Learning Optimal Controllers for Linear Systems With Multiplicative Noise via Policy Gradient

Benjamin Gravell, *Graduate Student Member, IEEE*, Peyman Mohajerin Esfahani, and Tyler Summers, *Member, IEEE*

***Abstract*—The linear quadratic regulator (LQR) problem has reemerged as an important theoretical benchmark for reinforcement learning-based control of complex dynamical systems with continuous state and action spaces. In contrast with nearly all recent work in this area, we consider multiplicative noise models, which are increasingly relevant because they explicitly incorporate inherent uncertainty and variation in the system dynamics and thereby improve robustness properties of the controller. Robustness is a critical and poorly understood issue in reinforcement learning; existing methods which do not account for uncertainty can converge to fragile policies or fail to converge at all. Additionally, intentional injection of multiplicative noise into learning algorithms can enhance robustness of policies, as observed in *ad hoc* work on domain randomization. Although policy gradient algorithms require optimization of a nonconvex cost function, we show that the multiplicative noise LQR cost has a special property called *gradient domination*, which is exploited to prove global convergence of policy gradient algorithms to the globally optimum control policy with polynomial dependence on problem parameters. Results are provided both in the model-known and model-unknown settings where samples of system trajectories are used to estimate policy gradients**

***Index Terms*—Gradient methods, noise, optimal control, reinforcement learning, stochastic systems, uncertain systems.**

I. Introduction

REINFORCEMENT learning-based control has recently achieved impressive successes in games [1] and simulators [2]. But these successes are significantly more challenging to translate to complex physical systems with continuous state and action spaces, safety constraints, and non-negligible operation and failure costs that demand data efficiency. An intense and growing research effort is creating a large array of models, algorithms, and heuristics for approaching the myriad of challenges arising from these systems. To complement a dominant trend of more computationally focused work, the canonical linear quadratic regulator (LQR) problem in control theory has reemerged as an important theoretical benchmark for learning-based control [3], [4]. Despite its long history, there remain fundamental open questions for LQR with unknown models, and a foundational understanding of learning in LQR problems can give insight into more challenging problems.

Almost all recent work on learning in LQR problems has utilized either deterministic or additive noise models [3]–[14], but, here, we consider *multiplicative noise models*. In control theory, multiplicative noise models have been studied almost as long as their deterministic and additive noise counterparts [15], [16], although this area is somewhat less developed and far less widely known. We believe the study of learning in LQR problems with multiplicative noise is important for three reasons. First, this class of models is much richer than deterministic or additive noise while still allowing exact solutions when models are known, which makes it a compelling additional benchmark [17]–[19]. Second, they explicitly incorporate model uncertainty and inherent stochasticity, thereby improving robustness properties of the controller. Robustness is a critical and poorly understood issue in reinforcement learning; existing methods that do not account for uncertainty can converge to fragile policies or fail to converge at all [18], [20], [21]. Additionally, intentional injection of multiplicative noise into learning algorithms is known to enhance robustness of policies from *ad hoc* work on domain randomization [22]. Third, in emerging difficult-to-model complex systems where learning-based control approaches are perhaps most promising, multiplicative noise models are increasingly relevant; examples include networked systems with noisy communication channels [23], [24], modern power networks with large penetration of intermittent renewables [25], [26], turbulent fluid flow [27], and neuronal brain networks [28].

A. Related Literature

Multiplicative noise LQR problems have been studied in control theory since the 1960s [15]. Since then, a line of research parallel to deterministic and additive noise has developed, including basic stability and stabilizability results [17], semidefinite programming formulations [29]–[31], robustness properties [16], [19], [32]–[34], and numerical algorithms [35].

Manuscript received April 23, 2020; accepted October 30, 2020. Date of publication November 10, 2020; date of current version November 4, 2021. This work was supported in part by the Army Research Office under Grant W911NF-17-1-0058 and in part by the United States Air Force Office of Scientific Research under Award FA2386-19-1-4073. Recommended by Associate Editor F. Pasqualetti. *(Corresponding author: Benjamin Gravell.)*

Benjamin Gravell and Tyler Summers are with the Control, Optimization, and Networks Lab., The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: bjgravell@gmail.com; tyler.summers@utdallas.edu).

Peyman Mohajerin Esfahani is with the Delft Center for Systems and Control, TU Delft, 2628 CD Delft, The Netherlands (e-mail: bjgravell@gmail.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2020.3037046

This line of research is less widely known perhaps because much of it studies continuous time systems, where the heavy machinery required to formalize stochastic differential equations is a barrier to entry for a broad audience. Multiplicative noise models are well-poised to offer data-driven model uncertainty representations and enhanced robustness in learning-based control algorithms and complex dynamical systems and processes. A related line of research that has seen recent activity is on learning optimal control of Markovian jump linear systems with unknown dynamics and noise distributions [36], [37], which, under certain assumptions, is a special case of the multiplicative noise system we analyze in this article.

In contrast to classical work on system identification and adaptive control, which has a strong focus on asymptotic results, more recent work has focused on nonasymptotic analysis using newly developed mathematical tools from statistics and machine learning. There remain fundamental open problems for learning in LQR problems, with several addressed only recently, including nonasymptotic sample complexity [4], [10], regret bounds [8], [11], [13], and algorithmic convergence [5]. Alternatives to reinforcement learning include other data-driven model-free optimal control schemes [38], [39] and those leveraging the behavioral framework [40], [41]. Subspace identification methods offer a model-based generalization to the output feedback setting [42].

B. Our Contributions

In Section II, we formulate the multiplicative noise LQR problem and motivate its study via a connection to robust stability. We then give several fundamental results for policy gradient algorithms on linear quadratic problems with multiplicative noise. Our main contributions are as follows, which can be viewed as a generalization of the recent results of Fazel *et al.* [5] for deterministic LQR to multiplicative noise LQR.

1) In Section III-A, we show that although the multiplicative noise LQR cost is generally nonconvex, it has a special property called *gradient domination*, which facilitates its optimization (Lemmas 3.1 and 3.3).
2) In particular, in Section IV, the gradient domination property is exploited to prove global convergence of three policy gradient algorithm variants (namely, exact gradient descent, "natural gradient descent," and Gauss–Newton/policy iteration) to the globally optimum control policy with a rate that depends polynomially on problem parameters (Theorems 4.1, 4.2, and 4.3).
3) Furthermore, in Section V, we show that a model-free policy gradient algorithm, where the gradient is estimated from trajectory data rather than computed from model parameters, also converges globally (with high probability) with an appropriate exploration scheme and sufficiently many samples (polynomial in problem data) (Theorem 5.1).

In comparison with the deterministic dynamics studied by [5], we make the following novel technical contributions.

1) We quantify the increase in computational burden of policy gradient methods due to the presence of multiplicative noise, which is evident from the bounds developed in Appendixes A and B. The noise acts to reduce the step size and thus convergence rate and increases the required number of samples and rollout length in the model-free setting.
2) A covariance dynamics operator $\mathcal{F}_K$ is established for multiplicative noise systems with a more complicated form than the deterministic case. This necessitated a more careful treatment and novel proof by induction and term matching argument in the proof of Lemma 7.4.
3) Several restrictions on the algorithmic parameters which are necessary for convergence, which were neglected by [5], are established and treated.
4) An important restriction on the support of the multiplicative noise distribution, which is naturally absent in [5], is established in the model-free setting.
5) A matrix Bernstein concentration inequality is stated explicitly and used to give explicit bounds on the algorithmic parameters in the model-free setting in terms of problem data.
6) We provide much more extensive numerical results and discussion, including an open-source code implementation and the use of backtracking line search.
7) When the multiplicative variances α_i, β_j are all zero, the assertions of Theorems 4.1, 4.2, 4.3, and 5.1 recover the same step sizes and convergence rates of the deterministic setting reported by [5].

Thus, policy gradient algorithms for the multiplicative noise LQR problem enjoy the same global convergence properties as deterministic LQR, while significantly enhancing the resulting controller's robustness to variations and inherent stochasticity in the system dynamics, as demonstrated by our numerical experiments in Section VI.

To the best of our knowledge, the present article is the first work to consider and obtain global convergence results using reinforcement learning algorithms for the multiplicative noise LQR problem. Our approach allows the explicit incorporation of a model uncertainty representation that significantly improves the robustness of the controller compared to deterministic and additive noise approaches.

II. Optimal Control of Linear Systems With Multiplicative Noise and Quadratic Costs

We consider the infinite-horizon linear quadratic regulator problem with multiplicative noise (LQRm)

$$\underset{\pi\in\Pi}{\text{minimize}} \quad C(\pi) := \underset{x_0,\{\delta_{ti}\},\{\gamma_{tj}\}}{\mathbb{E}} \sum_{t=0}^{\infty} (x_t^\intercal Q x_t + u_t^\intercal R u_t) \quad (1)$$

$$\text{subject to} \quad x_{t+1} = (A + \sum_{i=1}^{p} \delta_{ti} A_i) x_t + (B + \sum_{j=1}^{q} \gamma_{tj} B_j) u_t$$

where $x_t \in \mathbb{R}^n$ is the system state, $u_t \in \mathbb{R}^m$ is the control input, the initial state x_0 is distributed according to $\mathcal{P}_0$ with covariance $\Sigma_0 := \mathbb{E}_{x_0}[x_0 x_0^\intercal]$, $\Sigma_0 \succ 0$, and $Q \succ 0$ and $R \succ 0$. The dynamics are described by a dynamics matrix $A \in \mathbb{R}^{n\times n}$ and input matrix $B \in R^{n\times m}$ and incorporate multiplicative noise terms modeled by the i.i.d. (across time), zero-mean, mutually independent scalar random variables δ_{ti} and γ_{tj},

which have variances α_i and β_j, respectively. The matrices $A_i \in \mathbb{R}^{n\times n}$ and $B_i \in \mathbb{R}^{n\times m}$ specify how each scalar noise term affects the system dynamics and input matrices. Alternatively, suppose $\bar{A}$ and $\bar{B}$ are zero-mean random matrices with a joint covariance structure[1] over their entries governed by the covariance matrices $\Sigma_A := \mathbb{E}[\mathbf{vec}(\bar{A})\mathbf{vec}(\bar{A})^\intercal] \in \mathbb{R}^{n^2\times n^2}$ and $\Sigma_B := \mathbb{E}[\mathbf{vec}(\bar{B})\mathbf{vec}(\bar{B})^\intercal] \in \mathbb{R}^{nm\times nm}$. Then it suffices to take the variances α_i and β_j and matrices A_i and B_j as the eigenvalues and (reshaped) eigenvectors of Σ_A and Σ_B, respectively, after a projection onto a set of orthogonal real-valued vectors [43]. The goal is to determine a closed-loop state feedback policy π^* with $u_t = \pi^*(x_t)$ from a set Π of admissible policies which solves the optimization in (1).

We assume that the problem data A, B, α_i, A_i, β_j, and B_j permit existence and finiteness of the optimal value of the problem, in which case the system is called *mean-square stabilizable* and requires *mean-square stability* of the closed-loop system [17], [44]. The system in (1) is called *mean-square stable* if $\lim_{t\to\infty} \mathbb{E}_{x_0,\delta,\gamma}[x_t x_t^\intercal] = 0$ for any given initial covariance Σ_0, where, for brevity, we notate expectation with respect to the noises $\mathbb{E}_{\{\delta_{ti}\},\{\gamma_{tj}\}}$ as $\mathbb{E}_{\delta,\gamma}$. Mean-square stability is a form of robust stability, implying stability of the mean (i.e., $\lim_{t\to\infty} \mathbb{E}x_t = 0 \;\forall\, x_0$) as well as (in the absence of additive noise) almost-sure stability (i.e., $\lim_{t\to\infty} x_t = 0$ almost surely) [17]. Mean-square stability requires stricter and more complicated conditions than stabilizability of the nominal system (A, B) [17], which are discussed in the sequel. This essentially can limit the size of the multiplicative noise covariance [18], which can be viewed as a representation of uncertainty in the nominal system model or as inherent variation in the system dynamics.

A. Control Design With Known Models: Value Iteration

Dynamic programming can be used to show that the optimal policy π^* is linear state feedback $u_t = \pi^*(x_t) = K^* x_t$, where $K^* \in \mathbb{R}^{m\times n}$ denotes the optimal gain matrix. When the control policy is linear state feedback $u_t = \pi(x_t) = Kx_t$, with a very slight abuse of notation the cost becomes

$$C(K) = \mathbb{E}_{x_0,\{\delta_{ti}\},\{\gamma_{tj}\}} \sum_{t=0}^{\infty} x_t^\intercal (Q + K^\intercal R K) x_t.$$

Dynamic programming further shows that the resulting optimal cost is quadratic in the initial state, i.e., $C(K^*) = \mathbb{E}_{x_0} x_0^\intercal P x_0 = \mathrm{Tr}(P\Sigma_0)$, where $P \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix [21]. Note that the optimal controller does not directly observe the noise variables δ_{ti}, γ_{tj}. When the model parameters are known, there are several ways to compute the optimal feedback gains and corresponding optimal cost. The optimal cost is given by the solution of the *generalized* algebraic Riccati equation (GARE)

$$P = Q + A^\intercal P A + \sum_{i=1}^{p} \alpha_i A_i^\intercal P A_i - A^\intercal P B \left(R + B^\intercal P B + \sum_{j=1}^{q} \beta_j B_j^\intercal P B_j \right)^{-1} B^\intercal P A. \tag{2}$$

This is a special case of the GARE for optimal static output feedback given in [19] and can be solved via the value iteration

$$P_{k+1} = Q + A^\intercal P_k A + \sum_{i=1}^{p} \alpha_i A_i^\intercal P_k A_i - A^\intercal P_k B \left(R + B^\intercal P_k B + \sum_{j=1}^{q} \beta_j B_j^\intercal P_k B_j \right)^{-1} B^\intercal P_k A$$

with $P_0 = Q$, or via semidefinite programming formulations [29]–[31], or via more exotic iterations based on the Smith method and Krylov subspaces [45], [46]. The associated optimal gain matrix is

$$K^* = -\left(R + B^\intercal P B + \sum_{j=1}^{q} \beta_j B_j^\intercal P B_j \right)^{-1} B^\intercal P A.$$

It was verified in [17] that the existence of a positive definite solution to the GARE (2) is equivalent to mean-square stabilizability of the system, which depends on the problem data A, B, α_i, A_i, β_j, and B_j; in particular, mean-square stability generally imposes upper bounds on the variances α_i and β_j [18], though these may be infinite depending on the structure of A, B, A_i, B_j [17]. At a minimum, uniqueness and existence of a solution to the GARE (2) requires the standard conditions for uniqueness and existence of a solution to the standard ARE, namely of (A, B) stabilizable and $(A, Q^{1/2})$ detectable.

Although (approximate) value iteration can be implemented using sample trajectory data, policy gradient methods have been shown to be more effective for approximately optimal control of high-dimensional stochastic nonlinear systems, e.g., those arising in robotics [47]. This motivates our following analysis of the simpler case of stochastic linear systems wherein we show that policy gradient indeed facilitates a data-driven approach for learning optimal and robust policies.

B. Control Design With Known Models: Policy Gradient

Consider a fixed linear state feedback policy $u_t = Kx_t$. Defining the stochastic system matrices $\widetilde{A} = A + \sum_{i=1}^{p} \delta_{ti} A_i$, and $\widetilde{B} = B + \sum_{j=1}^{q} \gamma_{tj} B_j$, the (deterministic) nominal closed-loop state transition matrix $A_K = A + BK$, the stochastic closed-loop state transition matrix $\widetilde{A}_K = \widetilde{A} + \widetilde{B}K$, and the closed-loop state-cost matrix $Q_K = Q + K^\intercal R K$, the closed-loop dynamics become $x_{t+1} = \widetilde{A}_K x_t$. A gain K is mean-square stabilizing if the closed-loop system is mean-square stable. Denote the set of mean-square stabilizing K as $\mathcal{K}$. If $K \in \mathcal{K}$, then the cost can be written as $C(K) = \mathbb{E}_{x_0} x_0^\intercal P_K x_0 = \mathrm{Tr}(P_K \Sigma_0)$, where P_K is the unique positive semidefinite solution to the *generalized* Lyapunov equation

$$P_K = Q_K + A_K^\intercal P_K A_K + \sum_{i=1}^{p} \alpha_i A_i^\intercal P_K A_i + \sum_{j=1}^{q} \beta_j K^\intercal B_j^\intercal P_K B_j K. \tag{3}$$

[1]We assume $\bar{A}$ and $\bar{B}$ are independent for simplicity, but it is straightforward to include correlations between the entries of $\bar{A}$ and $\bar{B}$ into the model.

We define the state covariance matrices as $\Sigma_t := \mathbb{E}_{x_0,\delta,\gamma}[x_t x_t^\intercal]$ and the infinite-horizon aggregate state covariance matrix $\Sigma_K := \sum_{t=0}^{\infty} \Sigma_t$. If $K \in \mathcal{K}$, then Σ_K also satisfies a *dual* generalized Lyapunov equation

$$\Sigma_K = \Sigma_0 + A_K \Sigma_K A_K^\intercal + \sum_{i=1}^{p} \alpha_i A_i \Sigma_K A_i^\intercal + \sum_{j=1}^{q} \beta_j B_j K \Sigma_K K^\intercal B_j^\intercal. \quad (4)$$

Vectorization and Kronecker products can be used to convert (3) and (4) into systems of linear equations. Alternatively, iterative methods have been suggested for their solution [45], [46]. The state covariance dynamics are captured by two closed-loop finite-dimensional linear operators which operate on a symmetric matrix X

$$\mathcal{T}_K(X) := \mathop{\mathbb{E}}_{\delta,\gamma} \sum_{t=0}^{\infty} \widetilde{A}_K^t X \widetilde{A}_K^{\intercal t},$$

$$\mathcal{F}_K(X) := \mathop{\mathbb{E}}_{\delta,\gamma} \widetilde{A}_K X \widetilde{A}_K^\intercal = A_K X A_K^\intercal + \sum_{i=1}^{p} \alpha_i A_i X A_i^\intercal + \sum_{j=1}^{q} \beta_j B_j K X (B_j K)^\intercal.$$

Thus, $\mathcal{F}_K$ (without an argument) is a linear operator whose matrix representation is

$$\mathcal{F}_K := A_K \otimes A_K + \sum_{i=1}^{p} \alpha_i A_i \otimes A_i + \sum_{j=1}^{q} \beta_j (B_j K) \otimes (B_j K).$$

The Σ_t evolve according to the dynamics

$$\Sigma_{t+1} = \mathcal{F}_K(\Sigma_t) \quad \Leftrightarrow \quad \operatorname{vec}(\Sigma_{t+1}) = \mathcal{F}_K \operatorname{vec}(\Sigma_t).$$

We define the t-stage of $\mathcal{F}_K(X)$ as

$$\mathcal{F}_K^t(X) := \mathcal{F}_K(\mathcal{F}_K^{t-1}(X)) \text{ with } \mathcal{F}_K^0(X) = X$$

which gives the natural characterization

$$\Sigma_K = \mathcal{T}_K(\Sigma_0) = \sum_{t=0}^{\infty} \mathcal{F}_K^t(\Sigma_0). \quad (5)$$

We then have the following lemma.

Lemma 2.1 (Mean-Square Stability): A gain K is mean-square stabilizing if and only if the spectral radius $\rho(\mathcal{F}_K) < 1$.

Proof: Mean-square stability implies $\lim_{t\to\infty} \mathbb{E}[x_t x_t^\intercal] = 0$, which, for linear systems, occurs only when Σ_K is finite, which, by (5), is equivalent to $\rho(\mathcal{F}_K) < 1$. ■

Recalling the definition of $C(K)$ and (4), along with the basic observation that $K \notin \mathcal{K}$ induces infinite cost, gives the following characterization of the cost:

$$C(K) = \begin{cases} \operatorname{Tr}(Q_K \Sigma_K) = \operatorname{Tr}(P_K \Sigma_0) & \text{if } K \in \mathcal{K} \\ \infty & \text{otherwise.} \end{cases}$$

The evident fact that $C(K)$ is expressed as a closed-form function, up to a Lyapunov equation, of K leads to the idea of performing gradient descent on $C(K)$ (i.e., policy gradient) via the update $K \leftarrow K - \eta \nabla C(K)$ to find the optimal gain matrix. However, two properties of the LQR cost function $C(K)$ complicate a convergence analysis of gradient descent. First, $C(K)$ is extended valued since not all gain matrices provide closed-loop mean-square stability, and so it does not have (global) Lipschitz gradients. Second, and even more concerning, $C(K)$ is generally nonconvex in K (even for deterministic LQR problems, as observed by Fazel *et al.* [5]); so it is unclear if and when gradient descent converges to the global optimum or if it even converges at all. Fortunately, as in the deterministic case, we show that the multiplicative LQR cost possesses further key properties that enable proof of global convergence despite the lack of Lipschitz gradients and nonconvexity.

C. From Stochastic to Robust Stability

Additional motivation for designing controllers which stabilize a stochastic system in mean-square is to ensure robustness of stability of a nominal deterministic system to model parameter perturbations. Here we state a condition which guarantees robust deterministic stability for a perturbed deterministic system given mean-square stability of a stochastic single-state system with multiplicative noise where the noise variance and parameter perturbation size are related.

Example 2.2 (Robust Stability): Suppose the stochastic closed-loop system

$$x_{t+1} = (a + \delta_t) x_t \quad (6)$$

where a, x_t, δ_t are scalars with $\mathbb{E}[\delta_t^2] = \alpha$ is mean-square stable. Then, the perturbed deterministic system

$$x_{t+1} = (a + \phi) x_t \quad (7)$$

is stable for any constant perturbation $|\phi| \leq \sqrt{a^2 + \alpha} - |a|$.

Proof: By the bound on ϕ and triangle inequality, we have $\rho(a + \phi) = |a + \phi| \leq |a| + |\phi| \leq \sqrt{a^2 + \alpha}$. From Lemma 2.1, mean-square stability of (6) implies $\sqrt{\rho(\mathcal{F})} = \sqrt{a^2 + \alpha} < 1$ and, thus, $\rho(a + \phi) < 1$, proving stability of (7). ■

Although this is a simple example, it demonstrates that the robustness margin increases monotonically with the multiplicative noise variance. We also see that when $\alpha = 0$, the bound collapses so that no robustness is guaranteed, i.e., when $|a| \to 1$. This result can be extended to multiple states, inputs, and noise directions, but the resulting conditions become considerably more complex [19], [34]. We now proceed with developing methods for optimal control.

III. Gradient Domination and Other Properties of the Multiplicative Noise LQR Cost

In this section, we demonstrate that the multiplicative noise LQR cost function is *gradient dominated*, which facilitates optimization by gradient descent. Gradient dominated functions have been studied for many years in the optimization literature [48] and have recently been discovered in deterministic LQR problems by [5]. Proofs of the technical results are condensed here for brevity but are available in more verbose form in our article [49].

A. Multiplicative Noise LQR Cost Is Gradient Dominated

First, we give the expression for the policy gradient of the multiplicative noise LQR cost. For brevity, define

$$R_K := R + B^\intercal P_K B + \sum_{j=1}^{q} \beta_j B_j^\intercal P_K B_j$$

$$E_K := R_K K + B^\intercal P_K A.$$

Lemma 3.1 (Policy Gradient Expression):

The policy gradient is given by

$$\nabla_K C(K) = 2E_K\Sigma_K = 2(R_K K + B^\intercal P_K A)\Sigma_K.$$

Proof: Substituting the RHS of the generalized Lyapunov equation (3) into the cost $C(K) = \operatorname{Tr}(P_K\Sigma_0)$ yields

$$\begin{aligned} C(K) = {} & \operatorname{Tr}(Q_K\Sigma_0) + \operatorname{Tr}(A_K^\intercal P_K A_K\Sigma_0) \\ & + \operatorname{Tr}\left(\sum_{i=1}^{p} \alpha_i A_i^\intercal P_K A_i\Sigma_0\right) \\ & + \operatorname{Tr}\left(\sum_{j=1}^{q} \beta_j K^\intercal B_j^\intercal P_K B_j K\Sigma_0\right). \end{aligned}$$

Taking the gradient with respect to K and using the product rule and rules for matrix derivatives, we obtain

$$\begin{aligned} &\nabla_K C(K) = \nabla_K \operatorname{Tr}(P_K\Sigma_0) \\ &= \nabla_{\widetilde{K}}\Bigg[\operatorname{Tr}(Q_{\widetilde{K}}\Sigma_0) + \operatorname{Tr}(A_{\widetilde{K}}^\intercal P_K A_{\widetilde{K}}\Sigma_0) \\ &\quad + \operatorname{Tr}\left(\sum_{i=1}^{p} \alpha_i A_i^\intercal P_K A_i\Sigma_0\right) + \operatorname{Tr}\left(\sum_{j=1}^{q} \beta_j \widetilde{K}^\intercal B_j^\intercal P_K B_j \widetilde{K}\Sigma_0\right)\Bigg] \\ &\quad + \nabla_{\bar{K}}\Bigg[\operatorname{Tr}(A_K^\intercal P_{\bar{K}} A_K\Sigma_0) \\ &\quad + \operatorname{Tr}\left(\sum_{i=1}^{p} \alpha_i A_i^\intercal P_{\bar{K}} A_i\Sigma_0\right) + \operatorname{Tr}\left(\sum_{j=1}^{q} \beta_j K^\intercal B_j^\intercal P_{\bar{K}} B_j K\Sigma_0\right)\Bigg] \\ &= 2(R_K K + B^\intercal P_K A)\Sigma_0 + \nabla_{\bar{K}} \operatorname{Tr}(P_{\bar{K}}\mathcal{F}_K(\Sigma_0)) \\ &= 2(R_K K + B^\intercal P_K A)\Sigma_0 + \nabla_K \operatorname{Tr}(P_K\Sigma_1) \end{aligned}$$

where the tilde on $\widetilde{K}$ and overbar on $\overline{K}$ are used to denote the terms being differentiated. Applying this gradient formula recursively to the last term in the last line (namely $\nabla_{\bar{K}} \operatorname{Tr}(P_{\bar{K}}\Sigma_1)$) and recalling the definition of Σ_K completes the proof. See [49] for detailed intermediate steps. ■

For brevity, the gradient is implied to be with respect to the gains K in the rest of this article, i.e., ∇_K denoted by ∇. Now, we must develop some auxiliary results before demonstrating gradient domination. Throughout $\|Z\|$ and $\|Z\|_F$ are the spectral and Frobenius norms, respectively, of a matrix Z, and $\underline{\sigma}(Z)$ and $\overline{\sigma}(Z)$ are the minimum and maximum singular values of a matrix Z. The value function $V_K(x)$, evaluated at the initial condition of the process x_t (i.e., x_0), is defined as

$$V_K(x) := \mathbb{E}_{\delta,\gamma} \sum_{t=0}^{\infty} x_t^\intercal Q_K x_t \text{ given } x_0 = x$$

which relates to the cost as $C(K) = \mathbb{E}_{x_0} V_K(x_0)$. The advantage function is defined as

$$\mathcal{A}_K(x,u) := x^\intercal Q x + u^\intercal R u + \mathop{\mathbb{E}}_{\delta,\gamma} V_K(\widetilde{A}x + \widetilde{B}u) - V_K(x)$$

where the expectation is taken with respect to the variables $\widetilde{A}$ and $\widetilde{B}$. The advantage function can be thought of as the difference in cost ("advantage") when starting in state x of taking an action u for one step instead of the action generated by policy K. We also define the state, input, and cost sequences

$$\begin{aligned} \{x_t\}_{K,x} &:= \{x, \widetilde{A}_K x, \widetilde{A}_K^2 x, ..., \widetilde{A}_K^t x, ...\} \\ \{u_t\}_{K,x} &:= K\{x_t\}_{K,x} \\ \{c_t\}_{K,x} &:= \{x_t\}_{K,x}^\intercal Q_K \{x_t\}_{K,x}. \end{aligned}$$

Throughout the proofs, we will consider pairs of gains K and K' and their difference $\Delta := K' - K$.

Lemma 3.2 (Value Difference): Suppose K and K' generate the (stochastic) state, action, and cost sequences $\{x_t\}_{K,x}, \{u_t\}_{K,x}, \{c_t\}_{K,x}$, and $\{x_t\}_{K',x}, \{u_t\}_{K',x}, \{c_t\}_{K',x}$. Then the value difference and advantage satisfy

$$V_{K'}(x) - V_K(x) = \mathop{\mathbb{E}}_{\delta,\gamma} \sum_{t=0}^{\infty} \mathcal{A}_K\left(\{x_t\}_{K',x}, \{u_t\}_{K',x}\right)$$

$$\mathcal{A}_K(x, K'x) = 2x^\intercal \Delta^\intercal E_K x + x^\intercal \Delta^\intercal R_K \Delta x.$$

Proof: The proof follows the "cost-difference" lemma in [5] exactly substituting versions of value and cost functions, etc., which take expectation over the multiplicative noise. ■

Next, we see that the multiplicative noise LQR cost is gradient dominated.

Lemma 3.3 (Gradient Domination):

The LQR-with-multiplicative-noise cost $C(K)$ satisfies the gradient domination condition

$$C(K) - C(K^*) \le \frac{\|\Sigma_{K^*}\|}{4\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)^2} \|\nabla C(K)\|_F^2.$$

Proof: We start with the advantage expression

$$\begin{aligned} \mathcal{A}_K(x, K'x) &= 2x^\intercal \Delta^\intercal E_K x + x^\intercal \Delta^\intercal R_K \Delta x \\ &= 2\operatorname{Tr}[x x^\intercal \Delta^\intercal E_K] + \operatorname{Tr}[x x^\intercal \Delta^\intercal R_K \Delta]. \end{aligned}$$

Next, we rearrange and complete the square

$$\begin{aligned} \mathcal{A}_K(x, K'x) = {} & \operatorname{Tr}\left[x x^\intercal \left(\Delta^\intercal R_K \Delta + 2\Delta^\intercal E_K\right)\right] \\ = {} & \operatorname{Tr}\left[x x^\intercal (\Delta + R_K^{-1} E_K)^\intercal R_K (\Delta + R_K^{-1} E_K)\right] \\ & - \operatorname{Tr}\left[x x^\intercal E_K^\intercal R_K^{-1} E_K\right]. \end{aligned}$$

Since $R_K \succ 0$, we have

$$\mathcal{A}_K(x, K'x) \ge -\operatorname{Tr}\left[x x^\intercal E_K^\intercal R_K^{-1} E_K\right] \tag{8}$$

with equality only when $\Delta = -R_K^{-1} E_K$.

Let the state and control sequences associated with the optimal gain K^* be $\{x_t\}_{K^*,x}$ and $\{u_t\}_{K^*,x}$ respectively. We now obtain an upper bound for the cost difference by writing the cost difference in terms of the value function as

$$\begin{aligned} C(K) - C(K^*) &= \mathop{\mathbb{E}}_{x_0}\left[V(K, x_0)\right] - \mathop{\mathbb{E}}_{x_0}\left[V(K^*, x_0)\right] \\ &= \mathop{\mathbb{E}}_{x_0}\left[V(K, x_0) - V(K^*, x_0)\right]. \end{aligned}$$

Using the first part of the value-difference Lemma 3.2 and negating, we obtain

$$\begin{aligned} C(K) - C(K^*) &= -\mathop{\mathbb{E}}_{x_0}\left[\sum_{t=0}^{\infty} \mathcal{A}_K\left(\{x_t\}_{K^*,x}, \{u_t\}_{K^*,x}\right)\right] \\ &\le \mathop{\mathbb{E}}_{x_0}\left[\sum_{t=0}^{\infty} \operatorname{Tr}\left[\{x_t\}_{K^*,x}\{x_t\}_{K^*,x}^\intercal E_K^\intercal R_K^{-1} E_K\right]\right] \end{aligned}$$

$$= \operatorname{Tr}\left[\Sigma_{K^*} E_K^\intercal R_K^{-1} E_K\right]$$

where the second step used the advantage inequality in (8). Now using $|\operatorname{Tr}(YZ)| \le \|Y\| |\operatorname{Tr}(Z)|$, we obtain

$$\begin{aligned} C(K) - C(K^*) &\le \|\Sigma_{K^*}\| \operatorname{Tr}\left[E_K^\intercal R_K^{-1} E_K\right] \\ &\le \|\Sigma_{K^*}\| \|R_K^{-1}\| \operatorname{Tr}\left[E_K^\intercal E_K\right] \end{aligned} \tag{9}$$

where the first and second inequalities will be used later in the Gauss–Newton and gradient descent convergence proofs, respectively. Combining $\|R_K\| \ge \|R\| = \overline{\sigma}(R) \ge \underline{\sigma}(R)$ with $\|Z^{-1}\| \ge \|Z\|^{-1}$, we obtain

$$C(K) - C(K^*) \le \frac{\|\Sigma_{K^*}\|}{\underline{\sigma}(R)} \operatorname{Tr}\left[E_K^\intercal E_K\right] \tag{10}$$

which will be used later in the natural policy gradient descent convergence proof. Now we rearrange and substitute in the policy gradient expression $\frac{1}{2}\nabla C(K)(\Sigma_K)^{-1} = E_K$

$$\begin{aligned} C(K) - C(K^*) &\le \frac{\|\Sigma_{K^*}\|}{4\underline{\sigma}(R)} \operatorname{Tr}\left[(\nabla C(K)\Sigma_K^{-1})^\intercal (\nabla C(K)\Sigma_K^{-1})\right] \\ &\le \frac{\|\Sigma_{K^*}\|}{4\underline{\sigma}(R)} \|(\Sigma_K^{-1})^\intercal \Sigma_K^{-1}\| \operatorname{Tr}\left[\nabla C(K)^\intercal \nabla C(K)\right] \\ &\le \frac{\|\Sigma_{K^*}\|}{4\underline{\sigma}(R)\underline{\sigma}(\Sigma_K)^2} \operatorname{Tr}\left[\nabla C(K)^\intercal \nabla C(K)\right] \end{aligned}$$

where the last step used the definition and submultiplicativity of spectral norm. Using $\Sigma_K = \mathop{\mathbb{E}}_{x_0}[\sum_{t=0}^{\infty} x_t x_t^\intercal] \succeq \mathop{\mathbb{E}}_{x_0}[x_0 x_0^\intercal] = \Sigma_0 \Rightarrow \underline{\sigma}(\Sigma_K) < \underline{\sigma}(\Sigma_0)$ completes the proof. ■

The gradient domination property gives the following stationary point characterization.

Corollary 3.4: If $\nabla C(K) = 0$, then either $K = K^*$ or $\operatorname{rank}(\Sigma_K) < n$.

In other words, so long as Σ_K is full rank, stationarity is both necessary and sufficient for global optimality, as for convex functions. Note that it is not sufficient to just have multiplicative noise in the dynamics with a deterministic initial state x_0 to ensure that Σ_K is full rank. To see this, observe that if $x_0 = 0$ and $\Sigma_0 = 0$, then $\Sigma_K = 0$, which is clearly rank deficient. By contrast, additive noise is sufficient to ensure that Σ_K is full rank with a deterministic initial state x_0, although we will not consider this setting. Using a random initial state with $\Sigma_0 \succ 0$ ensures $\operatorname{rank}(\Sigma_K) = n$, and, thus, $\nabla C(K) = 0$ implies $K = K^*$.

Although the gradient of the multiplicative noise LQR cost is not globally Lipschitz continuous, it is locally Lipschitz continuous over any subset of $\mathcal{K}$. Gradient domination is then sufficient to show that policy gradient descent will converge to the optimal gains at a linear rate (a short proof for globally Lipschitz functions is given in [50]). We prove convergence of policy gradient to the optimum feedback gain by bounding the local Lipschitz constant in terms of the problem data, which bounds the maximum step size and the convergence rate.

B. Additional Setup Lemmas

Lemma 3.5 (Almost-Smoothness): The LQR-with-multiplicative-noise cost $C(K)$ satisfies the almost-smoothness expression

$$C(K') - C(K) = 2\operatorname{Tr}\left[\Sigma_{K'} \Delta^\intercal E_K\right] + \operatorname{Tr}\left[\Sigma_{K'} \Delta^\intercal R_K \Delta\right].$$

Proof: As in the gradient domination proof, we express the cost difference in terms of the advantage by taking expectation over the initial states to obtain

$$C(K') - C(K) = \mathop{\mathbb{E}}_{x_0}\left[\sum_{t=0}^{\infty} \mathcal{A}_K\left(\{x_t\}_{K',x}\,,\ \{u_t\}_{K',x}\right)\right].$$

From the value difference lemma for the advantage, we have

$$\mathcal{A}_K(x, K'x) = 2x^\intercal \Delta^\intercal E_K x + x^\intercal \Delta^\intercal R_K \Delta x.$$

Noting that $\{u_t\}_{K',x} = K'x$, we obtain $C(K') - C(K) =$

$$\mathop{\mathbb{E}}_{x_0}\left[\sum_{t=0}^{\infty} 2\{x_t\}_{K',x}^\intercal \Delta^\intercal E_K \{x_t\}_{K',x} + \{x_t\}_{K',x}^\intercal \Delta^\intercal R_K \Delta \{x_t\}_{K',x})\right].$$

Using the definition of $\Sigma_{K'}$ completes the proof. ■

Lemma 3.6 (Cost Bounds): We always have

$$\|P_K\| \le \frac{C(K)}{\underline{\sigma}(\Sigma_0)} \quad \text{and} \quad \|\Sigma_K\| \le \frac{C(K)}{\underline{\sigma}(Q)}.$$

Proof: The proof follows that in [5] exactly. ■

IV. Global Convergence of Policy Gradient in the Model-Based Setting

In this section, we show that the policy gradient algorithm and two important variants for multiplicative noise LQR converge globally to the optimal policy. In contrast with [5], the policies we obtain are robust to uncertainties and inherent stochastic variations in the system dynamics. We analyze three policy gradient algorithm variants as follows:

$$\begin{aligned} \text{Gradient:}\quad & K_{s+1} = K_s - \eta \nabla C(K_s) \\ \text{Natural Gradient:}\quad & K_{s+1} = K_s - \eta \nabla C(K_s)\Sigma_{K_s}^{-1} \\ \text{Gauss–Newton:}\quad & K_{s+1} = K_s - \eta R_{K_s}^{-1} \nabla C(K_s)\Sigma_{K_s}^{-1}. \end{aligned}$$

The more elaborate natural gradient and Gauss–Newton variants provide superior convergence rates and simpler proofs. A development of the natural policy gradient is given in [5] building on ideas from [51]. The Gauss–Newton step with step size $1/2$ is identical to policy iteration, first studied for deterministic LQR in [52]. This was extended to a model-free setting using policy iteration and Q-learning in [6]. For multiplicative noise LQR, we have the following results, which are not optimized for tightness; step sizes satisfying the bounds can become too small to be practically useful. Rather, our goal is to find algorithm settings that give guaranteed convergence. In practice, much less conservative constant and adaptive step sizes can be used, as shown in Section VI.

A. Gauss–Newton Descent

Theorem 4.1 (Gauss–Newton Convergence): Using the Gauss–Newton step

$$K_{s+1} = K_s - \eta R_{K_s}^{-1} \nabla C(K_s)\Sigma_{K_s}^{-1}$$

with step size $0 < \eta \leq \frac{1}{2}$ gives global convergence to the optimal gain matrix K^* at a linear rate described by

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \leq 1 - 2\eta \frac{\underline{\sigma}(\Sigma_0)}{\|\Sigma_{K^*}\|}.$$

Proof: The next-step gain matrix difference is

$$\Delta = K_{s+1} - K_s = -\eta R_{K_s}^{-1} \nabla C(K_s) \Sigma_{K_s}^{-1} = -2\eta R_{K_s}^{-1} E_{K_s}.$$

Using the almost-smoothness Lemma 3.5 and substituting in the next-step gain matrix difference, we obtain

$$\begin{aligned} &C(K_{s+1}) - C(K_s) \\ &= 2 \operatorname{Tr}\left[\Sigma_{K_{s+1}} \Delta^\intercal E_{K_s}\right] + \operatorname{Tr}\left[\Sigma_{K_{s+1}} \Delta^\intercal R_{K_s} \Delta\right] \\ &= 2 \operatorname{Tr}\left[\Sigma_{K_{s+1}} (-2\eta R_{K_s}^{-1} E_{K_s})^\intercal E_{K_s}\right] \\ &\quad + \operatorname{Tr}\left[\Sigma_{K_{s+1}} (-2\eta R_{K_s}^{-1} E_{K_s})^\intercal R_{K_s} (-2\eta R_{K_s}^{-1} E_{K_s})\right] \\ &= 4(-\eta + \eta^2) \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal R_{K_s}^{-1} E_{K_s}\right]. \end{aligned}$$

By hypothesis, we require $0 \leq \eta \leq \frac{1}{2}$; so we have

$$\begin{aligned} C(K_{s+1}) - C(K_s) &\leq -2\eta \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal R_{K_s}^{-1} E_{K_s}\right] \\ &\leq -2\eta \underline{\sigma}(\Sigma_{K_{s+1}}) \operatorname{Tr}\left[E_{K_s}^\intercal R_{K_s}^{-1} E_{K_s}\right] \\ &\leq -2\eta \underline{\sigma}(\Sigma_0) \operatorname{Tr}\left[E_{K_s}^\intercal R_{K_s}^{-1} E_{K_s}\right]. \end{aligned}$$

Recalling and substituting in (9), we obtain

$$C(K_{s+1}) - C(K_s) \leq -2\eta \frac{\underline{\sigma}(\Sigma_0)}{\|\Sigma_{K^*}\|} \left(C(K_s) - C(K^*)\right).$$

Adding $C(K_s) - C(K^*)$ to both sides and rearranging completes the proof. ■

B. Natural Policy Gradient Descent

Theorem 4.2 (Natural Policy Gradient Convergence): Using the natural policy gradient step

$$K_{s+1} = K_s - \eta \nabla C(K_s) \Sigma_{K_s}^{-1} \tag{11}$$

with step size $0 < \eta \leq c_{\text{npg}}$ where

$$c_{\text{npg}} := \frac{1}{2} \left(\|R\| + \left(\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2 \right) \frac{C(K_0)}{\underline{\sigma}(\Sigma_0)} \right)^{-1}$$

gives global convergence to the optimal gain matrix K^* at a linear rate described by

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \leq 1 - 2\eta \frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)}{\|\Sigma_{K^*}\|}.$$

Proof: First, we bound the one-step progress, where the step size depends explicitly on the current gain K_s. Using the update (11), the next-step gain matrix difference is

$$\Delta = K_{s+1} - K_s = -\eta \nabla C(K_s) \Sigma_{K_s}^{-1} = -2\eta E_{K_s}.$$

Using Lemma 3.5 and substituting, we obtain

$$\begin{aligned} &C(K_{s+1}) - C(K_s) \\ &= 2 \operatorname{Tr}\left[\Sigma_{K_{s+1}} \Delta^\intercal E_{K_s}\right] + \operatorname{Tr}\left[\Sigma_{K_{s+1}} \Delta^\intercal R_{K_s} \Delta\right] \\ &= 2 \operatorname{Tr}\left[\Sigma_{K_{s+1}} (-2\eta E_{K_s})^\intercal E_{K_s}\right] \\ &\quad + \operatorname{Tr}\left[\Sigma_{K_{s+1}} (-2\eta E_{K_s})^\intercal R_{K_s} (-2\eta E_{K_s})\right] \\ &= -4\eta \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal E_{K_s}\right] + 4\eta^2 \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal R_{K_s} E_{K_s}\right] \\ &\leq 4(-\eta + \eta^2 \|R_{K_s}\|) \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal E_{K_s}\right]. \end{aligned}$$

If we choose step size $0 < \eta \leq \frac{1}{2\|R_{K_s}\|}$, then

$$\begin{aligned} C(K_{s+1}) - C(K_s) &\leq -2\eta \operatorname{Tr}\left[\Sigma_{K_{s+1}} E_{K_s}^\intercal E_{K_s}\right] \\ &\leq -2\eta \underline{\sigma}(\Sigma_{K_{s+1}}) \operatorname{Tr}\left[E_{K_s}^\intercal E_{K_s}\right] \\ &\leq -2\eta \underline{\sigma}(\Sigma_0) \operatorname{Tr}\left[E_{K_s}^\intercal E_{K_s}\right]. \end{aligned}$$

Recalling and substituting (10), we obtain

$$\begin{aligned} &C(K_{s+1}) - C(K_s) \\ &\leq -2\eta \underline{\sigma}(\Sigma_0) \frac{\underline{\sigma}(R)}{\|\Sigma_{K^*}\|} \left(C(K_s) - C(K^*)\right). \end{aligned}$$

Adding $C(K_s) - C(K^*)$ to both sides and rearranging gives the one step progress bound

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \leq 1 - 2\eta \frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)}{\|\Sigma_{K^*}\|}. \tag{12}$$

Next, using the cost bound in Lemma 3.6, the triangle inequality, and submultiplicativity of spectral norm, we have

$$\begin{aligned} \frac{1}{\|R_K\|} &= \frac{1}{\|R + B^\intercal P_K B + \sum_{j=1}^{q} \beta_j B_j^\intercal P_K B_j\|} \\ &\geq \frac{1}{\|R\| + (\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2) \|P_K\|} \\ &\geq \frac{1}{\|R\| + (\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2) \frac{C(K)}{\underline{\sigma}(\Sigma_0)}}. \end{aligned}$$

Accordingly, choosing the step size as $0 < \eta \leq c_{\text{npg}}$ ensures that (12) holds at the first step. This ensures that $C(K_1) \leq C(K_0)$ which, in turn, ensures

$$\begin{aligned} \eta &\leq \frac{1}{\|R\| + (\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2) \frac{C(K_0)}{\underline{\sigma}(\Sigma_0)}} \\ &\leq \frac{1}{\|R\| + (\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2) \frac{C(K_1)}{\underline{\sigma}(\Sigma_0)}} \leq \frac{1}{\|R_{K_1}\|} \end{aligned}$$

which allows (12) to be applied at the next step as well. Proceeding inductively by applying (12) at each successive step completes the proof. ■

C. Policy Gradient Descent

Theorem 4.3 (Policy Gradient Convergence): Using the policy gradient step

$$K_{s+1} = K_s - \eta \nabla C(K_s)$$

with step size $0 < \eta \leq c_{\text{pg}}$ gives global convergence to the optimal gain matrix K^* at a linear rate described by

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \leq 1 - 2\eta \frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)^2}{\|\Sigma_{K^*}\|}$$

where c_{pg} is a polynomial in the problem data A, B, α_i, β_j, A_i, B_j, Q, R, Σ_0, K_0 given in the proof in Appendix A.

Proof: The proof is developed in Appendix A. ■

The proofs for these results explicitly incorporate the effects of the multiplicative noise terms δ_{ti} and γ_{tj} in the dynamics. For the policy gradient and natural policy gradient algorithms, we show explicitly how the maximum allowable step size depends on problem data and, in particular, on the multiplicative noise

Algorithm 1: Model-Free Policy Gradient Estimation.

Input: Gain matrix K, number of samples n_{sample}, rollout length ℓ, exploration radius r

1: **for** $i = 1, \ldots, n_{\text{sample}}$ **do**

2: Generate a sample gain matrix $\widehat{K}_i = K + U_i$, where U_i is drawn uniformly at random over matrices with Frobenius norm r;

3: Generate a sample initial state $x_0^{(i)} \sim \mathcal{P}_0$;

4: Simulate the closed-loop system for ℓ steps according to the stochastic dynamics in (1) starting from $x_0^{(i)}$ with $u_t^{(i)} = \widehat{K}_i x_t^{(i)}$, yielding the state sequence $\{x_t^{(i)}\}_{t=0}^{t=\ell}$;

5: Collect the empirical finite-horizon cost estimate $\widehat{C}_i := \sum_{t=0}^{\ell} x_t^{(i)\intercal}(Q + \widehat{K}_i^\intercal R \widehat{K}_i) x_t^{(i)}$;

6: **end for**

Output: Gradient estimate $\widehat{\nabla} C(K) := \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} \frac{mn}{r^2} \widehat{C}_i U_i$

terms. Compared to deterministic LQR, the multiplicative noise terms decrease the allowable step size and thereby decrease the convergence rate; specifically, the state-multiplicative noise increases the initial cost $C(K_0)$ and the norms of the covariance Σ_{K^*} and cost P_K, and the input-multiplicative noise also increases the denominator term $\|B\|^2 + \sum_{j=1}^{q} \beta_j \|B_j\|^2$. This means that the algorithm parameters for deterministic LQR in [5] may cause failure to converge on problems with multiplicative noise. Moreover, even the optimal policies for deterministic LQR may actually *destabilize* systems in the presence of small amounts of multiplicative noise uncertainty, indicating the possibility for a catastrophic lack of robustness; observe the results of the example in Section VI-A. The results and proofs also differ from that of [5] because the more complicated mean-square stability must be accounted for and because *generalized* Lyapunov equations must be solved to compute the gradient steps, which requires specialized solvers.

V. Global Convergence of Policy Gradient in the Model-Free Setting

The results in the previous section are model-based; the policy gradient steps are computed exactly based on knowledge of the model parameters. In the model-free setting, the policy gradient is estimated to arbitrary accuracy from sample trajectories with a sufficient number of sample trajectories n_{sample} of sufficiently long horizon length ℓ using gain matrices randomly selected from a Frobenius-norm ball around the current gain of sufficiently small exploration radius r. We show for multiplicative noise LQR that with a finite number of samples polynomial in the problem data, the model-free policy gradient algorithm still converges to the globally optimal policy, despite small perturbations on the gradient.

In the model-free setting, the policy gradient method proceeds as before except that at each iteration, Algorithm 1 is called to generate an estimate of the gradient via the zeroth-order optimization procedure described by Fazel *et al.* [5].

Theorem 5.1 (Model-Free Policy Gradient Convergence): Let ϵ and μ be a given small tolerance and probability, respectively, and N be the number of gradient descent steps taken. Suppose that the distribution of the initial states is bounded such that $x_0 \sim \mathcal{P}_0$ implies $\|x_0^i\| \leq L_0$ almost surely for any given realization x_0^i of x_0. Suppose additionally that the distribution of the multiplicative noises is bounded such that the following inequality is satisfied almost surely for any given realized sequence x_t^i of x_t with a positive scalar $z \geq 1$

$$\sum_{t=0}^{\ell-1} \left(x_t^{i\intercal} Q x_t^i + u_t^{i\intercal} R u_t^i \right) \leq z \mathop{\mathbb{E}}_{\delta,\gamma} \left[\sum_{t=0}^{\ell-1} \left(x_t^\intercal Q x_t + u_t^\intercal R u_t \right) \right]$$

under the closed-loop dynamics with any gain such that $C(K) \leq 2C(K_0)$. Suppose the step size η is chosen according to the restriction in Theorem 4.3 and at every iteration, the gradient is estimated according to the finite-horizon procedure in Algorithm 1 where the number of samples n_{sample}, rollout length ℓ, and exploration radius r are chosen according to the fixed polynomials of the problem data A, B, α_i, β_j, A_i, B_j, Q, R, Σ_0, K_0, L_0, and z which are all defined in the proofs in Appendix VIII. Then, with high probability of at least $1 - \mu$, performing gradient descent results in convergence to the global optimum over all N steps: at each step, either progress is made at the linear rate

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \leq 1 - \eta \frac{\underline{\sigma}(R) \underline{\sigma}(\Sigma_0)^2}{\|\Sigma_{K^*}\|}$$

or convergence has been attained with $C(K_s) - C(K^*) \leq \epsilon$.

Proof: The proof is developed in Appendix VIII. ■

From a sample complexity standpoint, it is notable that the number of samples n_{sample}, rollout length ℓ, and exploration radius r in Theorem 5.1 are polynomial in the problem data A, B, α_i, β_j, A_i, B_j, Q, R, Σ_0, $C(K_0)$. The constant z imposes a bound on the multiplicative noise, which is naturally absent in [5]. Note that $z \geq 1$ since any upper bound of a scalar distribution with finite support must be equal to or greater than the mean. In general, this implicitly requires the noises to have bounded support. Such an assumption is qualitatively the same as the condition imposed on the initial states. These assumptions are reasonable; in a practical setting with a physical system, the initial state and noise distributions will have finite support. There is no restriction on how large the support is, only that it is not unbounded. Also note that the rate is halved compared with the model-based case of Theorem 4.3; this is because the "other half" is consumed by the error between the estimated and true gradient.

VI. Numerical Experiments

In this section, we present results for three systems.

1) Section VI-A shows that "optimal" control that ignores actual multiplicative noise can lead to loss of mean-square stability.
2) Section VI-B shows the efficacy of the policy gradient algorithms on a networked system.
3) Section VI-C shows the increased difficulty of estimating the gradient from sample data in the presence of multiplicative noise.

All systems we consider permit a solution to the GARE (2). The bounds on the step size, number of rollouts, and rollout length given by the theoretical analysis can be rather conservative. For practicality, we selected the constant step size, number of rollouts, rollout length, and exploration radius according to a

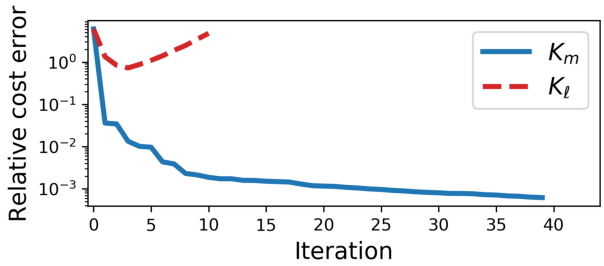

Fig. 1. Relative LQRm cost error $\frac{C(K)-C(K^*)}{C(K^*)}$ versus iteration during policy gradient descent on the four-state, one-input suspension example system.

grid search over reasonable values. Additionally, we investigated the use of backtracking line search to adaptively select the step size (see, e.g., [53]). Throughout the simulations, we computed the baseline optimal cost $C(K^*)$ by solving the GARE (2) to high precision via value iteration. Python code which implements the algorithms and generates the figures reported in this article can be found in the GitHub repository[2]. The code was run on a desktop PC with a quad-core Intel i7 6700K 4.0-GHz CPU, 16-GB RAM; no GPU computing was utilized.

A. Importance of Accounting for Multiplicative Noise

We first considered an open-loop mean-square unstable system with four states and one input representing an active two-mass suspension converted from continuous to discrete time using a standard bilinear transformation, with parameters

$$A = \begin{bmatrix} +0.261 & +0.315 & +0.093 & -0.008 \\ -2.955 & +0.261 & +0.373 & -0.033 \\ +1.019 & +0.255 & -0.853 & +0.011 \\ -3.170 & -0.793 & -4.902 & -0.146 \end{bmatrix}, B = \begin{bmatrix} 0.133 \\ 0.532 \\ 0.161 \\ 2.165 \end{bmatrix}$$

$$Q = I_4,\ R = I_1,\ [A_i]_{y,z} = \begin{cases} 1 \text{ if } z = i, \\ 0 \text{ otherwise,} \end{cases} \quad B_1 = \mathbf{1}_{4\times 1}$$

$$\{\alpha_i\} = \{0.017, 0.017, 0.017, 0.017\}, \qquad \beta_1 = 0.035.$$

We performed model-based policy gradient descent; at each iteration, gradients were calculated by solving generalized Lyapunov equations (3) and (4) using the problem data. The gains K_m and K_ℓ represent iterates during optimization of ("training" on) the LQRm and LQR cost (with the multiplicative noise variances set to zero), respectively. We performed the optimization starting from the same feasible initial gain, which was generated by perturbing the exact solution of the GARE such that the LQRm cost under the initial control was approximately 10 times that of the optimal control. The step size was chosen via backtracking line search. The optimization stopped once the Frobenius norm of the gradient fell below a small threshold. The plot in Fig. 1 shows the "testing" cost of the gains at each iteration evaluated on the LQRm cost (with multiplicative noise). From this figure, it is clear that K_m minimized the LQRm as desired. When there was high multiplicative noise, the noise-ignorant controller K_ℓ actually *destabilized* the system in the mean-square sense; this can be seen as the LQRm cost exploded upwards to infinity after iteration 10. In this sense, the multiplicative noise-aware optimization is generally safer and more robust than noise-ignorant optimization, and in examples like this is actually *necessary* for mean-square stabilization.

B. Policy Gradient Methods Applied to a Network

Many practical networked systems can be approximated by diffusion dynamics with losses and stochastic diffusion constants (edge weights) between nodes; examples include heat flow through uninsulated pipes, hydraulic flow through leaky pipes, information flow between processors with packet loss, electrical power flow between generators with resistant electrical power lines, etc. A derivation of the discrete-time dynamics of this system is given in [43]. We considered a particular four-state, four-input system, and open-loop mean-square stable with the following parameters:

$$A = \begin{bmatrix} 0.795 & 0.050 & 0.100 & 0.050 \\ 0.050 & 0.845 & 0.050 & 0.050 \\ 0.100 & 0.050 & 0.695 & 0.150 \\ 0.050 & 0.050 & 0.150 & 0.745 \end{bmatrix}$$

$$B = Q = R = \Sigma_0 = I_4$$

$$\{\alpha_i\} = \{0.005, 0.015, 0.010, 0.015, 0.005, 0.020\}$$

$$\{\beta_j\} = \{0.050, 0.150, 0.050, 0.100\}$$

$$[A_i]_{y,z} = \begin{cases} +1 \text{if} \{c_i = y \ \& \ d_i = y\} \text{or} \{c_i = z \ \& \ d_i = z\} \\ -1 \text{if} \{c_i = z \ \& \ d_i = y\} \text{or} \{c_i = y \ \& \ d_i = z\} \\ 0 \text{ otherwise.} \end{cases}$$

$$[B_j]_{y,z} = \begin{cases} +1 \text{ if } j = y = z \\ 0 \text{ otherwise.} \end{cases}$$

$$\{(c_i, d_i)\} = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}.$$

This system is open-loop mean-square stable; so we initialized the gains to all zeros for each trial. We performed policy optimization using the model-free gradient, and the model-based gradient, model-based natural gradient, and model-based Gauss-Newton step directions on 20 unique problem instances using two step size schemes.

1) Backtracking Line Search: Step sizes η were chosen adaptively at each iteration by backtracking line search with parameters $\alpha = 0.01$, $\beta = 0.5$ (see [53] for a description), except for Gauss–Newton which used the optimal constant step size of $1/2$. Model-free gradients and costs were estimated with 100 000 rollouts per iteration. We ran a fixed number, 20, of iterations chosen such that the final cost using model-free gradient descent was no more than 5% worse than optimal.

2) Constant Step Size: Step sizes were set to constants chosen as large as possible without observing infeasibility or divergence, which, on this problem instance, was $\eta = 5 \times 10^{-5}$ for gradient, $\eta = 2 \times 10^{-4}$ for natural gradient, and $\eta = 1/2$ for Gauss–Newton step directions. Model-free gradients were estimated with 1000 rollouts per iteration. We ran a fixed number, 20 000, of iterations chosen such that convergence was achieved with all step directions.

[2]Online available: https://github.com/TSummersLab/polgrad-multinoise/

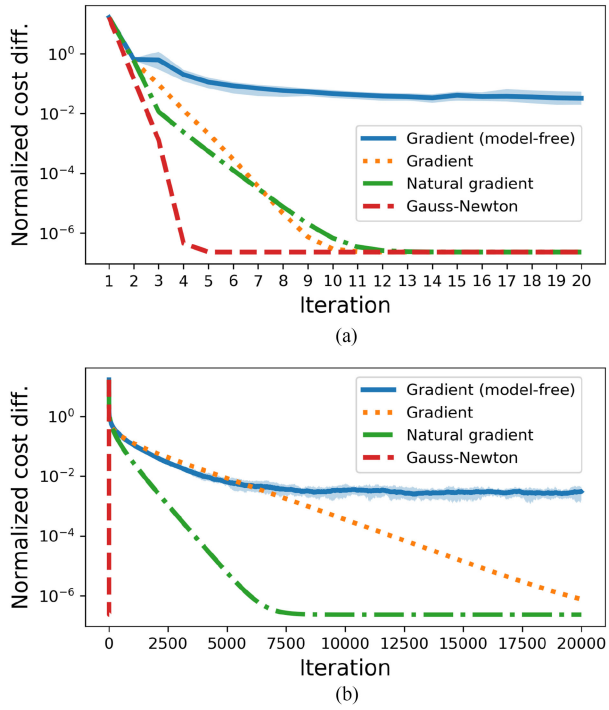

Fig. 2. Relative cost error $\frac{C(K)-C(K^*)}{C(K^*)}$ versus iteration during policy gradient methods on a four-state, four-input lossy diffusion network with multiplicative noise using (a) backtracking line search and (b) constant step sizes.

In both cases, sample gains were chosen for model-free gradient estimation with exploration radius $r=0.1$ and the rollout length was set to $\ell=20$. The plots in Fig. 2 show the relative cost over the iterations; for the model-free gradient descent, the bold centerline is the mean of all trials and the shaded region is between the 10th and 90th percentile of all trials. Using backtracking line search, it is evident that in terms of convergence, the Gauss–Newton step was extremely fast, and both the natural gradient and model-based gradient were slightly slower but still quite fast. The model-free policy gradient converged to a reasonable neighborhood of the minimum cost quickly but stagnated with further iterations; this is a consequence of the inherent gradient and cost estimation errors that arise due to random sampling and the multiplicative noise. Using constant step sizes, we were forced to take small steps due to the steepness of the cost function near the initial gains, slowing overall convergence using the gradient and natural gradient methods. Here we observed that Gauss–Newton again converged most quickly, followed by natural gradient and lastly the gradient methods. The smaller step size also allowed us to use far fewer samples in the model-free setting, where we observed somewhat faster initial cost decrease with eventual stagnation around 10^{-2}, or 1%, relative error, which represents excellent control performance. All algorithms exhibited convergence to the optimum, confirming the asserted theoretical claims.

C. Gradient Estimation

Multiplicative noise can significantly increase the variance and sample complexity of cost gradient estimates relative to the noiseless case, which is novelly reflected in the theoretical analysis for the number of rollouts and rollout length. To demonstrate this empirically, we evaluated the relative gradient estimation error versus number of rollouts for the system

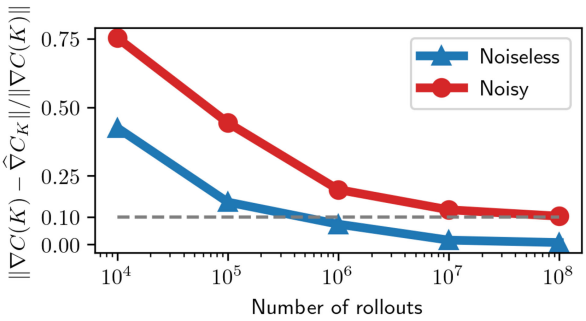

Fig. 3. Relative gradient estimation error versus number of rollouts for (13).

$$x_{t+1} = \left(\begin{bmatrix} 0.8 & 0.1 \\ 0.1 & 0.8 \end{bmatrix} + \delta_t \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} K \right) x_t \tag{13}$$

with $K = 0, Q = \Sigma_0 = I_2, R = 1$, $\delta_t \sim \mathcal{N}(0, 0.1)$, rollout length $l = 40$, exploration radius $r = 0.2$, averaged over 10 gradient estimates. The results are plotted in Fig. 3. To achieve the same gradient estimate error of 10%, the system with multiplicative noise required $200\times$ the number of rollout samples (10^8) as when there was no noise (5×10^5).

VII. Conclusion

We have shown that policy gradient methods in both model-based and model-free settings give global convergence to the globally optimal policy for LQR systems with multiplicative noise. These techniques are directly applicable for the design of robust controllers of uncertain systems and serve as a benchmark for data-driven control design. Our ongoing work is exploring ways of mitigating the relative sample inefficiency of model-free policy gradient methods by leveraging the special structure of LQR models and Nesterov-type acceleration, and exploring alternative system identification and adaptive control approaches. We are also investigating other methods of building robustness through $\mathcal{H}_\infty$ and dynamic game approaches. Another extension relevant to networked control systems is enforcing sparse structure constraints on the gain matrix via projected policy gradient as suggested in [54].

Appendix A
Model-Based Policy Gradient Descent

Throughout the proofs, please see the supplemental document [49] for additional details. The proof of convergence using gradient descent proceeds by establishing several technical lemmas, bounding the infinite-horizon covariance Σ_K, then using that bound to limit the step size, and finally obtaining a one-step bound on gradient descent progress and applying it inductively at each successive step.

We begin with a bound on the induced operator norm of $\mathcal{T}_K$.

Lemma A.1 ($\mathcal{T}_K$ Norm Bound): The following bound holds for any mean-square stabilizing K:

$$\|\mathcal{T}_K\| := \sup_X \frac{\|\mathcal{T}_K(X)\|}{\|X\|} \leq \frac{C(K)}{\underline{\sigma}(\Sigma_0)\underline{\sigma}(Q)}.$$

Proof: The proof follows that given in [5] using our definition of $\mathcal{T}_K$. ■

Lemma A.2 ($\mathcal{F}_K$ Perturbation): Consider a pair of mean-square stabilizing gain matrices K and K'. The following $\mathcal{F}_K$ perturbation bound holds:

$$\|\mathcal{F}_{K'} - \mathcal{F}_K\| \le 2\|A+BK\|\|B\|\|\Delta\| + h_B\|B\|\|\Delta\|^2$$

$$\text{where} \quad h_B := \|B\|^{-1}\left(\|B\|^2 + \sum_{j=1}^{q}\beta_j\|B_j\|^2\right).$$

Proof: Let $\Delta' = -\Delta$. For any matrix X, we have

$$\begin{aligned}(\mathcal{F}_K - \mathcal{F}_{K'})(X) &= \mathop{\mathbb{E}}_{\delta,\gamma}\left[\widetilde{A}_K X \widetilde{A}_K^\intercal - \widetilde{A}_{K'} X \widetilde{A}_{K'}^\intercal\right] \\ &= \mathop{\mathbb{E}}_{\delta,\gamma}\left[\widetilde{A}_K X(\widetilde{B}\Delta')^\intercal + (\widetilde{B}\Delta')X\widetilde{A}_K^\intercal - (\widetilde{B}\Delta')X(\widetilde{B}\Delta')^\intercal\right] \\ &= A_K X(B\Delta')^\intercal + (B\Delta')XA_K^\intercal - \mathop{\mathbb{E}}_{\gamma_{tj}}\left[(\widetilde{B}\Delta')X(\widetilde{B}\Delta')^\intercal\right] \\ &= A_K X(B\Delta')^\intercal + (B\Delta')XA_K^\intercal \\ &\quad - (B\Delta')X(B\Delta')^\intercal - \sum_{j=1}^{q}\beta_j(B_j\Delta')X(B_j\Delta')^\intercal. \end{aligned} \tag{14}$$

The operator norm $\|\mathcal{F}_{K'} - \mathcal{F}_K\|$ is

$$\|\mathcal{F}_{K'} - \mathcal{F}_K\| = \|\mathcal{F}_K - \mathcal{F}_{K'}\| = \sup_X \frac{\|(\mathcal{F}_K - \mathcal{F}_{K'})(X)\|}{\|X\|}.$$

Applying submultiplicativity of spectral norm to (14) and noting that $\|\Delta'\| = \|\Delta\|$ completes the proof. ■

Lemma A.3 ($\mathcal{T}_K$ Perturbation): If K and K' are mean-square stabilizing and $\|\mathcal{T}_K\|\|\mathcal{F}_{K'} - \mathcal{F}_K\| \le \frac{1}{2}$, then

$$\begin{aligned}\|(\mathcal{T}_{K'} - \mathcal{T}_K)(\Sigma)\| &\le 2\|\mathcal{T}_K\|\|\mathcal{F}_{K'} - \mathcal{F}_K\|\|\mathcal{T}_K(\Sigma)\| \\ &\le 2\|\mathcal{T}_K\|^2\|\mathcal{F}_{K'} - \mathcal{F}_K\|\|\Sigma\|.\end{aligned}$$

Proof: The proof follows [5] using our modified definitions of $\mathcal{T}_K$ and $\mathcal{F}_K$. ■

Lemma A.4 (Σ_K Trace Bound): If $\rho(\mathcal{F}_K) < 1$, then

$$\operatorname{Tr}(\Sigma_K) \ge \frac{\underline{\sigma}(\Sigma_0)}{1-\rho(\mathcal{F}_K)}.$$

Proof: We have by (5) that

$$\operatorname{Tr}(\Sigma_K) = \operatorname{Tr}(\mathcal{T}_K(\Sigma_0)) = \sum_{t=0}^{\infty}\operatorname{Tr}(\mathcal{F}_K^t(\Sigma_0)).$$

Since $\Sigma_0 \succeq \underline{\sigma}(\Sigma_0)I$, we know the tth term satisfies the inequality $\mathcal{F}_K^t(\Sigma_0) \succeq \underline{\sigma}(\Sigma_0)\mathcal{F}_K^t(I)$; so we have

$$\operatorname{Tr}(\Sigma_K) \ge \underline{\sigma}(\Sigma_0)\sum_{t=0}^{\infty}\operatorname{Tr}(\mathcal{F}_K^t(I)). \tag{15}$$

We have a generic inequality for a sum of n matrices M_i

$$\begin{aligned}\operatorname{Tr}\left[\sum_i^n M_i M_i^\intercal\right] &= \sum_i^n \operatorname{Tr}\left[M_i M_i^\intercal\right] = \sum_i^n \|M_i\|_F^2 \\ &= \sum_i^n \|M_i \otimes M_i\|_F \ge \left\|\sum_i^n M_i \otimes M_i\right\|_F\end{aligned} \tag{16}$$

where the last step is due to the triangle inequality. Recalling the definitions of $\mathcal{F}_K^t(I)$ and $\mathcal{F}_K^t$, we see they are of the form of the LHS and RHS in (16) with all terms matched between $\mathcal{F}_K^t(I)$ and $\mathcal{F}_K^t$ so that the inequality in (16) holds; this can be seen by starting with $t = 1$ and incrementing t up by 1 which will give $(1+p+q)^t$ terms which are all matched. Thus

$$\operatorname{Tr}[\mathcal{F}_K^t(I)] \ge \|\mathcal{F}_K^t\|_F \ge \rho(\mathcal{F}_K)^t.$$

Continuing from (15), we have

$$\operatorname{Tr}(\Sigma_K) \ge \underline{\sigma}(\Sigma_0)\sum_{t=0}^{\infty}\rho(\mathcal{F}_K)^t.$$

By hypothesis $\rho(\mathcal{F}_K) < 1$, and taking the sum of the geometric series completes the proof. ■

Lemma A.5 (Σ_K Perturbation): If K is mean-square stabilizing and $\|\Delta\| \le h_\Delta(K)$ where $h_\Delta(K)$ is the polynomial

$$h_\Delta(K) := \frac{\underline{\sigma}(Q)\underline{\sigma}(\Sigma_0)}{4h_B C(K)\left(\|A_K\|+1\right)}$$

then the associated state covariance matrices satisfy

$$\|\Sigma_{K'} - \Sigma_K\| \le 4\left(\frac{C(K)}{\underline{\sigma}(Q)}\right)^2\frac{\|B\|(\|A_K\|+1)}{\underline{\sigma}(\Sigma_0)}\|\Delta\| \le \frac{C(K)}{\underline{\sigma}(Q)}.$$

Proof: First, since K is mean-square stabilizing and $\|\Delta\| \le h_\Delta(K)$, then K' is also mean-square stabilizing. This follows from an analogous argument in [5] by characterizing mean-square stability in terms of $\rho(\mathcal{F}_K)$ rather than $\rho(A_K)$ and using Lemma 7.4. The rest of the proof follows [5] by using the condition on $\|\Delta\|$, $\|\Sigma_K\| \ge \underline{\sigma}(\Sigma_0)$, and Lemmas 3.6, 7.1, and 7.3. Details are available in [49]. ■

Now we bound the one step progress of policy gradient where we allow the step size to depend explicitly on the current gain matrix iterate K_s.

Lemma A.6 (Gradient Descent, One-Step): Using the policy gradient step update $K_{s+1} = K_s - \eta\nabla C(K_s)$ with step size

$$0 < \eta \le \frac{1}{16}\min\left\{\frac{\left(\frac{\underline{\sigma}(Q)\underline{\sigma}(\Sigma_0)}{C(K)}\right)^2}{h_B\|\nabla C(K)\|(\|A_K\|+1)}, \frac{\underline{\sigma}(Q)}{C(K)\|R_K\|}\right\}$$

gives the one step progress bound

$$\frac{C(K_{s+1}) - C(K^*)}{C(K_s) - C(K^*)} \le 1 - 2\eta\frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)^2}{\|\Sigma_{K^*}\|}.$$

Proof: The gradient update yields $\Delta = -2\eta E_{K_s}\Sigma_{K_s}$. Putting this into Lemma 3.5 gives

$$\begin{aligned}&C(K_{s+1}) - C(K_s) \\ &= 2\operatorname{Tr}\left[\Sigma_{K_{s+1}}\Delta^\intercal E_{K_s}\right] + \operatorname{Tr}\left[\Sigma_{K_{s+1}}\Delta^\intercal R_{K_s}\Delta\right] \\ &\le -4\eta\operatorname{Tr}\left[\Sigma_{K_s}\Sigma_{K_s}E_{K_s}^\intercal E_{K_s}\right] \\ &\quad + 4\eta\frac{\|\Sigma_{K_{s+1}} - \Sigma_{K_s}\|}{\underline{\sigma}(\Sigma_{K_s})}\operatorname{Tr}\left[\Sigma_{K_s}^\intercal E_{K_s}^\intercal E_{K_s}\Sigma_{K_s}\right] \\ &\quad + 4\eta^2\|\Sigma_{K_{s+1}}\|\|R_{K_s}\|\operatorname{Tr}\left[\Sigma_{K_s}\Sigma_{K_s}E_{K_s}^\intercal E_{K_s}\right] \\ &\le -\eta\left(1 - \frac{\|\Sigma_{K_{s+1}} - \Sigma_{K_s}\|}{\underline{\sigma}(\Sigma_0)} - \eta\|\Sigma_{K_{s+1}}\|\|R_{K_s}\|\right) \\ &\quad \times 4\frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)^2}{\|\Sigma_{K^*}\|}(C(K_s) - C(K^*))\end{aligned}$$

where the last step is due to $\underline{\sigma}(\Sigma_0) \le \underline{\sigma}(\Sigma_{K_s})$ and Lemma 3.3. Note that the assumed condition on the step size ensures that the

gain matrix difference satisfies the condition for Lemma A.5 as follows:

$$\begin{aligned}\|\Delta\| &= \eta\|\nabla C(K_s)\| \\ &\le \frac{1}{16}\left(\frac{\underline{\sigma}(Q)\underline{\sigma}(\Sigma_0)}{C(K_s)}\right)^2 \frac{\|\nabla C(K_s)\|}{h_B\|\nabla C(K_s)\|(\|A_{K_s}\|+1)} \\ &\le \frac{1}{4}\left(\frac{\underline{\sigma}(Q)\underline{\sigma}(\Sigma_0)}{C(K_s)}\right)^2 \frac{1}{h_B(\|A_{K_s}\|+1)} \le h_\Delta(K)\end{aligned}$$

where the last inequality is due to Lemma 3.6. Thus, we can indeed apply Lemma A.5, by which we have

$$\frac{\|\Sigma_{K_{s+1}}-\Sigma_{K_s}\|}{\underline{\sigma}(\Sigma_0)} \le \frac{4C(K_s)^2}{\underline{\sigma}(Q)^2\underline{\sigma}(\Sigma_0)^2}\|B\|(\|A_{K_s}\|+1)\|\Delta\| \le \frac{1}{4}$$

where the last inequality is due to using the substitution $\|\Delta\| = \eta\|\nabla C(K_s)\|$ and the hypothesized condition on η. Using this and Lemma 3.6, we have

$$\begin{aligned}\|\Sigma_{K_{s+1}}\| &\le \|\Sigma_{K_{s+1}}-\Sigma_{K_s}\| + \|\Sigma_{K_s}\| \\ &\le \frac{\underline{\sigma}(\Sigma_0)}{4} + \frac{C(K_s)}{\underline{\sigma}(Q)} \le \frac{\|\Sigma_{K_{s+1}}\|}{4} + \frac{C(K_s)}{\underline{\sigma}(Q)}.\end{aligned}$$

Solving for $\|\Sigma_{K_{s+1}}\|$ gives $\|\Sigma_{K_{s+1}}\| \le \frac{4}{3}\frac{C(K_s)}{\underline{\sigma}(Q)}$; so

$$\begin{aligned}&1 - \frac{\|\Sigma_{K_{s+1}}-\Sigma_{K_s}\|}{\underline{\sigma}(\Sigma_0)} - \eta\|\Sigma_{K_{s+1}}\|\|R_{K_s}\| \\ &\ge 1 - \frac{1}{4} - \eta\frac{4}{3}\frac{C(K_s)}{\underline{\sigma}(Q)}\|R_{K_s}\| \ge 1 - \frac{1}{4} - \frac{4}{3}\cdot\frac{1}{16} = \frac{2}{3} \ge \frac{1}{2}\end{aligned}$$

where the second-to-last inequality used the hypothesized condition on η. Therefore

$$\frac{C(K_{s+1})-C(K_s)}{C(K_s)-C(K^*)} \le -2\eta\frac{\underline{\sigma}(R)\underline{\sigma}(\Sigma_0)^2}{\|\Sigma_{K^*}\|}.$$

Adding 1 to both sides completes the proof. ■

Lemma A.7 (Cost Difference Lower Bound): The following cost difference inequality holds:

$$C(K) - C(K^*) \ge \frac{\underline{\sigma}(\Sigma_0)}{\|R_K\|}\operatorname{Tr}(E_K^\intercal E_K).$$

Proof: The proof follows that for an analogous condition located in the gradient domination lemma in [5]. ■

Lemma A.8: The following inequalities hold:

$$\|\nabla C(K)\| \le \|\nabla C(K)\|_F \le h_1(K) \quad \text{and} \quad \|K\| \le h_2(K)$$

where $h_0(K)$, $h_1(K)$, and $h_2(K)$ are the polynomials

$$h_0(K) := \sqrt{\frac{\|R_K\|(C(K)-C(K^*))}{\underline{\sigma}(\Sigma_0)}},$$

$$h_1(K) := 2\frac{C(K)h_0(K)}{\underline{\sigma}(Q)}, \quad h_2(K) := \frac{h_0(K)+\|B^\intercal P_K A\|}{\underline{\sigma}(R)}.$$

Proof: The proof follows [5] with R_K defined here. ■

We now give the parameter and proof of global convergence of policy gradient descent in Theorem 4.3.

Theorem A.9 (Policy Gradient Convergence): Consider the assumptions and notations of Theorem 4.3 and define

$$c_{\text{pg}} := \frac{1}{16}\min\left\{\frac{\left(\frac{\underline{\sigma}(Q)\underline{\sigma}(\Sigma_0)}{C(K_0)}\right)^2}{h_B\overline{h_1}(\|A\|+\overline{h_2}\|B\|+1)}, \frac{\underline{\sigma}(Q)}{C(K_0)\overline{\|R_K\|}}\right\}$$

$$\overline{h_1} := \max_K h_1(K) \text{ subject to } C(K) \le C(K_0)$$

$$\overline{h_2} := \max_K h_2(K) \text{ subject to } C(K) \le C(K_0),$$

$$\overline{\|R_K\|} := \max_K \|R_K\| \text{ subject to } C(K) \le C(K_0).$$

Then the claim of Theorem 4.3 holds.

Proof: We have by Weyl's inequality for singular values [49], submultiplicativity of spectral norm, and Lemma 7.8 that

$$\begin{aligned}&\|B\|\|\nabla C(K)\|(\|A+BK\|+1) \\ &\quad\le \|B\|\|\nabla C(K)\|(\|A\|+\|B\|\|K\|+1) \\ &\quad\le \|B\|h_1(K)(\|A\|+\|B\|h_2(K)+1).\end{aligned}$$

Thus, by choosing $0 < \eta \le c_{\text{pg}}$, we satisfy the requirements for Lemma 7.6 at $s = 1$, which implies that progress is made at $s = 1$, i.e., that $C(K_1) \le C(K_0)$ according to the rate in Lemma 7.6. Proceeding inductively and applying Lemma 7.6 at each step completes the proof. ■

Remark A.10: The quantities $\overline{h_1}$, $\overline{h_2}$, and $\overline{\|R_K\|}$ may be upper bounded by quantities that depend only on problem data and $C(K_0)$, e.g., using the cost bounds in Lemma 3.6, which we omit for brevity; so a conservative minimum step size η may be computed exactly.

APPENDIX B
MODEL-FREE POLICY GRADIENT DESCENT

This lemma shows that $C(K)$ and Σ_K can be estimated with arbitrarily high accuracy as the rollout length ℓ increases.

Lemma B.1 (Approximating $C(K)$ and Σ_K With Infinitely Many Finite Horizon Rollouts): Suppose K gives finite $C(K)$. Define the finite-horizon estimates

$$\Sigma_K^{(\ell)} := \mathbb{E}\left[\sum_{i=0}^{\ell-1} x_i x_i^\intercal\right], \; C^{(\ell)}(K) := \mathbb{E}\left[\sum_{i=0}^{\ell-1} x_i^\intercal Q x_i + u_i^\intercal R u_i\right]$$

where expectation is with respect to $x_0, \{\delta_{ti}\}, \{\gamma_{tj}\}$. Let ϵ be an arbitrary small constant. Then the following hold:

$$\ell \ge \overline{h_\ell}(\epsilon) := \frac{n\cdot C^2(K)}{\epsilon\underline{\sigma}(\Sigma_0)\underline{\sigma}^2(Q)} \Rightarrow \|\Sigma_K^{(\ell)} - \Sigma_K\| \le \epsilon$$

$$\ell \ge h_\ell(\epsilon) := \overline{h_\ell}(\epsilon)\|Q_K\| \quad \Rightarrow |C^{(\ell)}(K) - C(K)| \le \epsilon.$$

Proof: The proof follows [5] exactly using suitably modified definitions of $C(K)$, $\mathcal{T}_K$, $\mathcal{F}_K$. ■

Next we bound cost and gradient perturbations in terms of gain matrix perturbations and problem data. Using the same restriction as in Lemma A.5, we have Lemmas B.2 and B.3.

Lemma B.2 ($C(K)$ Perturbation): If $\|\Delta\| \le h_\Delta(K)$, then the cost difference is bounded as

$$|C(K') - C(K)| \le h_{\text{cost}}(K)C(K)\|\Delta\|$$

where $h_{\text{cost}}(K)$ is the polynomial

$$\begin{aligned}h_{\text{cost}}(K) := \frac{4\operatorname{Tr}(\Sigma_0)\|R\|}{\underline{\sigma}(\Sigma_0)\underline{\sigma}(Q)}\Bigg(&\|K\| + \frac{h_\Delta(K)}{2} + \|B\|\|K\|^2 \\ &\times (\|A_K\|+1)\frac{C(K)}{\underline{\sigma}(\Sigma_0)\underline{\sigma}(Q)}\Bigg).\end{aligned}$$

Proof: The proof follows [5] using suitably modified definitions of $C(K)$, $\mathcal{T}_K$, $\mathcal{F}_K$; however, compared with [5], we terminate the proof bound earlier so as to avoid a degenerate bound in the case of $K = 0$, and we also correct typographical errors. Note that $\|\Delta\|$ has a more restrictive upper bound due to the multiplicative noise. ■

Lemma B.3 ($\nabla C(K)$ Perturbation): If $\|\Delta\| \leq h_\Delta(K)$, then the policy gradient difference is bounded as

$$\|\nabla C(K') - \nabla C(K)\| \leq h_{\text{grad}}(K)\|\Delta\|$$

$$\text{and } \|\nabla C(K') - \nabla C(K)\|_F \leq h_{\text{grad}}(K)\|\Delta\|_F$$

$$\text{where } h_{\text{grad}}(K) :=$$

$$4\left(\frac{C(K)}{\underline{\sigma}(Q)}\right)\left[\|R\| + \|B\|\left(\|A\| + h_B(\|K\| + h_\Delta(K))\right)\right.$$

$$\left.\times\left(\frac{h_{\text{cost}}(K)C(K)}{\text{Tr}(\Sigma_0)}\right) + h_B\|B\|\left(\frac{C(K)}{\underline{\sigma}(\Sigma_0)}\right)\right]$$

$$+ 8\left(\frac{C(K)}{\underline{\sigma}(Q)}\right)^2\left(\frac{\|B\|(\|A_K\|+1)}{\underline{\sigma}(\Sigma_0)}\right)h_0(K).$$

Proof: The proof generally follows [5] using Lemmas A.5, B.2, and 7.7 with R_K and E_K modified appropriately, with details available in [49]. ■

As in [5], in the model-free setting, we apply Frobenius-norm ball smoothing to the cost. Let $\mathbb{S}_r$ be the uniform distribution over all matrices with Frobenius norm r (the boundary of the ball), and $\mathbb{B}_r$ be the uniform distribution over all matrices with Frobenius norm at most r (the entire ball). The smoothed cost is

$$C_r(K) = \mathbb{E}_{U\sim\mathbb{B}_r}[C(K+U)]$$

where U is a random matrix with the same dimensions as K and Frobenius norm r. The following lemma shows that the gradient of the smoothed function can be estimated just with an oracle of the function value.

Lemma B.4 (Zeroth-Order Gradient Estimation): The gradient of the smoothed cost is related to the unsmoothed cost by

$$\nabla C_r(K) = \frac{mn}{r^2}\mathbb{E}_{U\sim\mathbb{S}_r}[C(K+U)U].$$

Proof: The result is proved in [55, Lemma 2.1]. ■

Lemma B.4 shows that the gradient of the smoothed cost can be found exactly with infinitely many infinite-horizon rollouts. Much of the remaining proofs goes toward showing that the error between the gradient of the smoothed cost and the unsmoothed cost, the error due to using finite-horizon rollouts, and the error due to using finitely many rollouts can all be bounded by polynomials of the problem data. As noted by [5] the reason for smoothing in a Frobenius norm ball rather than over a Gaussian distribution is to ensure stability and finiteness of the cost of every gain within the smoothing domain, although now in the multiplicative noise case, we must be even more restrictive about our choice of perturbation on K because we require not only mean stability but also mean-square stability.

We now give a Bernstein inequality for random matrices; this allows us to bound the difference between the sample average of a random matrix and its expectation.

Lemma B.5 (Matrix Bernstein Inequality):

Let $\{Z_i\}_{i=1}^N$ be a set of N independent random matrices of dimension $d_1 \times d_2$ with $\mathbb{E}[Z_i] = Z$, $\|Z_i - Z\| \leq R_Z$ almost surely, and maximum variance $\max(\|\mathbb{E}(Z_iZ_i^\intercal) - ZZ^\intercal\|, \|\mathbb{E}(Z_i^\intercal Z_i) - Z^\intercal Z\|) \leq \sigma_Z^2$, and sample average $\widehat{Z} := \frac{1}{N}\sum_{i=1}^N Z_i$. Let a small tolerance $\epsilon \geq 0$ and small probability $0 \leq \mu \leq 1$ be given. If

$$N \geq \frac{2\min(d_1,d_2)}{\epsilon^2}\left(\sigma_Z^2 + \frac{R_Z\epsilon}{3\sqrt{\min(d_1,d_2)}}\right)\log\left[\frac{d_1+d_2}{\mu}\right]$$

then $\mathbb{P}[\|\widehat{Z} - Z\|_F \leq \epsilon] \geq 1-\mu$.

Proof: The lemma follows readily from the matrix Bernstein inequality in [56] by simple variable substitutions, rearrangement, and the bound $\|M\|_F \leq \sqrt{\min(d_1,d_2)}\|M\|$. ■

Lemma B.6 (Estimating $\nabla C(K)$ With Finitely Many Infinite-Horizon Rollouts): Given an arbitrary tolerance ϵ and probability μ, suppose the exploration radius r is chosen as

$$r \leq h_r\left(\frac{\epsilon}{2}\right) := \min\left\{h_\Delta(K), \frac{1}{h_{\text{cost}}(K)}, \frac{\epsilon}{2h_{\text{grad}}(K)}\right\}$$

and the number of samples n_{sample} of $U_i \sim \mathbb{S}_r$ is chosen as

$$n_{\text{sample}} \geq h_{\text{sample}}\left(\frac{\epsilon}{2},\mu\right)$$

$$:= \frac{8\min(m,n)}{\epsilon^2}\left(\sigma_{\widehat{\nabla}}^2 + \frac{R_{\widehat{\nabla}}\epsilon}{6\sqrt{\min(m,n)}}\right)\log\left[\frac{m+n}{\mu}\right]$$

$$R_{\widehat{\nabla}} := \frac{2mnC(K)}{r} + \frac{\epsilon}{2} + h_1(K)$$

$$\sigma_{\widehat{\nabla}}^2 := \left(\frac{2mnC(K)}{r}\right)^2 + \left(\frac{\epsilon}{2} + h_1(K)\right)^2.$$

Then, with high probability of at least $1-\mu$, the estimate

$$\hat{\nabla}C(K) = \frac{1}{n_{\text{sample}}}\sum_{i=1}^{n_{\text{sample}}}\frac{mn}{r^2}C\left(K+U_i\right)U_i$$

satisfies the error bound $\|\hat{\nabla}C(K) - \nabla C(K)\|_F \leq \epsilon$.

Proof: First note that $\|K' - K\|_F = \|\Delta\|_F = \|U\|_F = r$. We break the difference between estimated and true gradient $\hat{\nabla}C(K) - \nabla C(K)$ into two terms as

$$\left(\nabla C_r(K) - \nabla C(K)\right) + \left(\hat{\nabla}C(K) - \nabla C_r(K)\right). \tag{17}$$

Since $r \leq h_\Delta(K)$, we see that Lemmas B.2 and B.3 hold. By enforcing the bound $r \leq \frac{1}{h_{\text{cost}}(K)}$, by Lemma B.2 and noting that $\|\Delta\| \leq \|\Delta\|_F$, we have

$$|C(K+U) - C(K)| \leq C(K) \to C(K+U) \leq 2C(K). \tag{18}$$

This ensures stability of the system under the perturbed gains so that $C(K+U)$ is well-defined. For the first term $\nabla C_r(K) - \nabla C(K)$, by enforcing $r \leq \frac{\epsilon}{2h_{\text{grad}}(K)}$, by Lemma B.3, we have

$$\|\nabla C(K+U) - \nabla C(K)\|_F \leq \frac{\epsilon}{2}.$$

Since $\nabla C_r(K)$ is the expectation of $\nabla C(K+U)$, by the triangle inequality, we have

$$\|\nabla C_r(K) - \nabla C(K)\|_F \leq \frac{\epsilon}{2}. \tag{19}$$

For the second term $\hat{\nabla}C(K) - \nabla C_r(K)$, we work toward using the matrix Bernstein inequality and adopt the notation of the associated lemma. First note that by Lemma B.4, we have $Z := \nabla C_r(K) = \mathbb{E}[\hat{\nabla}C(K)]$. Each individual sample $Z_i := (\frac{mn}{r^2})C(K+U_i)U_i$ has the bounded Frobenius norm

$$\|Z_i\|_F = \left\|\left(\frac{mn}{r^2}\right)C(K+U_i)U_i\right\|_F = \frac{mnC(K+U_i)\|U_i\|_F}{r^2}$$

$$= \frac{mnC(K+U_i)r}{r^2} = \frac{mnC(K+U_i)}{r} \leq \frac{2mnC(K)}{r}.$$

Next, from (19) and Lemma 7.8, we have

$$\|Z\|_F = \|\nabla C_r(K)\|_F \le \frac{\epsilon}{2} + \|\nabla C(K)\|_F \le \frac{\epsilon}{2} + h_1(K).$$

So by the triangle inequality, each sample difference has the bounded Frobenius norm

$$\|Z_i - Z\|_F \le \|Z_i\|_F + \|Z\|_F \le \frac{2mnC(K)}{r} + \frac{\epsilon}{2} + h_1(K).$$

Using (18) and $\|U_i\|_F \le r$, the variance of the differences is likewise bounded as

$$\begin{aligned}\|\mathbb{E}(Z_i Z_i^\intercal) - ZZ^\intercal\| &\le \|\mathbb{E}(Z_i Z_i^\intercal)\|_F + \|ZZ^\intercal\|_F \\ &\le \max_{Z_i} (\|Z_i\|_F)^2 + \|Z\|_F^2 \\ &\le \left(\frac{2mnC(K)}{r}\right)^2 + \left(\frac{\epsilon}{2} + h_1(K)\right)^2.\end{aligned}$$

An identical argument holds for $\|\mathbb{E}(Z_i^\intercal Z_i) - Z^\intercal Z\|$; so the assumed choice of $\sigma_{\widehat{\nabla}}^2$ is valid. Thus, using the assumed number of samples $n_{\text{sample}} \ge h_{\text{sample}}$ satisfies the condition of the matrix Bernstein inequality, and thus with high probability of at least $1-\mu$, we have

$$\|\hat{\nabla} C(K) - \mathbb{E}[\hat{\nabla} C(K)]\|_F = \|\hat{\nabla} C(K) - \nabla C_r(K)\|_F \le \frac{\epsilon}{2}.$$

Adding the bounds on the two terms in (17) and using the triangle inequality completes the proof. ■

Lemma B.7 (Estimating $\nabla_K C(K)$ With Finitely Many Finite-Horizon Rollouts): Given an arbitrary tolerance ϵ and probability μ, suppose the exploration radius r is chosen as

$$r \le h_r\left(\frac{\epsilon}{4}\right) = \min\left\{h_\Delta(K), \frac{1}{h_{\text{cost}}(K)}, \frac{\epsilon}{4h_{\text{grad}}(K)}\right\}$$

and the rollout length ℓ is chosen as

$$\ell \ge h_\ell\left(\frac{r\epsilon}{4mn}\right) = \frac{4mn^2C^2(K)\left(\|Q\| + \|R\|\|K\|^2\right)}{r\epsilon\underline{\sigma}(\Sigma_0)\underline{\sigma}^2(Q)}.$$

Suppose that the distribution of the initial states is such that $x_0 \sim \mathcal{P}_0$ implies $\|x_0^i\| \le L_0$ almost surely for any given realization x_0^i of x_0. Suppose additionally that the multiplicative noises are distributed such that the following bound is satisfied almost surely under the closed-loop dynamics with any gain $K + U_i$ where $\|U_i\| \le r$ for any given realized sequence x_t^i of x_t with a positive scalar $z \ge 1$

$$\sum_{t=0}^{\ell-1}\left(x_t^{i\intercal} Q x_t^i + u_t^{i\intercal} R u_t^i\right) \le z\mathop{\mathbb{E}}_{\delta,\gamma}\left[\sum_{t=0}^{\ell-1}\left(x_t^\intercal Q x_t + u_t^\intercal R u_t\right)\right].$$

Suppose the number n_{sample} of $U_i \sim \mathbb{S}_r$ is chosen as

$$\begin{aligned}n_{\text{sample}} &\ge h_{\text{sample,trunc}}\left(\frac{\epsilon}{4}, \mu, \frac{L_0^2}{\underline{\sigma}(\Sigma_0)}, z\right) \\ &:= \frac{32\min(m,n)}{\epsilon^2}\left(\sigma_\nabla^2 + \frac{R_{\widehat{\nabla}}\epsilon}{12\sqrt{\min(m,n)}}\right)\log\left[\frac{m+n}{\mu}\right]\end{aligned}$$

where

$$R_{\widetilde{\nabla}} := \frac{2mnzL_0^2C(K)}{r\underline{\sigma}(\Sigma_0)} + \frac{\epsilon}{2} + h_1(K)$$

$$\sigma_{\widetilde{\nabla}}^2 := \left(\frac{2mnzL_0^2C(K)}{r\underline{\sigma}(\Sigma_0)}\right)^2 + \left(\frac{\epsilon}{2} + h_1(K)\right)^2.$$

The finite-horizon estimate of the cost is defined as

$$\hat{C}(K + U_i) := \sum_{t=0}^{\ell-1}\left(x_t^{i\intercal} Q x_t^i + u_t^{i\intercal} R u_t^i\right)$$

under the closed-loop dynamics with gain $K + U_i$. Then with high probability of at least $1-\mu$, the estimated gradient

$$\widetilde{\nabla} C(K) := \frac{1}{n_{\text{sample}}}\sum_{i=1}^{n_{\text{sample}}}\frac{mn}{r^2}\hat{C}(K + U_i)U_i$$

satisfies the error bound $\|\widetilde{\nabla} C(K) - \nabla C(K)\|_F \le \epsilon$.

Proof: Similar to before, we break the difference between estimated and true gradient into three terms as

$$\widetilde{\nabla} C(K) - \nabla C(K) = (\widetilde{\nabla} - \nabla') + (\nabla' - \hat{\nabla}) + (\hat{\nabla} - \nabla)$$

$$\text{where}\quad \nabla' C(K) = \frac{1}{n_{\text{sample}}}\sum_{i=1}^{n_{\text{sample}}}\frac{mn}{r^2}C^{(\ell)}(K + U_i)U_i$$

and $\hat{\nabla} C(K)$ is defined as in Lemma B.6. The third term is handled by Lemma B.6. Note that since $\|x_0^i\| \le L_0$, we have $\underline{\sigma}(\Sigma_0) \le \overline{\sigma}(\Sigma_0) \le L_0^2$ so $\frac{L_0^2}{\underline{\sigma}(\Sigma_0)} \ge 1$. Similarly, $z \ge 1$, and thus

$$h_{\text{sample,trunc}}\left(\frac{\epsilon}{4}, \mu, \frac{L_0^2}{\underline{\sigma}(\Sigma_0)}, z\right) \ge h_{\text{sample}}\left(\frac{\epsilon}{4}, \mu\right).$$

Therefore, the choice of r and n_{sample} satisfy the conditions of Lemma B.6; so with high probability of at least $1-\mu$

$$\|\hat{\nabla} C(K) - \nabla C(K)\|_F \le \frac{\epsilon}{4}. \tag{20}$$

For the second term, by using the choices $\ell \ge h_\ell(\frac{r\epsilon}{4mn})$ and $C(K + U_i) \le 2C(K)$, Lemma B.1 holds and implies that

$$\left\|C^{(\ell)}(K + U_i) - C(K + U_i)\right\|_F \le \frac{r\epsilon}{4mn}.$$

By the triangle inequality, submultiplicativity, and $\|U_i\|_F \le r$

$$\begin{aligned}&\left\|\frac{1}{n_{\text{sample}}}\sum_{i=1}^{n_{\text{sample}}}\frac{mn}{r^2}\left[C^{(\ell)}(K + U_i) - C(K + U_i)\right]U_i\right\|_F \\ &= \|\nabla'_K C(K) - \hat{\nabla} C(K)\|_F \le \frac{\epsilon}{4}.\end{aligned} \tag{21}$$

For the first term, $\|x_0^i\| \le L_0$ implies $\frac{L_0^2}{\underline{\sigma}(\Sigma_0)}\Sigma_0 \succeq x_0^i x_0^{i\intercal}$. Applying this to the cost, summing over time and using the assumed restriction on the multiplicative noise, we have

$$\begin{aligned}\frac{2zL_0^2C(K)}{\underline{\sigma}(\Sigma_0)} &\ge \frac{zL_0^2}{\underline{\sigma}(\Sigma_0)}C(K + U_i) \\ &\ge z\mathop{\mathbb{E}}_{\delta,\gamma}\left[\sum_{t=0}^{\infty}\left(x_t^{i\intercal} Q x_t^i + u_t^{i\intercal} R u_t^i\right)\right] \\ &\ge \sum_{t=0}^{\ell-1}\left(x_t^{i\intercal} Q x_t^i + u_t^{i\intercal} R u_t^i\right).\end{aligned}$$

Using this and an argument identical to Lemma B.6, each sample $Z_i := (\frac{mn}{r^2})\hat{C}(K + U_i)U_i$ has bounded Frobenius norm

$$\|Z_i\|_F = \left\|\left(\frac{mn}{r^2}\right)\hat{C}(K + U_i)U_i\right\|_F \le \frac{2mnzL_0^2C(K)}{r\underline{\sigma}(\Sigma_0)}.$$

By (20) and (21), we have for $Z := \mathbb{E}[\widetilde{\nabla} C(K)] = \nabla'_K C(K)$

$$\begin{aligned}\|Z\|_F = \|\nabla'_K C(K)\|_F &\le \frac{\epsilon}{4} + \|\hat{\nabla} C(K)\|_F \\ &\le \frac{\epsilon}{4} + \frac{\epsilon}{4} + \|\nabla_K C(K)\|_F \le \frac{\epsilon}{2} + h_1(K).\end{aligned}$$

Using arguments identical to Lemma B.6, we obtain the bounds on the sample difference $R_{\widetilde{\nabla}} = \|Z_i - Z\|_F$ and variance $\sigma_{\widetilde{\nabla}}^2$

given in the assumption. Thus, the polynomial $h_{\text{sample,trunc}}$ is large enough so that the matrix Bernstein inequality implies

$$\|\widetilde{\nabla} C(K) - \nabla'_K C(K)\|_F \leq \frac{\epsilon}{4}$$

with high probability $1-\mu$. Adding the three terms together and using the triangle inequality completes the proof. ■

We now give the parameters and proof of high-probability global convergence in Theorem 5.1.

Theorem B.8 (Model-Free Policy Gradient Convergence): Consider the assumptions and notations of Theorem 5.1 where the number of samples n_{sample}, rollout length ℓ, and exploration radius r are chosen according to the fixed quantities

$$r \geq h_{r,\text{GD}} := h_r\left(\frac{\epsilon'}{4}\right), \quad \ell \geq h_{\ell,\text{GD}} := h_\ell\left(\frac{r\epsilon'}{4mn}\right)$$

$$n_{\text{sample}} \geq h_{\text{sample,GD}} := h_{\text{sample,trunc}}\left(\frac{\epsilon'}{4}, \frac{\mu}{N}, \frac{L_0^2}{\underline{\sigma}(\Sigma_0)}, z\right)$$

where

$$\epsilon' := \min\left\{\frac{\underline{\sigma}(\Sigma_0)^2\underline{\sigma}(R)}{\|\Sigma_{K^*}\|\, C(K_0)\overline{h_{\text{cost}}}} \cdot \epsilon, \frac{\overline{h_\Delta}}{\eta}\right\}$$

$$\overline{h_{\text{cost}}} := \max_K\ h_{\text{cost}}(K) \text{ subject to } C(K) \leq 2C(K_0)$$

$$\overline{h_\Delta} := \min_K\ h_\Delta(K) \text{ subject to } C(K) \leq 2C(K_0).$$

Then the claim of Theorem 5.1 holds.

Proof: The proof follows [5] using the polynomials defined in our theorem. The last part of the proof is the same as in Theorem 4.3. As noted by [5], the monotonic decrease in the function value during gradient descent and the choice of exploration radius r are sufficient to ensure that all cost values encountered throughout the entire algorithm are bounded by $2C(K_0)$, ensuring that all polynomial quantities used are bounded as well. We also require $\epsilon' \leq \frac{\overline{h_\Delta}}{\eta}$ in order for $\|\Delta\| = \eta\|\widetilde{\nabla} C(K) - \nabla_K C(K))\|$ to satisfy the condition of Lemma B.2, which was neglected by [5]. ■

Remark B.9: As in Remark 7.10, the quantities $\overline{h_{\text{cost}}}$ and $\overline{h_\Delta}$ may be upper (lower) bounded by quantities that depend on problem data and $C(K_0)$; so a conservative minimum exploration radius r, number of rollouts n_{sample}, and rollout length ℓ can be computed exactly in terms of problem data. Looking back across the terms that feed into the step size, number of rollouts, rollout length, and exploration radius, we see $C(K)$, $\|\Sigma_K\|$, $\|P_K\|$, and $\|B\|^2 + \sum_{j=1}^q \beta_j \|B_j\|^2$ are necessarily greater with state- and/or input-dependent multiplicative noise, and, thus, the algorithmic parameters are worsened by the noise.

REFERENCES

[1] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[2] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, 2015, Art. no. 529.

[3] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annu. Rev. Control, Robot., Autonomous Syst.*, vol. 2, pp. 253–279, 2018.

[4] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," 2017, *arXiv:1710.01688*.

[5] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. 35th Int. Conf. Mach. Learning*, in Proceedings of Machine Learning Research, vol. 80, Jul. 10–15, 2018, pp. 1467–1476.

[6] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. Amer. Control Conf.*, vol. 3, 1994, pp. 3475–3479.

[7] C.-N. Fiechter, "PAC adaptive control of linear systems," in *Proc. 10th Annu. Conf. Comput. Learning Theory*, in COLT '97, New York, NY, USA: ACM, 1997, pp. 72–80.

[8] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.

[9] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

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

[11] M. Abeille and A. Lazaric, "Improved regret bounds for thompson sampling in linear quadratic control problems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–9.

[12] J. Umenberger and T. B. Schön, "Learning convex bounds for linear quadratic control policy synthesis," in *Proc. Adv. Neural Inform. Process. Syst.*, 2018, pp. 9561–9572.

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

[14] H. K. Venkataraman and P. J. Seiler, "Recovering robustness in model-free reinforcement learning," 2018, *arXiv:1810.09337*.

[15] W. M. Wonham, "Optimal stationary control of a linear system with state-dependent noise," *SIAM J. Control*, vol. 5, no. 3, pp. 486–500, 1967.

[16] T. Damm, *Rational Matrix Equations in Stochastic Control*, vol. 297. Berlin, Germany: Springer, 2004.

[17] J. L. Willems and J. C. Willems, "Feedback stabilizability for stochastic systems with state and control dependent noise," *Automatica*, vol. 12, no. 3, pp. 277–283, 1976.

[18] M. Athans, R. Ku, and S. Gershwin, "The uncertainty threshold principle: Some fundamental limitations of optimal decision making under dynamic uncertainty," *IEEE Trans. Autom. Control*, vol. AC-22, no. 3, pp. 491–495, Jun. 1977.

[19] D. Bernstein, "Robust static and dynamic output-feedback stabilization: Deterministic and stochastic perspectives," *IEEE Trans. Autom. Control*, vol. AC-32, no. 12, pp. 1076–1084, Dec. 1987.

[20] R. Ku and M. Athans, "Further results on the uncertainty threshold principle," *IEEE Trans. Autom. Control*, vol. 22, no. 5, pp. 866–868, Oct. 1977.

[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Scientific, 1995.

[22] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.

[23] P. Antsaklis and J. Baillieul, "Special issue on technology of networked control systems," *Proc. IEEE*, vol. 95, no. 1, pp. 5–8, Jan. 2007.

[24] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proc. IEEE*, vol. 95, no. 1, pp. 138–162, Jan. 2007.

[25] J. M. Carrasco *et al.*, "Power-electronic systems for the grid integration of renewable energy sources: A survey," *IEEE Trans. Ind. Electron.*, vol. 53, no. 4, pp. 1002–1016., Jun. 2006.

[26] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič, "Foundations and challenges of low-inertia systems," in *Proc. Power Syst. Comput. Conf.*, 2018, pp. 1–25.

[27] J. L. Lumley, *Stochastic Tools in Turbulence*. Chelmsford, MA, USA: Courier Corporation, 2007.

[28] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature Neurosci.*, vol. 20, no. 3, 2017, Art. no. 340.

[29] L. El Ghaoui, "State-feedback control of systems with multiplicative noise via linear matrix inequalities," *Syst. Control Lett.*, vol. 24, no. 3, pp. 223–228, 1995.

[30] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1994.

[31] W. Li, E. Todorov, and R. E. Skelton, "Estimation and control of systems with multiplicative noise via linear matrix inequalities," in *Proc. Amer. Control Conf.,* 2005, pp. 1811–1816.

[32] D. Hinrichsen and A. J. Pritchard, "Stochastic H^∞," *SIAM J. Control Optim.*, vol. 36, no. 5, pp. 1504–1538, 1998.

[33] B. Bamieh and M. Filo, "An input-output approach to structured stochastic uncertainty," 2018, *arXiv:1806.07473*.
[34] B. Gravell, P. Esfahani, and T. Summers, "Robust control design for linear systems via multiplicative noise," in *Proc. 18th IFAC World Congress*, to be published.
[35] P. Benner and T. Damm, "Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems," *SIAM J. Control Optim.*, vol. 49, no. 2, pp. 686–711, 2011.
[36] M. Schuurmans, P. Sopasakis, and P. Patrinos, "Safe learning-based control of stochastic jump linear systems: A distributionally robust approach," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 6498–6503.
[37] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for Markovian jump linear systems," in *Proc. IEEE Amer. Control Conf.*, 2020, pp. 2882–2887.
[38] G. R. Gonçalves da Silva, A. S. Bazanella, C. Lorenzini, and L. Campestrini, "Data-driven LQR control design," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 180–185, Jan. 2019.
[39] G. Baggio, V. Katewa, and F. Pasqualetti, "Data-driven minimum-energy controls for linear systems," *IEEE Control Syst. Lett.*, vol. 3, no. 3, pp. 589–594, Jul. 2019.
[40] T. Maupong and P. Rapisarda, "Data-driven control: A behavioral approach," *Syst. Control Lett.*, vol. 101, pp. 37–43, 2017.
[41] C. D. Persis and P. Tesi, "On persistency of excitation and formulas for data-driven control," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 873–878.
[42] J.-N. Juang and R. S. Pappa, "An eigensystem realization algorithm for modal parameter identification and model reduction," *J. Guidance Control Dyn.*, vol. 8, no. 5, pp. 620–627, 1985.
[43] Y. Xing, B. Gravell, X. He, K. H. Johansson, and T. Summers, "Linear system identification under multiplicative noise from multiple trajectory data," in *Proc. IEEE 2020 Amer. Control Conf.*, 2020, pp. 5157–5261.
[44] F. Kozin, "A survey of stability of stochastic systems," *Automatica*, vol. 5, no. 1, pp. 95–112, 1969.
[45] G. Freiling and A. Hochhaus, "Properties of the solutions of rational matrix difference equations," *Comput. Math. Appl.*, vol. 45, no. 6, pp. 1137–1154, 2003.
[46] I. G. Ivanov, "Properties of Stein (Lyapunov) iterations for solving a general Riccati equation," *Nonlinear Anal.: Theory, Methods Appl.*, vol. 67, no. 4, pp. 1155–1166, 2007.
[47] J. Peters and S. Schaal, "Policy gradient methods for robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 2219–2225.
[48] B. Polyak, "Gradient methods for the minimisation of functionals," *USSR Comput. Math. Math. Phys.*, vol. 3, no. 4, pp. 864–878, 1963.
[49] B. Gravell, P. M. Esfahani, and T. Summers, "Learning robust control for LQR systems with multiplicative noise via policy gradient," 2019, *arXiv:1905.13547*.
[50] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition," in *Machine Learning and Knowledge Discovert in Databases*. Cham, Switzerland: Springer International Publishing, 2016, pp. 795–811.
[51] S. M. Kakade, "A natural policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1531–1538.
[52] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Trans. Autom. Control*, vol. AC-16, no. 4, pp. 382–384, Aug. 1971.
[53] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[54] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," 2019, *arXiv:1907.08921*.
[55] A. D. Flaxman, A. T. Kalai, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms*, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2005, pp. 385–394.
[56] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, Aug 2012.

Benjamin Gravell (Graduate Student Member, IEEE) received the B.S. degree in mechanical engineering from The University of Texas at Dallas, Richardson, TX, USA, in 2017. He is currently working toward the Ph.D. degree in mechanical engineering with a specialization in control of dynamical systems at The University of Texas at Dallas.

His research interests are in data-driven optimal and robust control of uncertain dynamical systems with applications to networked robotics.

Peyman Mohajerin Esfahani received the B.Sc. and M.Sc. degrees in electrical engineering at Sharif University of Technology, Iran, in 2005 and 2007, respectively, and the PhD degree in control from ETH Zurich, Switzerland, in 2014.

He is currently an Assistant Professor with the Delft Center for Systems and Control, Delft University of Technology, The Netherlands. Prior to joining TU Delft, he held several research appointments at EPFL, ETH Zurich, and MIT between 2014 and 2016. His research interests include theoretical and practical aspects of decision-making problems in uncertain and dynamic environments, with applications to control and security of large-scale and distributed systems.

Dr. Esfahani was one of the three finalists for the Young Researcher Prize in Continuous Optimization awarded by the Mathematical Optimization Society in 2016, and a recipient of the 2016 George S. Axelby Outstanding Paper Award from the IEEE Control Systems Society. He also received the ERC Starting Grant and the INFORMS Frederick W. Lanchester Prize in 2020.

Tyler Summers (Member, IEEE) received the B.S. degree in mechanical engineering from Texas Christian University, Fort Worth, TX, USA, in 2004, and the M.S. and Ph.D. degrees in aerospace engineering with emphasis on feedback control theory from the University of Texas at Austin, Austin, TX, USA, in 2007 and 2010, respectively.

He is an Assistant Professor of Mechanical Engineering with an affiliate appointment in Electrical Engineering with The University of Texas at Dallas, Dallas, TX, USA. Prior to joining UT Dallas, he was an ETH Postdoctoral Fellow with the Automatic Control Laboratory, ETH Zurich, Zurich, Switzerland, from 2011 to 2015. He was a Fulbright Postgraduate Scholar with the Australian National University, Canberra, Australia from 2007 to 2008. His research interests are in feedback control and optimization in complex dynamical networks, with applications to electric power networks and distributed robotics.