - \nabla C(K)$. By (D.11), $\|e(K)\|_F \le \mathcal{E}_K$, and the update can be written as

$$K' = K - \eta\big(\nabla C(K) + e(K)\big).$$

Since both K and K' are stabilizing, Lemma B.2 applies. The remainder of the argument is the same inexact-gradient calculation as in the proof of Lemma B.3, with $e(K)$ replacing $\nabla C_H(K) - \nabla C(K)$. In particular, the first condition in (D.8) gives

$$\frac{\|\Sigma_{K'} - \Sigma_K\|}{\sigma_{\min}(W)} \le \frac{1}{4},$$

and hence

$$\|\Sigma_{K'}\| \le \frac{4C(K)}{3\sigma_{\min}(Q)}.$$

The second condition in (D.8) then gives

$$\eta \|\Sigma_{K'}\| \|R + B^\top P_K B\| \le \frac{1}{8}.$$

Consequently, the calculation in Lemma B.3 yields

$$C(K') - C(K) \le -\frac{\eta}{4}\|\nabla C(K)\|_F^2 + \eta\left(\frac{16C(K)^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}\right)\|e(K)\|_F^2.$$

Finally, applying the gradient-domination inequality in Lemma B.1 and using $\|e(K)\|_F \le \mathcal{E}_K$ gives (D.9). □

Uniform parameter choices. We now choose the parameters in Lemmas D.1 and D.2 uniformly over the sublevel set

$$\mathcal{K}_0 := \{K \in \mathcal{K} : C(K) \le 2C_0\}.$$

By Lemma C.1, define

$$\overline{h}^f_{\text{cost}} := \sup_{K \in \mathcal{K}_0} h^f_{\text{cost}}(K), \qquad \overline{h}^f_{\text{grad}} := \sup_{K \in \mathcal{K}_0} h^f_{\text{grad}}(K), \qquad \overline{h}^a_{\text{cost}} := \sup_{K \in \mathcal{K}_0} h^a_{\text{cost}}(K).$$

These constants are finite and polynomially bounded in C_0 and the system parameters.

We first specify a uniform error budget for the inexact gradient step. Set

$$\Theta_0 := \frac{64C_0^2}{9\sigma_{\min}^2(Q)\sigma_{\min}^2(W)} + \frac{1}{4}, \qquad L_{\text{mf}} := \Theta_0 \frac{\|\Sigma_{K^\star}\|}{\sigma_{\min}^2(W)\sigma_{\min}(R)},$$

and define

$$e_0 := \left(\frac{C_0 \sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|\Theta_0}\right)^{1/2}.$$

Thus, with

$$\gamma := \eta \frac{\sigma_{\min}^2(W)\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|},$$

we have

$$\eta\Theta_0 e_0^2 = \gamma C_0. \tag{D.13}$$

Recall the uniform bounds A_0, S_0, G_0 introduced in Appendix B. Since $\nabla C_H(K) \to \nabla C(K)$, the bound defining G_0 also yields $\|\nabla C(K)\|_F \le G_0$ on $\mathcal{K}_0$. We may therefore set

$$\eta_0^{\mathrm{mf}} := \min\left\{ \frac{1}{16}\left(\frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{2C_0}\right)^2 \frac{1}{\|B\|(G_0+e_0)(1+A_0)},\ \frac{3\sigma_{\min}(Q)}{64C_0S_0},\ \frac{\|\Sigma_{K^\star}\|}{2\sigma_{\min}^2(W)\sigma_{\min}(R)} \right\}.$$

Then, for every $K \in \mathcal{K}_0$, whenever $\|\widehat{\nabla}(K) - \nabla C(K)\|_F \le e_0$, any $0 < \eta \le \eta_0^{\mathrm{mf}}$ satisfies the step-size condition of Lemma D.2.

We next choose the perturbation radius and horizon so that the smoothing and finite-horizon biases are uniformly small. For $K \in \mathcal{K}_0$, we have

$$C(K) \le 2C_0, \qquad \|A - BK\| \le A_0.$$

Moreover, for $H \ge 2$,

$$C_H(K) \ge \frac{H-1}{H}\operatorname{Tr}(QW) \ge \frac{1}{2}\operatorname{Tr}(QW), \qquad C(K) \ge \operatorname{Tr}(QW).$$

Hence the pointwise radius condition $r \le r_K$ in Lemma D.1 is guaranteed by $r \le r_{\mathrm{loc}}$, where

$$r_{\mathrm{loc}} := \min\left\{ \frac{\sigma_{\min}(Q)\sigma_{\min}(W)}{8C_0\|B\|(1+A_0)},\ 1,\ \frac{\operatorname{Tr}(QW)}{8\overline{h}^f_{\mathrm{cost}}},\ \frac{\operatorname{Tr}(QW)}{\overline{h}^a_{\mathrm{cost}}} \right\}.$$

For a target accuracy $\epsilon > 0$, define

$$r_0(\epsilon) := \min\left\{ r_{\mathrm{loc}},\ \frac{1}{4\overline{h}^f_{\mathrm{grad}}}\sqrt{\frac{\epsilon}{L_{\mathrm{mf}}}} \right\},$$

and

$$H_0(\epsilon) := \max\left\{ 2,\ 4\overline{L}_{\mathrm{grad}}\sqrt{\frac{L_{\mathrm{mf}}}{\epsilon}},\ \frac{2L_2}{\epsilon} \right\}.$$

Then $r \le r_0(\epsilon)$ and $H \ge H_0(\epsilon)$ imply

$$\overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H} \le \frac{1}{2}\sqrt{\frac{\epsilon}{L_{\mathrm{mf}}}}, \qquad \frac{L_2}{H} \le \frac{\epsilon}{2}.$$

Consequently, if $\zeta \le e_0/2$ and $\epsilon \le L_{\mathrm{mf}} e_0^2$, then

$$\zeta + \overline{h}^f_{\mathrm{grad}} r + \frac{\overline{L}_{\mathrm{grad}}}{H} \le e_0. \tag{D.14}$$

It remains to choose the sample size uniformly. By Lemma 3.2, for every $K \in \mathcal{K}_0$,

$$C_H(K) \le C(K) + \frac{L_{\mathrm{cost}}(K)}{H} \le 2C_0 + \overline{L}_{\mathrm{cost}} =: \overline{C}_H.$$

Also, since the constants ν_K and α_K in Lemma D.1 are polynomially bounded in $C(K)$ and the system and noise parameters, we may define

$$\overline{\nu} := \sup_{K \in \mathcal{K}_0} \nu_K, \qquad \overline{\alpha} := \sup_{K \in \mathcal{K}_0} \alpha_K.$$

These quantities are finite and polynomially bounded in C_0 and the system and noise parameters.

For $N \geq 1$, set

$$m_0(\zeta, \delta, N, r) := \left\lceil c\left(D \log 5 + \log \frac{N}{\delta}\right) \max\left\{ \frac{D^2 \overline{C}_H^2}{r^2\zeta^2}, \frac{D^2\overline{\nu}^2}{r^2\zeta^2}, \frac{D\overline{\alpha}}{r\zeta} \right\} \right\rceil.$$

Then, conditional on any $K_n \in \mathcal{K}_0$, $m \geq m_0(\zeta, \delta, N, r)$ allows Lemma D.1 to be applied with failure probability δ/N.

Lemma D.3. *Suppose*

$$0 < \zeta \leq \frac{e_0}{2}, \qquad 0 < \epsilon \leq L_{\mathrm{mf}} e_0^2, \qquad 0 < \delta < 1,$$

and let $N \geq 1$. If

$$0 < \eta \leq \eta_0^{\mathrm{mf}}, \qquad 0 < r \leq r_0(\epsilon), \qquad H \geq H_0(\epsilon), \qquad m \geq m_0(\zeta, \delta, N, r),$$

then, with probability at least $1 - \delta$,

$$K_n \in \mathcal{K}_0, \qquad n = 0, \ldots, N.$$

Proof. We argue by induction. The claim holds for $n = 0$ since $K_0 \in \mathcal{K}_0$.

Let $\mathcal{G}_n$ denote the sigma-field generated by all perturbations and rollout randomness used before iteration n, and suppose $K_n \in \mathcal{K}_0$. Conditional on $\mathcal{G}_n$, the policy K_n is fixed and the randomness used at iteration n is independent of the past. Hence, by the definition of $m_0(\zeta, \delta, N, r)$ and Lemma D.1,

$$\|\widehat{\nabla}_n - \nabla C_H^r(K_n)\|_F \leq \zeta \tag{D.15}$$

with conditional probability at least $1 - \delta/N$.

On the event (D.15), Lemma C.1, Lemma 3.3, and (D.14) yield

$$\begin{aligned} \|\widehat{\nabla}_n - \nabla C(K_n)\|_F &\leq \zeta + h_{\mathrm{grad}}^f(K_n) r + \frac{L_{\mathrm{grad}}(K_n)}{H} \\ &\leq \zeta + \overline{h}_{\mathrm{grad}}^f r + \frac{\overline{L}_{\mathrm{grad}}}{H} \leq e_0. \end{aligned}$$

Therefore, since $\eta \leq \eta_0^{\mathrm{mf}}$, Lemma D.2 implies that K_{n+1} is stabilizing and

$$C(K_{n+1}) - C(K^\star) \leq (1 - \gamma)\big(C(K_n) - C(K^\star)\big) + \eta \Theta_0 e_0^2.$$

Using (D.13), this becomes

$$C(K_{n+1}) - C(K^\star) \leq (1 - \gamma)\big(C(K_n) - C(K^\star)\big) + \gamma C_0. \tag{D.16}$$

Since $C(K_n) \leq 2C_0$,

$$\begin{aligned} C(K_{n+1}) - C(K^\star) &\leq (1 - \gamma)(2C_0 - C(K^\star)) + \gamma C_0 \\ &= 2C_0 - C(K^\star) - \gamma\big(C_0 - C(K^\star)\big) \\ &\leq 2C_0 - C(K^\star), \end{aligned}$$

where we used $C_0 \geq C(K^\star)$. Thus $C(K_{n+1}) \leq 2C_0$, and together with stability this gives $K_{n+1} \in \mathcal{K}_0$.

Applying the induction step for $n = 0, \ldots, N-1$ and taking a union bound over the N conditional failure events shows that

$$K_n \in \mathcal{K}_0, \qquad n = 0, \ldots, N,$$

with probability at least $1 - \delta$. □