# Safe-PG-LQR: Provably Safe, Convergent and Optimal Model-Free Linear Quadratic Regulator with Hard Constraints

**Vipul K. Sharma**
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
sharm697@purdue.edu

**Wesley A. Suttle**
U.S. Army Research Laboratory
Adelphi, Maryland 20783
suttlewes@gmail.com

**S. Sivaranjani**
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
sseetha@purdue.edu

## Abstract

The linear quadratic regulator (LQR) is a canonical optimal control problem where recent work has established global convergence guarantees for model-free policy gradient methods. However, designing model-free algorithms that enforce hard constraints throughout learning remains an open challenge. Existing approaches are either partially model-based, enforce soft constraints through Lagrangian relaxations, or project pretrained policies onto the constraint set, sacrificing hard constraint satisfaction or convergence guarantees of the underlying algorithms. In this work, we develop principled model-free policy gradient methods for LQR problems with hard constraints. Our key idea is to parameterize the policy class using a truncated stochastic policy so that hard constraints are satisfied by construction, then perform policy optimization entirely within this class. We prove that the resulting LQR objective over the proposed truncated policy class satisfies a weak gradient-dominance property in both the undiscounted and discounted settings. We leverage this property to develop model-free policy-gradient algorithms that maintain zero violation of the encoded hard constraints at each iteration, and converge to an approximately globally optimal policy within the proposed safe policy class.

## 1 Introduction

The Linear Quadratic Regulator (LQR) is a canonical optimal control problem [24, 7, 11] that has emerged as a key benchmark to analyze the optimization and convergence properties of reinforcement learning (RL) methods in continuous control settings. Specifically, recent work has shown that model-free policy gradient algorithms applied to LQR enjoy global convergence guarantees despite the nonconvexity of the policy optimization landscape [20]. Despite these advances, existing convergence guarantees for model-free LQR do not extend to settings with hard safety constraints. RL-based works typically address safety through constrained or risk-sensitive formulations, such as constrained Markov decision processes (CMDPs) [3], wherein constraint violations are controlled in expectation or with high probability and hard constraint satisfaction during learning is not guaranteed. In addition, while these approaches are well suited to discrete or episodic settings, they do not ensure pointwise hard constraint satisfaction along all system trajectories, particularly in continuous-state,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

infinite-horizon control problems. In parallel, the control theory literature has developed a rich set of tools for enforcing hard constraints in continuous-time and continuous-state systems, including control barrier functions, reachability-based methods, and projection-based safety filters [10, 6]. These approaches provide strong feasibility guarantees, often ensuring zero constraint violation by construction. However, they typically rely on modifying or projecting a nominal controller at each time step, which alters the optimization landscape and breaks the optimality and convergence guarantees of the underlying algorithms.

**Main Contributions.** We establish theoretical foundations and derive model-free algorithms for constrained LQR, while providing simultaneous guarantees of hard constraint satisfaction, and convergence to approximate global optimality in both the discounted and undiscounted settings. Our contributions are as follows:

- **Stochastic wrappers.** We introduce a stochastic wrapper around the deterministic linear state-feedback control in LQR to enable safe exploration in model-free learning. The construction is tailored to infinite-horizon LQR with hard constraints, and forms the foundation for well-defined policy gradients and convergence analysis in this setting.
- **Hard-constrained policy parameterization.** To enforce hard state-action constraints, we restrict the support of the stochastic wrapper to a state-dependent constraint set, yielding a class of *truncated stochastic policies* that satisfy constraints by construction. Unlike CMDP-based or projection-based approaches, safety is enforced directly at the policy level throughout learning, without any post-hoc alteration of control actions. We also derive exact gradients for this truncated policy class, allowing us to develop model-free policy gradient algorithms.
- **Well-posedness under truncated policy exploration.** Truncation induces state-dependent, biased, and temporally correlated noise processes. We show that, under mild regularity conditions, the resulting system admits well-defined infinite-horizon costs and gradients, despite the loss of a uniform quadratic structure and independent exploration noise. We accomplish this by leveraging the geometry of the feasible domain and stability properties of LQR to bound the accumulation of truncation-induced noise.
- **Gradient dominance for constrained LQR.** We establish gradient dominance properties for the constrained LQR objective over the truncated policy class for both undiscounted and discounted infinite-horizon settings. This necessitates the development of new mathematical machinery to control the cost difference between optimal deterministic and stochastic hard-constrained policies by explicitly characterizing and bounding truncation-induced bias terms.
- **Provably safe, convergent, and globally optimal algorithms.** Leveraging these gradient dominance results under hard-constrained policies, we develop model-free policy gradient algorithms with convergence, sample complexity, and approximate global optimality guarantees that simultaneously guarantee hard constraint satisfaction at every iteration. Establishing these guarantees is nontrivial, as it requires proving local Lipschitzness of the truncated objective and controlling the long-horizon effects of truncation-induced noise in the absence of model knowledge.

**Related Work. (i) Policy Optimization for LQR.** Global convergence guarantees for model-free policy optimization in unconstrained LQR have been established via policy gradient analyses [20], with extensions to derivative-free methods and refined complexity analyses [23, 12, 13, 27, 16]. However, these analyses critically rely on gradient dominance under the unconstrained stabilizing policy class and thus do not extend to constrained LQR. **(ii) Safe RL.** Safety in RL is often addressed through constrained MDPs [3] or risk-sensitive objectives [9] and primal-dual policy optimization methods [1, 8, 29, 18]. Such cost- and risk-constrained formulations have also been applied to LQR, where constraint satisfaction is enforced in expectation or probability [44, 45, 38, 42]. While these approaches provide statistically meaningful safety guarantees in expectation, they do not guarantee pointwise satisfaction of deterministic state, input, or state-action coupled constraints over infinite horizons. A recent line of works study hard instantaneous constraint enforcement in RL by using custom neural network optimization layers [30], generalized reduced gradient based policy updates [19] for neural network policies, or optimizing directly over truncated policy parameterizations [36]. The first two works rely on action projection and do not consider convergence guarantees, while the latter only preserves convergence for the underlying RL algorithm updates, and does not establish convergence or global optimality guarantees for the resulting constrained continuous-control problem. Here, we construct truncated constraint-preserving policies that are tailored to LQR, which helps establish convergence and global optimality guarantees. **(iii) Safe learning-based control.** Hard state and input constraints in continuous control are typically handled using frameworks such as control barrier functions (CBFs), including uncertain and risk-sensitive formulations, reachability analysis,

and safety filters [10, 6, 5, 39, 34, 14, 22]. However, these approaches are typically model-based, and enforce constraints by projecting a nominal control action onto the safe set through quadratic programming (QP), which can break the convergence of the underlying learning algorithms [36]. Constrained LQR has also been studied, either via MPC-like finite-horizon QP-based approximations in model-based settings [33], or through system identification followed by robust control synthesis [17, 26] rather than fully model-free policy optimization. These works cannot typically handle state-action coupled constraints such as CBFs that frequently arise in safety-critical control. In contrast to the above literature, we directly address constrained LQR in a fully model-free, infinite-horizon setting, and enforce hard constraints throughout learning by embedding safety directly into the policy class. This approach allows us to establish gradient dominance and convergence to approximate global optimality in the constrained LQR setting, all while maintaining hard safety guarantees.

## 2 Problem Formulation

We begin by formalizing the infinite-horizon LQR problem under hard constraints and introduce a new class of safe policies that are suitable for model-free optimization. Specifically, we define truncated stochastic LQR policies that enforce constraints by construction while enabling exploration, and reformulate the constrained LQR problem as an unconstrained optimization problem over this policy class. We also derive policy gradient expressions for our truncated policies, which form the basis for the analysis in subsequent sections.

### 2.1 LQR under Hard Constraints

We consider the constrained infinite-horizon LQR problem, termed the *safe LQR problem*, given by

$$\begin{aligned} \underset{\{u_t\}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t)\right] \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t, \quad x_0 \sim \mathcal{D}, \\ & G(x_t, u_t) \leq 0, \end{aligned} \tag{1}$$

where $x_t \in \mathcal{X} \subseteq \mathbb{R}^n$ denotes the state in the operating domain on which the admissible-control map is defined, and $u_t \in \mathcal{U} \subseteq \mathbb{R}^m$ denotes the control input at time $t$. The matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ denote the system and input matrices, the initial condition $x_0$ is drawn from a distribution $\mathcal{D}$ over $\mathbb{R}^n$, matrices $Q \geq 0, R > 0$ denote the cost weighting matrices, and $\gamma \in (0, 1]$ is a discount factor, with $\gamma = 1$ representing the undiscounted setting. Here $G(x, u) \leq 0$, where $G(x, u) : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^g$ is a vector-valued function, represents the time-varying state-action coupled hard constraints, termed as *safety constraints*. For a given state $x \in \mathcal{X}$, we define the *safe control set* to be the set of all admissible control inputs $u \in \mathcal{U}$ satisfying the pointwise hard constraints encoded by $G$ at state $x$, given by

$$H(x) = \{u \in \mathcal{U} \mid G(x, u) \leq 0\}. \tag{2}$$

**Assumption 2.1.** For every $x \in \mathcal{X}$, the safe control set $H(x)$ in (2) is a convex subset of $\mathbb{R}^m$.

*Remark* 2.2. The hard constraints in (1) capture a broad class, including action constraints, state constraints encoded through admissible-control sets, and state-action coupled constraints such as CBF-type constraints [33, 17, 6]. When $G(x, u) \leq 0$ is chosen to encode a valid CBF-type admissibility condition, enforcing $u_t \in H(x_t)$ recovers the corresponding forward-invariance guarantee. Our framework is entirely agnostic to how $H(x)$ is constructed: once $H(x)$ is specified, the truncated policy enforces the resulting admissible-control condition $u_t \in H(x_t)$ by construction. Assumption 2.1 holds when the resulting admissible-control set is convex [21, 40, 4], or can be enforced via convex inner approximations [43, 15].

Before introducing safe LQR policies, we make a few technical assumptions to ensure that integrals over $H(x)$ are well-defined and that constraint-satisfying stabilizing controllers exist.

**Assumption 2.3.** There exist $b_l, b_h > 0$ such that $b_l \leq \mu(H(x)) \leq b_h$, for all $x \in \mathcal{X}$, where $\mu$ denotes the Lebesgue measure over $\mathbb{R}^m$.

Assumption 2.3 excludes degenerate cases in which the safe control set collapses to a measure-zero set. For safety constraints such as standard affine-in-control CBFs, this assumption holds when their intersection with bounded actuator constraints has nonempty interior in the control space.

### 2.2 Stochastic Wrappers for Safe LQR Policies

We next introduce a class of safe stochastic LQR policies that enable exploration in a model-free setting while enforcing hard constraints by construction, providing the first major contribution of the paper. The key idea is to introduce a stochastic wrapper, defined below, through a Gaussian nominal sampling distribution and then restrict its support to the safe control set. The use of the stochastic wrapper enables a model-free approach to the constrained LQR problem via use of the policy gradient theorem, guarantees safety by construction, and, as we show in subsequent sections, allows us to derive powerful gradient dominance and convergence guarantees. However, unlike standard stochastic policy parameterizations, the resulting exploration noise is state-dependent, truncated, and temporally correlated, placing our setting outside the scope of existing policy gradient analyses for LQR.

**Constraint-Preserving Stochastic Wrappers.** We recall that, for the unconstrained LQR problem, the optimal control policy is a deterministic linear state-feedback law obtained by solving a Riccati equation [7]. Such a policy need not satisfy the hard constraints $G(x, u) \leq 0$, and its computation requires knowledge of the system matrices $A$ and $B$ in (1). In practice, however, such system knowledge may not be available, making it desirable to learn an optimal policy through interaction with and exploration of the system. One way to enable exploration in deterministic, model-free control is to perturb a deterministic nominal policy with additive stochastic noise. In our LQR setting, we achieve this by sampling control inputs from a Gaussian distribution centered at the deterministic nominal map $c_K(x)$, which is not necessarily safe. Such Gaussian exploration schemes have been explored in model-free policy gradient methods for unconstrained LQR and more general deterministic control problems to facilitate gradient estimation [20, 28, 35], but their analysis does not extend to constrained settings. We define a stochastic LQR policy, with parameter $K$, with density

$$\pi_K(u \mid x_t) = (2\pi)^{-\frac{m}{2}} \det(\Sigma_t)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}[u - c_K(x_t)]^\top \Sigma_t^{-1}[u - c_K(x_t)]\right), \tag{3}$$

termed a *stochastic wrapper*, with mean $c_K(x_t)$ and covariance $\Sigma_t = \sigma_t^2 I_m$, where $I_m$ is the identity matrix and $\sigma_t > 0$ controls the exploration magnitude at time $t$. Notice that sampling a control input $u \sim \pi_K(\cdot \mid x_t)$ is thus equivalent to applying the deterministic nominal control $c_K(x_t)$ plus zero-mean Gaussian noise with covariance $\Sigma_t$.

**Assumption 2.4.** The variance of the exploration noise decays as $\sigma_t^2 = \frac{\sigma^2}{(t+1)^2}$, with $\sigma > 0$.

This assumption will be used to control the cumulative contribution of exploration noise over the infinite horizon, and establish finite second-moment bounds for the resulting closed-loop trajectories in subsequent analyses. While the policy $\pi_K(\cdot|x_t)$ enables exploration and derivation of policy gradient expressions, it does not, in general, satisfy the hard constraints in (1). To enforce safety, we adopt a truncation mechanism similar to [36], where the infinite support of the Gaussian policy $\pi_K(\cdot|x_t)$ is truncated to the safe control set $H(x_t)$. More precisely, given a policy $\pi_K$, we consider the truncated Gaussian distribution $\pi_K^H(\cdot|x_t)$ defined by

$$\pi_K^H(u|x_t) = \begin{cases} \dfrac{\pi_K(u|x_t)}{\pi_K(H(x_t)|x_t)}, & u \in H(x_t), \\ 0, & u \notin H(x_t), \end{cases} \tag{4}$$

where $\pi_K(H(x_t)|x_t) = \int_{H(x_t)} \pi_K(u|x_t)\, du$. Note that this *truncated policy* $\pi_K^H(\cdot|x_t)$ has a well-defined density function under Assumption 2.3, and restricts control values to the safe control set $H(x_t)$. This truncation therefore enforces hard constraints directly at the policy level, rather than through projection or safety filtering, ensuring that all sampled control inputs are feasible by design.

**Assumption 2.5** (Linear Safe Policy Class)**.** Let $\mathcal{S}_{\mathcal{K}}^{\gamma} = \begin{cases} \mathcal{K}^{\gamma}, & \gamma \in (0,1), \\ \mathcal{K}, & \gamma = 1, \end{cases}$ where $\mathcal{K} := \{K \in \mathbb{R}^{m\times n} : \rho(A_K) < 1\}$, $\mathcal{K}^{\gamma} := \{K \in \mathbb{R}^{m\times n} : \sqrt{\gamma}\rho(A_K) < 1\}$, with $A_K := A + BK$, and where $\rho(\cdot)$ denotes the spectral radius. There exists a nonempty set $\mathcal{K}_H^{\gamma} \subseteq \mathcal{S}_{\mathcal{K}}^{\gamma}$ with nonempty relative interior such that, for every $K \in \mathcal{K}_H^{\gamma}$, the truncated policy $\pi_K^H$ satisfies $\mu_K^H(x) := \mathbb{E}_{u\sim\pi_K^H(\cdot|x)}[u] = Kx$, for every state $x$ reachable under the closed-loop dynamics induced by $\pi_K^H$ from $x_0 \sim \mathcal{D}$.

Assumption 2.5 imposes a linear feedback structure on the conditional mean of the safe truncated policy along the closed-loop trajectories considered in the analysis. Consequently, all optimization, convergence, and optimality claims are with respect to $K \in \mathcal{K}_H^{\gamma}$. Here, $c_K(x)$ specifies the Gaussian

center before truncation, while $Kx$ denotes the conditional mean after truncation; one way to choose $c_K(x)$ is through the moment-matching problem discussed in Remark 7.3. This condition holds in many standard constrained-regulation settings where the desired linear mean remains strictly feasible for the safe control set, including state-dependent box, ellipsoidal, and polyhedral admissible-control sets. Additionally, while the optimal safe controller for (1) may in general be nonlinear, most LQR and constrained LQR works, e.g., [17], consider structured linear policy parameterizations. We similarly restrict attention to a truncated stochastic policy class whose post-truncation mean is linear along the closed-loop trajectories considered in the analysis. However, unlike existing constrained LQR works, our focus is on model-free convergence and optimality guarantees while maintaining hard constraint satisfaction by construction. Even in this policy class, obtaining theoretical guarantees in a model-free setting is highly nontrivial, as detailed in the following analysis. Finally, we impose the following regularity conditions on the deterministic nominal map $c_K$ to ensure that the associated score functions and policy gradients are well-behaved.

**Assumption 2.6** (Nominal Map Regularity)**.** For every $K \in \mathcal{K}_H^\gamma$, $c_K : \mathcal{X} \to \mathbb{R}^m$ is measurable in $x$ and twice continuously differentiable in $K$, with $K$-derivatives denoted by $D_K c_K(x)$ and $D_K^2 c_K(x)$. For every compact $\mathcal{C} \subset \mathcal{K}_H^\gamma$, there exist constants $M_{\mathcal{C}}^c, M_{\mathcal{C}}^m < \infty$ such that, for all $K \in \mathcal{C}$ and all $x$ reachable under the closed-loop trajectories considered in the analysis, $\|D_K c_K(x)\| + \|D_K^2 c_K(x)\| \le M_{\mathcal{C}}^c$. Moreover, the induced truncated moments $m_K^H(x) := \mathbb{E}_{u\sim\pi_K^H(\cdot|x)}[u]$, $S_K^H(x) := \mathbb{E}_{u\sim\pi_K^H(\cdot|x)}[uu^\top]$ are twice continuously differentiable in $K$, and their first and second $K$-derivatives are uniformly bounded by $M_{\mathcal{C}}^m$ over all $K \in \mathcal{C}$ and all reachable $x$ considered in the analysis.

Assumption 2.6 is a local regularity condition imposed on the operating domain and reachable states considered in the analysis, and rules out pathological choices of $c_K$ and degenerate truncation geometries for which the truncated moments change nonsmoothly with the policy parameter.

To aid analysis, we decompose control inputs from the truncated policy into a deterministic mean and a zero-mean stochastic term, isolating the effect of safe exploration on the closed-loop dynamics. Along closed-loop trajectories generated by any $K \in \mathcal{K}_H^\gamma$, for $u \sim \pi_K^H(\cdot|x)$, we can write

$$u = \mu_K^H(x) + \eta^H = Kx + \eta^H, \qquad \mu_K^H(x) := \mathbb{E}_{y\sim\pi_K^H(\cdot|x)} y, \tag{5}$$

$$\eta^H := u - \mu_K^H(x). \tag{6}$$

By construction, $\mathbb{E}[\eta^H \mid x] = 0$.

**Proposition 2.7.** *Suppose Assumptions 2.3 and 2.5 hold. Furthermore, let $\mathcal{F}^H$ denote the pre-decision filtration generated by the initial state and past truncated exploration noises, defined as*

$$\mathcal{F}_0^H := \sigma(x_0), \qquad \mathcal{F}_i^H := \sigma(x_0, \eta_0^H, \dots, \eta_{i-1}^H), \quad i \ge 1. \tag{7}$$

*Then $\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H] = 0$ for all $i \ge 0$, and $\mathbb{E}[\eta_i^H (\eta_j^H)^\top] = 0$ for all $i \ne j$.*

*See proof in Section 7.3.* Proposition 2.7 establishes the martingale-difference property of the truncated exploration noise and the vanishing of its cross-time second moments, which will be used in subsequent analysis.

### 2.3 Policy Gradients for Safe LQR

We now show that the constrained LQR problem admits a policy gradient formulation when restricted to the truncated stochastic policy class introduced in Section 2.2. Since hard constraints are enforced implicitly through the support of the policy, the original constrained optimal control problem can be reformulated as an unconstrained optimization over the control parameter $K$, thus enabling model-free policy optimization while maintaining hard constraints by construction. Specifically, we rewrite (1) over the hard-constraint satisfying truncated policy class $\pi_K^H(\cdot|x_t)$ as the following problem:

$$\begin{aligned} \underset{K \in \mathcal{K}_H^\gamma}{\text{minimize}} \quad & J_\pi^H(K) = \mathbb{E}_{x_0\sim\mathcal{D}, u_t\sim\pi_K^H}\Big[\sum_{t=0}^{\infty} \gamma^t (x_t^\top Q x_t + u_t^\top R u_t)\Big] \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t, \end{aligned} \tag{8}$$

In the optimization problem (8), constraint satisfaction is guaranteed implicitly through the policy parameterization and no explicit constraints other than the system dynamics appear in the formulation.

Our goal is to solve problem (8) in a model-free manner, that is, without knowledge of matrices $A$ and $B$. The objective $J_\pi^H(K)$ depends on the unknown system dynamics only through sampled trajectories, and can therefore be estimated from data without access to $A$ and $B$. Further, since $J_\pi^H(K)$ is defined entirely through the truncated stochastic policy $\pi_K^H(\cdot|x_t)$ with controller gain $K$ as the optimization parameter, we can derive policy gradient expressions that can be evaluated using only sampled state and control trajectories, thus enabling model-free algorithms.

# 3 Undiscounted Safe LQR

We begin by considering the undiscounted hard-constrained LQR problem ($\gamma = 1$) under the truncated stochastic policy class introduced in Section 2. The undiscounted setting is particularly natural in control, as closed-loop stability of the system is classically sufficient to ensure that the infinite-horizon cost is well-defined, without the need for an explicit discount factor. However, in the hard-constrained setting considered here, closed-loop stability is no longer sufficient to ensure well-posedness as truncation induces state-dependent and correlated noise, necessitating a more careful analysis. In this section, we establish that the undiscounted cost and its policy gradient are well-defined under the truncated policy class and that the resulting objective satisfies a weak gradient dominance property. Building on these results, we develop a model-free policy gradient algorithm and show that it converges to an approximately globally optimal safe policy, with explicit bounds on the optimality gap and sample complexity. Together with the discounted analysis in Section 4, these results form the main theoretical contributions of the paper.

## 3.1 Well-Posedness and Optimality Difference

We now analyze the undiscounted infinite-horizon objective induced by our truncated stochastic policies and establish regularity properties of the objective. Specifically, we show that the expected cost remains finite and admits a decomposition into the deterministic LQR objective and additional terms arising from truncated exploration. This explicitly characterizes how persistent, state-dependent noise enters the infinite-horizon cost and provides the foundation for the gradient dominance and convergence analysis developed in subsequent subsections. A key technical challenge is controlling the accumulation of stochastic exploration effects over an infinite horizon. We achieve this by combining the decay condition in Assumption 2.4 with the geometric properties of the admissible control set, which together allow us to bound the second moments of the induced noise terms. These bounds play a central role in establishing well-defined costs, gradients, and convergence guarantees.

We begin by establishing that the infinite-horizon cost $J_\pi^H(K)$ induced by the truncated policy is well-defined, starting from stability properties of the corresponding deterministic policy. Considering the safe deterministic policy $\mu_K^H(x) = Kx$ with $\gamma = 1$, standard LQR theory ensures closed-loop stability and boundedness of the corresponding cost, defined as $J_\mu^H(K) = \mathbb{E}_{x_0} V_K(x_0)$, where $V_K(x_0) = \sum_{i=0}^\infty x_i^T Q x_i + u_i^T R u_i = x_0^\top P_K x_0$ and $P_K$ satisfies the discrete-time Lyapunov equation for the closed-loop system given by $P_K = A_K^\top P_K A_K + Q_K^R$, where $Q_K^R := Q + K^T R K$. This deterministic cost serves as a reference against which we quantify the additional contribution introduced by truncated stochastic exploration. Finally, we define vector $\bar{\eta}_t^H$ and matrix $\mathcal{C}_t$ by

$$\bar{\eta}_{t+1}^H := \begin{bmatrix} {\eta_t^H}^\top & \dots & {\eta_0^H}^\top \end{bmatrix}^\top, \bar{\eta}_0^H := 0_{m\times 1}, \tag{9}$$

$$\mathcal{C}_{t+1} := \begin{bmatrix} B & A_K B & A_K^2 B & \cdots & A_K^t B \end{bmatrix}, \mathcal{C}_0 := 0_{m\times mn}. \tag{10}$$

We now have the following result establishing the relationship between the deterministic cost and the stochastic cost under the truncated policy.

**Theorem 3.1.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then,*

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K), \quad \nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta(K), \tag{11}$$

*where* $\alpha_1(K) := \sum_{i=0}^\infty \mathbb{E}_{\pi_K^H, x_0}\left[{\eta_i^H}^\top R \eta_i^H\right]$, $\alpha_2(K) := \sum_{i=0}^\infty \mathbb{E}_{\pi_K^H, x_0}\left[(\mathcal{C}_i \bar{\eta}_i^H)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\right]$ *and* $\Delta(K) := \alpha_1(K) + \alpha_2(K)$.

*See proof in Section 7.4.* We now show that the cost under the truncated policy and its gradient are well-defined under mild assumptions.

**Theorem 3.2.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then, the undiscounted cost* $J_\pi^H(K)$ *and its gradient* $\nabla_K J_\pi^H(K)$ *are well-defined over the policy class* $\pi_K^H$.

*See proof in Section 7.4.* In contrast to the mild assumptions required above, analyses based on unstructured dynamics and policy classes typically require strong global assumptions, such as ergodicity and log-Lipschitz regularity of the induced Markov process and policy parameterization [28, 36]. We now state another key result that characterizes the optimality gap between the truncated stochastic cost and the deterministic LQR cost.

**Theorem 3.3** (Optimality Difference)**.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Further, let* $K_\pi^* \in \arg\min_{K \in \mathcal{K}_H^\gamma} J_\pi^H(K)$ *and* $K_\mu^* \in \arg\min_{K \in \mathcal{K}_H^\gamma} J_\mu^H(K)$ *with respective costs given by* $J_\pi^{H^*}$ *and* $J_\mu^{H^*}$*. Then, with* $\beta^* := \Delta(K_\mu^*)$*, we have* $0 \leq J_\pi^{H^*} - J_\mu^{H^*} \leq \beta^*$.

*See proof in Section 7.4.* Here, $\beta^*$ represents the inherent *"price of safety"* arising from sampling control inputs from a truncated stochastic policy rather than the deterministic linear mean policy. This suboptimality reflects the necessary safety-optimality trade-off introduced by safe exploration.

### 3.2 Convergence, Optimality, and Sample Complexity

We now combine the structural results established in the preceding subsections to derive convergence guarantees for the model-free policy gradient method in Algorithm 1 introduced in Section 7.1.1, which can be understood as a specific instantiation of the classic undiscounted policy gradient update [37]. In particular, leveraging the well-posedness of the undiscounted cost, the explicit optimality gap induced by truncation, and the regularity of the policy gradient, we show that policy gradient updates of the form $K_{t+1} = K_t - \zeta_t \nabla J_\pi^H(K_t)$, with step-size $\zeta_t$, converge to a neighborhood of the globally optimal safe controller. Furthermore, we explicitly characterize the radius of this neighborhood in terms of the truncation-induced bias, yielding approximate global optimality guarantees together with convergence and sample complexity bounds. In unconstrained LQR, gradient dominance relates suboptimality to the norm of the policy gradient [20]. In our hard-constrained setting, truncation introduces an intrinsic optimality gap, and we therefore establish a relaxed gradient dominance condition that explicitly accounts for this. By Lemma 7.20, the policy gradient is bounded on the sublevel sets considered below. To simplify constants, we use the normalized convention $G_\mathcal{L} = 1$; the general case is obtained by absorbing $G_\mathcal{L}$ into $\alpha^H$ and $\alpha_\gamma^H$ in the following results.

**Theorem 3.4** (Weak Gradient Dominance)**.** *Let* $\mathcal{L}_0 := \left\{K \in \mathcal{K}_H^\gamma : J_\pi^H(K) \leq J_\pi^H(K_0)\right\}$*. Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold, and that the conditions of Lemma 7.20 hold. Then, for all* $K \in \mathcal{L}_0$*,*

$$J_\pi^H(K) - J_\pi^{H^*} \leq \alpha^H \left\|\nabla_K J_\pi^H(K)\right\| + \beta^H, \tag{12}$$

*where* $\alpha^H := \alpha\left(1 + 2\sup_{K \in \mathcal{L}_0}\{\|\nabla_K \Delta(K)\|\}\right)$, $\alpha = \frac{\|\Sigma_K^*\|}{\sigma_{min}(\Sigma_K)^2 \sigma_{min}(R)}$, $\Sigma_K = \mathbb{E}_{x_0}\left[\sum_{t=0}^\infty x_t x_t^\top\right]$ *and* $\beta^H := \sup_{K \in \mathcal{L}_0}\left\{\Delta(K) + \alpha \left\|\nabla_K \Delta(K)\right\|^2\right\}$.

*See proof in Section 7.4.* This result establishes a weak gradient dominance condition for the truncated LQR objective. Unlike in the unconstrained setting, suboptimality cannot be driven to zero due to the intrinsic optimality gap characterized in Theorem 3.3. Instead, we show that the norm of the policy gradient controls suboptimality up to an additive bias term determined by truncation. This condition is sufficient to enable convergence of our policy gradient algorithm, while explicitly accounting for the price of safety induced by hard constraints.

We now leverage this relaxed gradient dominance condition to establish convergence and sample complexity guarantees for model-free policy gradient updates applied to the truncated LQR objective. To analyze the convergence of model-free policy gradient updates, we require a standard regularity assumption on the stochastic gradient estimator used by the algorithm.

**Assumption 3.5** (Gradient Estimator)**.** The gradient estimator $\widehat{\nabla_K J_\pi^H(K)}$, computed with batch size $N > 0$, satisfies

$$\mathbb{E}\left[\widehat{\nabla_K J_\pi^H(K)}\right] = \nabla_K J_\pi^H(K), \quad \mathbb{E}\left[\left\|\widehat{\nabla_K J_\pi^H(K)} - \nabla_K J_\pi^H(K)\right\|_F^2\right] \leq \frac{V}{N}. \tag{13}$$

Assumption 3.5 is standard in policy gradient analyses for learning deterministic policies with stochastic policy gradients [28, 35], and allows us to control sampling-induced error. We are now ready to establish convergence and sample complexity guarantees for safe undiscounted LQR.

**Theorem 3.6.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, 2.6 and 3.5 hold. Further, suppose that Algorithm 1 runs for* $T > 0$ *iterations with updates* $\{K_i\}_{i=0}^T \subset \mathcal{K}_H^\gamma$*, batch size* $N > 0$ *and step-size*

*$\zeta$ following* $\zeta \leq \min\left\{\frac{1}{l^{mx}(T)}, \frac{(\alpha^H)^2}{c_0}, \left(\frac{N(\alpha^H)^2}{l^{mx}(T)V}\right)^{1/3}\right\}$, *where* $c_t := \max\{0, J_\pi^H(K_t) - J_\pi^{H^*} - \beta^H\}$ *and* $l^{mx}(T) := \max\{l(K_0), \ldots, l(K_T)\}$. *Then, we have*

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \beta^H + \beta^* + c_T\left(1 - \sqrt{\frac{\zeta^3 l^{mx}(T)V}{4(\alpha^H)^2 N}}\right)^T + \sqrt{\frac{l^{mx}(T)V\zeta(\alpha^H)^2}{N}}. \quad (14)$$

*Consequently, to guarantee* $\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \epsilon + \beta^H + \beta^*$, *for* $\epsilon > 0$ *and step size* $\zeta = \frac{\epsilon^2 N}{4\, l^{mx}(T)V(\alpha^H)^2}$ , *the sample complexity follows* $NT \leq \frac{16(\alpha^H)^4 l^{mx}(T)V}{\epsilon^3}\log\frac{c_T}{\epsilon}$.

*See proof in Section 7.4.* This result establishes convergence and sample complexity guarantees for Algorithm 1 under hard constraints enforced via truncation. Specifically, Algorithm 1 converges to an $\epsilon$-optimal policy up to the intrinsic optimality gap characterized in Theorem 3.3, with sample complexity $NT \leq \mathcal{O}(\frac{1}{\epsilon^3}\log\frac{1}{\epsilon})$. The constants in the stepsize condition are used for the theoretical guarantee and can be absorbed into a tuned or diminishing stepsize in practice. This shows that the proposed model-free approach achieves approximate global optimality with polynomial sample complexity, despite the presence of truncation-induced stochasticity over an infinite horizon.

# 4 Discounted Safe LQR

We now consider the discounted ($\gamma \in (0,1)$) safe LQR problem with hard constraints under the truncated stochastic policy class of Section 2. The discounted setting is popular in the standard RL literature due to its practical utility and ease of analysis, but existing works on safe RL in the general discounted setting provide safety and convergence to local minima while failing to guarantee global optimality. We address this gap by first establishing gradient dominance of the discounted objective and then proving convergence to near global optimality, all while satisfying hard safety constraints due to the special structure of our truncated policy class. In addition, we explicitly quantify the suboptimality that arises due to the presence of hard constraints in our approach to the constrained LQR problem. The core technical challenge lies in handling the discount factor, which alters the manner in which exploration noises accumulate in the convergence analysis.

## 4.1 Well-Posedness and Optimality Difference

We first establish that the discounted cost is well-defined and provide bounds characterizing the suboptimality gap between our optimal truncated stochastic policies and an optimal deterministic linear controller. In the discounted framework, the cost function emphasizes short-term performance over infinite-horizon stability, yet the truncation procedure required for safe exploration introduces additional noise-dependent terms during the learning process. It is thus necessary to verify that the discounted cost $J_\pi^H(K)$ remains well-defined to ensure that the gradient-based learning procedure we propose below is justified. Considering the safe deterministic policy $\mu_K^H(x) = Kx$, we define the corresponding cost as $J_\mu^H(K) = \mathbb{E}_{x_0} V_K(x_0)$, where $V_K(x_0) = \sum_{i=0}^\infty \gamma^i \left(x_i^T Q x_i + u_i^T R u_i\right) = x_0^\top P_K^\gamma x_0$, and $P_K$ satisfies the discrete-time Lyapunov equation $P_K^\gamma = Q_K^R + \gamma(A_K)^\top P_K^\gamma A_K$. With this definition in hand, we provide policy gradient expressions for $\nabla J_\pi^H(K)$ in terms of $\pi_K^H$ and $\mu_K^H$.

**Theorem 4.1.** *Suppose Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then, we have*

$$J_\pi^H(K) = J_\mu^H(K) + \Delta_\gamma(K), \quad \nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta_\gamma(K), \quad (15)$$

*where* $\alpha_1^\gamma(K) := \sum_{t=0}^\infty \gamma^t \mathbb{E}_{\pi_K^H, x_0}\left[\eta_t^{H^\top} R \eta_t^H\right]$, $\alpha_2^\gamma(K) := \sum_{t=0}^\infty \gamma^t \mathbb{E}_{\pi_K^H, x_0}\left[(\mathcal{C}_t\bar{\eta}_t^H)^\top Q_K^R \mathcal{C}_t \bar{\eta}_t^H\right]$ *and* $\Delta_\gamma(K) := \alpha_1^\gamma(K) + \alpha_2^\gamma(K)$.

*See proof in Section 7.5.* Just as in the undiscounted case, the gradient under safe exploration directly depends on the deterministic cost gradient up to an additional term that accounts for the accumulation of discounted infinite-horizon noises. Now that we are equipped with expressions for $J_\pi^H(K)$ and $\nabla J_\pi^H(K)$, we establish that both are well-defined.

**Theorem 4.2.** *Suppose Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then the discounted cost* $J_\pi^H(K)$ *and its gradient* $\nabla_K J_\pi^H(K)$ *are well-defined over the policy class* $\pi_K^H$.

*See proof in Section 7.5.* We complete this subsection by characterizing the suboptimality gap between optimal truncated stochastic policies and optimal deterministic linear controllers. This result is essential to the global optimality and convergence results of the next section.

**Theorem 4.3** (Optimality Difference)**.** *Suppose Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Furthermore, let* $K_\pi^* \in \arg\min_{K \in \mathcal{K}_H^\gamma} J_\pi^H(K)$ *and* $K_\mu^* \in \arg\min_{K \in \mathcal{K}_H^\gamma} J_\mu^H(K)$ *Then, with* $\beta_\gamma^* := \Delta_\gamma(K_\mu^*)$*, we have* $0 \leq J_\pi^{H^*} - J_\mu^{H^*} \leq \beta_\gamma^*$.

*See proof in Section 7.5.* This result characterizes the suboptimality gap introduced due to safe exploration, explicitly quantifying the trade-off between enforcing safety through truncation and achieving global optimality. Specifically, as a result of the truncation procedure proposed in Section 2, Theorem 4.3 shows that optimal policies from the class of safe truncated policies achieve within $\beta_\gamma^*$ of optimal deterministic linear policies.

### 4.2 Convergence, Optimality, and Sample Complexity

We now move on to establishing weak gradient dominance of the discounted objective as well as convergence and sample complexity results for Algorithm 2 introduced in Section 7.1.2, which is of the general form $K_{t+1} = K_t - \zeta_t \nabla J_\pi^H(K_t)$, where $\nabla J_\pi^H(K_t)$ is a discounted policy gradient estimate [25]. Taken together, our gradient dominance and convergence results imply that searching over the class of truncated policies $\pi_K^H$ is guaranteed to achieve within a neighborhood of a globally optimal solution to the discounted LQR problem with hard constraints. We emphasize that, though the gradient dominance and convergence results of this section are similar to those for the undiscounted setting, significantly different techniques are required to handle the presence of discounting on the noise-dependent terms in the analysis. This is reflected in the different constant definitions in the final bounds. We start by establishing weak gradient dominance of $J_\pi^H(K)$.

**Theorem 4.4** (Weak Gradient Dominance)**.** *Let* $\mathcal{L}_{0,\gamma} := \left\{K \in \mathcal{K}_H^\gamma : J_\pi^H(K) \leq J_\pi^H(K_0)\right\}$*. Suppose that Assumptions 2.1, 2.3, 2.5, and 2.6 hold, and that the conditions of Lemma 7.20 hold. Then, for all* $K \in \mathcal{L}_{0,\gamma}$*,*

$$J_\pi^H(K) - J_\pi^{H^*} \leq \alpha_\gamma^H \left\|\nabla_K J_\pi^H(K)\right\| + \beta_\gamma^H, \tag{16}$$

*where* $\alpha_\gamma^H := \alpha_\gamma \left(1 + 2\sup_{K \in \mathcal{L}_{0,\gamma}} \{\|\nabla_K \Delta_\gamma(K)\|\}\right)$, $\alpha_\gamma = \frac{\|\Sigma_{K^*}^\gamma\|}{4\sigma_{\min}(R)\sigma_{min}(\Sigma_K^\gamma)^2}$, *and* $\Sigma_K^\gamma = \mathbb{E}_{x_0}\left[\sum_{t=0}^\infty \gamma^t x_t x_t^\top\right]$ *and* $\beta_\gamma^H := \sup_{K \in \mathcal{L}_{0,\gamma}} \left\{\Delta_\gamma(K) + \alpha_\gamma \left\|\nabla_K \Delta_\gamma(K)\right\|^2\right\}$.

*See proof in Section 7.5.* This result shows that, if we find $K$ such that $\left\|\nabla J_\pi^H(K)\right\|$ is close to zero, then we are guaranteed that the policy $\pi_K^H$ has achieved the $\beta_\gamma^H$-neighborhood of a global optimum of $J_\pi^H(K)$. This powerful property implies that any policy gradient descent procedure that reaches a first-order stationary point of the objective $J_\pi^H(K)$ has in fact reached a neighborhood of a global optimum of $J_\pi^H(K)$. This motivates the convergence and sample complexity results up next.

**Theorem 4.5.** *Suppose Assumptions 2.1, 2.3, 2.5, and 2.6 hold, and let Assumption 3.5 hold for the discounted cost. Run Algorithm 2 for* $T > 0$ *iterations with updates* $\{K_i\}_{i=0}^T \subset \mathcal{K}_H^\gamma$*, batch size* $N > 0$ *and step-size* $\zeta$ *satisfying* $\zeta \leq \min\{\frac{1}{l_\gamma^{mx}(T)}, \frac{(\alpha_\gamma^H)^2}{c_0}, \left(\frac{N(\alpha_\gamma^H)^2}{l_\gamma^{mx}(T)V}\right)^{1/3}\}$*, where* $c_t := \max\{0, J_\pi^H(K_t) - J_\pi^{H^*} - \beta_\gamma^H\}$ *and* $l_\gamma^{mx}(T) := \max\{l_\gamma(K_0), \ldots, l_\gamma(K_T)\}$*. We then have*

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \beta_\gamma^H + \beta_\gamma^* + c_T\left(1 - \sqrt{\frac{\zeta^3 l_\gamma^{mx}(T)V}{4(\alpha_\gamma^H)^2 N}}\right)^T + \sqrt{\frac{l_\gamma^{mx}(T)V\zeta(\alpha_\gamma^H)^2}{N}}. \tag{17}$$

*Furthermore, for* $\epsilon > 0$ *and* $\zeta = \frac{\epsilon^2 N}{4\, l_\gamma^{mx}(T)V(\alpha_\gamma^H)^2}$ *the sample complexity needed to guarantee* $\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \epsilon + \beta_\gamma^H + \beta_\gamma^*$ *is* $NT \leq \frac{16(\alpha_\gamma^H)^4 l_\gamma^{mx}(T)V}{\epsilon^3} \log\frac{c_T}{\epsilon}$.

*See proof in Section 7.5.* This result explicitly characterizes the convergence rate and sample complexity to achieve near-global optimality, reaching an $\epsilon + \beta_\gamma^H + \beta_\gamma^*$-optimal policy with sample complexity $NT \leq \mathcal{O}(\frac{1}{\epsilon^3}\log\frac{1}{\epsilon})$. This also demonstrates that the price of guaranteeing safety is the small, constant residual gap $\beta_\gamma^H$ from Theorem 4.4.

*Remark* 4.6. In contrast to the proofs of Theorem 3.4 and 3.6 for the undiscounted case, our proofs for the discounted setting do not assume the decaying variance in the stochastic wrappers in Assumption 2.4. Instead, we utilize the discount factor to ensure well-definedness of the terms $\alpha_1^\gamma(K)$ and $\alpha_2^\gamma(K)$. This enables us to establish convergence and optimality results in the discounted setting without restricting safe exploration to time-decaying noise.

## 5 Numerical Example

We consider the following normalized battery control model:

$$x_{k+1} = x_k + u_k, \qquad x_k \in [-1, 1], \qquad |u_k| \leq \frac{1}{2}, \tag{18}$$

where $x_k$ is the battery state of charge centered at its midpoint and scaled to $[-1, 1]$, and $u_k$ denotes the control input. We use $x_0 \sim \mathrm{Unif}[-1, 1]$, $Q = 1$, and $R = 4$. For this example, we require the normalized state of charge to satisfy $x_k \in [-1, 1]$ for all $k \geq 0$. To ensure forward invariance of this safe state-of-charge interval $[-1, 1]$, we use the nonlinear control barrier function $h(x) = 1 - x^2$ and impose the condition $h(x_{k+1}) \geq \frac{3}{4}h(x_k)$, yielding the safety constraint $x_{k+1}^2 \leq \frac{1}{4} + \frac{3}{4}x_k^2$. Together with the actuator constraint $|u_k| \leq 1/2$, this induces the following nonlinear state-dependent safe-control set:

$$H(x) = \left[-\frac{1}{2}, \frac{1}{2}\right] \cap \left[-x - \sqrt{\frac{1}{4} + \frac{3}{4}x^2}, -x + \sqrt{\frac{1}{4} + \frac{3}{4}x^2}\right]. \tag{19}$$

This example combines bounded actuators with nonlinear state-dependent barrier constraints. We further show in Remark 7.4 that this example jointly satisfies all the assumptions on the safe-set 2.1,2.3, exploration 2.4, policy-class 2.5, and center-regularity 2.6 on an explicit compact gain region.

We compare our undiscounted safe LQR Algorithm 1 with (a) safe policy gradient using projection (Projection-PG), which projects unconstrained policies onto $H(x)$, and (b) Lagrangian-based policy gradient (Lagrangian-PG), which penalizes actuator and barrier violations. We choose $\sigma = 0.2$, $K_0 = -0.1$, $T = 300$, $N = 64$ and $M = 100$. We evaluate the learned policies from each implementation on $5000$ fresh trajectories with $200$ time steps each, using the original quadratic cost. We provide further details regarding the implementation in Appendix 7.2.

Table 1: Battery control example: mean $\pm$ sample standard deviation over ten independent training runs. Reference cost is evaluated numerically for cost-differenc column. Violation rates are percentages of transitions and an unsafe trajectory contains at least one violating transition.

| Algorithm | Learned gain $K_T$ | Estimated cost | Cost-difference from reference | Unsafe test traj. (%) |
|---|---|---|---|---|
| Safe-PG-LQR | $-0.32 \pm 0.002$ | $1.21 \pm 0.014$ | $0.03 \pm 0.014$ | $0$ |
| Projection-PG | $-0.31 \pm 0.003$ | $1.27 \pm 0.01$ | $0.09 \pm 0.014$ | $0$ |
| Lagrangian-PG | $-0.33 \pm 0.004$ | $1.30 \pm 0.01$ | $0.12 \pm 0.01$ | $10.89 \pm 0.44$ |

From Table 1, we can observe that Safe-PG-LQR achieves the lowest cost among the evaluated methods, with zero safety violations in the test trajectories. Projection-PG also exhibits no test violations, but incurs a higher mean cost of $1.27$, compared with $1.21$ for Safe-PG-LQR. Lagrangian-PG incurs the highest mean cost, $1.30$, and approximately $11\%$ of its test trajectories contain at least one constraint violation. Safe-PG-LQR also achieves the smallest empirical cost difference from the numerical reference, $0.03$, compared with $0.09$ for Projection-PG and $0.12$ for Lagrangian-PG.

## 6 Conclusion

We address hard constraint satisfaction in model-free RL for the infinite horizon LQR problem. By parameterizing the policy class through a truncated stochastic wrapper, we enforce hard constraints by construction throughout learning. We establish that the resulting LQR objective over this policy class satisfies a weak gradient dominance property in both undiscounted and discounted infinite-horizon settings, enabling well-defined model-free policy gradient algorithms with provable convergence, sample complexity guarantees, and near-global optimality. While our analysis focuses on constrained LQR with linear feedback structure, it provides a foundation for nontrivial future extensions to nonlinear policy classes and settings such as LQG control, where preserving hard safety, convergence, and optimality guarantees remains challenging.

## References


[1] Joshua Achiam et al. "Constrained policy optimization". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 22–31.

[2] Andrea Agazzi and Craig Xu Chen. "Policy gradient convergence in discounted LQR". In: 2020. URL: https://api.semanticscholar.org/CorpusID:227106893.

[3] Eitan Altman. *Constrained Markov Decision Processes*. Routledge, 2021.

[4] Aaron D Ames, Jessy W Grizzle, and Paulo Tabuada. "Control barrier function based quadratic programs with application to adaptive cruise control". In: *53rd IEEE conference on decision and control*. IEEE. 2014, pp. 6271–6278.

[5] Aaron D Ames et al. "Control barrier function based quadratic programs for safety critical systems". In: *IEEE Transactions on Automatic Control* 62.8 (2016), pp. 3861–3876.

[6] Aaron D Ames et al. "Control barrier functions: Theory and applications". In: *2019 18th European Control Conference*. 2019, pp. 3420–3431.

[7] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[8] Qinbo Bai et al. "Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 4. 2022, pp. 3682–3689.

[9] Vivek Borkar and Rahul Jain. "Risk-constrained Markov decision processes". In: *IEEE Transactions on Automatic Control* 59.9 (2014), pp. 2574–2579.

[10] Lukas Brunke et al. "Safe learning in robotics: From learning-based control to safe reinforcement learning". In: *Annual Review of Control, Robotics, and Autonomous Systems* 5 (2022), pp. 411–444.

[11] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. Routledge, 2018.

[12] Jingjing Bu and Mehran Mesbahi. "Global convergence of policy gradient algorithms for indefinite least squares stationary optimal control". In: *IEEE Control Systems Letters* 4.3 (2020), pp. 638–643.

[13] Jingjing Bu et al. "LQR through the lens of first order methods: Discrete-time case". In: *arXiv preprint arXiv:1907.08921* (2019).

[14] Jason J Choi et al. "Robust control barrier–value functions for safety-critical control". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 6814–6821.

[15] Hongkai Dai and Frank Permenter. "Convex synthesis and verification of control-Lyapunov and barrier functions with input constraints". In: *2023 American Control Conference (ACC)*. IEEE. 2023, pp. 4116–4123.

[16] Sarah Dean et al. "On the sample complexity of the linear quadratic regulator". In: *Foundations of Computational Mathematics* 20.4 (2020), pp. 633–679.

[17] Sarah Dean et al. "Safely learning to control the constrained linear quadratic regulator". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 5582–5588.

[18] Dongsheng Ding et al. "Natural policy gradient primal-dual method for constrained markov decision processes". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8378–8390.

[19] Shutong Ding et al. "Reduced policy optimization for continuous control with hard constraints". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 38642–38667.

[20] Maryam Fazel et al. "Global convergence of policy gradient methods for the linear quadratic regulator". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1467–1476.

[21] Federica Ferraguti et al. "Safety and efficiency in robotics: The control barrier functions approach". In: *IEEE Robotics & Automation Magazine* 29.3 (2022), pp. 139–151.

[22] Antoine Girard. "Reachability of uncertain linear systems using zonotopes". In: *International Workshop on Hybrid Systems: Computation and Control*. Springer. 2005, pp. 291–305.

[23] Bin Hu et al. "Toward a theoretical foundation of policy optimization for learning control policies". In: *Annual Review of Control, Robotics, and Autonomous Systems* 6.1 (2023), pp. 123–158.

[24] Rudolf E Kalman. "On the general theory of control systems". In: *Proceedings first international conference on automatic control, Moscow, USSR*. 1960, pp. 481–492.

[25] V. Konda. "Actor-Critic Algorithms". PhD thesis. Massachusetts Institute of Technology, 2002.

[26] Yingying Li et al. "Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees". In: *arXiv preprint arXiv:2111.00411* (2021).

[27] Dhruv Malik et al. "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems". In: *Journal of Machine Learning Research* 21.21 (2020), pp. 1–51.

[28] Alessandro Montenegro et al. "Learning Optimal Deterministic Policies with Stochastic Policy Gradients". In: *Forty-first International Conference on Machine Learning*. 2024.

[29] Santiago Paternain et al. "Constrained reinforcement learning has zero duality gap". In: *Advances in Neural Information Processing Systems* 32 (2019).

[30] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. "Optlayer-practical constrained optimization for deep reinforcement learning in the real world". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 6236–6243.

[31] Chang-han Rhee and Peter W Glynn. "Unbiased estimation with square root convergence for SDE models". In: *Operations Research* 63.5 (2015), pp. 1026–1043.

[32] Adrien Saumard and Jon A Wellner. "Log-concavity and strong log-concavity: a review". In: *Statistics surveys* 8 (2014), p. 45.

[33] Pierre OM Scokaert and James B Rawlings. "Constrained linear quadratic regulation". In: *IEEE Transactions on automatic control* 43.8 (2002), pp. 1163–1169.

[34] Vipul K Sharma and S Sivaranjani. "Safe Control Design Through Risk-Tunable Control Barrier Functions". In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. 2023, pp. 4393–4398.

[35] Vipul K Sharma, Wesley A Suttle, and S Sivaranjani. "Globally Optimal Policy Gradient Algorithms for Reinforcement Learning with PID Control Policies". In: *Advances in Neural Information Processing Systems* 39 (2025).

[36] Wesley A Suttle et al. "Sampling-based safe reinforcement learning for nonlinear dynamical systems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 4420–4428.

[37] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation". In: *Advances in neural information processing systems* 12 (1999).

[38] Anastasios Tsiamis et al. "Risk-constrained linear-quadratic regulators". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 3040–3047.

[39] Wei Xiao and Calin Belta. "High-order control barrier functions". In: *IEEE Transactions on Automatic Control* 67.7 (2021), pp. 3655–3662.

[40] Bin Xu and Koushil Sreenath. "Safe teleoperation of dynamic uavs through control barrier functions". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 7848–7855.

[41] Kaiqing Zhang et al. "Convergence and iteration complexity of policy gradient method for infinite-horizon reinforcement learning". In: *2019 58th Conference on Decision and Control*. 2019, pp. 7415–7422.

[42] Kaiqing Zhang et al. "Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity". In: *Advances in neural information processing systems* 34 (2021), pp. 2949–2964.

[43] Feiran Zhao and Keyou You. "Policy Gradient Methods for the Cost-Constrained LQR: Strong Duality and Global Convergence". In: *IEEE Transactions on Automatic Control* (2025).

[44] Feiran Zhao, Keyou You, and Tamer Başar. "Global convergence of policy gradient primal–dual methods for risk-constrained LQRs". In: *IEEE Transactions on Automatic Control* 68.5 (2023), pp. 2934–2949.

[45] Feiran Zhao, Keyou You, and Tamer Başar. "Infinite-horizon risk-constrained linear quadratic regulator with average cost". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 390–395.

# 7 Appendix

## Appendix index


7 Appendix . . . 13
7.1 Algorithms . . . 15
Lemma 7.1: Score function of truncated safe policy . . . 15
7.1.1 Undiscounted Model-Free Safe LQR Algorithm . . . 15
7.1.2 Discounted Model-Free Safe LQR Algorithm . . . 15
7.1.3 General Remarks . . . 16
Remark 7.2: Gradient Estimator . . . 16
Remark 7.3: Computing the Gaussian center . . . 16
7.2 Battery Example: Feasibility and Experimental Setup . . . 18
Remark 7.4: Joint Feasibility . . . 18
7.3 Common Calculations and Helper Results . . . 20
Proof of Proposition 2.7 . . . 20
Corollary 7.5 . . . 20
Proof of Corollary 7.5 . . . 20
Proof for Lemma 7.1 . . . 21
Lemma 7.6: Second Moment Bound for Truncated Noise . . . 22
Proof of Lemma 7.6 . . . 22
Lemma 7.7: Weighted Long-horizon Noise . . . 23
Proof of Lemma 7.7 . . . 23
Lemma 7.8: Smoothness of the Lyapunov Matrices . . . 24
Proof of Lemma 7.8 . . . 24
7.4 Undiscounted Safe LQR . . . 26
Lemma 7.9: Cost Difference between Stochastic and Deterministic Safe LQR . . . 26
Proof of Lemma 7.9 . . . 26
Lemma 7.10: Helper Lemma . . . 28
Proof of Lemma 7.10 . . . 28
Lemma 7.11: Helper Lemma . . . 31
Proof of Lemma 7.11 . . . 31
Proof of Theorem 3.1 . . . 32
Lemma 7.12 . . . 32
Proof of Lemma 7.12 . . . 32
Lemma 7.13: Regularity of $\alpha_1$ . . . 32
Proof of Lemma 7.13 . . . 32
Lemma 7.14 . . . 33
Proof of Lemma 7.14 . . . 33
Lemma 7.15: Regularity of $\Delta$ . . . 33
Proof of Lemma 7.15 . . . 34
Proof on Theorem 3.2 . . . 34
Lemma 7.16: Lemma 5, [27] . . . 34
Lemma 7.17 . . . 34
Proof of Lemma 7.17 . . . 34
Proof of Theorem 3.3 . . . 35
Lemma 7.18 . . . 35
Proof of Lemma 7.18 . . . 35
Lemma 7.19: Boundedness of sublevel constants . . . 36
Proof of Lemma 7.19 . . . 36
Lemma 7.20: Sublevel gradient bound . . . 36
Proof of Lemma 7.20 . . . 36
Proof of Theorem 3.4 . . . 37
Proof of Theorem 3.6 . . . 37
7.5 Discounted Safe LQR . . . 39
Lemma 7.21: Policy Gradient . . . 39
Proof of Lemma 7.21 . . . 39
Lemma 7.22: Helper Lemma . . . 39
Proof of Lemma 7.22 . . . 39
Lemma 7.23: Gradient Dominance . . . 41
Proof of Lemma 7.23 . . . 41
Lemma 7.24: Cost Difference between Stochastic and Deterministic Safe LQR . . . 42
Proof of Lemma 7.24 . . . 42
Proof of Theorem 4.1 . . . 47
Lemma 7.25 . . . 47

Proof of Lemma 7.25 . . . 47
Lemma 7.26 . . . 48
Proof of Lemma 7.26 . . . 48
Lemma 7.27: Regularity of $\alpha_1^\gamma$ . . . 48
Proof of Lemma 7.27 . . . 49
Lemma 7.28: Regularity of $\Delta_\gamma$ . . . 49
Proof of Lemma 7.28 . . . 49
Proof on Theorem 4.2 . . . 49
Lemma 7.29: Lemma 5, [27] . . . 49
Lemma 7.30 . . . 50
Proof of Lemma 7.30 . . . 50
Proof of Theorem 4.3 . . . 50
Lemma 7.31: Helper Result . . . 50
Proof of Lemma 7.31 . . . 51
Proof of Theorem 4.4 . . . 51
Proof of Theorem 4.5 . . . 52
7.6 Broader Impacts . . . 53

In this section, we first describe our proposed model-free safe LQR algorithm by adopting standard policy gradient theorems in the undiscounted and discounted frameworks. Then, in subsequent subsections, we provide proofs and helper results for each of our theoretical claims. We begin with claims and helper results that are common to both the undiscounted and discounted frameworks. Afterwards, we establish helper results and proofs for gradient dominance, optimality, and convergence in the undiscounted framework, which require us to bound the undiscounted infinite-horizon accumulation of truncated noise. Finally, we prove analogous theoretical guarantees for the discounted framework, where we do not assume the variance decay condition on the safe truncated exploration noise. For simplicity, unless otherwise stated, we use $\mathbb{E}$ to denote the expectation taken with respect to the appropriate random variables.

### 7.1 Algorithms

We provide model-free algorithms to obtain safe LQR policies in both the undiscounted and discounted settings. Specifically, to enable application of standard policy gradient theorems [37, 25], we derive the score function of the truncated policy with respect to the policy parameter $K$ in the following result.

**Lemma 7.1** (Score function of truncated safe policy)**.** *Suppose Assumptions 2.3, 2.5, and 2.6 hold. Then, it follows that $\pi_K^H(u|x) > 0$ and $\pi_K^H(u|x)$ is differentiable in $K$, for all $x \in \mathcal{X}$ and $u \in H(x)$. Let $D_K c_K(x)[\Delta K]$ denote the directional derivative of $c_K(x)$ at $K$ in the direction $\Delta K$, and let $D_K c_K(x)^*$ denote its adjoint with respect to the Frobenius inner product. Then the corresponding score function is given by*

$$\nabla_K \log \pi_K^H(u|x) = D_K c_K(x)^* \left[\Sigma^{-1}(u - m_H(x))\right], \tag{20}$$

*where*

$$m_H(x) = \frac{\int_{H(x)} u \, \pi_K(u|x) du}{\pi_K(H(x)|x)} = \mu_K^H(x). \tag{21}$$

*For states satisfying the linear-mean condition in Assumption 2.5, $m_H(x) = \mu_K^H(x) = Kx$, and hence*

$$\nabla_K \log \pi_K^H(u|x) = D_K c_K(x)^* \left[\Sigma^{-1}(u - Kx)\right]. \tag{22}$$

*See proof in Section 7.3.* In contrast to unstructured truncated policy parameterizations [36], the score function in Lemma 7.1 admits a structured form in which the effect of truncation is captured through the truncated-policy mean, while the dependence on the policy parameter enters through the deterministic nominal map $c_K$.

For the rest of this subsection, we define the per-stage cost as $\ell(x, u) = x^\top Q x + u^\top R u$, use $t$ to denote policy updates, $i$ for independent episodes, and $k$ for steps within an episode. Furthermore, we use $N$ for the number of episodes per update and $M$ for the rollout length. We start each episode with an independently sampled $x_0^{(i)} \sim \mathcal{D}$, where $\mathcal{D}$ is the initial-state distribution.

#### 7.1.1 Undiscounted Model-Free Safe LQR Algorithm

We now give a model-free safe policy-gradient algorithm for $\gamma = 1$ that does not require the system matrices $A$ and $B$. We use $\Sigma_k = \sigma^2 (k+1)^{-2} I_m$, at $k^{th}$ step within an episode, and choose the center $c_{K,k}(x)$ so that the Gaussian truncated to $H(x)$ has post-truncation mean as $Kx$. First, from Lemma 7.1, the score function can be written as

$$S_{K,k}(x, u) := \nabla_K \log \pi_{K,k}^H(u \mid x) = D_K c_{K,k}(x)^* \left[\Sigma_k^{-1}(u - Kx)\right]. \tag{23}$$

Then, we provide a model-free gradient estimator construction as

$$\widehat{g}_{1,M,N,t} := \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{M-1} S_{K_t,k}(x_k^{(i)}, u_k^{(i)}) \widehat{Q}_k^{(i)}, \quad \text{where} \quad \widehat{Q}_k^{(i)} := \sum_{r=k}^{M-1} \ell(x_r^{(i)}, u_r^{(i)}). \tag{24}$$

We use this estimator in (24) and finally present the undiscounted safe LQR Algorithm 1.

#### 7.1.2 Discounted Model-Free Safe LQR Algorithm

We now build the safe LQR policy gradient algorithm for the discounted case, i.e. $0 < \gamma < 1$. We use the truncated policy with $\Sigma = \sigma^2 I_m$ and moment-matched center $c_K(x)$ from the previous section.

**Algorithm 1 Safe-PG-LQR:** Undiscounted finite-rollout update

**Input:** $K_0$, update budget $T$, rollout length $M$, batch size $N$, stepsize $\zeta$, exploration scale $\sigma$
**Output:** Final gain $K_T$
**for** $t = 0, \dots, T-1$ **do**
  **for** $i = 1, \dots, N$ **do**
    Independently draw $x_0^{(i)} \sim \mathcal{D}$
    **for** $k = 0, \dots, M-1$ **do**
      Set $\Sigma_k \leftarrow \sigma^2(k+1)^{-2} I_m$
      Compute the moment-matched center $c_{K_t,k}(x_k^{(i)})$
      Truncate $\mathcal{N}(c_{K_t,k}(x_k^{(i)}), \Sigma_k)$ to $H(x_k^{(i)})$ to form $\pi^H_{K_t,k}(\cdot \mid x_k^{(i)})$
      Draw $u_k^{(i)} \sim \pi^H_{K_t,k}(\cdot \mid x_k^{(i)})$
      Record $\ell_k^{(i)} \leftarrow \ell(x_k^{(i)}, u_k^{(i)})$
      Compute $S_k^{(i)} \leftarrow S_{K_t,k}(x_k^{(i)}, u_k^{(i)})$ using (23)
      Apply $u_k^{(i)}$ and observe $x_{k+1}^{(i)}$
    **end for**
    Set $\widehat{Q}_M^{(i)} \leftarrow 0$
    **for** $k = M-1, \dots, 0$ **do**
      $\widehat{Q}_k^{(i)} \leftarrow \ell_k^{(i)} + \widehat{Q}_{k+1}^{(i)}$
    **end for**
    $\widehat{G}_{1,M}^{(i)} \leftarrow \sum_{k=0}^{M-1} S_k^{(i)} \widehat{Q}_k^{(i)}$
  **end for**
  $\widehat{g}_{1,M,N,t} \leftarrow N^{-1} \sum_{i=1}^{N} \widehat{G}_{1,M}^{(i)}$
  $K_{t+1} \leftarrow K_t - \zeta \widehat{g}_{1,M,N,t}$
**end for**

Then with $S_K(x,u) = D_K c_K(x)^*[\Sigma^{-1}(u - Kx)]$, a model-free policy gradient estimator can be written as, similar to (24), as

$$\widehat{g}_{\gamma,M,N,t} := \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{M-1} \gamma^k S_{K_t}(x_k^{(i)}, u_k^{(i)}) \widehat{Q}_{\gamma,k}^{(i)}, \quad \text{where} \quad \widehat{Q}_{\gamma,k}^{(i)} := \sum_{r=k}^{M-1} \gamma^{r-k} \ell(x_r^{(i)}, u_r^{(i)}). \tag{25}$$

We use the gradient estimator in (25) and finally present the discounted safe LQR algorithm as 2.

### 7.1.3 General Remarks

We now briefly comment on the model-free policy gradient estimator constructions as follows.

*Remark* 7.2 (Gradient Estimator). We note that finite-horizon estimators, such as (24), represent an instance of model-free estimator design, which may be biased in general. At the same time, we can also construct unbiased model-free estimators using a generalized version of the randomized-horizon framework of [31], for $\gamma = 1$, or [41], for $0 < \gamma < 1$, under suitable conditions. However, the construction of such unbiased model-free gradient estimators is beyond the scope of this paper, where we instead focus on developing principled model-free policy gradient methods for LQR problems under hard time-varying state-action coupled constraints.

*Remark* 7.3 (Computing the Gaussian center). The center $c_K(x)$ in Algorithms 1 and 2 is chosen so that the truncated policy has conditional mean $Kx$. For fixed $K$, $x$, covariance $\Sigma$, and safe set $H(x)$, this amounts to solving the moment-matching equation

$$\frac{\int_{H(x)} u \exp\!\big(-\frac{1}{2}(u-c)^\top \Sigma^{-1}(u-c)\big)\, du}{\int_{H(x)} \exp\!\big(-\frac{1}{2}(u-c)^\top \Sigma^{-1}(u-c)\big)\, du} = Kx \tag{26}$$

for $c = c_K(x)$. Equivalently, with $\theta = \Sigma^{-1} c$, define

$$A_x(\theta) := \log \int_{H(x)} \exp\!\left(\theta^\top u - \frac{1}{2} u^\top \Sigma^{-1} u\right) du. \tag{27}$$

**Algorithm 2 Safe-PG-LQR-$\gamma$:** Discounted finite-rollout update

**Input:** $K_0$, update budget $T$, rollout length $M$, batch size $N$, stepsize $\zeta$, exploration scale $\sigma$, discount factor $\gamma \in (0,1)$
**Output:** Final gain $K_T$
**for** $t = 0, \dots, T-1$ **do**
  **for** $i = 1, \dots, N$ **do**
    Independently draw $x_0^{(i)} \sim \mathcal{D}$
    **for** $k = 0, \dots, M-1$ **do**
      Compute the moment-matched center $c_{K_t}(x_k^{(i)})$
      Truncate $\mathcal{N}(c_{K_t}(x_k^{(i)}), \sigma^2 I_m)$ to $H(x_k^{(i)})$ to form $\pi_{K_t}^H(\cdot \mid x_k^{(i)})$
      Draw $u_k^{(i)} \sim \pi_{K_t}^H(\cdot \mid x_k^{(i)})$
      Record $\ell_k^{(i)} \leftarrow \ell(x_k^{(i)}, u_k^{(i)})$
      Compute $S_k^{(i)} \leftarrow S_{K_t}(x_k^{(i)}, u_k^{(i)})$ using Lemma 7.1
      Apply $u_k^{(i)}$ and observe $x_{k+1}^{(i)}$
    **end for**
    Set $\widehat{Q}_{\gamma,M}^{(i)} \leftarrow 0$
    **for** $k = M-1, \dots, 0$ **do**
      $\widehat{Q}_{\gamma,k}^{(i)} \leftarrow \ell_k^{(i)} + \gamma \widehat{Q}_{\gamma,k+1}^{(i)}$
    **end for**
    $\widehat{G}_{\gamma,M}^{(i)} \leftarrow \sum_{k=0}^{M-1} \gamma^k S_k^{(i)} \widehat{Q}_{\gamma,k}^{(i)}$
  **end for**
  $\widehat{g}_{\gamma,M,N,t} \leftarrow N^{-1} \sum_{i=1}^{N} \widehat{G}_{\gamma,M}^{(i)}$
  $K_{t+1} \leftarrow K_t - \zeta \widehat{g}_{\gamma,M,N,t}$
**end for**

Then $\nabla_\theta A_x(\theta)$ is the mean of the truncated Gaussian, so the moment-matching condition can be solved from $\nabla_\theta A_x(\theta) = Kx$, or equivalently by minimizing the convex objective $A_x(\theta) - \theta^\top Kx$.

Thus, when $H(x)$ is full-dimensional and $Kx$ lies in its relative interior, $c_K(x)$ can be obtained using standard smooth root-finding or convex optimization methods such as damped Newton, trust-region, or quasi-Newton methods, with the required moments evaluated analytically when available or numerically otherwise. For simple bounded safe sets, the computation further simplifies. For box constraints with diagonal covariance, the moment-matching equations decouple across coordinates. For intersections of affine-in-control CBF constraints with bounded actuator constraints, $H(x)$ is a bounded polyhedral set, and the same convex moment-matching formulation applies. Sampling from $\pi_K^H(\cdot|x)$ can be performed using standard truncated-Gaussian samplers; for box constraints with diagonal covariance, this reduces to coordinatewise truncated normal sampling, whereas for bounded polyhedral $H(x)$, hit-and-run or other log-concave sampling methods may be used.

We illustrate the above moment-matching construction in the scalar case. For a given state $x$ and an episode step $k$, we consider $U \sim \mathcal{N}(c, \sigma_k^2)$ for an arbitrary center $c$, with $H(x) = [a(x), b(x)]$ and $a(x) < b(x)$. We further define $\alpha = (a(x) - c)/\sigma_k$, $\beta = (b(x) - c)/\sigma_k$, and $Z = \Phi(\beta) - \Phi(\alpha)$, where $\phi$ and $\Phi$ are the standard normal density and CDF. Then, the conditional mean and variance are given as

$$\begin{aligned} m_H(x;c) &:= \mathbb{E}[U \mid U \in H(x)] = c + \sigma_k \frac{\phi(\alpha) - \phi(\beta)}{Z}, \\ v_H(x;c) &:= \mathrm{Var}(U \mid U \in H(x)) = \sigma_k^2 \left[ 1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{Z} - \left( \frac{\phi(\alpha) - \phi(\beta)}{Z} \right)^2 \right]. \end{aligned} \tag{28}$$

For the log-partition function $A_x$ above with $\Sigma = \sigma_k^2$, $m_H(x;c) = A_x'(c/\sigma_k^2)$. Thus the matched center $c_{K,k}(x)$ solves $m_H(x; c_{K,k}(x)) = Kx$, giving conditional second moment $(Kx)^2 + v_H(x; c_{K,k}(x))$. We write the matched center as $c_K(x)$.

### 7.2 Battery Example: Feasibility and Experimental Setup

We first verify the structural conditions used in our analysis for the battery model in Section 5, and then describe its numerical implementation.

*Remark* 7.4 (Joint Feasibility). For the battery example in Section 5, we choose the safe and stabilizing policy class as $\mathcal{K}_H^\gamma = [-0.48, -0.02]$ for $\gamma \in (0, 1]$. We note that, for every gain $K$ in this class $\mathcal{K}_H^\gamma$, the scalar closed loop dynamics $1 + K \in [0.52, 0.98]$ is Schur stable. Furthermore, the barrier constraint $(x + u)^2 \leq \frac{1}{4} + \frac{3}{4}x^2$ makes the desired safe set for state-of-charge, given by $[-1, 1]$, forward invariant under the battery dynamics.

For every $x \in [0, 1]$, all actions in $[-1/2, 0]$ satisfy both the actuator and the barrier conditions, and similarly, for every $x \in [-1, 0]$, all actions in $[0, 1/2]$ satisfy both constraints. Since $H(x)$ is an intersection of two intervals from (19), it is convex. The above inclusions and the actuator bound imply $1/2 \leq \mu(H(x)) \leq 1$, so $H(x)$ is also full-dimensional. Furthermore, from $\mathcal{K}_H^\gamma$ and the barrier condition, we have

$$|Kx| \leq 0.48, \qquad \tfrac{1}{4} + \tfrac{3}{4}x^2 - (1 + K)^2 x^2 \geq 0.0396.$$

Thus, for every $x \in [-1, 1]$ and $K \in \mathcal{K}_H^\gamma$, the desired policy mean $Kx$ lies strictly inside the safe-action interval $H(x)$, with a distance of at least $0.0198 =: \rho$ from each endpoint.

We next establish existence and uniqueness of the Gaussian center $c_{K,k}(x)$. For fixed $x$ and $k$, let $U \sim \mathcal{N}(c, \sigma_k^2)$ and $m_H(x; c) := \mathbb{E}[U \mid U \in H(x)]$. From (19), we can write $H(x)$ as $H(x) = [a(x), b(x)]$. Then we have

$$\frac{\partial m_H(x; c)}{\partial c} = \frac{\mathrm{Var}(U \mid U \in H(x))}{\sigma_k^2} > 0,$$
$$\lim_{c \to -\infty} m_H(x; c) = a(x), \qquad \lim_{c \to +\infty} m_H(x; c) = b(x).$$

Since $a(x) < Kx < b(x)$, continuity and strict monotonicity imply that there exists a unique center $c_{K,k}(x)$ satisfying $m_H(x; c_{K,k}(x)) = Kx$. For fixed positive variance, the implicit-function theorem gives smooth dependence on $K$, with uniform bounds for the first two center and moment derivatives over $\mathcal{K}_H^\gamma \times [-1, 1]$.

As $\sigma_k \to 0$, Gaussian tail bounds and monotonicity of $m_H(x; c)$ place the matched centers inside the $\rho/2$-neighborhood of $Kx$, uniformly over $\mathcal{K}_H^\gamma \times [-1, 1]$, for all sufficiently large $k$. Since $Kx$ is at least $\rho$ distance away from each endpoint of $H(x) = [a(x), b(x)]$, the matched center satisfies

$$\begin{aligned}\min\{c_{K,k}(x) - a(x),\, b(x) - c_{K,k}(x)\} &\geq \rho - |c_{K,k}(x) - Kx| \\ &\geq \rho/2\end{aligned}$$

for all sufficiently large $k$, uniformly over $\mathcal{K}_H^\gamma \times [-1, 1]$. With $S_{K,k}^H(x) := \mathbb{E}_{\pi_{K,k}^H}[u^2 \mid x]$, we now use the scalar moment identities in (28) from Remark 7.3, evaluated at $c = c_{K,k}(x)$. Since $m_H(x; c_{K,k}(x)) = Kx$ and $\partial_c m_H(x; c) = v_H(x; c)/\sigma_k^2$, Gaussian tail bounds and implicit differentiation give constants $C > 0$, $p \geq 0$, and an integer $k_0 \geq 1$ such that the following estimate holds for every $k \geq k_0$, uniformly over $K \in \mathcal{K}_H^\gamma$ and $x \in [-1, 1]$:

$$\begin{aligned}\sup_{K \in \mathcal{K}_H^\gamma, x \in [-1,1]} \sum_{j=0}^{2} \Big( \big|\partial_K^j (c_{K,k}(x) - Kx)\big| + \big|\partial_K^j (S_{K,k}^H(x) - (Kx)^2 - \sigma_k^2)\big| \Big) \\ \leq C \sigma_k^{-p} \exp\left(-\frac{\rho^2}{8\sigma_k^2}\right) \longrightarrow 0.\end{aligned}$$

Since $\sigma_k^2 \to 0$, the center and its first two $K$-derivatives converge uniformly to $Kx, x, 0$, while the second moment and its first two $K$-derivatives converge uniformly to $(Kx)^2, 2Kx^2, 2x^2$.

For each $k \geq 0$, define

$$B_k := \sup_{\substack{K \in \mathcal{K}_H^\gamma \\ x \in [-1,1]}} \sum_{j=1}^{2} \left( |\partial_K^j c_{K,k}(x)| + |\partial_K^j S_{K,k}^H(x)| \right).$$

For each fixed $0 \le k < k_0$, continuity of these derivatives on the compact set $\mathcal{K}_H^\gamma \times [-1, 1]$ gives $B_k < \infty$.

The preceding uniform convergence gives $B_{\text{tail}} := \sup_{k \ge k_0} B_k < \infty$. Hence

$$M_{reg} := \max\{1, B_0, \dots, B_{k_0-1}, B_{\text{tail}}\} < \infty.$$

Thus the first two $K$-derivatives of the center and second moment are bounded by $M_{reg}$ for every $k$, $K$, and $x$ in the stated domain. The matched mean is $Kx$, with derivatives $x$ and $0$, which are also bounded by $M_{reg}$. This verifies Assumption 2.6.

For $0 < \gamma < 1$, the variance is constant and positive, so the fixed-variance compactness argument applies directly. Together with $\mathbb{E}[x_0^2] = 1/3$ and $Q = 1, R = 4$, this example therefore satisfies the safe-set conditions in Assumptions 2.1 and 2.3, the linear safe-policy condition in Assumption 2.5, and the center/moment regularity condition in Assumption 2.6. For $\gamma = 1$, the schedule $\sigma_k^2 = \sigma^2/(k+1)^2$ also satisfies Assumption 2.4.

**Safe-policy implementation.** The experiment uses $\gamma = 1$ and the parameters in Section 5. At time $k$ in each episode, given the state $x_k$ and current gain $K_t$, we compute $c_{K_t,k}(x_k)$ by solving the moment-matching equation in Remark 7.3 using Newton iterations. We use mean-residual tolerance $2 \times 10^{-11}$, at most 60 iterations, and a final residual check of $10^{-8}$. Actions are sampled using the inverse truncated-normal CDF. The policy score is evaluated using Lemma 7.1, with the required conditional moments calculated from the truncated-normal formulas.

**Gradient estimation and updates.** We implement the finite-rollout update of Algorithm 1 with a baseline. At each stage, the baseline for an episode is the average sampled cost of $N$ episodes. We subtract this baseline before multiplying by the policy score and then average the resulting episode gradients. With exact score evaluation, this preserves the expected finite-horizon gradient because the baseline uses other independent episodes and the score has zero conditional mean.

**Baselines.** Projection-PG draws a Gaussian action with mean $Kx_k$ and variance $\sigma_k^2$, projects the draw onto $H(x_k)$, and uses the score of the original Gaussian draw with the resulting quadratic costs. Lagrangian-PG executes the unprojected Gaussian sample and uses the training stage cost

$$\ell(x_k, u_k) + \lambda_t \left[ \left(|u_k| - \tfrac{1}{2}\right)_+^2 + \left(x_{k+1}^2 - \tfrac{1}{4} - \tfrac{3}{4}x_k^2\right)_+^2 \right], \qquad (z)_+ := \max\{z, 0\}.$$

The coefficient $\lambda_t$ is held fixed within update $t$. It starts at $\lambda_0 = 10$ and increases after each update by 10 times the average of the bracketed penalty over the $NM$ sampled transitions.

**Evaluation and safety metrics.** Each final gain $K_T$ is evaluated using its method's stochastic policy described above on 5000 fresh trajectories of length 200. For each episode, we restart the exploration schedule $\sigma_k^2 = \sigma^2/(k+1)^2$ at $k = 0$. The reported cost is the sample mean of

$$\sum_{k=0}^{199} \ell(x_k, u_k).$$

This uses the original quadratic cost for all methods, excluding the Lagrangian training penalty. We report means and sample standard deviations across the ten independent training runs. A safety violation is recorded at time $k$ when

$$|u_k| - \tfrac{1}{2} > 10^{-10} \quad \text{or} \quad x_{k+1}^2 - \tfrac{1}{4} - \tfrac{3}{4}x_k^2 > 10^{-10}.$$

**Numerical reference.** We numerically search over gains in $\mathcal{K}_H^\gamma = [-0.48, -0.02]$ and evaluate the corresponding truncated stochastic policies to obtain a reference cost of approximately 1.18. Table 1 subtracts this same numerical reference from each method's evaluated cost. Safe-PG-LQR achieves the smallest cost difference among the compared methods, consistent with its lowest evaluated cost.

### 7.3 Common Calculations and Helper Results

In what follows, we use the subscripts $i$ and $j$, instead of $t$ and $t'$, for the truncated noises. We begin by proving the zero-mean and cross-time orthogonality properties of $\eta_i^H$.

*Proof of Proposition 2.7.* Fix $K \in \mathcal{K}_H^\gamma$. By Assumption 2.5,

$$\mu_K^H(x) = Kx, \qquad \forall x \in \mathcal{X}.$$

Thus, for $u_i \sim \pi_K^H(\cdot|x_i)$,

$$u_i = Kx_i + \eta_i^H, \qquad \eta_i^H := u_i - \mu_K^H(x_i).$$

The closed-loop dynamics are

$$x_{i+1} = Ax_i + Bu_i = (A + BK)x_i + B\eta_i^H = A_K x_i + B\eta_i^H.$$

Iterating gives

$$x_i = A_K^i x_0 + \mathcal{C}_i \bar{\eta}_i^H,$$

so $x_i$ is $\mathcal{F}_i^H$-measurable. Since the policy samples $u_i$ from $\pi_K^H(\cdot|x_i)$ using only the current state $x_i$,

$$\mathbb{E}[u_i \mid x_i, \mathcal{F}_i^H] = \mathbb{E}_{u\sim\pi_K^H(\cdot|x_i)}[u] = \mu_K^H(x_i) = Kx_i.$$

Therefore,

$$\mathbb{E}[\eta_i^H \mid x_i, \mathcal{F}_i^H] = \mathbb{E}[u_i - Kx_i \mid x_i, \mathcal{F}_i^H] = 0.$$

Using the tower property,

$$\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H] = \mathbb{E}\big[\mathbb{E}[\eta_i^H \mid x_i, \mathcal{F}_i^H] \mid \mathcal{F}_i^H\big] = 0.$$

For $j < i$, $\eta_j^H$ is $\mathcal{F}_i^H$-measurable, hence

$$\mathbb{E}[\eta_i^H(\eta_j^H)^\top] = \mathbb{E}\big[\mathbb{E}[\eta_i^H(\eta_j^H)^\top \mid \mathcal{F}_i^H]\big] = \mathbb{E}\big[\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H](\eta_j^H)^\top\big] = 0.$$

The case $j > i$ follows by symmetry after transposition. □

We now prove a helper Lemma as a corollary of Proposition 2.7 which is helpful for the proofs in later subsections.

**Corollary 7.5.** *Suppose Assumptions 2.3 and 2.5 hold. Then, the truncated noises $\eta_i^H$ and $\bar{\eta}_i^H$ from* (6) *and* (9) *follow*

$$\mathbb{E}\left[\eta_i^H | \bar{\eta}_i^H\right] = 0, \forall i \geq 1, \tag{29}$$

$$\mathbb{E}\left[\bar{\eta}_i^H | x_0\right] = 0, \forall i \geq 0, x_0 \sim \mathcal{D}. \tag{30}$$

*Proof of Corollary 7.5.* From (9), we know that

$$\bar{\eta}_i^H = \begin{bmatrix} \eta_{i-1}^H \\ \eta_{i-2}^H \\ \vdots \\ \eta_0^H \end{bmatrix}.$$

Thus, with (7), $\bar{\eta}_i^H$ is $\mathcal{F}_i^H$-measurable. Hence, by the tower property and Proposition 2.7,

$$\mathbb{E}[\eta_i^H \mid \bar{\eta}_i^H] = \mathbb{E}\big[\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H] \mid \bar{\eta}_i^H\big] = 0. \tag{31}$$

For the second claim, fix any $j < i$. Since $\sigma(x_0) \subseteq \mathcal{F}_j^H$, the tower property gives

$$\mathbb{E}[\eta_j^H \mid x_0] = \mathbb{E}\big[\mathbb{E}[\eta_j^H \mid \mathcal{F}_j^H] \mid x_0\big] = 0. \tag{32}$$

Therefore,

$$\mathbb{E}[\bar{\eta}_i^H \mid x_0] = \begin{bmatrix} \mathbb{E}[\eta_{i-1}^H \mid x_0] \\ \mathbb{E}[\eta_{i-2}^H \mid x_0] \\ \vdots \\ \mathbb{E}[\eta_0^H \mid x_0] \end{bmatrix} = 0. \tag{33}$$

□

We next prove Lemma 7.1, which gives the score function of the truncated safe policy $\pi_K^H$.

*Proof for Lemma 7.1.* Recall from $\pi_K^H$'s definition in (4) that, for $u \in H(x)$,

$$\begin{aligned}\log \pi_K^H(u|x) &= \log \frac{\pi_K(u|x)}{\pi_K(H(x)|x)}, \\ &= \log \pi_K(u|x) - \log \pi_K(H(x)|x).\end{aligned}$$

Taking the operator $\nabla_K$ on both sides, we get

$$\begin{aligned}&\nabla_K \log \pi_K^H(u|x) \\ &= \nabla_K \log \pi_K(u|x) - \nabla_K \log \pi_K(H(x)|x), \\ &= \frac{\nabla_K \pi_K(u|x)}{\pi_K(u|x)} - \frac{\nabla_K \pi_K(H(x)|x)}{\pi_K(H(x)|x)}.\end{aligned} \tag{34}$$

We recall the policy $\pi_K$ from (3), given as

$$\pi_K(u|x) = (2\pi)^{-\frac{m}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}[u-c_K(x)]^\top \Sigma^{-1}[u-c_K(x)]}. \tag{35}$$

From (35), the gradient of $\pi_K$ with respect to $K$ is given by

$$\nabla_K \pi_K(u|x) = \pi_K(u|x) D_K c_K(x)^* \left[\Sigma^{-1}(u - c_K(x))\right].$$

Thus, we know the first term $\frac{\nabla_K \pi_K(u|x)}{\pi_K(u|x)}$ can be written as

$$\frac{\nabla_K \pi_K(u|x)}{\pi_K(u|x)} = D_K c_K(x)^* \left[\Sigma^{-1}(u - c_K(x))\right]. \tag{36}$$

For the second term $\frac{\nabla_K \pi_K(H(x)|x)}{\pi_K(H(x)|x)}$, we begin with recalling the definition of $\pi_K(H(x)|x)$ as

$$\pi_K(H(x)|x) = \int_{H(x)} \pi_K(u|x) du.$$

Applying the $\nabla_K$ operator on both sides and using Assumption 2.6 to interchange differentiation and integration, we get

$$\begin{aligned}\nabla_K \pi_K(H(x)|x) &= \int_{H(x)} \nabla_K \pi_K(u|x) du \\ &= \int_{H(x)} \pi_K(u|x) D_K c_K(x)^* \left[\Sigma^{-1}(u - c_K(x))\right] du.\end{aligned}$$

Therefore,

$$\frac{\nabla_K \pi_K(H(x)|x)}{\pi_K(H(x)|x)} = D_K c_K(x)^* \left[\Sigma^{-1}\left(\frac{\int_{H(x)} u \pi_K(u|x) du}{\pi_K(H(x)|x)} - c_K(x)\right)\right].$$

From the definition $\pi_K(H(x)|x) = \int_{H(x)} \pi_K(u|x) du$ and $m_H(x)$ from (21), we can further write

$$\frac{\nabla_K \pi_K(H(x)|x)}{\pi_K(H(x)|x)} = D_K c_K(x)^* \left[\Sigma^{-1}(m_H(x) - c_K(x))\right]. \tag{37}$$

Then combining (36) and (37), we rewrite $\nabla_K \log \pi_K^H(u|x)$ from (34) as

$$\begin{aligned}\nabla_K \log \pi_K^H(u|x) &= D_K c_K(x)^* \left[\Sigma^{-1}(u - c_K(x))\right] - D_K c_K(x)^* \left[\Sigma^{-1}(m_H(x) - c_K(x))\right] \\ &= D_K c_K(x)^* \left[\Sigma^{-1}(u - m_H(x))\right].\end{aligned}$$

Finally, under Assumption 2.5, $m_H(x) = \mu_K^H(x) = Kx$, and therefore

$$\nabla_K \log \pi_K^H(u|x) = D_K c_K(x)^* \left[\Sigma^{-1}(u - Kx)\right].$$

□

Now, we state another critical helper Lemma, that obtains an upper bound on the second moment of the truncated noises $\eta_i^H$, in terms of the variance of untruncated noises. This result enables us to guarantee that the infinite horizon accumulation of truncated noises is bounded.

**Lemma 7.6** (Second Moment Bound for Truncated Noise)**.** *Suppose that Assumptions 2.1, 2.3, and 2.5 hold. Then the second moment of the truncated noise $\eta_i^H$ satisfies*

$$\mathbb{E}\left[\eta_i^H \eta_i^{H^\top}\right] \preceq \sigma_i^2 I_m. \tag{38}$$

*Proof of Lemma 7.6.* Recall from the definition of $\pi_K(\cdot|x_i)$ in 3 that the pre-truncation stochastic wrapper is Gaussian with deterministic nominal map $c_K(x_i)$ and covariance $\Sigma_i = \sigma_i^2 I_m$. Also, recall from the definition of $\pi_K^H$ in 4 that it is based on the Gaussian policy $\pi_K$. We know that this base policy's density $\pi_K$ can be written as

$$\pi_K(u|x_i) \propto \exp\left(-\frac{1}{2}(u - c_K(x_i))^\top \Sigma_i^{-1}(u - c_K(x_i))\right).$$

Suppose

$$\phi_i(u) := \frac{1}{2}(u - c_K(x_i))^\top \Sigma_i^{-1}(u - c_K(x_i)).$$

Then, with $\nabla^2 \phi_i = \Sigma_i^{-1}$, we know that $\phi_i$ is strongly convex. Thus, it follows that $\pi_K(u|x_i)$ is strongly log-concave in $u$ for every $x_i$. Now, from the definition in 4 under Assumption 2.3, the truncated density follows

$$\pi_K^H(u|x_i) \propto \pi_K(u|x_i)\mathbf{1}_{\{u \in H(x_i)\}}, \tag{39}$$

where the indicator is given as

$$1_{H(x_i)}(u) = \begin{cases} 1, & u \in H(x_i), \\ 0, & u \notin H(x_i). \end{cases} \tag{40}$$

The above relation between the truncated and untruncated policies holds because $\pi_K(H(x_i)|x_i) > 0$ from Assumption 2.3. Since $H(x_i)$ is convex, we define its convex indicator function as

$$I_{H(x_i)}(u) = \begin{cases} 0, & u \in H(x_i), \\ \infty, & u \notin H(x_i). \end{cases} \tag{41}$$

Note that both indicator functions are related as follows:

$$1_{H(x_i)}(u) = e^{-I_{H(x_i)}(u)}. \tag{42}$$

Further, the convex indicator function is convex in $u$. Therefore, $1_{H(x_i)}(u)$ is log-concave in $u$. From (39), $\pi_K^H$ is a product of $\pi_K$ and $1_{H(x_i)}$:

$$\pi_K^H(u|x_i) \propto \exp\left(-\phi_i(u)\right) \times \exp\left(-I_{H(x_i)}(u)\right), \tag{43}$$

$$= \exp\left(-\phi_i(u) - I_{H(x_i)}(u)\right), \tag{44}$$

$$= \exp(-\phi_i^H(u)), \tag{45}$$

where $\phi_i^H(u) := \phi_i(u) + I_{H(x_i)}(u)$, which is a sum of two convex functions, and hence is convex. Therefore, the truncated density remains log-concave on its support $H(x_i)$. Now, from the Brascamp–Lieb (Poincaré) inequality for strongly log-concave measures with convex support [32], we obtain

$$\mathrm{Cov}\big(u_i \sim \pi_K^H(\cdot|x_i) \mid x_i\big) \preceq \Sigma_i = \sigma_i^2 I_m.$$

From (6), we know that $\eta_i^H = u_i - \mu_K^H(x_i)$. Therefore, conditional on $x_i$, $\eta_i^H$ is centered around zero by construction, and its conditional second moment equals the covariance of $u_i \sim \pi_K^H(\cdot|x_i)$:

$$\mathbb{E}\Big[\eta_i^H \eta_i^{H^\top} \mid x_i\Big] = \mathrm{Cov}\big(u_i \sim \pi_K^H(\cdot|x_i) \mid x_i\big) \preceq \sigma_i^2 I_m.$$

Taking expectation over $x_i$ gives

$$\mathbb{E}\left[\eta_i^H \eta_i^{H^\top}\right] \preceq \sigma_i^2 I_m.$$

□

Note that the bound above depends only on the covariance of the pre-truncation Gaussian wrapper and the convexity of the truncation set and does not require the Gaussian center $c_K(x_i)$ to coincide with the post-truncation mean $Kx_i$.

We present another helper result on a weighted expression of concatenated noise $\bar{\eta}_i^H$, for $i \geq 1$. This result is helpful in characterizing the infinite horizon accumulation of truncated noises.

**Lemma 7.7** (Weighted Long-horizon Noise)**.** *Suppose that Assumptions 2.3 and 2.5 hold. Then, it follows that*

$$\mathcal{C}_i \mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] \mathcal{C}_i^\top = \sum_{j=0}^{i-1} A_K^j B \Sigma_{i-1-j}^H B^\top A_K^{j^\top}, \tag{46}$$

*where*

$$\Sigma_j^H := \mathbb{E}[\eta_j^H (\eta_j^H)^\top]. \tag{47}$$

*Proof of Lemma 7.7.* Recall from (9) that we can write $\bar{\eta}_i^H \bar{\eta}_i^{H^\top}$ as

$$\begin{aligned}
\bar{\eta}_i^H \bar{\eta}_i^{H^\top} &= \begin{bmatrix} \eta_{i-1}^H \\ \eta_{i-2}^H \\ \vdots \\ \eta_0^H \end{bmatrix} \begin{bmatrix} \eta_{i-1}^{H^\top} & \eta_{i-2}^{H^\top} & \dots & \eta_0^{H^\top} \end{bmatrix}, \\
&= \begin{bmatrix} \eta_{i-1}^H \eta_{i-1}^{H^\top} & \eta_{i-1}^H \eta_{i-2}^{H^\top} & \dots & \eta_{i-1}^H \eta_0^{H^\top} \\ \eta_{i-2}^H \eta_{i-1}^{H^\top} & \eta_{i-2}^H \eta_{i-2}^{H^\top} & \dots & \eta_{i-2}^H \eta_0^{H^\top} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_0^H \eta_{i-1}^{H^\top} & \eta_0^H \eta_{i-2}^{H^\top} & \dots & \eta_0^H \eta_0^{H^\top} \end{bmatrix}.
\end{aligned}$$

Therefore, we can also write

$$\begin{aligned}
\mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] &= \mathbb{E}\begin{bmatrix} \eta_{i-1}^H \eta_{i-1}^{H^\top} & \eta_{i-1}^H \eta_{i-2}^{H^\top} & \dots & \eta_{i-1}^H \eta_0^{H^\top} \\ \eta_{i-2}^H \eta_{i-1}^{H^\top} & \eta_{i-2}^H \eta_{i-2}^{H^\top} & \dots & \eta_{i-2}^H \eta_0^{H^\top} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_0^H \eta_{i-1}^{H^\top} & \eta_0^H \eta_{i-2}^{H^\top} & \dots & \eta_0^H \eta_0^{H^\top} \end{bmatrix}, \\
&= \begin{bmatrix} \mathbb{E}\eta_{i-1}^H \eta_{i-1}^{H^\top} & \mathbb{E}\eta_{i-1}^H \eta_{i-2}^{H^\top} & \dots & \mathbb{E}\eta_{i-1}^H \eta_0^{H^\top} \\ \mathbb{E}\eta_{i-2}^H \eta_{i-1}^{H^\top} & \mathbb{E}\eta_{i-2}^H \eta_{i-2}^{H^\top} & \dots & \mathbb{E}\eta_{i-2}^H \eta_0^{H^\top} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\eta_0^H \eta_{i-1}^{H^\top} & \mathbb{E}\eta_0^H \eta_{i-2}^{H^\top} & \dots & \mathbb{E}\eta_0^H \eta_0^{H^\top} \end{bmatrix}.
\end{aligned}$$

However, from Proposition 2.7, we know that for $l \neq k$

$$\mathbb{E}\left[\eta_l^H \eta_k^{H^\top}\right] = 0.$$

Therefore, we can further simplify the expression $\mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right]$ as

$$\mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] = \begin{bmatrix} \mathbb{E}\left[\eta_{i-1}^H \eta_{i-1}^{H^\top}\right] & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbb{E}\left[\eta_{i-2}^H \eta_{i-2}^{H^\top}\right] & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbb{E}\left[\eta_0^H \eta_0^{H^\top}\right] \end{bmatrix}, \tag{48}$$

$$= \begin{bmatrix} \Sigma_{i-1}^H & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_{i-2}^H & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_0^H \end{bmatrix}, \tag{49}$$

where $\Sigma_j^H := \mathbb{E}\left[\eta_j^H \eta_j^{H^\top}\right]$. Then with (10), we now write the expression for $\mathcal{C}_i \mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] \mathcal{C}_i^\top$ as

$$\mathcal{C}_i \mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] \mathcal{C}_i^\top = \mathcal{C}_i \begin{bmatrix} \Sigma_{i-1}^H & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_{i-2}^H & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_0^H \end{bmatrix} \begin{bmatrix} B^\top \\ (A_K B)^\top \\ \vdots \\ (A_K^{i-1} B)^\top \end{bmatrix}, \tag{50}$$

$$= \begin{bmatrix} B & A_K B & \dots & A_K^{i-1} B \end{bmatrix} \begin{bmatrix} \Sigma_{i-1}^H B^\top \\ \Sigma_{i-2}^H B^\top A_K^\top \\ \vdots \\ \Sigma_0^H B^\top A_K^{i-1^\top} \end{bmatrix}, \tag{51}$$

$$= \sum_{j=0}^{i-1} A_K^j B \Sigma_{i-1-j}^H B^\top A_K^{j^\top}. \tag{52}$$

□

Now we establish an underlying smoothness property of the Lyapunov matrices for both discounted and undiscounted frameworks.

**Lemma 7.8** (Smoothness of the Lyapunov Matrices)**.** *The matrices $P_K^\gamma$ and $P_K$ are $\mathcal{C}^\infty$-function on every open subset of $\mathcal{K}^\gamma$ and $\mathcal{K}$, respectively.*

*Proof of Lemma 7.8.* We recall that

$$G_K^\gamma = \begin{cases} P_K^\gamma, & \gamma \in (0,1), \\ P_K, & \gamma = 1, \end{cases} \tag{53}$$

with $K$ following

$$K \in \begin{cases} \mathcal{K}^\gamma, & \gamma \in (0,1), \\ \mathcal{K}, & \gamma = 1. \end{cases} \tag{54}$$

Further, recall the set $\mathcal{S}_\mathcal{K}^\gamma$ for $\gamma \in (0,1]$, as

$$\mathcal{S}_\mathcal{K}^\gamma = \begin{cases} \mathcal{K}^\gamma, & \gamma \in (0,1), \\ \mathcal{K}, & \gamma = 1. \end{cases} \tag{55}$$

Then, we know that $G_K^\gamma$, for every $K \in \mathcal{S}_\mathcal{K}^\gamma$, satisfies

$$G_K^\gamma = Q + K^\top R K + \gamma (A + BK)^\top G_K^\gamma (A + BK) \tag{56}$$

We apply the vectorization operator $\text{vec}(\cdot)$, which stacks the columns of a matrix into a vector. Using the identity $\text{vec}(AXB) = (B^\top \otimes A)\text{vec}(X)$, we rewrite the equation as:

$$\text{vec}(G_K^\gamma) = \text{vec}(Q + K^\top R K) + \gamma \left(A_K^\top \otimes A_K^\top\right) \text{vec}(G_K^\gamma). \tag{57}$$

Let $\mathcal{A}_K = \sqrt{\gamma}(A + BK)$. The equation becomes:

$$\left(I - \mathcal{A}_K^\top \otimes \mathcal{A}_K^\top\right) \text{vec}(G_K^\gamma) = \text{vec}(Q_K) \tag{58}$$

$$\text{vec}(G_K^\gamma) = (M(K))^{-1} v(K), \tag{59}$$

where

$$M(K) := I_{n^2} - \mathcal{A}_K^\top \otimes \mathcal{A}_K^\top, \tag{60}$$

$$v(K) := \text{vec}(Q + K^\top R K). \tag{61}$$

First, since $A + BK$ is a linear and therefore smooth function of each $K$. Consequently, the Kronecker product $\mathcal{A}_K^\top \otimes \mathcal{A}_K^\top$, and hence $M(K)$, is also a smooth function of $K$. Further, we note that the spectral radius $\rho(\mathcal{A}_K) < 1$, for each $K \in \mathcal{S}_\mathcal{K}^\gamma$. This implies that the eigenvalues of $\mathcal{A}_K^\top \otimes \mathcal{A}_K^\top$ strictly have magnitude less than 1. Therefore, the matrix $M(K) = I_{n^2} - \mathcal{A}_K^\top \otimes \mathcal{A}_K^\top$ is strictly

non-singular and invertible. Therefore, the map $M(K) \mapsto (M(K))^{-1}$ is also a $\mathcal{C}^\infty$ on the set of invertible matrices, for each open subset of the set $\mathcal{S}_{\mathcal{K}}^{\gamma}$.

Then, since $Q + K^T RK$ is a smooth function of $K$, $v(K)$ is also a smooth function. Finally, we conclude that $K \mapsto \text{vec}(G_K^\gamma)$ is a composition of smooth functions, and hence $\text{vec}(G_K^\gamma)$ is a $\mathcal{C}^\infty$ function on each open subsets of $\mathcal{S}_{\mathcal{K}}^{\gamma}$. Consequently, we conclude that both $P_K$ and $P_K^\gamma$ are $\mathcal{C}^\infty$ functions over open subset of $\mathcal{K}$ and $\mathcal{K}^\gamma$, respectively. □

We now shift our attention to proofs for theoretical results in the undiscounted framework.

### 7.4 Undiscounted Safe LQR

We begin with proving a helper Lemma, that relates the undiscounted stochastic constrained LQR cost with deterministic constrained LQR cost.

**Lemma 7.9** (Cost Difference between Stochastic and Deterministic Safe LQR)**.** *Suppose that Assumptions 2.1, 2.3, 2.4, and 2.5 hold. Then, the cost $J_\pi^H(K)$ satisfies*

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K), \tag{62}$$

*where*

$$\Delta(K) := \alpha_1(K) + \alpha_2(K), \tag{63}$$

$$\alpha_1(K) := \sum_{i=0}^{\infty} \mathbb{E}_{\pi_K^H, x_0} \left[ {\eta_i^H}^\top R \eta_i^H \right] < \infty, \tag{64}$$

$$\alpha_2(K) := \sum_{i=0}^{\infty} \mathbb{E}_{\pi_K^H, x_0} \left[ (\mathcal{C}_i \bar{\eta}_i^H)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H \right] < \infty. \tag{65}$$

*Proof of Lemma 7.9.* Recall that with safe deterministic policy $\mu_K^H(x) = Kx$ with $\gamma = 1$, standard LQR theory ensures closed-loop stability and boundedness of the corresponding cost, defined as

$$J_\mu^H(K) = \mathbb{E}_{x_0} V_K(x_0), \tag{66}$$

where

$$V_K(x_0) = \sum_{i=0}^{\infty} x_i^T Q x_i + u_i^T R u_i = x_0^\top P_K x_0, \tag{67}$$

and $P_K$ satisfies the discrete-time Lyapunov equation for the closed-loop system given by $P_K = A_K^\top P_K A_K + Q_K^R$,

$$P_K = A_K^\top P_K A_K + Q_K^R, \tag{68}$$

where $Q_K^R = Q + K^T R K$. Then, with $u_t = \mu_K^H(x_t) + \eta_t^H$ and $\mu_K^H(x_t) = Kx_t$, $J_\pi^H(K)$ can be written, from 8 with $\gamma = 1$, as

$$J_\pi^H(K) = \mathbb{E}_{\{\eta_i^H\}, x_0} V_K(x_0) = \mathbb{E}_{\{\eta_i^H\}, x_0} \sum_{i=0}^{\infty} x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R (Kx_i + \eta_i^H).$$

Since each of the summand is nonnegative, we can apply Tonelli's theorem and interchange the expectation and the infinite sum operators as

$$J_\pi^H(K) = \sum_{i=0}^{\infty} \mathbb{E}_{\{\eta_i^H\}, x_0} \left[ x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R (Kx_i + \eta_i^H) \right]$$

Simplifying, we get

$$\begin{aligned} J_\pi^H(K) &= \sum_{i=0}^{\infty} \mathbb{E} \left[ x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R (Kx_i + \eta_i^H) \right], \\ &= \sum_{i=0}^{\infty} \mathbb{E} \big[ x_i^\top (Q + K^\top R K) x_i + 2 (Kx_i)^\top R \eta_i^H + {\eta_i^H}^\top R \eta_i^H \big], \\ &= \sum_{i=0}^{\infty} \mathbb{E} \left[ x_i^\top Q_K^R x_i + 2 (Kx_i)^\top R \eta_i^H + {\eta_i^H}^\top R \eta_i^H \right]. \end{aligned}$$

Then, note from $x_{i+1} = Ax_i + Bu_i$, $u_i = \mu_K^H(x_i) + \eta_i^H$ and $\mu_K^H(x_i) = Kx_i$, we can write

$$x_i = A_K^i x_0 + \mathcal{C}_i \bar{\eta}_i^H.$$

Thus, $J_\pi^H(K)$ can be written as

$$\begin{aligned}
J_\pi^H(K) =& \sum_{i=0}^{\infty} \mathbb{E}\big[(A_K^i x_0 + \mathcal{C}_i \bar{\eta}_i^H)^\top Q_K^R (A_K^i x_0 + \mathcal{C}_i \bar{\eta}_i^H) + 2(K A_K^i + K\mathcal{C}_i \bar{\eta}_i^H)^\top R \eta_i^H x_0 \\
&\qquad + {\eta_i^H}^\top R \eta_i^H\big] \\
=& \sum_{i=0}^{\infty} \mathbb{E}\big[(A_K^i x_0)^\top Q_K^R A_K^i x_0 + 2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H + {\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H \\
&\qquad + 2(K A_K^i x_0)^\top R \eta_i^H 2(K\mathcal{C}_i \bar{\eta}_i^H)^\top R \eta_i^H + {\eta_i^H}^\top R \eta_i^H\big] \\
=& \sum_{i=0}^{\infty} \mathbb{E}\big[(A_K^i x_0)^\top Q_K^R A_K^i x_0\big] + \sum_{i=0}^{\infty} \mathbb{E}\big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H + {\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H \\
&\qquad + 2(K A_K^i x_0)^\top R \eta_i^H + 2(K\mathcal{C}_i \bar{\eta}_i^H)^\top R \eta_i^H + {\eta_i^H}^\top R \eta_i^H\big].
\end{aligned}$$

However, from (66), we know that

$$J_\mu^H(K) = \mathbb{E}_{x_0} \sum_{i=0}^{\infty} x_i^\top Q x_i + \mu_K^H(x_i)^\top R \mu_K^H(x_i). \tag{69}$$

Since $\mu_K^H(x_i) = K x_i$, we get $x_i = A_K^i x_0$. Therefore, by Tonelli's theorem, we can interchange the expectation and the infinite sum operators and write

$$\begin{aligned}
J_\mu^H(K) &= \mathbb{E}_{x_0} \sum_{i=0}^{\infty} (A_K^i x_0)^\top Q_K^R (A_K^i x_0), \\
&= \sum_{i=0}^{\infty} \mathbb{E}_{x_0} \left[(A_K^i x_0)^\top Q_K^R (A_K^i x_0)\right].
\end{aligned}$$

Therefore, we can write

$$\begin{aligned}
J_\pi^H(K) =& J_\mu^H(K) + \sum_{i=0}^{\infty} \mathbb{E}\big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H + 2(K\mathcal{C}_i \bar{\eta}_i^H)^\top R \eta_i^H \\
&\qquad + 2(K A_K^i x_0)^\top R \eta_i^H + {\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H + {\eta_i^H}^\top R \eta_i^H\big].
\end{aligned}$$

Now, we define the following terms from the above cost expressions as

$$T_1 = \sum_{i=0}^{\infty} \mathbb{E}\Big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big], \tag{70}$$

$$T_2 = \sum_{i=0}^{\infty} \mathbb{E}\Big[2(K A_K^i x_0)^\top R \eta_i^H\Big], \tag{71}$$

$$T_3 = \sum_{i=0}^{\infty} \mathbb{E}\Big[2(K\mathcal{C}_i \bar{\eta}_i^H)^\top R \eta_i^H\Big], \tag{72}$$

$$T_4 = \sum_{i=0}^{\infty} \mathbb{E}\Big[{\eta_i^H}^\top R \eta_i^H\Big], \tag{73}$$

$$T_5 = \sum_{i=0}^{\infty} \mathbb{E}\Big[{\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big]. \tag{74}$$

Therefore, we can write

$$J_\pi^H(K) = J_\mu^H(K) + T_1 + T_2 + T_3 + T_4 + T_5. \tag{75}$$

We first prove two interim helper lemmas. We begin with $T_4$ and $T_5$.

**Lemma 7.10** (Helper Lemma)**.** *The terms $T_4$ and $T_5$ are upper bounded and have the following values:*

$$\alpha_2(K) = T_5 = \sum_{i=0}^{\infty} \mathbb{E}\Big[\bar{\eta}_i^{H^\top} \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big] < \infty,$$
$$\alpha_1(K) = T_4 = \sum_{i=0}^{\infty} \mathbb{E}\Big[\eta_i^{H^\top} R \eta_i^H\Big] < \infty.$$

*Proof.* We will begin with analyzing the term $T_4$ and then move on to the term $T_5$.
**Analysis for** $T_4$: First, we take the positive semidefinite matrix $R$ out of $T_4$'s expression and write

$$T_4 \leq \lambda_{\max}(R) \sum_{i=0}^{\infty} \mathbb{E}\eta_i^{H^\top} \eta_i^H.$$

Recall that for any random variable $X$, it holds that

$$\mathbb{E}[X^\top X] = \mathbb{E}[\text{Tr}(XX^\top)] = \text{Tr}(\mathbb{E}[XX^\top]),$$

thus, we can write

$$T_4 \leq \lambda_{\max}(R) \sum_{i=0}^{\infty} \text{Tr}(\mathbb{E}\eta_i^H \eta_i^{H^\top}).$$

Then, from Lemma 7.6 and Assumption 2.4, we can write

$$\begin{aligned} T_4 &\leq \sum_{i=0}^{\infty} \lambda_{\max}(R) \text{Tr}\left(\frac{\sigma^2}{(i+1)^2} I_m\right), \\ &= \sum_{i=0}^{\infty} \lambda_{\max}(R) \frac{m\sigma^2}{(i+1)^2}, \\ &= \lambda_{\max}(R) \frac{m\pi^2}{6}, \\ &< \infty. \end{aligned}$$

From (64) and $T_4$'s boundedness, we get

$$\begin{aligned} \alpha_1(K) &= T_4 \\ &< \infty. \end{aligned}$$

We shift our attention to $T_5$.
**Analysis for** $T_5$: Recall that

$$T_5 = \sum_{i=0}^{\infty} \mathbb{E}\Big[\bar{\eta}_i^{H^\top} \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big].$$

We invoke the properties of the trace operator and write the following:

$$\begin{aligned} T_5 &= \sum_{i=0}^{\infty} \mathbb{E}\left[\text{Tr}\Big(\bar{\eta}_i^{H^\top} \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big)\right], \\ &= \sum_{i=0}^{\infty} \mathbb{E}\left[\text{Tr}\Big(Q_K^R \mathcal{C}_i \bar{\eta}_i^H \bar{\eta}_i^{H^\top} \mathcal{C}_i^\top\Big)\right], \\ &= \sum_{i=0}^{\infty} \text{Tr}\Big(Q_K^R \mathcal{C}_i \mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] \mathcal{C}_i^\top\Big). \end{aligned} \tag{76}$$

For notational clarity, let

$$\Sigma_i^H(K) := \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right].$$

| **Indices** | $t=0$ | $t=1$ | $t=2$ | $t=3$ |
|---|---|---|---|---|
| $i=1$ | $B^\top Q_K^R B\Sigma_0^H(K)$ | | | |
| $i=2$ | $B^\top A_K^\top Q_K^R A_K B\Sigma_0^H(K)$ | $B^\top Q_K^R B\Sigma_1^H(K)$ | | |
| $i=3$ | $B^\top (A_K^2)^\top Q_K^R A_K^2 B\Sigma_0^H(K)$ | $B^\top A_K^\top Q_K^R A_K B\Sigma_1^H(K)$ | $B^\top Q_K^R B\Sigma_2^H(K)$ | |
| $i=4$ | $B^\top (A_K^3)^\top Q_K^R A_K^3 B\Sigma_0^H(K)$ | $B^\top (A_K^2)^\top Q_K^R A_K^2 B\Sigma_1^H(K)$ | $B^\top A_K^\top Q_K^R A_K B\Sigma_2^H(K)$ | $B^\top Q_K^R B\Sigma_3^H(K)$ |

Table 2: Expansion of dual sum in (77)

From Lemma 7.7, we write $T_5$ as

$$T_5 = \sum_{i=1}^{\infty} \mathrm{Tr}\Big( Q_K^R \sum_{j=0}^{i-1} A_K^j B\Sigma_{i-1-j}^H(K) B^\top {A_K^j}^\top \Big).$$

Using a property of the trace operator, we can further write

$$\begin{aligned} T_5 &= \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} \mathrm{Tr}\Big( Q_K^R A_K^j B\Sigma_{i-1-j}^H(K) B^\top {A_K^j}^\top \Big) \\ &= \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} \mathrm{Tr}\Big( B^\top {A_K^j}^\top Q_K^R A_K^j B\Sigma_{i-1-j}^H(K) \Big). \end{aligned} \tag{77}$$

Now, we simplify the above dual sum. From Table 2, recall that the columns go over the index $t$ and the rows go over $i$. Then, for every $t \geq 0$, the summation across rows, that is from $i=0$ to $i \to \infty$, yields

$$C_t := \sum_{j=0}^{\infty} \mathrm{Tr}\left( B^\top {A_K^j}^\top Q_K^R A_K^j B\Sigma_t^H(K) \right). \tag{78}$$

Furthermore, note that $T_5$ can be written as the sum

$$\begin{aligned} T_5 &= \sum_{t=0}^{\infty} C_t, \\ &= \sum_{t=0}^{\infty}\sum_{j=0}^{\infty} \mathrm{Tr}\left( B^\top {A_K^j}^\top Q_K^R A_K^j B\Sigma_t^H(K) \right). \end{aligned}$$

For simplicity, we equivalently write $t=i$. Then we can write $T_5$ as

$$T_5 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathrm{Tr}\Big( B^\top {A_K^j}^\top Q_K^R A_K^j B\Sigma_i^H(K) \Big).$$

Now we take the trace operator outside the internal infinite sum, as

$$\begin{aligned} T_5 &= \sum_{i=0}^{\infty} \mathrm{Tr}\Big( \sum_{j=0}^{\infty} \left( B^\top {A_K^j}^\top Q_K^R A_K^j B \right) \Sigma_i^H(K) \Big), \\ &= \sum_{i=0}^{\infty} \mathrm{Tr}\Big( B^\top \left( \sum_{j=0}^{\infty} \left( {A_K^j}^\top Q_K^R A_K^j \right) \right) B\Sigma_i^H(K) \Big). \end{aligned}$$

Note that the trace operator can be taken out only if the limit

$$\mathrm{Tr}\left( \lim_{M\to\infty} \sum_{j=0}^{M} \left( B^\top {A_K^j}^\top Q_K^R A_K^j B \right) \Sigma_i^H(K) \right)$$

exists. In what follows, we show that this limit indeed exists and that the swap of operators is justified. From (68) and $Q_K^R = Q + K^\top R K$, we have

$$A_K^\top P_K A_K - P_K + Q_K^R = \mathbf{0}.$$

Equivalently, we can rearrange $P_K$, and then multiply it with $A_K^\top$ from the left and $A_K$ from the right side, and again write $P_K$ as

$$\begin{aligned} P_K &= A_K^\top P_K A_K + Q_K^R, \\ A_K^\top P_K A_K &= {A_K^2}^\top P_K A_K^2 + A_K^\top Q_K^R A_K, \\ P_K &= {A_K^2}^\top P_K A_K^2 + A_K^\top Q_K^R A_K + Q_K^R. \end{aligned}$$

We again multiply $P_K$'s expression with $A_K^\top$ and $A_K$ appropriately to obtain

$$\begin{aligned} A_K^\top P_K A_K &= {A_K^3}^\top P_K A_K^3 + {A_K^2}^\top Q_K^R A_K^2 + A_K^\top Q_K^R A_K, \\ P_K &= {A_K^3}^\top P_K A_K^3 + {A_K^2}^\top Q_K^R A_K^2 + A_K^\top Q_K^R A_K + Q_K^R. \end{aligned}$$

Repeating the same steps, we can further obtain

$$P_K = \lim_{l\to\infty} {A_K^l}^\top P_K A_K^l + \sum_{j=0}^{\infty} \left( {A_K^j}^\top Q_K^R A_K^j \right). \tag{79}$$

However, since $A_K$ is a Schur matrix, we know that

$$\lim_{l\to\infty} {A_K^l}^\top P_K A_K^l = \mathbf{0}.$$

Thus, we can finally write

$$P_K = \sum_{j=0}^{\infty} \left( {A_K^j}^\top Q_K^R A_K^j \right). \tag{80}$$

We now shift our focus back to $T_5$ and substitute the inner infinite sum with $P_K$ as

$$T_5 = \sum_{i=0}^{\infty} \text{Tr}\Big( B^\top P_K B \Sigma_i^H(K) \Big). \tag{81}$$

Using inequality $\text{Tr}(M_1 M_2) \le \lambda_{max}(M_1)\text{Tr}(M_2)$ for positive semidefinite matrix $M_1$, we can obtain the following upper bound on $T_5$ as

$$\begin{aligned} T_5 &\le \lambda_{max}(B^\top P_K B) \sum_{i=0}^{\infty} \text{Tr}\Big( \Sigma_i^H(K) \Big), \\ &\le \lambda_B^P \sum_{i=0}^{\infty} \text{Tr}\Big( \mathbb{E}\left[ \eta_i^H {\eta_i^H}^\top \right] \Big), \end{aligned}$$

where

$$\lambda_B^P := \lambda_{max}(B^\top P_K B). \tag{82}$$

Then, from Lemma 7.6 and Assumption 2.4, we get

$$\begin{aligned} T_5 &\le \lambda_B^P \sum_{i=0}^{\infty} \text{Tr}\left( \frac{\sigma^2}{(i+1)^2} I_m \right), \\ &= \lambda_B^P \sum_{i=0}^{\infty} \frac{m\sigma^2}{(i+1)^2}, \\ &= \lambda_B^P \frac{m\pi^2}{6}, \\ &< \infty. \end{aligned}$$

Finally, from (65) and $T_5$'s boundedness we get

$$\begin{aligned} \alpha_2(K) &= T_5, \\ &< \infty. \end{aligned}$$

Note that $\alpha_2(K)$ is a function of $K$ through $P_K$, $Q_K^R$, $\mathcal{C}_i$, and the truncated-noise law induced by $\pi_K^H$. □

**Lemma 7.11** (Helper Lemma)**.** *The terms $T_1$, $T_2$ and $T_3$ are 0.*

*Proof.* It suffices to show that each summand in $T_1, T_2$ and $T_3$ has zero expectation. First, consider $T_1$. For each $i \geq 0$,

$$\mathbb{E}[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H] = \mathbb{E}\big[\mathbb{E}[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H \mid x_0]\big]. \tag{83}$$

Since $A_K^i x_0$ is measurable with respect to $x_0$, the internal expectation can be written as

$$\mathbb{E}[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H \mid x_0] = 2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \mathbb{E}[\bar{\eta}_i^H \mid x_0]. \tag{84}$$

Now, from Corollary 7.5,

$$\mathbb{E}[\bar{\eta}_i^H \mid x_0] = 0. \tag{85}$$

Therefore every summand in $T_1$ is zero, and hence

$$T_1 = 0.$$

Second, consider $T_2$. For each $i \geq 0$,

$$\mathbb{E}[2(KA_K^i x_0)^\top R \eta_i^H] = \mathbb{E}\big[\mathbb{E}[2(KA_K^i x_0)^\top R \eta_i^H \mid x_0]\big]. \tag{86}$$

Since $KA_K^i x_0$ is measurable with respect to $x_0$,

$$\mathbb{E}[2(KA_K^i x_0)^\top R \eta_i^H \mid x_0] = 2(KA_K^i x_0)^\top R \mathbb{E}[\eta_i^H \mid x_0]. \tag{87}$$

Using Proposition 2.7 and the tower property,

$$\mathbb{E}[\eta_i^H \mid x_0] = \mathbb{E}\big[\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H] \mid x_0\big] = 0. \tag{88}$$

Therefore every summand in $T_2$ is zero, and hence

$$T_2 = 0.$$

Finally consider $T_3$. From (7), for each $i \geq 0$, $\bar{\eta}_i^H$ is $\mathcal{F}_i^H$-measurable. Hence $K\mathcal{C}_i\bar{\eta}_i^H$ is also $\mathcal{F}_i^H$-measurable. Therefore,

$$\mathbb{E}[2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R \eta_i^H] = \mathbb{E}\big[\mathbb{E}[2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R \eta_i^H \mid \mathcal{F}_i^H]\big]. \tag{89}$$

Inside the conditional expectation,

$$\mathbb{E}[2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R \eta_i^H \mid \mathcal{F}_i^H] = 2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R \mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H]. \tag{90}$$

By Proposition 2.7,

$$\mathbb{E}[\eta_i^H \mid \mathcal{F}_i^H] = 0. \tag{91}$$

Therefore every summand in $T_3$ is zero, and hence

$$T_3 = 0.$$

Combining the three identities gives

$$T_1 = T_2 = T_3 = 0. \tag{92}$$

□

With these two helper Lemma, from (75), we can finally obtain

$$J_\pi^H(K) = J_\mu^H(K) + \alpha_1(K) + \alpha_2(K) = J_\mu^H(K) + \Delta(K). \tag{93}$$

□

Based on the above helper Lemma, we can conclude this key result relating the derivatives undiscounted stochastic and deterministic constrained LQR costs.

*Proof of Theorem 3.1.* From Lemma 7.9, we have

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K).$$

Under Assumption 2.6, together with the local regularity results for $\alpha_1(K)$ and $\alpha_2(K)$, the term $\Delta(K) = \alpha_1(K) + \alpha_2(K)$ is differentiable in $K$. Therefore, differentiating the cost decomposition gives

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta(K).$$

□

Given that the additional terms in the Lemma 7.9 and Theorem 3.1 is $\alpha_2(K)$, we characterize its upper bound and smoothness in helper Lemmas below.

**Lemma 7.12.** *Suppose that Assumptions 2.1, 2.3, 2.5, 2.4, and 2.6 hold. Then, the term $\alpha_2(K)$ can be expressed as*

$$\alpha_2(K) = \mathrm{Tr}(P_K M_{\alpha_2}(K)), \tag{94}$$

*where*

$$M_{\alpha_2}(K) := B\Sigma^H(K)B^\top, \qquad \Sigma^H(K) := \sum_{i=0}^{\infty} \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right].$$

*Moreover, $M_{\alpha_2}(K)$ is locally differentiable in $K$ with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$.*

*Proof of Lemma 7.12.* From (81), we have

$$\alpha_2(K) = T_5 = \sum_{i=0}^{\infty} \mathrm{Tr}\left(B^\top P_K B \Sigma_i^H(K)\right).$$

Therefore,

$$\begin{aligned}
\alpha_2(K) &= \mathrm{Tr}\left(B^\top P_K B \sum_{i=0}^{\infty} \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right]\right) \\
&= \mathrm{Tr}\left(P_K B \Sigma^H(K) B^\top\right) \\
&= \mathrm{Tr}\left(P_K M_{\alpha_2}(K)\right).
\end{aligned}$$

The summability of $\Sigma^H(K)$ follows from Lemma 7.6 and Assumption 2.4. Moreover, the truncated-noise law depends on $K$ through the deterministic nominal map $c_K$. By Assumption 2.6, this dependence is locally regular in $K$, and the second-moment bounds provide the domination needed to differentiate the infinite sum term by term. Hence $M_{\alpha_2}(K)$ is locally differentiable in $K$ with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$.

Lastly, $P_K$ is a $\mathcal{C}^\infty$ function over every open subset of the stabilizing set by Lemma 7.8. In addition, the truncated-noise law depends on $K$ through $c_K$, and this dependence is locally smooth by Assumption 2.6. Together with the summability bounds above, this implies that $\alpha_2(K)$ is differentiable on $\mathcal{K}_H^\gamma$, with locally controlled derivatives. □

We now prove key helper results on the regularity of $\alpha_1$ and $\alpha_2$ in what follows.

**Lemma 7.13** (Regularity of $\alpha_1$)**.** *Suppose Assumptions 2.3, 2.4, 2.5, and 2.6 hold. Then $\nabla_K \alpha_1(K)$ exists and is locally Lipschitz on $\mathcal{K}_H^\gamma$, with Lipschitz constant $l_{\alpha_1}(K)$ as:*

$$\|\nabla_K \alpha_1(K') - \nabla_K \alpha_1(K)\| \le l_{\alpha_1}(K) \|K' - K\|. \tag{95}$$

*Proof.* Using

$$\eta_i^{H^\top} R \eta_i^H = \mathrm{Tr}\left(R \eta_i^H \eta_i^{H^\top}\right),$$

we can write

$$\alpha_1(K) = \mathrm{Tr}\left(R \sum_{i=0}^{\infty} \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right]\right).$$

For each finite $i$, the map

$$K \mapsto \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right]$$

is differentiable because the truncated policy depends smoothly on $K$ through the deterministic nominal map $c_K$ by Assumption 2.6, and the closed-loop recursion depends smoothly on $K$ through $A_K = A + BK$. Moreover, Assumption 2.6 gives uniform compact-set bounds on the first and second $K$-derivatives of $c_K$. Together with Lemma 7.6 and Assumption 2.4, these bounds provide a summable dominating sequence for the derivatives of the summands. Therefore, differentiation can be interchanged with the expectation and the infinite sum by dominated convergence. Hence $\nabla_K \alpha_1(K)$ exists and is continuous. On any compact subset $\mathcal{C} \subset \mathcal{K}_H^\gamma$, the same compact-set derivative bounds give

$$\sup_{K \in \mathcal{C}} \left\|\nabla_K^2 \alpha_1(K)\right\| < \infty.$$

Thus, by the mean-value inequality, $\nabla_K \alpha_1(K)$ is locally Lipschitz. Consequently, for $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, there exists $l_{\alpha_1}(K) < \infty$ such that

$$\|\nabla_K \alpha_1(K') - \nabla_K \alpha_1(K)\| \le l_{\alpha_1}(K) \|K' - K\|.$$

□

**Lemma 7.14.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then, for stabilizing parameters $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, the gradient $\nabla_K \alpha_2(K)$ is locally Lipschitz with Lipschitz constant $l_{\alpha_2}(K)$ as*

$$\|\nabla_K \alpha_2(K') - \nabla_K \alpha_2(K)\| \le l_{\alpha_2}(K) \|K' - K\|. \tag{96}$$

*Proof of Lemma 7.14.* From Lemma 7.12, we have

$$\alpha_2(K) = \mathrm{Tr}(P_K M_{\alpha_2}(K)).$$

Therefore, the full differential is

$$d\alpha_2 = \mathrm{Tr}((dP_K) M_{\alpha_2}(K)) + \mathrm{Tr}(P_K \, dM_{\alpha_2}(K)). \tag{97}$$

We first consider the term $\mathrm{Tr}((dP_K) M_{\alpha_2}(K))$. Let $Y_K$ be the solution of the Lyapunov equation

$$Y_K = A_K Y_K A_K^\top + M_{\alpha_2}(K). \tag{98}$$

Following the same Lyapunov-adjoint calculation as in the deterministic LQR smoothness proof from Lemma 7.8, the first term in (97) gives

$$\mathrm{Tr}((dP_K) M_{\alpha_2}(K)) = \left\langle 2E_K Y_K^\top, dK \right\rangle, \tag{99}$$

where

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

For the second term in (97), define $G_{\alpha_2}(K)$ as the unique matrix satisfying

$$\langle G_{\alpha_2}(K), dK \rangle = \mathrm{Tr}(P_K \, D_K M_{\alpha_2}(K)[dK]), \qquad \forall dK. \tag{100}$$

Such a matrix exists by the Riesz representation theorem under the Frobenius inner product. Combining the two terms yields

$$\nabla_K \alpha_2(K) = 2E_K Y_K^\top + G_{\alpha_2}(K). \tag{101}$$

By Assumption 2.6 and Lemma 7.12, the map $K \mapsto M_{\alpha_2}(K)$ is locally differentiable with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$. Moreover, $K \mapsto P_K$, $K \mapsto E_K$, and the Lyapunov solution $K \mapsto Y_K$ are locally smooth on the stabilizing set Lemma 7.8. Hence both terms in (101) are locally Lipschitz. Therefore, for stabilizing $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, there exists $l_{\alpha_2}(K) < \infty$ such that

$$\|\nabla_K \alpha_2(K') - \nabla_K \alpha_2(K)\| \le l_{\alpha_2}(K) \|K' - K\|.$$

□

**Lemma 7.15** (Regularity of $\Delta$)**.** *Suppose Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then $\nabla_K \Delta(K)$ exists and is locally Lipschitz on $\mathcal{K}_H^\gamma$. In particular, for stabilizing $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$,*

$$\|\nabla_K \Delta(K') - \nabla_K \Delta(K)\| \le l_\Delta(K) \|K' - K\|,$$

*where one may take $l_\Delta(K) = l_{\alpha_1}(K) + l_{\alpha_2}(K)$.*

*Proof.* The claim follows directly from $\Delta(K) = \alpha_1(K) + \alpha_2(K)$, Lemma 7.13, and Lemma 7.14. □

We are now ready to prove the well-definedness of the stochastic constrained LQR cost and its gradient.

*Proof on Theorem 3.2.* Fix any $K \in \mathcal{K}_H^\gamma$. Since $K \in \mathcal{K}_H^\gamma \subseteq S_K^\gamma$, the closed-loop matrix $A_K = A + BK$ is stabilizing. Hence the deterministic LQR cost

$$J_\mu^H(K) = \mathbb{E}_{x_0}[x_0^\top P_K x_0]$$

is finite, where $P_K$ satisfies

$$P_K = A_K^\top P_K A_K + Q_K^R, \qquad Q_K^R := Q + K^\top R K.$$

Moreover, standard Lyapunov sensitivity arguments imply that $J_\mu^H(K)$ and $\nabla_K J_\mu^H(K)$ are well-defined for stabilizing $K$. By Lemma 7.9,

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K).$$

The terms $\alpha_1(K)$ and $\alpha_2(K)$ are finite by Lemma 7.9, and therefore $\Delta(K) < \infty$. Hence $J_\pi^H(K)$ is well-defined. For the gradient, Theorem 3.1 gives

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta(K).$$

The first term is well-defined by deterministic LQR smoothness, and the second term is well-defined by Lemma 7.15. Therefore, $\nabla_K J_\pi^H(K)$ is well-defined over the truncated policy class. □

Before finally stating the Lipschitzness of stochastic gradient $\nabla_K J_\pi^H(K)$, we recall this result for undiscounted LQR cost.

**Lemma 7.16** (Lemma 5, [27])**.** *For any stabilizing policy parameters $K, K'$ satisfying the property $\|K' - K\| \le \epsilon(K)$, there exist constant $l(K)$ such that the undiscounted gradient $\nabla_K J_\mu^H(K)$ satisfies the following*

$$\left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K)\right\| \le \bar{l}(K)\left\|K' - K\right\|. \tag{102}$$

With the well-definedness of the truncated objective established, we next prove local smoothness of the stochastic constrained LQR objective.

**Lemma 7.17.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then, for any stabilizing policy parameter pairs $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, there exists a Lipschitz constant $l(K)$ such that*

$$\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| \le l(K)\left\|K' - K\right\|. \tag{103}$$

*Proof of Lemma 7.17.* From Lemma 7.16, we know that

$$\left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K)\right\| \le \bar{l}(K)\left\|K' - K\right\|.$$

Further, from Theorem 3.1, we know that

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta(K).$$

Therefore,

$$\begin{aligned}
&\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| \\
&= \left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K) + \nabla_K \Delta(K') - \nabla_K \Delta(K)\right\| \\
&\le \left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K)\right\| + \left\|\nabla_K \Delta(K') - \nabla_K \Delta(K)\right\| \\
&\le \bar{l}(K)\left\|K' - K\right\| + \left\|\nabla_K \Delta(K') - \nabla_K \Delta(K)\right\|.
\end{aligned}$$

Using Lemma 7.15, we can further write

$$\begin{aligned}
\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| &\le \bar{l}(K)\left\|K' - K\right\| + l_\Delta(K)\left\|K' - K\right\| \\
&= l(K)\left\|K' - K\right\|,
\end{aligned}$$

where $l(K) = \bar{l}(K) + l_\Delta(K)$. □

This Lipschitzness established above plays a central role in ensuring that local variations of the objective are well controlled over the stabilizing set $K$, despite the presence of truncated stochastic exploration.

With this Lipschitz property of the gradient proven, we turn our focus to characterizing the suboptimality that is introduced due exploration.

*Proof of Theorem 3.3.* By Lemma 7.9,

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K),$$

where $\Delta(K) \geq 0$. Therefore,

$$J_\pi^{H^*} = \min_{K \in \mathcal{K}_H^\gamma} \left( J_\mu^H(K) + \Delta(K) \right) \geq \min_{K \in \mathcal{K}_H^\gamma} J_\mu^H(K) = J_\mu^{H^*}.$$

This proves the lower bound. For the upper bound, evaluate $J_\pi^H$ at $K_\mu^*$:

$$J_\pi^{H^*} \leq J_\pi^H(K_\mu^*) = J_\mu^H(K_\mu^*) + \Delta(K_\mu^*) = J_\mu^{H^*} + \Delta(K_\mu^*).$$

Thus,

$$0 \leq J_\pi^{H^*} - J_\mu^{H^*} \leq \Delta(K_\mu^*) = \beta^*.$$

□

**Lemma 7.18.** *Suppose that Assumptions 2.1, 2.3, 2.4, 2.5, and 2.6 hold. Then the cost $J_\pi^H(K)$ satisfies*

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} \leq & \alpha \left\| \nabla_K J_\pi^H(K) \right\|^2 + \alpha \left\| \nabla_K \Delta(K) \right\|^2 \\ & + 2\alpha \left\| \nabla_K \Delta(K) \right\| \left\| \nabla_K J_\pi^H(K) \right\| + \Delta(K). \end{aligned}$$

*Proof of Lemma 7.18.* By Lemma 7.9,

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K),$$

and by Theorem 3.3,

$$J_\pi^{H^*} - J_\mu^{H^*} \geq 0.$$

Therefore,

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} &= J_\mu^H(K) + \Delta(K) - J_\pi^{H^*} \\ &= J_\mu^H(K) - J_\mu^{H^*} + \Delta(K) + J_\mu^{H^*} - J_\pi^{H^*} \\ &\leq J_\mu^H(K) - J_\mu^{H^*} + \Delta(K), \end{aligned}$$

Using the deterministic LQR gradient dominance condition from [20] for $K \in \mathcal{K}_H^\gamma \subseteq \mathcal{S}_\mathcal{K}^\gamma$, we get

$$J_\mu^H(K) - J_\mu^* \leq \alpha \left\| \nabla_K J_\mu^H(K) \right\|^2, \tag{104}$$

where $J_\mu^*$ denotes the unconstrained deterministic LQR optimum, and

$$\alpha = \frac{\|\Sigma_K^*\|}{\sigma_{min}(\Sigma_K)^2 \sigma_{min}(R)}, \tag{105}$$

$$\Sigma_K = \mathbb{E}_{x_0} \left[ \sum_{t=0}^{\infty} x_t x_t^\top \right]. \tag{106}$$

Here, $\Sigma_K^*$ denotes the state covariance associated with the unconstrained deterministic LQR optimum. Since the constrained deterministic optimum satisfies $J_\mu^{H^*} \geq J_\mu^*$, it follows that

$$J_\mu^H(K) - J_\mu^{H^*} \leq \alpha \left\| \nabla_K J_\mu^H(K) \right\|^2. \tag{107}$$

From Theorem 3.1,

$$\nabla_K J_\mu^H(K) = \nabla_K J_\pi^H(K) - \nabla_K \Delta(K).$$

Thus,

$$\left\|\nabla_K J_\mu^H(K)\right\| \le \left\|\nabla_K J_\pi^H(K)\right\| + \left\|\nabla_K \Delta(K)\right\|.$$

Combining these inequalities gives

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} &\le \alpha \left(\left\|\nabla_K J_\pi^H(K)\right\| + \left\|\nabla_K \Delta(K)\right\|\right)^2 + \Delta(K) \\ &\le \alpha \left\|\nabla_K J_\pi^H(K)\right\|^2 + \alpha \left\|\nabla_K \Delta(K)\right\|^2 \\ &\quad + 2\alpha \left\|\nabla_K \Delta(K)\right\| \left\|\nabla_K J_\pi^H(K)\right\| + \Delta(K). \end{aligned}$$

□

**Lemma 7.19** (Boundedness of sublevel constants)**.** *Suppose $\mathbb{E}[x_0 x_0^\top] \succ 0$. In the undiscounted setting, let*

$$\mathcal{L}_0 := \left\{K \in \mathcal{K}_H^\gamma : J_\pi^H(K) \le J_\pi^H(K_0)\right\}.$$

*Assume that $\overline{\mathcal{L}_0}$ is contained in the relative interior of $\mathcal{K}_H^\gamma$. Then*

$$\sup_{K \in \mathcal{L}_0} \Delta(K) < \infty, \qquad \sup_{K \in \mathcal{L}_0} \left\|\nabla_K \Delta(K)\right\| < \infty.$$

*Similarly, in the discounted setting, let*

$$\mathcal{L}_{0,\gamma} := \left\{K \in \mathcal{K}_H^\gamma : J_\pi^H(K) \le J_\pi^H(K_0)\right\}.$$

*Assume that $\overline{\mathcal{L}_{0,\gamma}}$ is contained in the relative interior of $\mathcal{K}_H^\gamma$. Then*

$$\sup_{K \in \mathcal{L}_{0,\gamma}} \Delta_\gamma(K) < \infty, \qquad \sup_{K \in \mathcal{L}_{0,\gamma}} \left\|\nabla_K \Delta_\gamma(K)\right\| < \infty.$$

*Proof.* We prove the undiscounted claim; the discounted proof is identical with $\Delta$ replaced by $\Delta_\gamma$. From the cost decomposition,

$$J_\pi^H(K) = J_\mu^H(K) + \Delta(K), \qquad \Delta(K) \ge 0.$$

Thus, for every $K \in \mathcal{L}_0$,

$$J_\mu^H(K) \le J_\pi^H(K) \le J_\pi^H(K_0).$$

Hence,

$$\mathcal{L}_0 \subseteq \left\{K \in S_K^\gamma : J_\mu^H(K) \le J_\pi^H(K_0)\right\}.$$

Under $\mathbb{E}[x_0 x_0^\top] \succ 0$, standard LQR coercivity implies that the deterministic LQR sublevel set on the right-hand side is compact. Since $\overline{\mathcal{L}_0}$ is contained in the relative interior of $\mathcal{K}_H^\gamma$, the functions $\Delta(K)$ and $\nabla_K \Delta(K)$ are continuous on a compact set containing $\overline{\mathcal{L}_0}$ by Lemmas 7.13, 7.14, and 7.15. Therefore, the stated suprema are finite. The discounted claim follows similarly from the discounted decomposition and Lemmas 7.27, 7.25, and 7.28. □

**Lemma 7.20** (Sublevel gradient bound)**.** *Let $\mathcal{L} \subset \mathcal{K}_H^\gamma$ be a sublevel set satisfying the conditions of Lemma 7.19. Then there exists a constant $G_\mathcal{L} < \infty$ such that*

$$\sup_{K \in \mathcal{L}} \left\|\nabla_K J_\pi^H(K)\right\| \le G_\mathcal{L}.$$

*Proof of Lemma 7.20.* By Lemma 7.19, $\mathcal{L}$ is contained in a compact subset of the relative interior of $\mathcal{K}_H^\gamma$. For the undiscounted case, Theorem 3.1 gives

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta(K).$$

The deterministic LQR gradient $\nabla_K J_\mu^H(K)$ is continuous on the stabilizing set, and Lemma 7.15 implies that $\nabla_K \Delta(K)$ is locally Lipschitz, hence continuous, on compact subsets of $\mathcal{K}_H^\gamma$. Therefore, $\nabla_K J_\pi^H(K)$ is continuous on a compact set containing $\mathcal{L}$, and the desired bound follows. The discounted case follows analogously by replacing $\Delta(K)$ with $\Delta_\gamma(K)$ and using Lemma 7.28. □

*Proof of Theorem 3.4.* Recall from Lemma 7.18 that

$$J_\pi^H(K) - J_\pi^{H^*} \leq \alpha \left\|\nabla_K J_\pi^H(K)\right\|^2 + \alpha \left\|\nabla_K \Delta(K)\right\|^2 + 2\alpha \left\|\nabla_K \Delta(K)\right\| \left\|\nabla_K J_\pi^H(K)\right\| + \Delta(K). \tag{108}$$

By Lemma 7.20, the policy gradient is bounded on $\mathcal{L}_0$. Under the normalized convention $G_{\mathcal{L}_0} = 1$, we have

$$\left\|\nabla_K J_\pi^H(K)\right\|^2 \leq \left\|\nabla_K J_\pi^H(K)\right\|. \tag{109}$$

Combining (108) and (109), we get

$$J_\pi^H(K) - J_\pi^{H^*} \leq (\alpha + 2\alpha \left\|\nabla_K \Delta(K)\right\|) \left\|\nabla_K J_\pi^H(K)\right\| + \alpha \left\|\nabla_K \Delta(K)\right\|^2 + \Delta(K).$$

Taking suprema over $K \in \mathcal{L}_0$, whose finiteness follows from Lemma 7.19, we obtain

$$J_\pi^H(K) - J_\pi^{H^*} \leq \alpha^H \left\|\nabla_K J_\pi^H(K)\right\| + \beta^H,$$

with $\alpha^H$ and $\beta^H$ as stated in Theorem 3.4. □

*Proof of Theorem 3.6.* We first define

$$\bar{J}_\pi^{H^*} := J_\pi^{H^*} + \beta^H.$$

Then, from Theorem 3.4, we have

$$\frac{1}{\alpha^H} \max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\} \leq \left\|\nabla_K J_\pi^H(K_t)\right\|. \tag{110}$$

Using the local Lipschitz property of $\nabla_K J_\pi^H(K)$ from Lemma 7.17, we get

$$J_\pi^H(K_{t+1}) \leq J_\pi^H(K_t) + \left\langle \nabla_K J_\pi^H(K_t), K_{t+1} - K_t \right\rangle + \frac{l(K_t)}{2} \left\|K_{t+1} - K_t\right\|_F^2. \tag{111}$$

Since Algorithm 1 updates

$$K_{t+1} = K_t - \zeta_t \widehat{\nabla_K J_\pi^H(K_t)},$$

we can write

$$J_\pi^H(K_{t+1}) \leq J_\pi^H(K_t) - \zeta_t \left\langle \nabla_K J_\pi^H(K_t), \widehat{\nabla_K J_\pi^H(K_t)} \right\rangle + \frac{l(K_t)}{2} \zeta_t^2 \left\| \widehat{\nabla_K J_\pi^H(K_t)} \right\|_F^2. \tag{112}$$

Let $\mathcal{F}_t$ denote the filtration generated by the iterates and sampling randomness up to iteration $t$. For compactness, write

$$g_t := \nabla_K J_\pi^H(K_t), \qquad \widehat{g}_t := \widehat{\nabla_K J_\pi^H(K_t)}.$$

Taking conditional expectation with respect to $\mathcal{F}_t$, and using Assumption 3.5, gives

$$\begin{aligned}\mathbb{E}\left[J_\pi^H(K_{t+1}) \mid \mathcal{F}_t\right] &\leq J_\pi^H(K_t) - \zeta_t \left\langle g_t, \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] \right\rangle + \frac{l(K_t)}{2} \zeta_t^2 \mathbb{E}\left[\left\|\widehat{g}_t\right\|_F^2 \mid \mathcal{F}_t\right] \\ &= J_\pi^H(K_t) - \zeta_t \left\|g_t\right\|_F^2 + \frac{l(K_t)}{2} \zeta_t^2 \mathbb{E}\left[\left\|\widehat{g}_t\right\|_F^2 \mid \mathcal{F}_t\right]. \end{aligned} \tag{113}$$

Moreover, by the Frobenius mean-square bound in Assumption 3.5,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_t\right\|_F^2 \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\left\|g_t + (\widehat{g}_t - g_t)\right\|_F^2 \mid \mathcal{F}_t\right] \\ &= \left\|g_t\right\|_F^2 + \mathbb{E}\left[\left\|\widehat{g}_t - g_t\right\|_F^2 \mid \mathcal{F}_t\right] \\ &\leq \left\|g_t\right\|_F^2 + \frac{V}{N}. \end{aligned} \tag{114}$$

Substituting (114) into (113), we obtain

$$\mathbb{E}\left[J_\pi^H(K_{t+1}) \mid \mathcal{F}_t\right] \leq J_\pi^H(K_t) - \zeta_t \left(1 - \frac{l(K_t)\zeta_t}{2}\right) \left\|g_t\right\|_F^2 + \frac{l(K_t)}{2} \zeta_t^2 \frac{V}{N}. \tag{115}$$

Since $\zeta_t \leq 1/l^{mx}(T)$, we have

$$1 - \frac{l(K_t)\zeta_t}{2} \geq \frac{1}{2}.$$

In what follows, we omit the conditional expectation for simplicity of exposition. Using (110), we get

$$\begin{aligned}\mathbb{E}\left[J_\pi^H(K_{t+1})\right] - \bar{J}_\pi^{H^*} \leq J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \frac{\zeta_t}{2(\alpha^H)^2}\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}^2 \\ + \frac{l^{mx}(T)}{2}\zeta_t^2\frac{V}{N}. \end{aligned} \tag{116}$$

Then, using Jensen's inequality on two convex functions, first on the square function and then on the $\max$ function, we get the following

$$\begin{aligned}\mathbb{E}\left[\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}^2\right] &\geq \mathbb{E}\left[\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}\right]^2, \\ &\geq \max\{0, \mathbb{E}\left[J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\right]\}^2. \end{aligned} \tag{117}$$

Now, we establish the following key inequality based on (116) and (117), as

$$\rho_{t+1} \leq \rho_t - \frac{\zeta_t}{2(\alpha^H)^2}\max\{0, \rho_t\}^2 + \frac{l(K_t)}{2}\zeta_t^2\frac{V}{N}, \tag{118}$$

where $\rho_t := \mathbb{E}_t\left[J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\right]$.
By substituting $L_2 = l^{mx}(T)$, we follow the proof of Theorem 6.1 in [28] to obtain

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - \bar{J}_\pi^{H^*} \leq c_T\left(1 - \sqrt{\frac{\zeta^3 l^{mx}(T)V}{4(\alpha^H)^2 N}}\right)^T + \sqrt{\frac{l^{mx}(T)V\zeta(\alpha^H)^2}{N}},$$

or, from $\bar{J}_\pi^{H^*} = J_\pi^{H^*} + \beta^H$, we can also write

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\pi^{H^*} \leq \beta^H + c_T\left(1 - \sqrt{\frac{\zeta^3 l^{mx}(T)V}{4(\alpha^H)^2 N}}\right)^T + \sqrt{\frac{l^{mx}(T)V\zeta(\alpha^H)^2}{N}}.$$

Finally, from Theorem 3.3, we can further write

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \beta^H + \beta^* + c_T\left(1 - \sqrt{\frac{\zeta^3 l^{mx}(T)V}{4(\alpha^H)^2 N}}\right)^T + \sqrt{\frac{l^{mx}(T)V\zeta(\alpha^H)^2}{N}}.$$

The remainder of this proof follows along the lines of [28, Theorem 6.1]. □

### 7.5 Discounted Safe LQR

We begin with characterizing global optimality for the nonconvex $J_\mu^H(K)$ over $K \in \mathcal{K}^\gamma$ by providing the gradient dominance condition, similar to [2].

**Lemma 7.21** (Policy Gradient)**.** *For the cost $J_\mu^H(K)$, with $K \in \mathcal{K}^\gamma$, defined above in* (127)*, we have the following gradient*

$$\nabla J_\mu^H(K) = 2E_K^\gamma \Sigma_K^\gamma, \tag{119}$$

*where*

$$E_K^\gamma = (R + \gamma B^\top P_K^\gamma B)K + \gamma B^\top P_K^\gamma A, \tag{120}$$

$$\Sigma_K^\gamma = \mathbb{E}_{x_0}\left[\sum_{t=0}^{\infty} \gamma^t x_t x_t^\top\right]. \tag{121}$$

*Proof of Lemma 7.21.* From the cost to go, $V_K(x_0)$'s definition in (127), we can write the following expansion

$$V_K(x_0) = x_0^\top (Q + K^\top R K) x_0 + \gamma V_K(A_K x_0)$$

Applying the gradient operator $\nabla_K$ to both sides, we get

$$\nabla_K V_K(x_0) = 2RKx_0x_0^\top + 2\gamma B^\top P_K^\gamma A_K x_0 x_0^\top + \gamma \nabla_K V_K(x_1)|_{x_1 = A_K x_0}$$

Similarly, we can obtain the following gradient expression for the next state $x_1 = A_K x_0$, as

$$\nabla_K V_K(x_1) = 2RKx_1x_1^\top + 2\gamma B^\top P_K^\gamma A_K x_1 x_1^\top + \gamma \nabla_K V_K(x_2)|_{x_2 = A_K x_1}$$

Thus, iteratively substituting the expressions for $\nabla_K V(x_t, K)$, we get

$$\nabla_K V_K(x_0) = 2(RK + \gamma B^\top P_K^\gamma A_K) \sum_{t=0}^{\infty} \gamma^t x_t x_t^\top$$

Therefore, we can finally obtain

$$\begin{aligned}
\nabla J_\mu^H(K) =& \mathbb{E}_{x_0}\left[\nabla_K V_K(x_0)\right] \\
=& 2\left((R + \gamma B^\top P_K^\gamma B)K + \gamma B^\top P_K^\gamma A\right)\Sigma_K^\gamma \\
=& 2E_K^\gamma \Sigma_K^\gamma.
\end{aligned}$$

□

With the gradient of the deterministic constrained LQR cost evaluated, we now prove a helper Lemma.

**Lemma 7.22** (Helper Lemma)**.** *Suppose that for parameters $K$ and $K'$ with , we have the state-action trajectories $(x_t, u_t)$ and $(x'_t, u'_t)$ respectively, and the same initial condition as $x_0 = x'_0 = x$, then we have*

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

*where the advantage of the policy parameter $K'$ over the parameter $K$ i.e. $A_K(x, K'x)$, with $\Delta K := K' - K$, can be written as*

$$\begin{aligned}
A_K(x, K'x) =& x^\top (\Delta K)^\top (R + \gamma B^\top P_K^\gamma B)\Delta K x \\
&+ 2x^\top (\Delta K)^\top E_K^\gamma x.
\end{aligned}$$

*Proof of Lemma 7.22.* We define the per step cost, with $u'_t = K'x'_t$, as $c'_t := c(x'_t, u'_t) = x'^{\top}_t Q x'_t + u'^{\top}_t R u'_t$. Then we can establish the following $V_{K'}(x) - V_K(x)$

$$\begin{aligned}
&= \sum_{t=0}^{\infty} \gamma^t c'_t - V_K(x), \\
&= \sum_{t=0}^{\infty} \gamma^t (c'_t - V_K(x'_t) + V_K(x'_t)) - V_K(x), \\
&= \sum_{t=0}^{\infty} \gamma^t (c'_t - V_K(x'_t)) + \gamma^t V_K(x'_t) - V_K(x'_0), \\
&= \sum_{t=0}^{\infty} \gamma^t (c'_t + \gamma V_K(x'_{t+1}) - V_K(x'_t)), \\
&= \sum_{t=0}^{\infty} \gamma^t A_K(x'_t, u'_t).
\end{aligned}$$

Second, recall that $A_K = A + BK$. With $u' = K'x$, we know that $Q_K(x, u') = x^{\top}(X + K'^{\top} R K')x + \gamma V_K((A + BK')x)$. Then, we can write

$$\begin{aligned}
A_K(x, K'x) &= Q_K(x, K'x) - V_K(x) \\
&= x^{\top}(Q + K'^{\top} R K')x + \gamma V_K(A_{K'}x) - V_K(x)
\end{aligned}$$

At this point, recall that $V_K(A_{K'}x) = x^{\top} A_{K'}^{\top} P_K^{\gamma} A_{K'} x$. Therefore, we can get

$$\begin{aligned}
A_K(x, K'x) &= x^{\top}(Q + K'^{\top} R K')x + \gamma x^{\top} A_{K'}^{\top} P_K^{\gamma} A_{K'} x - V_K(x) \\
&= x^{\top}(Q + (K' - K + K)^{\top} R (K' - K + K))x + \\
&\quad \gamma x^{\top}(A_{K'} + B(-K + K))^{\top} P_K^{\gamma}(A_{K'} + B(-K + K))x - V_K(x) \\
&= x^{\top}(Q + (K + \Delta K)^{\top} R (K + \Delta K))x + \\
&\quad \gamma x^{\top}(A + BK' + B(-K + K))^{\top} P_K^{\gamma}(A + BK' + B(-K + K))x - V_K(x), \\
&= x^{\top}(Q + K^{\top} R K + 2\Delta K^T R K + \Delta K^{\top} R \Delta K)x - V_K(x) + \\
&\quad \gamma x^{\top}(A + BK + B(\Delta K))^{\top} P_K^{\gamma}(A + BK + B(\Delta K))x
\end{aligned}$$

Or, it can also be written as

$$\begin{aligned}
A_K(x, K'x) &= x^{\top}(Q + K^{\top} R K)x + 2x^{\top} \Delta K^T R K x + x^{\top} \Delta K^{\top} R \Delta K x - V_K(x) \\
&\quad + \gamma x^{\top}(A + BK + B(\Delta K))^{\top} P_K^{\gamma}(A + BK + B(\Delta K))x, \\
&= x^{\top}(Q + K^{\top} R K)x + 2x^{\top} \Delta K^T R K x \\
&\quad + \gamma x^{\top}(A_K + B(\Delta K))^{\top} P_K^{\gamma}(A_K + B(\Delta K))x - V_K(x) + x^{\top} \Delta K^{\top} R \Delta K x, \\
&= x^{\top}(Q + K^{\top} R K)x + 2x^{\top} \Delta K^T R K x + \gamma x^{\top} A_K^{\top} P_K^{\gamma} A_K^{\top} x + x^{\top} \Delta K^{\top} R \Delta K x \\
&\quad + 2\gamma x^{\top} A_K^{\top} P_K^{\gamma} B(\Delta K)x + \gamma x^{\top} \Delta K^{\top} B^{\top} P_K^{\gamma} B \Delta K x - V_K(x)
\end{aligned}$$

Since $V_K(x) = x^{\top}(Q + K^{\top} R K)x + \gamma x^{\top} A_K^{\top} P_K^{\gamma} A_K x$, we can write

$$\begin{aligned}
A_K(x, K'x) &= 2x^{\top} \Delta K^T R K x + x^{\top} \Delta K^{\top} R \Delta K x + 2\gamma x^{\top} A_K^{\top} P_K^{\gamma} B(\Delta K)x \\
&\quad + \gamma x^{\top} \Delta K^{\top} B^{\top} P_K^{\gamma} B \Delta K x \\
&= \gamma x^{\top} \Delta K^{\top} B^{\top} P_K^{\gamma} B \Delta K x + x^{\top} \Delta K^{\top} R \Delta K x + 2\gamma x^{\top} \Delta K^{\top} B^{\top} P_K^{\gamma} A_K x \\
&\quad + 2x^{\top} \Delta K^T R K x \\
&= x^{\top}(\Delta K)^{\top}(R + \gamma B^{\top} P_K^{\gamma} B)(\Delta K)x + 2x^{\top}(\Delta K)^{\top}(R K + \gamma B^{\top} P_K^{\gamma} A_K)x \\
&= x^{\top}(\Delta K)^{\top}(R + \gamma B^{\top} P_K^{\gamma} B)(\Delta K)x + 2x^{\top}(\Delta K)^{\top} E_K^{\gamma} x,
\end{aligned}$$

completing the second claim. □

We are now ready to provide the gradient dominance result of the safe deterministic constrained LQR as follows.

**Lemma 7.23** (Gradient Dominance)**.** *Let $K^*$ denote the unconstrained deterministic discounted LQR optimum, and let $J_\mu^* := J_\mu^H(K^*)$. Then, with $K \in \mathcal{K}^\gamma$,*

$$J_\mu^H(K) - J_\mu^* \leq \alpha_\gamma \left\| \nabla_K J_\mu^H(K) \right\|^2, \tag{122}$$

*where*

$$\alpha_\gamma = \frac{\|\Sigma_{K^*}^\gamma\|}{4\sigma_{\min}(R)\sigma_{min}(\Sigma_K^\gamma)^2}.$$

*Consequently, since the constrained deterministic optimum satisfies $J_\mu^{H^*} \geq J_\mu^*$, we also have*

$$J_\mu^H(K) - J_\mu^{H^*} \leq \alpha_\gamma \left\| \nabla_K J_\mu^H(K) \right\|^2. \tag{123}$$

*Proof of Lemma 7.23.* Recall from Lemma 7.22 that

$$\begin{aligned} A_K(x, K'x) &= x^\top (\Delta K)^\top (R + \gamma B^\top P_K^\gamma B)(\Delta K)x + 2x^\top (\Delta K)^\top E_K^\gamma x, \\ &= \text{Tr}(x^\top (\Delta K)^\top (R + \gamma B^\top P_K^\gamma B)(\Delta K)x) + 2\text{Tr}(x^\top (\Delta K)^\top E_K^\gamma x). \end{aligned}$$

Then, using cyclic properties of trace operator, we can write

$$A_K(x, K'x) = \text{Tr}(xx^\top (\Delta K)^\top R_B^P (\Delta K)) + 2\text{Tr}(xx^\top (\Delta K)^\top E_K^\gamma),$$

where $R_B^P := R + \gamma B^\top P_K^\gamma B$. After some algebraic manipulations, we can obtain

$$\begin{aligned} A_K(x, K'x) = \text{Tr}(xx^\top (\Delta K + R_B^{P^{-1}} E_K^\gamma)^\top R_B^P (\Delta K + R_B^{P^{-1}} E_K^\gamma)) \\ - \text{Tr}(xx^\top (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma) \end{aligned}$$

It trivially follows that

$$A_K(x, K'x) \geq -\text{Tr}\left( xx^\top (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma \right),$$

with equality at $K' = K - R_B^{P^{-1}} E_K^\gamma$. Now recall, that $V(K^*) - V(K) = \sum_{t=0}^\infty \gamma^t A_K(x_t^*, u_t^*)$ from Lemma 7.22. Then, we can get

$$\begin{aligned} J_\mu^H(K) - J_\mu^H(K^*) &= -\mathbb{E}\left[ \sum_{t=0}^\infty \gamma^t A_K(x_t^*, u_t^*) \right] \\ &\leq \mathbb{E}\left[ \sum_{t=0}^\infty \gamma^t \text{Tr}\left( x_t^* x_t^{*\top} (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma \right) \right] \\ &\leq \text{Tr}\left( \mathbb{E} \sum_{t=0}^\infty \gamma^t x_t^* x_t^{*\top} (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma \right) \\ &\leq \text{Tr}\left( \Sigma_{K^*}^\gamma (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma \right) \\ &\leq \|\Sigma_{K^*}^\gamma\| \, \text{Tr}\left( (E_K^\gamma)^\top R_B^{P^{-1}} E_K^\gamma \right). \end{aligned}$$

Using cyclic property of the trace operator, we can further write

$$\begin{aligned} J_\mu^H(K) - J_\mu^H(K^*) &\leq \|\Sigma_{K^*}^\gamma\| \, \text{Tr}\left( R_B^{P^{-1}} E_K^\gamma (E_K^\gamma)^\top \right) \\ &\leq \|\Sigma_{K^*}^\gamma\| \left\| R_B^{P^{-1}} \right\| \text{Tr}\left( E_K^\gamma (E_K^\gamma)^\top \right) \end{aligned}$$

With the property $\left\| M^{-1} \right\|_2 = \frac{1}{\sigma_{min}(M)}$ for any nontrivial matrix $M$ with $\sigma_{min}(M) \neq 0$, we can write

$$\begin{aligned} J_\mu^H(K) - J_\mu^H(K^*) &\leq \|\Sigma_{K^*}^\gamma\| \, \text{Tr}\left( R_B^{P^{-1}} E_K^\gamma (E_K^\gamma)^\top \right) \\ &\leq \|\Sigma_{K^*}^\gamma\| \left\| R_B^{P^{-1}} \right\| \text{Tr}\left( E_K^\gamma (E_K^\gamma)^\top \right) \\ &= \frac{\|\Sigma_{K^*}^\gamma\|}{\sigma_{min}(R_B^P)} \text{Tr}\left( E_K^\gamma (E_K^\gamma)^\top \right). \end{aligned}$$

In our case, given that $R_B^P$ is a positive definite matrix, we know that $\sigma_{min}(R_B^P) > 0$.

Since $\nabla_K J_\mu^H(K) = 2E_K^\gamma \Sigma_K^\gamma$ from Lemma 7.21, we can write $E_K^\gamma = \frac{1}{2}(\nabla_K J_\mu^H(K))(\Sigma_K^\gamma)^{-1}$, given $(\Sigma_K^\gamma)^{-1}$ exists. Furthermore, since we know that $\sigma_{min}(R_B^P) = \sigma_{min}(R + \gamma B^\top P_K^\gamma B) \geq \sigma_{min}(R)$ for the positive definite matrix $R$, we can obtain

$$J_\mu^H(K) - J_\mu^H(K^*) \leq \frac{\|\Sigma_{K^*}^\gamma\|}{\sigma_{\min}(R)} \times \text{Tr}\left(\frac{1}{4}(\nabla_K J_\mu^H(K))(\Sigma_K^\gamma)^{-1}{(\Sigma_K^\gamma)^{-1}}^\top(\nabla_K J_\mu^H(K))^\top\right)$$

Using the cyclic property of trace operator again, we get

$$J_\mu^H(K) - J_\mu^H(K^*) \leq \frac{\|\Sigma_{K^*}^\gamma\|}{4\sigma_{\min}(R)} \times \text{Tr}\left((\Sigma_K^\gamma)^{-1}{(\Sigma_K^\gamma)^{-1}}^\top(\nabla_K J_\mu^H(K))^\top \nabla_K J_\mu^H(K)\right)$$

Taking out the matrices $(\Sigma_K^\gamma)^{-1}{(\Sigma_K^\gamma)^{-1}}^\top$ outside the trace and then applying the equality $\left\|M^{-1}\right\|_2 = \frac{1}{\sigma_{min}(M)}$, we get

$$J_\mu^H(K) - J_\mu^H(K^*) \leq \frac{\|\Sigma_{K^*}^\gamma\|}{4\sigma_{\min}(R)\sigma_{min}(\Sigma_K^\gamma)^2} \times \text{Tr}\left((\nabla_K J_\mu^H(K))^\top \nabla_K J_\mu^H(K)\right).$$

Finally, we obtain the desired gradient dominance condition with respect to the unconstrained deterministic discounted LQR optimum:

$$J_\mu^H(K) - J_\mu^* \leq \alpha_\gamma \left\|\nabla_K J_\mu^H(K)\right\|^2,$$

where $J_\mu^* = J_\mu^H(K^*)$, $K^*$ denotes the unconstrained deterministic discounted LQR optimum, and

$$\alpha_\gamma = \frac{\|\Sigma_{K^*}^\gamma\|}{4\sigma_{\min}(R)\sigma_{min}(\Sigma_K^\gamma)^2}.$$

Since the constrained deterministic optimum satisfies $J_\mu^{H^*} \geq J_\mu^*$, it follows that

$$J_\mu^H(K) - J_\mu^{H^*} \leq \alpha_\gamma \left\|\nabla_K J_\mu^H(K)\right\|^2.$$

□

We have now proven the gradient dominance characterizing global optimality for the deterministic discounted LQR under hard constraints. We now move on to proving helper lemmas for stochastic LQR costs over safe truncated policies. First, we begin with relationship between the discounted stochastic constrained LQR cost and the deterministic constrained LQR.

**Lemma 7.24** (Cost Difference between Stochastic and Deterministic Safe LQR)**.** *Suppose that Assumptions 2.1, 2.3, and 2.5 hold. Then the cost $J_\pi^H(K)$ satisfies the property*

$$J_\pi^H(K) = J_\mu^H(K) + \Delta_\gamma(K),$$

*where*

$$\alpha_1^\gamma(K) := \sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{\{\eta_i^H\},x_0}\left[{\eta_i^H}^\top R \eta_i^H\right] < \infty, \tag{124}$$

$$\alpha_2^\gamma(K) := \sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{\{\eta_i^H\},x_0}\left[(\mathcal{C}_i \bar{\eta}_i^H)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\right] < \infty, \tag{125}$$

$$\Delta_\gamma(K) := \alpha_1^\gamma(K) + \alpha_2^\gamma(K). \tag{126}$$

*Proof of Lemma 7.24.* Recall that $u_t = \mu_K^H(x_t) + \eta_t^H$ and safe $\mu_K^H(x_t) = Kx_t$. Then, $J_\pi^H(K)$ can be written as

$$J_\pi^H(K) = \mathbb{E}_{\{\eta_i^H\},x_0} \sum_{i=0}^{\infty} \gamma^i \left(x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R(Kx_i + \eta_i^H)\right).$$

Since each of the summand is nonnegative, we can apply Tonelli's theorem and interchange the expectation and the infinite sum operators as

$$J_\pi^H(K) = \sum_{i=0}^{\infty} \mathbb{E}_{\{\eta_i^H\},x_0}\left[\gamma^i \left(x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R(Kx_i + \eta_i^H)\right)\right]$$

For simplicity, we denote $\mathbb{E}_{\{\eta_i^H\},x_0}$ as $\mathbb{E}$ with expectation taken over appropriate random variables. Then, we get

$$\begin{aligned} J_\pi^H(K) &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\left[x_i^\top Q x_i + (Kx_i + \eta_i^H)^\top R(Kx_i + \eta_i^H)\right], \\ &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[x_i^\top (Q + K^\top R K) x_i + 2(Kx_i)^\top R\eta_i^H + {\eta_i^H}^\top R\eta_i^H\big], \\ &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\left[x_i^\top Q_K^R x_i + 2(Kx_i)^\top R\eta_i^H + {\eta_i^H}^\top R\eta_i^H\right]. \end{aligned}$$

Then, note from $x_{i+1} = Ax_i + Bu_i$, $u_i = \mu_K^H(x_i) + \eta_i^H$ and $\mu_K^H(x_i) = Kx_i$, we can write

$$x_i = A_K^i x_0 + \mathcal{C}_i \bar{\eta}_i^H.$$

Thus, $J_\pi^H(K)$ can be written as

$$\begin{aligned} J_\pi^H(K) =& \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[(A_K^i x_0 + \mathcal{C}_i\bar{\eta}_i^H)^\top Q_K^R (A_K^i x_0 + \mathcal{C}_i\bar{\eta}_i^H) \\ &\qquad + 2(KA_K^i x_0 + K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H + {\eta_i^H}^\top R\eta_i^H\big] \\ =& \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[(A_K^i x_0)^\top Q_K^R A_K^i x_0 + 2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H \\ &\qquad + {\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H + 2(KA_K^i x_0)^\top R\eta_i^H \\ &\qquad + 2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H + {\eta_i^H}^\top R\eta_i^H\big] \\ =& \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[(A_K^i x_0)^\top Q_K^R A_K^i x_0\big] \\ &+ \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H + {\bar{\eta}_i^H}^\top \mathcal{C}_i^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H \\ &\qquad + 2(KA_K^i x_0)^\top R\eta_i^H + 2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H + {\eta_i^H}^\top R\eta_i^H\big]. \end{aligned}$$

Recall that with a safe deterministic policy $\mu_K^H(x) = Kx$, we define the corresponding cost as

$$J_\mu^H(K) = \mathbb{E}_{x_0} V_K(x_0), \tag{127}$$

where $V_K(x_0) = \sum_{i=0}^{\infty} \gamma^i \left(x_i^T Q x_i + u_i^T R u_i\right) = x_0^\top P_K^\gamma x_0$, and $P_K^\gamma$ satisfies the discrete-time Lyapunov equation

$$P_K^\gamma = Q_K^R + \gamma (A_K)^\top P_K^\gamma A_K. \tag{128}$$

Therefore, from (127), we know that

$$J_\mu^H(K) = \mathbb{E}_{x_0} \sum_{i=0}^{\infty} \gamma^i \left(x_i^\top Q x_i + \mu_K^H(x_i)^\top R \mu_K^H(x_i)\right).$$

Since $\mu_K^H(x_i) = Kx_i$, we get $x_i = A_K^i x_0$. Therefore, by Tonelli's theorem, we can interchange the expectation and the infinite sum operators and write

$$\begin{aligned} J_\mu^H(K) &= \mathbb{E}_{x_0} \sum_{i=0}^{\infty} \gamma^i (A_K^i x_0)^\top Q_K^R (A_K^i x_0), \\ &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{x_0}\left[(A_K^i x_0)^\top Q_K^R (A_K^i x_0)\right]. \end{aligned}$$

Therefore, we can write

$$J_\pi^H(K) = J_\mu^H(K) + \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H + 2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H + 2(KA_K^i x_0)^\top R\eta_i^H + \bar{\eta}_i^{H^\top}\mathcal{C}_i^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H + \eta_i^{H^\top} R\eta_i^H\big].$$

Now, we define the following terms from the above cost expressions as

$$T_1^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\Big], \tag{129}$$

$$T_2^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[2(KA_K^i x_0)^\top R\eta_i^H\Big], \tag{130}$$

$$T_3^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H\Big], \tag{131}$$

$$T_4^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[\eta_i^{H^\top} R\eta_i^H\Big], \tag{132}$$

$$T_5^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[\bar{\eta}_i^{H^\top}\mathcal{C}_i^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H\Big]. \tag{133}$$

Therefore, we can write

$$J_\pi^H(K) = J_\mu^H(K) + T_1^\gamma + T_2^\gamma + T_3^\gamma + T_4^\gamma + T_5^\gamma. \tag{134}$$

Consider $T_1^\gamma, T_2^\gamma$ and $T_3^\gamma$. With martingale-difference arguments used in Lemma 7.11, we know that

$$\mathbb{E}[2(A_K^i x_0)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H] = 0, \tag{135}$$

$$\mathbb{E}[2(KA_K^i x_0)^\top R\eta_i^H] = 0, \tag{136}$$

$$\mathbb{E}[2(K\mathcal{C}_i\bar{\eta}_i^H)^\top R\eta_i^H] = 0. \tag{137}$$

Multiplying these zero expectations by $\gamma^i$ preserves the identities and yields

$$T_1^\gamma = T_2^\gamma = T_3^\gamma = 0. \tag{138}$$

We now analyze $T_4^\gamma$ and $T_5^\gamma$. For $T_4^\gamma$, using Lemma 7.6, we obtain

$$\begin{aligned} T_4^\gamma &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[\eta_i^{H^\top} R\eta_i^H\Big] \\ &\le \lambda_{\max}(R) \sum_{i=0}^{\infty} \gamma^i \operatorname{Tr}\left(\mathbb{E}\left[\eta_i^H \eta_i^{H^\top}\right]\right) \\ &\le \lambda_{\max}(R) \sum_{i=0}^{\infty} \gamma^i \operatorname{Tr}\left(\sigma^2 I_m\right) \\ &= \frac{\lambda_{\max}(R) m\sigma^2}{1-\gamma} < \infty. \end{aligned} \tag{139}$$

Recall from (124) that $\alpha_1^\gamma(K) = T_4^\gamma$, then we also conclude

$$\alpha_1^\gamma(K) < \infty.$$

Further, recall from (133),

$$T_5^\gamma = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\Big[\bar{\eta}_i^{H^\top}\mathcal{C}_i^\top Q_K^R \mathcal{C}_i\bar{\eta}_i^H\Big].$$

| **Indices** | $t=0$ | $t=1$ | $t=2$ |
|---|---|---|---|
| $i=1$ | $\gamma B^\top Q_K^R B \Sigma_0^H(K)$ | | |
| $i=2$ | $\gamma^2 B^\top A_K^\top Q_K^R A_K B \Sigma_0^H(K)$ | $\gamma^2 B^\top Q_K^R B \Sigma_1^H(K)$ | |
| $i=3$ | $\gamma^3 B^\top (A_K^2)^\top Q_K^R A_K^2 B \Sigma_0^H(K)$ | $\gamma^3 B^\top A_K^\top Q_K^R A_K B \Sigma_1^H(K)$ | $\gamma^3 B^\top Q_K^R B \Sigma_2^H(K)$ |

Table 3: Expansion of dual sum in (141)

We invoke the properties of the trace operator and write

$$\begin{aligned} T_5^\gamma &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\left[\operatorname{Tr}\left(\bar{\eta}_i^{H^\top} \mathcal{C}_i^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\right)\right] \\ &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}\left[\operatorname{Tr}\left(Q_K^R \mathcal{C}_i \bar{\eta}_i^H \bar{\eta}_i^{H^\top} \mathcal{C}_i^\top\right)\right] \\ &= \sum_{i=0}^{\infty} \gamma^i \operatorname{Tr}\left(Q_K^R \mathcal{C}_i \mathbb{E}\left[\bar{\eta}_i^H \bar{\eta}_i^{H^\top}\right] \mathcal{C}_i^\top\right). \end{aligned} \tag{140}$$

For notational clarity, let

$$\Sigma_i^H(K) := \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right].$$

From Lemma 7.7, we write $T_5^\gamma$ as

$$T_5^\gamma = \sum_{i=1}^{\infty} \gamma^i \operatorname{Tr}\left(Q_K^R \sum_{j=0}^{i-1} A_K^j B \Sigma_{i-1-j}^H(K) B^\top A_K^{j^\top}\right).$$

Using linearity and the cyclic property of the trace operator, we can further write

$$\begin{aligned} T_5^\gamma &= \sum_{i=1}^{\infty} \gamma^i \sum_{j=0}^{i-1} \operatorname{Tr}\left(Q_K^R A_K^j B \Sigma_{i-1-j}^H(K) B^\top A_K^{j^\top}\right) \\ &= \sum_{i=1}^{\infty} \gamma^i \sum_{j=0}^{i-1} \operatorname{Tr}\left(B^\top A_K^{j^\top} Q_K^R A_K^j B \Sigma_{i-1-j}^H(K)\right). \end{aligned} \tag{141}$$

Now, we simplify the above dual sum. From Table 3, recall that the columns go over the index $t$ and the rows go over $i$. Then, for every $t \geq 0$, the summation across rows, that is from $i = 0$ to $i \to \infty$, yields

$$C_t := \gamma^{t+1} \sum_{j=0}^{\infty} \gamma^j \operatorname{Tr}\left(B^\top A_K^{j^\top} Q_K^R A_K^j B \Sigma_t^H(K)\right).$$

Furthermore, note that $T_5^\gamma$ can be written as the sum

$$\begin{aligned} T_5^\gamma &= \sum_{t=0}^{\infty} C_t \\ &= \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{j=0}^{\infty} \gamma^j \operatorname{Tr}\left(B^\top A_K^{j^\top} Q_K^R A_K^j B \Sigma_t^H(K)\right). \end{aligned} \tag{142}$$

For simplicity, we equivalently write $t = i$. Then we can write $T_5^\gamma$ as

$$T_5^\gamma = \gamma \sum_{i=0}^{\infty} \gamma^i \sum_{j=0}^{\infty} \gamma^j \operatorname{Tr}\left(B^\top A_K^{j^\top} Q_K^R A_K^j B \Sigma_i^H(K)\right). \tag{143}$$

Now we take the trace operator outside the internal infinite sum, as

$$\begin{aligned} T_5^\gamma &= \gamma \sum_{i=0}^{\infty} \gamma^i \operatorname{Tr}\left(\sum_{j=0}^{\infty} \gamma^j \left(B^\top A_K^{j^\top} Q_K^R A_K^j B\right) \Sigma_i^H(K)\right) \\ &= \gamma \sum_{i=0}^{\infty} \gamma^i \operatorname{Tr}\left(B^\top \left(\sum_{j=0}^{\infty} \gamma^j A_K^{j^\top} Q_K^R A_K^j\right) B \Sigma_i^H(K)\right). \end{aligned} \tag{144}$$

Note that the trace operator can be taken out only if the limit

$$\operatorname{Tr}\left(\lim_{M\to\infty}\sum_{j=0}^{M}\gamma^j\left(B^\top {A_K^j}^\top Q_K^R A_K^j B\right)\Sigma_i^H(K)\right)$$

exists. In what follows, we show that this limit indeed exists as $B^\top P_K^\gamma B\Sigma_i^H(K)$, and that such a swap of operators is justified. We now recall from (128) and $Q_K^R = Q + K^\top RK$ that

$$\gamma A_K^\top P_K^\gamma A_K - P_K^\gamma + Q_K^R = \mathbf{0}.$$

Equivalently,

$$\begin{aligned}
P_K^\gamma &= \gamma A_K^\top P_K^\gamma A_K + Q_K^R,\\
\gamma A_K^\top P_K^\gamma A_K &= \gamma^2 {A_K^2}^\top P_K^\gamma A_K^2 + \gamma A_K^\top Q_K^R A_K,\\
P_K^\gamma &= \gamma^2 {A_K^2}^\top P_K^\gamma A_K^2 + \gamma A_K^\top Q_K^R A_K + Q_K^R.
\end{aligned}$$

Repeating the same steps, we obtain

$$P_K^\gamma = \lim_{l\to\infty}\gamma^l {A_K^l}^\top P_K^\gamma A_K^l + \sum_{j=0}^{\infty}\gamma^j {A_K^j}^\top Q_K^R A_K^j.$$

Since $\sqrt{\gamma}A_K$ is a Schur matrix, we know that

$$\lim_{l\to\infty}\gamma^l {A_K^l}^\top P_K^\gamma A_K^l = \mathbf{0}.$$

Thus,

$$P_K^\gamma = \sum_{j=0}^{\infty}\gamma^j {A_K^j}^\top Q_K^R A_K^j.$$

We now shift our focus back on $T_5^\gamma$ and substitute the inner infinite sum with $P_K^\gamma$ as

$$T_5^\gamma = \gamma\sum_{i=0}^{\infty}\gamma^i \operatorname{Tr}\left(B^\top P_K^\gamma B\Sigma_i^H(K)\right). \tag{145}$$

Using (145), the trace inequality, and Lemma 7.6, we obtain

$$\begin{aligned}
T_5^\gamma &= \gamma\sum_{i=0}^{\infty}\gamma^i \operatorname{Tr}\left(B^\top P_K^\gamma B\Sigma_i^H(K)\right)\\
&\le \gamma\lambda_{\max}(B^\top P_K^\gamma B)\sum_{i=0}^{\infty}\gamma^i \operatorname{Tr}(\Sigma_i^H(K))\\
&\le \gamma\lambda_{\max}(B^\top P_K^\gamma B)\sum_{i=0}^{\infty}\gamma^i m\sigma^2\\
&= \frac{\gamma}{1-\gamma}\lambda_{\max}(B^\top P_K^\gamma B)m\sigma^2 < \infty.
\end{aligned}$$

Lastly, recall from (125) that $\alpha_2^\gamma(K) = T_5^\gamma$, then we have

$$\alpha_2^\gamma(K) < \infty.$$

Finally, from (134), we can write

$$J_\pi^H(K) = J_\mu^H(K) + \alpha_1^\gamma(K) + \alpha_2^\gamma(K) = J_\mu^H(K) + \Delta_\gamma(K). \tag{146}$$

□

With the proof of above helper Lemma, we are now ready to prove one of the key results.

*Proof of Theorem 4.1.* From Lemma 7.24, we know that

$$J_\pi^H(K) = J_\mu^H(K) + \alpha_1^\gamma(K) + \alpha_2^\gamma(K).$$

Taking the gradient operator $\nabla_K$ on both sides, we get

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \alpha_1^\gamma(K) + \nabla_K \alpha_2^\gamma(K).$$

Equivalently, with $\Delta_\gamma(K) := \alpha_1^\gamma(K) + \alpha_2^\gamma(K)$, we obtain

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta_\gamma(K).$$

□

We prove the following helper lemma, characterizing the smoothness and upper-bound of the term $\alpha_2^\gamma$ appearing in Theorem 4.1.

**Lemma 7.25.** *Suppose that Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then, the term $\alpha_2^\gamma(K)$ from Theorem 4.1 can be expressed as*

$$\alpha_2^\gamma(K) = \gamma \operatorname{Tr}\left(P_K^\gamma M_{\alpha_2^\gamma}(K)\right), \tag{147}$$

*where*

$$M_{\alpha_2^\gamma}(K) := B\Sigma_\gamma^H(K)B^\top, \qquad \Sigma_\gamma^H(K) := \sum_{i=0}^\infty \gamma^i \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right].$$

*Moreover, $M_{\alpha_2^\gamma}(K)$ is locally differentiable in $K$ with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$. Finally, for every $K \in \mathcal{K}_H^\gamma$,*

$$\alpha_2^\gamma(K) \le \frac{\gamma}{1-\gamma}\lambda_{\max}(B^\top P_K^\gamma B) m\sigma^2. \tag{148}$$

*Proof of Lemma 7.25.* Recall the term $\alpha_2^\gamma(K)$ from Theorem 4.1:

$$\alpha_2^\gamma(K) = \sum_{i=0}^\infty \gamma^i \mathbb{E}\left[(\mathcal{C}_i \bar{\eta}_i^H)^\top Q_K^R \mathcal{C}_i \bar{\eta}_i^H\right].$$

Then, from (133) in Lemma 7.24, we know that

$$\alpha_2^\gamma(K) = T_5^\gamma.$$

Using (145), we can write

$$\begin{aligned}
\alpha_2^\gamma(K) &= \gamma \sum_{i=0}^\infty \gamma^i \operatorname{Tr}\left(B^\top P_K^\gamma B \Sigma_i^H(K)\right) \\
&= \gamma \operatorname{Tr}\left(B^\top P_K^\gamma B \sum_{i=0}^\infty \gamma^i \Sigma_i^H(K)\right) \\
&= \gamma \operatorname{Tr}\left(P_K^\gamma B \Sigma_\gamma^H(K) B^\top\right) \\
&= \gamma \operatorname{Tr}\left(P_K^\gamma M_{\alpha_2^\gamma}(K)\right).
\end{aligned}$$

Moreover, by Lemma 7.6 with $\sigma_i = \sigma, i \ge 0$,

$$\begin{aligned}
\alpha_2^\gamma(K) &\le \gamma \lambda_{\max}(B^\top P_K^\gamma B) \sum_{i=0}^\infty \gamma^i \operatorname{Tr}\left(\mathbb{E}\left[\eta_i^H \eta_i^{H^\top}\right]\right) \\
&\le \gamma \lambda_{\max}(B^\top P_K^\gamma B) \sum_{i=0}^\infty \gamma^i m\sigma^2 \\
&= \frac{\gamma}{1-\gamma}\lambda_{\max}(B^\top P_K^\gamma B) m\sigma^2.
\end{aligned}$$

Finally, the truncated-noise law depends on $K$ through the deterministic nominal map $c_K$. By Assumption 2.6, this dependence is locally regular in $K$, and the discounted summability above provides the domination needed to differentiate the infinite sum term by term. Hence $M_{\alpha_2^\gamma}(K)$ is locally differentiable in $K$ with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$.

Lastly, since $P_K^\gamma$ is a $\mathcal{C}^\infty$ function over open subsets of $\mathcal{K}^\gamma$ from Lemma 7.8, its dependence on $K$ is locally smooth. The remaining dependence of $\alpha_2^\gamma(K)$ on $K$ enters through the truncated-noise law induced by $\pi_K^H$, equivalently through $\Sigma_\gamma^H(K)$. By Assumption 2.6, this policy-induced dependence is locally regular in $K$. The discounted summability and Lemma 7.6 justify differentiating the infinite series term by term. Hence, $\nabla_K \alpha_2^\gamma(K)$ exists and is locally Lipschitz on $\mathcal{K}_H^\gamma$. □

We now establish and prove local Lipschitzness of $\nabla_K \alpha_2^\gamma(K)$, which is another helper result for obtaining local Lipschitzness of discounted stochastic LQR cost defined over hard constraint satisfying stochastic policies.

**Lemma 7.26.** *Suppose that Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then, for stabilizing parameters $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, the gradient $\nabla_K \alpha_2^\gamma(K)$ is locally Lipschitz with Lipschitz constant $l_{\alpha_2^\gamma}(K)$ as*

$$\|\nabla_K \alpha_2^\gamma(K') - \nabla_K \alpha_2^\gamma(K)\| \le l_{\alpha_2^\gamma}(K)\,\|K' - K\|\,. \tag{149}$$

*Proof of Lemma 7.26.* From Lemma 7.25, we have

$$\alpha_2^\gamma(K) = \gamma \operatorname{Tr}\Big(P_K^\gamma M_{\alpha_2^\gamma}(K)\Big)\,.$$

Therefore, the full differential is

$$d\alpha_2^\gamma = \gamma \operatorname{Tr}\Big((dP_K^\gamma) M_{\alpha_2^\gamma}(K)\Big) + \gamma \operatorname{Tr}\Big(P_K^\gamma\, dM_{\alpha_2^\gamma}(K)\Big)\,. \tag{150}$$

We first consider the term $\gamma \operatorname{Tr}((dP_K^\gamma) M_{\alpha_2^\gamma}(K))$. Let $Y_K^\gamma$ be the solution of the discounted Lyapunov equation

$$Y_K^\gamma = \gamma A_K Y_K^\gamma A_K^\top + M_{\alpha_2^\gamma}(K). \tag{151}$$

Following the same discounted Lyapunov-adjoint calculation as in the deterministic LQR smoothness proof, the first term in (150) gives

$$\gamma \operatorname{Tr}\Big((dP_K^\gamma) M_{\alpha_2^\gamma}(K)\Big) = \big\langle 2\gamma E_K^\gamma (Y_K^\gamma)^\top, dK\big\rangle\,, \tag{152}$$

where

$$E_K^\gamma = (R + \gamma B^\top P_K^\gamma B)K + \gamma B^\top P_K^\gamma A.$$

For the second term in (150), define $G_{\alpha_2^\gamma}(K)$ as the unique matrix satisfying

$$\Big\langle G_{\alpha_2^\gamma}(K), dK\Big\rangle = \operatorname{Tr}\Big(P_K^\gamma D_K M_{\alpha_2^\gamma}(K)[dK]\Big)\,, \qquad \forall dK. \tag{153}$$

Such a matrix exists by the Riesz representation theorem under the Frobenius inner product. Combining the two terms yields

$$\nabla_K \alpha_2^\gamma(K) = 2\gamma E_K^\gamma (Y_K^\gamma)^\top + \gamma G_{\alpha_2^\gamma}(K). \tag{154}$$

By Assumption 2.6 and Lemma 7.25, the map $K \mapsto M_{\alpha_2^\gamma}(K)$ is locally differentiable with locally Lipschitz derivative on compact subsets of $\mathcal{K}_H^\gamma$. Moreover, $K \mapsto P_K^\gamma$, $K \mapsto E_K^\gamma$, and the discounted Lyapunov solution $K \mapsto Y_K^\gamma$ are locally smooth on the discounted stabilizing set. Hence both terms in (154) are locally Lipschitz. Therefore, for stabilizing $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, there exists $l_{\alpha_2^\gamma}(K) < \infty$ such that

$$\|\nabla_K \alpha_2^\gamma(K') - \nabla_K \alpha_2^\gamma(K)\| \le l_{\alpha_2^\gamma}(K)\,\|K' - K\|\,.$$

□

**Lemma 7.27** (Regularity of $\alpha_1^\gamma$)**.** *Suppose Assumptions 2.3, 2.5, and 2.6 hold. Then $\nabla_K \alpha_1^\gamma(K)$ exists and is locally Lipschitz on $\mathcal{K}_H^\gamma$, with Lipschitz constant $l_{\alpha_1^\gamma}(K)$ as:*

$$\|\nabla_K \alpha_1^\gamma(K') - \nabla_K \alpha_1^\gamma(K)\| \le l_{\alpha_1^\gamma}(K)\,\|K' - K\|\,. \tag{155}$$

*Proof.* Using

$$\eta_i^{H^\top} R \eta_i^H = \mathrm{Tr}\Big(R \eta_i^H \eta_i^{H^\top}\Big),$$

we can write

$$\alpha_1^\gamma(K) = \mathrm{Tr}\left(R \sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right]\right).$$

For each finite $i$, the map

$$K \mapsto \mathbb{E}_{\pi_K^H, x_0}\left[\eta_i^H \eta_i^{H^\top}\right]$$

is differentiable because the truncated policy depends smoothly on $K$ through $c_K$ by Assumption 2.6, and the closed-loop recursion depends smoothly on $K$ through $A_K = A + BK$. The discount factor, together with Lemma 7.6 and the compact-set derivative bounds in Assumption 2.6, provides a summable dominating sequence for the derivatives of the summands. Therefore, differentiation can be interchanged with the expectation and the infinite sum by dominated convergence. Hence $\nabla_K \alpha_1^\gamma(K)$ exists and is continuous. On any compact subset $\mathcal{C} \subset \mathcal{K}_H^\gamma$, the same bounds applied to second derivatives give

$$\sup_{K \in \mathcal{C}} \left\|\nabla_K^2 \alpha_1^\gamma(K)\right\| < \infty.$$

Thus, by the mean-value inequality, $\nabla_K \alpha_1^\gamma(K)$ is locally Lipschitz. Consequently, for $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$, there exists $l_{\alpha_1^\gamma}(K) < \infty$ such that

$$\left\|\nabla_K \alpha_1^\gamma(K') - \nabla_K \alpha_1^\gamma(K)\right\| \le l_{\alpha_1^\gamma}(K) \left\|K' - K\right\|.$$

□

**Lemma 7.28** (Regularity of $\Delta_\gamma$)**.** *Suppose Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then $\nabla_K \Delta_\gamma(K)$ exists and is locally Lipschitz on $\mathcal{K}_H^\gamma$. In particular, for stabilizing $K, K' \in \mathcal{K}_H^\gamma$ satisfying $\|K' - K\| \le \epsilon(K)$,*

$$\left\|\nabla_K \Delta_\gamma(K') - \nabla_K \Delta_\gamma(K)\right\| \le l_{\Delta_\gamma}(K) \left\|K' - K\right\|,$$

*where one may take $l_{\Delta_\gamma}(K) = l_{\alpha_1^\gamma}(K) + l_{\alpha_2^\gamma}(K)$.*

*Proof.* The claim follows from

$$\Delta_\gamma(K) = \alpha_1^\gamma(K) + \alpha_2^\gamma(K),$$

Lemma 7.26 and Lemma 7.27. □

With the discounted residual terms shown to be finite and locally regular, we are ready to prove the well-definedness of the stochastic constrained LQR cost and its gradient.

*Proof on Theorem 4.2.* From Lemma 7.24, we know that

$$J_\pi^H(K) = J_\mu^H(K) + \Delta_\gamma(K).$$

Given that $J_\mu^H(K)$ is well-defined for all $\mu_K^H(x) = Kx$, $K \in \mathcal{K}^\gamma$, and that $\Delta_\gamma(K) < \infty$, we can conclude that $J_\pi^H(K)$ is well-defined for all $K \in \mathcal{K}_H^\gamma$.
Now, we shift focus to $\nabla_K J_\pi^H(K)$. Given that $\nabla_K J_\mu^H(K)$ is well-defined from Lemma 7.21, and that $\nabla_K \Delta_\gamma(K)$ is well-defined from Lemma 7.28, we conclude that $\nabla_K J_\pi^H(K)$ is also well-defined. □

Before proving a key local Lipschitzness result, we recall a result from [27], as follows.

**Lemma 7.29** (Lemma 5, [27])**.** *For any stabilizing policy parameters $K, K' \in \mathcal{K}^\gamma$ satisfying the property $\|K' - K\| \le \epsilon(K)$, there exist constant $l(K)$ such that the gradient $\nabla_K J_\mu^H(K)$ satisfies the following*

$$\left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K)\right\| \le \bar{l}_\gamma(K) \left\|K' - K\right\|.$$

We are now ready to prove much anticipated local Lipschitzness result.

**Lemma 7.30.** *Suppose Assumptions 2.3 and 2.5 hold. Then, for any policy parameters $K, K' \in \mathcal{K}^\gamma$ satisfying $\|K' - K\| \leq \epsilon(K)$, there exist a Lipschitz constant $l_\gamma(K)$ such that the gradients follow*

$$\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| \leq l_\gamma(K) \left\|K' - K\right\|. \tag{156}$$

By proving Lipschitzness, we guarantee that small changes in the controller gain $K$ lead to bounded changes in the gradient, enabling us to utilize standard policy gradient results.

*Proof of Lemma 7.30.* From Lemma 7.29, we know that

$$\left\|\nabla_K J_\mu^H(K') - \nabla_K J_\mu^H(K)\right\| \leq \bar{l}_\gamma(K) \left\|K' - K\right\|.$$

Further, from Theorem 4.1, we know that

$$\nabla_K J_\pi^H(K) = \nabla_K J_\mu^H(K) + \nabla_K \Delta_\gamma(K).$$

So we can write

$$\left\|\nabla_K J_\pi^H(K') - \nabla_K \Delta_\gamma(K') - \nabla_K J_\pi^H(K) + \nabla_K \Delta_\gamma(K)\right\| \leq \bar{l}_\gamma(K) \left\|K' - K\right\|. \tag{157}$$

However, from the property of norms, we also know that

$$\begin{aligned}&\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| - \left\|\nabla_K \Delta_\gamma(K') - \nabla_K \Delta_\gamma(K)\right\| \\ &\leq \left\|\nabla_K J_\pi^H(K') - \nabla_K \Delta_\gamma(K') - \nabla_K J_\pi^H(K) + \nabla_K \Delta_\gamma(K)\right\|. \end{aligned}\tag{158}$$

Then, from Lemma 7.29, we can obtain

$$\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| \leq \bar{l}_\gamma(K) \left\|K' - K\right\| + \left\|\nabla_K \Delta_\gamma(K') - \nabla_K \Delta_\gamma(K)\right\|. \tag{159}$$

With Lemma 7.28, we can further write

$$\begin{aligned}\left\|\nabla_K J_\pi^H(K') - \nabla_K J_\pi^H(K)\right\| &\leq \bar{l}_\gamma(K) \left\|K' - K\right\| + l_{\Delta_\gamma}(K) \left\|K' - K\right\| \\ &= l_\gamma(K) \left\|K' - K\right\|, \end{aligned}\tag{160}$$

where

$$l_\gamma(K) = \bar{l}_\gamma(K) + l_{\Delta_\gamma}(K).$$

□

We are ready to prove weak gradient dominance result characterizing suboptimality introduced due to exploration in our model-free methods. We begin by proving a key result on the aforementioned suboptimality, as follows.

*Proof of Theorem 4.3.* By Lemma 7.24,

$$J_\pi^H(K) = J_\mu^H(K) + \Delta_\gamma(K),$$

where $\Delta_\gamma(K) \geq 0$. Therefore,

$$J_\pi^{H^*} = \min_{K \in \mathcal{K}_H^\gamma} \left(J_\mu^H(K) + \Delta_\gamma(K)\right) \geq \min_{K \in \mathcal{K}_H^\gamma} J_\mu^H(K) = J_\mu^{H^*}.$$

This proves the lower bound. For the upper bound, evaluate $J_\pi^H$ at $K_\mu^*$:

$$J_\pi^{H^*} \leq J_\pi^H(K_\mu^*) = J_\mu^H(K_\mu^*) + \Delta_\gamma(K_\mu^*) = J_\mu^{H^*} + \Delta_\gamma(K_\mu^*).$$

Thus,

$$0 \leq J_\pi^{H^*} - J_\mu^{H^*} \leq \Delta_\gamma(K_\mu^*) = \beta_\gamma^*.$$

□

Before the main weak gradient dominance result, we establish and prove another helper Lemma.

**Lemma 7.31** (Helper Result)**.** *Suppose that Assumptions 2.1, 2.3, 2.5, and 2.6 hold. Then,*

$$\begin{aligned}J_\pi^H(K) - J_\pi^{H^*} \leq &\alpha_\gamma \left\|\nabla_K J_\pi^H(K)\right\|^2 + \alpha_\gamma \left\|\nabla_K \Delta_\gamma(K)\right\|^2 \\ &+ 2\alpha_\gamma \left\|\nabla_K \Delta_\gamma(K)\right\| \left\|\nabla_K J_\pi^H(K)\right\| + \Delta_\gamma(K).\end{aligned}$$

*Proof of Lemma 7.31.* By Lemma 7.24, we note that

$$J_\pi^H(K) = J_\mu^H(K) + \Delta_\gamma(K),$$

and by Theorem 4.3, we note

$$0 \leq J_\pi^{H^*} - J_\mu^{H^*}.$$

Therefore, we can write

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} &= J_\mu^H(K) + \Delta_\gamma(K) - J_\pi^{H^*} \\ &= J_\mu^H(K) - J_\mu^{H^*} + \Delta_\gamma(K) + J_\mu^{H^*} - J_\pi^{H^*} \\ &\leq J_\mu^H(K) - J_\mu^{H^*} + \Delta_\gamma(K), \end{aligned}$$

Using Lemma 7.23,

$$J_\mu^H(K) - J_\mu^{H^*} \leq \alpha_\gamma \left\| \nabla_K J_\mu^H(K) \right\|^2.$$

From Theorem 4.1,

$$\nabla_K J_\mu^H(K) = \nabla_K J_\pi^H(K) - \nabla_K \Delta_\gamma(K).$$

Thus,

$$\left\| \nabla_K J_\mu^H(K) \right\| \leq \left\| \nabla_K J_\pi^H(K) \right\| + \left\| \nabla_K \Delta_\gamma(K) \right\|.$$

Combining these inequalities gives

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} &\leq \alpha_\gamma \left( \left\| \nabla_K J_\pi^H(K) \right\| + \left\| \nabla_K \Delta_\gamma(K) \right\| \right)^2 + \Delta_\gamma(K) \\ &\leq \alpha_\gamma \left\| \nabla_K J_\pi^H(K) \right\|^2 + \alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\|^2 \\ &\quad + 2\alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\| \left\| \nabla_K J_\pi^H(K) \right\| + \Delta_\gamma(K). \end{aligned}$$

□

We are now ready to prove a key result characterizing suboptimlaity in the discounted stochastic constrained LQR cost.

*Proof of Theorem 4.4.* Recall from Lemma 7.31 that

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} \leq & \alpha_\gamma \left\| \nabla_K J_\pi^H(K) \right\|^2 + \alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\|^2 \\ & + 2\alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\| \left\| \nabla_K J_\pi^H(K) \right\| + \Delta_\gamma(K). \end{aligned} \tag{161}$$

By Lemma 7.20, the policy gradient is bounded on $\mathcal{L}_{0,\gamma}$. Under the normalized convention $G_{\mathcal{L}_{0,\gamma}} = 1$, we have

$$\left\| \nabla_K J_\pi^H(K) \right\|^2 \leq \left\| \nabla_K J_\pi^H(K) \right\|. \tag{162}$$

Combining (161) and (162), we can write

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} \leq & \left( \alpha_\gamma + 2\alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\| \right) \left\| \nabla_K J_\pi^H(K) \right\| \\ & + \alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\|^2 + \Delta_\gamma(K). \end{aligned}$$

Taking suprema over $K \in \mathcal{L}_{0,\gamma}$, whose finiteness follows from Lemma 7.19, we obtain

$$\begin{aligned} J_\pi^H(K) - J_\pi^{H^*} \leq & \alpha_\gamma \left( 1 + 2 \sup_{K \in \mathcal{L}_{0,\gamma}} \left\{ \left\| \nabla_K \Delta_\gamma(K) \right\| \right\} \right) \left\| \nabla_K J_\pi^H(K) \right\| \\ & + \sup_{K \in \mathcal{L}_{0,\gamma}} \left\{ \Delta_\gamma(K) + \alpha_\gamma \left\| \nabla_K \Delta_\gamma(K) \right\|^2 \right\}. \end{aligned}$$

This gives the result with the definitions of $\alpha_\gamma^H$ and $\beta_\gamma^H$. □

We now prove the global convergence result for our discounted framework.

*Proof of Theorem 4.5.* We first define $\bar{J}_\pi^{H^*} := J_\pi^{H^*} + \beta_\gamma^H$. Then, following weak gradient dominance from Theorem 4.4, we have

$$\frac{1}{\alpha_\gamma^H} \max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\} \leq \left\|\nabla_K J_\pi^H(K_t)\right\|. \tag{163}$$

The cost difference $J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*}$ can also be written as

$$J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*} = J_\pi^H(K_{t+1}) - J_\pi^H(K_t) + J_\pi^H(K_t) - \bar{J}_\pi^{H^*}.$$

We focus on $J_\pi^H(K_{t+1}) - J_\pi^H(K_t)$ in the interim, and write, using the local Lipschitz property from Lemma 7.30 and second order Taylor's expansion, the following

$$J_\pi^H(K_{t+1}) - J_\pi^H(K_t) \leq \langle \nabla_K J_\pi^H(K_t), K_{t+1} - K_t \rangle + \frac{l_\gamma(K_t)}{2} \|K_{t+1} - K_t\|_F^2.$$

Thus, we write

$$J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*} \leq J_\pi^H(K_t) - \bar{J}_\pi^{H^*} + \langle \nabla_K J_\pi^H(K_t), K_{t+1} - K_t \rangle + \frac{l_\gamma(K_t)}{2} \|K_{t+1} - K_t\|_F^2.$$

However, with the update rule $K_{t+1} = K_t - \zeta_t \widehat{\nabla_K J_\pi^H(K_t)}$, we can write

$$\begin{aligned} J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*} \leq & J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \zeta_t \langle \nabla_K J_\pi^H(K_t), \widehat{\nabla_K J_\pi^H(K_t)} \rangle \\ & + \frac{l_\gamma(K_t)}{2} \zeta_t^2 \left\| \widehat{\nabla_K J_\pi^H(K_t)} \right\|_F^2. \end{aligned}$$

Let $\mathcal{F}_t = \sigma(K_l, \zeta_l; 0 \leq l \leq t)$ denote the $\sigma-$algebra generated by all randomness in the system up to time step $t$, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. Thence, we take expectation on the foregoing as

$$\begin{aligned} & \mathbb{E}_t \left[ J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*} \right] \\ & \leq \mathbb{E}_t \left[ J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \zeta_t \langle \nabla_K J_\pi^H(K_t), \widehat{\nabla_K J_\pi^H(K_t)} \rangle + \frac{l_\gamma(K_t)}{2} \zeta_t^2 \left\| \widehat{\nabla_K J_\pi^H(K_t)} \right\|_F^2 \right]. \end{aligned}$$

From Assumption 3.5, we know that

$$\mathbb{E}_t \left[ \widehat{\nabla_K J_\pi^H(K_t)} \right] = \nabla_K J_\pi^H(K_t),$$

and

$$\mathbb{E}_t \left[ \left\| \widehat{\nabla_K J_\pi^H(K_t)} - \nabla_K J_\pi^H(K_t) \right\|_F^2 \right] \leq \frac{V}{N}.$$

Therefore,

$$\begin{aligned} & \mathbb{E}_t \left[ \left\| \widehat{\nabla_K J_\pi^H(K_t)} \right\|_F^2 \right] \\ & = \mathbb{E}_t \left[ \left\| \nabla_K J_\pi^H(K_t) + \widehat{\nabla_K J_\pi^H(K_t)} - \nabla_K J_\pi^H(K_t) \right\|_F^2 \right] \\ & = \left\| \nabla_K J_\pi^H(K_t) \right\|_F^2 + \mathbb{E}_t \left[ \left\| \widehat{\nabla_K J_\pi^H(K_t)} - \nabla_K J_\pi^H(K_t) \right\|_F^2 \right] \\ & \leq \left\| \nabla_K J_\pi^H(K_t) \right\|_F^2 + \frac{V}{N}. \end{aligned}$$

Thus, we get

$$\begin{aligned} \mathbb{E}_t \left[ J_\pi^H(K_{t+1}) - \bar{J}_\pi^{H^*} \right] \leq & J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \zeta_t \left\| \nabla_K J_\pi^H(K_t) \right\|_F^2 \\ & + \frac{l_\gamma(K_t)}{2} \zeta_t^2 \left( \left\| \nabla_K J_\pi^H(K_t) \right\|_F^2 + \frac{V}{N} \right) \\ = & J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \zeta_t \left( 1 - \frac{l_\gamma(K_t)}{2} \zeta_t \right) \left\| \nabla_K J_\pi^H(K_t) \right\|_F^2 + \frac{l_\gamma(K_t)}{2} \zeta_t^2 \frac{V}{N}. \end{aligned}$$

Since $\zeta_t \leq 1/l_\gamma^{mx}(T)$, we have

$$1 - \frac{l_\gamma(K_t)}{2}\zeta_t \geq \frac{1}{2}.$$

Therefore, using (163), we get

$$\mathbb{E}\left[J_\pi^H(K_{t+1})\right] - \bar{J}_\pi^{H^*} \leq J_\pi^H(K_t) - \bar{J}_\pi^{H^*} - \frac{\zeta_t}{2(\alpha_\gamma^H)^2}\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}^2 + \frac{l_\gamma^{mx}(T)}{2}\zeta_t^2\frac{V}{N}. \tag{164}$$

Then, using Jensen's inequality on two convex functions, first on the square function and then on the $\max$ function, we get the following

$$\begin{aligned}\mathbb{E}\left[\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}^2\right] &\geq \mathbb{E}\left[\max\{0, J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\}\right]^2, \\ &\geq \max\{0, \mathbb{E}\left[J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\right]\}^2. \end{aligned} \tag{165}$$

Now, we establish the following key inequality based on (164) and (165), as

$$\rho_{t+1} \leq \rho_t - \frac{\zeta_t}{2(\alpha_\gamma^H)^2}\max\{0, \rho_t\}^2 + \frac{l_\gamma(K_t)}{2}\zeta_t^2\frac{V}{N}, \tag{166}$$

where $\rho_t := \mathbb{E}_t\left[J_\pi^H(K_t) - \bar{J}_\pi^{H^*}\right]$.
By substituting $L_2 = l_\gamma^{mx}(T)$, we follow the proof of Theorem 6.1 in [28] to obtain

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - \bar{J}_\pi^{H^*} \leq c_T\left(1 - \sqrt{\frac{\zeta^3 l_\gamma^{mx}(T)V}{4(\alpha_\gamma^H)^2 N}}\right)^T + \sqrt{\frac{l_\gamma^{mx}(T)V\zeta(\alpha_\gamma^H)^2}{N}},$$

or, from $\bar{J}_\pi^{H^*} = J_\pi^{H^*} + \beta_\gamma^H$, we can also write

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\pi^{H^*} \leq \beta_\gamma^H + c_T\left(1 - \sqrt{\frac{\zeta^3 l_\gamma^{mx}(T)V}{4(\alpha_\gamma^H)^2 N}}\right)^T + \sqrt{\frac{l_\gamma^{mx}(T)V\zeta(\alpha_\gamma^H)^2}{N}}.$$

Finally, from Theorem 4.3, we can further write

$$\mathbb{E}\left[J_\pi^H(K_T)\right] - J_\mu^{H^*} \leq \beta_\gamma^H + \beta_\gamma^* + c_T\left(1 - \sqrt{\frac{\zeta^3 l_\gamma^{mx}(T)V}{4(\alpha_\gamma^H)^2 N}}\right)^T + \sqrt{\frac{l_\gamma^{mx}(T)V\zeta(\alpha_\gamma^H)^2}{N}}.$$

The remainder of this proof follows along the lines of the proof of Theorem 6.1 in [28]. □

### 7.6 Broader Impacts

This work provides a theoretical framework for globally convergent and optimal policy gradient algorithms to design hard constrained LQR control policies, and does not have any direct societal impact.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in our abstract and introduction accurately reflect the paper's scope and contributions. All the theoretical results are aligned with the claims made in the abstract and introduction.

   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations behind our theoretical contributions, based on underlying assumptions.

   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the assumptions behind the theoretical results in Sections 2, 3 and 4. The detailed proofs of all results are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and Appendix 7.2 specify the model, methods, training, evaluation procedure, and numerical calculations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This paper does not provide public code repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and evaluation details are given in Section 5 and Appendix 7.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports means and standard deviations over ten independent runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our numerical experiment requires no specialized hardware. We report simulation budgets, but omit detailed hardware specifications and execution times.

   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics, and ensured that our paper conforms to these regulations.

   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper is primarily theoretical, and does not have any immediate societal impact. We discuss this in Appendix.

    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not pose any such risk.

    Guidelines:
    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:
    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper is theoretical in contributions and does not release new assets.

    Guidelines:
    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper has theoretical contributions and neither involve crowd-sourcing nor human subjects.

    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper has theoretical contributions and neither involve crowd-sourcing nor human subjects.

    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We did not use LLMs in any of theoretical contributions of the paper.

    Guidelines:
    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.