2025 IEEE 64th Conference on Decision and Control (CDC)
December 10-12, 2025. Rio de Janeiro, Brazil

Policy Gradient for LQR with Domain Randomization

Tesshu Fujinami[1], Bruce D. Lee, Nikolai Matni, George J. Pappas

2025 IEEE 64th Conference on Decision and Control (CDC) | 979-8-3315-2627-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CDC57313.2025.11312642

***Abstract*— Domain randomization (DR) enables sim-to-real transfer by training controllers on a distribution of simulated environments, with the goal of achieving robust performance in the real world. Although DR is widely used in practice and is often solved using simple policy gradient (PG) methods, understanding of its theoretical guarantees remains limited. Toward addressing this gap, we provide the first convergence analysis of PG methods for domain-randomized linear quadratic regulator (LQR). We show that PG converges globally to the minimizer of a finite-sample approximation of the DR objective under suitable bounds on the heterogeneity of the sampled systems. We also quantify the sample-complexity associated with achieving a small performance gap between the sample-average and population-level objectives. Additionally, we propose and analyze a discount-factor annealing algorithm that obviates the need for an initial jointly stabilizing controller, which may be challenging to find. Empirical results support our theoretical findings and highlight promising directions for future work, including risk-sensitive DR formulations and stochastic PG algorithms.**

I. Introduction

Domain randomization (DR) has emerged as a dominant paradigm to enable transfer of policies optimized in simulation to the real world by randomizing simulator parameters during training [1–3]. In doing so, just as with robust control, DR accounts for discrepancies between the model used in simulation to synthesize a policy and the system that it is deployed on. However, unlike conventional robust control approaches, DR minimizes an average control objective over the uncertainty in the system rather than a worst case objective. Since DR does not solely focus on optimizing the worst-case performance, it can result in less conservative controller performance while still ensuring robust stability with high probability. Furthermore, DR can be easily implemented via first order methods. This makes it straightforward to incorporate into a wide variety of reinforcement learning (RL) schemes and to benefit from the increasing availability of parallel computation.

Despite the ease with which DR can be implemented using first order methods, ensuring convergence of these methods remains a critical challenge, with practitioners relying upon complex scheduling of various hyperparameters in the optimization procedure [3]. Motivated by this challenge, we rigorously study the convergence of policy gradient (PG) methods for DR in the setting of the linear quadratic regulator (LQR).

[1] All authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania. Emails: {ftesshu, brucele, nmatni, pappasg}@seas.upenn.edu

A. Related Work

a) Domain Randomization: Domain randomization, introduced by Tobin et al. [1], is widely used for enabling *sim-to-real transfer*. By randomizing simulator parameters during training, it aims to produce policies robust to simulator variations, thereby enabling transfer to the real world. The sampling distribution may be a design variable in the synthesis problem [4], or come from a learning procedure [5], e.g. as the posterior distribution of a Bayesian identification procedure. Recently, the approach has been applied in areas like autonomous racing [6] and robotic control [2, 3]. The control community has used similar randomized approaches to achieve high-probability guarantees (e.g. scenario optimization and related methods [7–10]). Recent work has explored generalization of DR in discrete Markov Decision Processes [11, 12] and for continuous control [5]. However, finding optimization procedures which minimize the DR objective remains a challenge, with many empirical studies relying on heuristic approaches like curriculum learning [13]. In this work, we rigorously study the convergence of PG for minimizing the DR objective for LQR.

b) Policy Gradient: Global convergence of PG methods was first established for LQR by [14], relying upon a condition known as *gradient dominance* to demonstrate that driving the gradient to zero suffices to ensure convergence to the optimal solution. Subsequently, the results have been extended to a multitude of settings including mixed $\mathcal{H}_2/\mathcal{H}_\infty$ [15], Markovian jump systems [16], and output estimation [17]. Hu et al. [18] provide a comprehensive survey on convergence studies of policy optimization methods for continuous control. There are two works closely related to DR: Gravell et al. [19] and Wang et al. [20]. Gravell et al. [19] study a related setting with multiplicative noise on the system, which resembles DR but with variations drawn independently at each time step. Consequently, the resulting optimal controller exhibits robustness primarily in an average sense, rather than explicitly accounting for structured uncertainty as in DR. Wang et al. [20] study a federated approach to minimize the average LQR cost over a collection of systems; however, they prove convergence only up to a gap defined in terms of the heterogeneity between the systems. In contrast, we prove that as long as the systems are sufficiently similar, an approximate gradient domination condition suffices to ensure convergence of PG methods to the globally optimal solution. Furthermore, while these works share a common assumption of initial stabilizing controller, we remove this limitation by extending a discount annealing method for stabilization in the single-system setting [21]. We focus exclusively on model-

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

based analysis; the extension of our results to the model-free setting involves routine extension of the analysis for the variance of gradient estimates from prior work [14].

B. Contributions

We study the convergence of PG methods applied to the domain randomized LQR objective and establish the following results:

- **DR Objective Approximation:** We propose optimizing a sample average approximation of the DR objective defined in terms of a finite number of sampled systems. We characterize the excess cost incurred by the solution to the approximate problem on the original objective.
- **Policy Gradient Convergence:** We prove that a first order PG algorithm converges to the globally optimal solution for the sample average approximation of the DR objective. This is the first result showing that PG converges to the optimal solution for domain randomized LQR, removing the heterogeneity bias that appears in prior work [20].
- **Joint Stabilization:** We propose a first order gradient descent algorithm based on discount factor annealing which provably converges to a controller stabilizing a collection of dynamical systems, extending prior work [21] to the multi-system setting.

The above contributions establish the convergence properties of first order gradient-based algorithms applied to the DR objective for continuous control. Given the prevalence of DR for the practice of learning-enabled robotic control and the engineering challenges involved in achieving convergent optimization procedures for the DR objective, we view this result as an important step towards the analysis and design of reliable RL algorithms for sim-to-real transfer. Motivated by this, we include a number of empirical results highlighting limitations and open questions regarding the analysis of this work. These include the generalization of the DR objective to risk metrics other than the expectation and the convergence of stochastic PG algorithms.

II. Problem Formulation

We consider a fully observed linear time-invariant system defined by an unknown parameter $\theta \in \mathbb{R}^{d_\theta}$ with state $x_t \in \mathbb{R}^{d_{\mathsf{x}}}$, control input $u_t \in \mathbb{R}^{d_{\mathsf{u}}}$, and disturbance $w_t \in \mathbb{R}^{d_{\mathsf{x}}}$. The state of the system evolves as

$$x_{t+1} = A(\theta)x_t + B(\theta)u_t + w_t,$$

where $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ is a zero-mean $i.i.d.$ disturbance.[1] For a fixed parameter θ, we define the infinite horizon LQR cost function in terms of positive definite matrices $Q \succeq I$ and $R = I$ as

$$J(K, \theta) \triangleq \limsup_{T\to\infty} \frac{1}{T} \mathbf{E}^K \left[\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \middle| \theta \right],$$

where the superscript K indicates that the expectation is taken under the feedback policy $u_t = Kx_t$.[2] The goal of DR is to determine a controller K which minimizes the expectation of the LQR objective over a distribution of systems, as in prior work on sampling-based control [10]. Such a distribution can be hand-designed to capture parametric uncertainty, or learned through Bayesian system identification [5]. In particular, we let Θ be a random variable with density p_Θ. The DR problem is then to find the controller $K^\star_{DR} \triangleq \operatorname{argmin}_K J_{DR}(K)$, with

$$J_{DR}(K) \triangleq \mathbf{E}_\Theta\, J(K, \Theta). \tag{1}$$

To solve this problem, we propose applying gradient descent to a finite sample approximation to objective (1). We use M independently sampled systems $\theta_1, \ldots, \theta_M \sim p_\Theta$ to construct such a finite-sample approximation:

$$J_{SA}(K) \triangleq \frac{1}{M} \sum_{i=1}^{M} J(K, \theta_i), \tag{2}$$

and apply PG methods to find its minimizer. In particular, we start from a controller K_0, and iteratively update as

$$K_{n+1} = K_n - \frac{\alpha}{M} \sum_{i=1}^{M} \nabla_K J(K_n, \theta_i), \tag{3}$$

where α is a fixed stepsize and, as shown in Fazel et al. [14], the gradient $\nabla_K J(K, \theta_i)$ for system i is given by

$$\nabla_K J(K, \theta_i) \triangleq E^i_K \Sigma^i_K. \tag{4}$$

In the above expression, Σ^i_K is the steady state covariance of system i under controller K, and E^i_K can be written in terms of the Lyapunov equation defining the LQR cost of applying controller K to system i:

$$\begin{aligned}
\Sigma^i_K &\triangleq \texttt{dlyap}(A(\theta_i) + B(\theta_i)K, \Sigma_w) \\
E^i_K &\triangleq 2(R + B(\theta_i)^\top P^i_K B(\theta_i))K + B(\theta_i)^\top P^i_K A(\theta_i) \\
P^i_K &\triangleq \texttt{dlyap}(A(\theta_i) + B(\theta_i)K, Q + KRK^\top),
\end{aligned}$$

where $\texttt{dlyap}(\mathrm{X}, \mathrm{Y})$ denotes the unique positive semidefinite solution P to the equation $X^T P X - P = Y$.

The primary technical challenge in proving convergence of the above scheme to the minimizer of (1) is that the approach to show gradient domination by Fazel et al. [14] does not generalize to the sample average LQR cost for $M \geq 2$. Consequently, we introduce an additional assumption which ensures the support of the distribution p_Θ is suitably small.

Assumption II.1 (Heterogeneity bound)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^{d_\theta}$ be such that $p_\Theta(\theta) = 0$ for $\theta \notin \mathcal{S}$, and that for all $\theta_1, \theta_2 \in \mathcal{S}$*

$$\left\| \begin{bmatrix} A(\theta_1) & B(\theta_1) \end{bmatrix} - \begin{bmatrix} A(\theta_2) & B(\theta_2) \end{bmatrix} \right\| \leq \varepsilon_{\mathsf{HET}},$$

where

$$\varepsilon_{\mathsf{HET}} \leq \frac{1}{5e5\tau_B(\theta)\operatorname{tr}\left(P^\theta_{K(\theta)}\right)^{11/2}}$$

[1] Generalization to noise covariance $I \neq \Sigma_w \succ 0$ may be achieved by scaling $A(\theta) \leftarrow \Sigma_w^{1/2} A(\theta)$ and $B(\theta) \leftarrow \Sigma_w^{1/2} B(\theta)$.

[2] Generalization to $Q \not\succeq I$ $R \neq I$ can be achieved by rescaling the cost by $1/\lambda_{\min}(Q)$ and the input as $B(\theta) \leftarrow B(\theta)R^{-1/2}$.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

for all $\theta \in \mathcal{S}$ and where $\tau_B \triangleq \max\{1, \sup_{\theta\in\mathcal{S}} \|B(\theta)\|\}$.[3] *Here $K(\theta)$ is the LQR controller for system θ, and $P^{\theta}_{K(\theta)}$ its corresponding cost matrix from the Riccati equation.*

The bound on $\varepsilon_{\mathsf{HET}}$ is a sufficient condition for our analysis to hold. It scales inversely with $\|B(\theta)\|$ and the trace of $P^{\theta}_{K(\theta)}$. In particular, if the systems are too heterogeneous, our analysis cannot guarantee convergence. Given this condition, we demonstrate the convergence of PG methods. The proof consists of three parts. We first show that PG methods converge to the minimizer of the sample average objective $J_{SA}(K)$ starting from an initial controller K_0 which simultaneously stabilizes $\theta_1, \ldots, \theta_M$ and achieves a sufficiently small cost $J_{SA}(K_0)$. We then introduce a discount annealing scheme to achieve convergence from an arbitrary initial controller. Finally, we characterize the discrepancy between the empirical objective (2) and its population counterpart (1).

III. Policy Gradient Convergence

We now prove convergence of the gradient update (3) to the minimizer of the sample average LQR objective (2). Denote this minimizer by $K^{\star}_{SA} \triangleq \operatorname{argmin}_K J_{SA}(K)$. To show convergence, we impose an additional assumption that the initial controller simultaneously stabilizes all systems.

Assumption III.1 (Simultaneous Stabilization)**.** *We assume that $K_0 \in \mathcal{K}$ where*

$$\mathcal{K} := \{K\colon \rho(A(\theta_i) + B(\theta_i)K) < 1,\ i = 1, \ldots, M\}.$$

To achieve sufficient conditions to ensure that (3) converges starting from such a K_0, we introduce a definition that characterizes the cost of a controller on a particular sample in terms of the sample average cost of that controller.

Definition III.1. *(Scenario Boundedness) Let $B, \mathbf{s} \in \mathbb{R}$. We call the collection of scenarios $\theta_1, \ldots, \theta_M$ $(B, \mathbf{s})$-*bounded *if for all K satisfying $J_{SA}(K) \le B J_{SA}(K^{\star}_{SA})$, it holds that $J(K, \theta_i) \le \mathbf{s} J_{SA}(K)$.*

Any collection of scenarios is $(B, \mathbf{s})$-bounded with $B = \infty$ and $\mathbf{s} = M$ by the fact that the individual cost can never exceed the average scaled by the number of scenarios. However, if the scenarios are sufficiently close, we can determine a $(B, \mathbf{s})$-stability condition such that $\mathbf{s}$ does not scale with the number of scenarios. This corresponds to the case where one single controller can perform well in multiple scenarios. One such condition is provided in Lemma III.5.

Under the above definition, we show that the sample average objective is coercive [18, Definition 1] with a smoothness bound defined in terms of the average cost of the initial iterate K_0 on the systems $\theta_1, \ldots, \theta_M$. This will in turn be used to ensure convergence of (3) to a fixed point. Note that for any θ_i, the LQR cost $J(K, \theta_i)$ is coercive, twice continuously differentiable, and L_i-smooth over any sublevel set $\mathcal{K}^i_{\mathbf{s}\zeta} \triangleq \{K \in \mathcal{K}\colon J(K, \theta_i) \le \mathbf{s}\zeta\}$ for $\zeta < \infty$ [18, Lemma 1] for L_i a polynomial in system parameters. This enables us to establish the following lemma.

[3]For a matrix X, the spectral norm is denoted $\|X\|$ and the Frobenius norm is denoted $\|X\|_F$.

Lemma III.1 (Sum of LQR Costs is Coercive)**.** *Suppose that the collection $(\theta_1, \ldots, \theta_M)$ is $(B, \mathbf{s})$-bounded. Let $\mathcal{K}_\zeta \triangleq \{K \in \mathcal{K}\colon J_{SA}(K) \le \zeta\}$. It holds that $\mathcal{K}_\zeta \subseteq \cap_{i=1}^M \mathcal{K}^i_{\mathbf{s}\zeta}$ for any $\zeta \le B$. Furthermore, it follows that $J_{SA}(K)$ is coercive, twice continuously differentiable, and L-smooth over $\mathcal{K}_\zeta$ where $L = \frac{\sum_{i=1}^M L_i}{M}$.*

Proof. By Definition III.1, it holds that for any $i \in \{1, \ldots, M\}$, the cost is bounded as $J(K, \theta_i) \le \mathbf{s} J_{SA}(K) \le \mathbf{s}\zeta$ for all $K \in \mathcal{K}_\zeta$. Consequently, $\mathcal{K}_\zeta \subseteq \cap_{i=1}^M \mathcal{K}^i_{\mathbf{s}\zeta}$. Now, since $M J_{SA}(K) \ge J(K, \theta^i)$ for $i = 1, \cdots, M$, when one of $J(K, \theta^i) \to \infty$ for any sequence $\{K^l\}_{l=1}^{\infty}$ if $\|K^l\|_2 \to \infty$ or K^l converges to an element of $\partial\mathcal{K}$, $J_{SA}(K) \to \infty$ also holds. Thus $J_{SA}(K)$ is coercive. By the fact that $J(K, \theta_i)$ is twice differentiable and L_i-smooth over $\mathcal{K}^i_{\mathbf{s}\zeta}$, it holds by the mean value theorem that

$$\begin{aligned} &J(K', \theta_i) - J(K, \theta_i) \\ &\le (K' - K)^\top \nabla_K J(K, \theta_i) + \frac{L_i}{2}\|K' - K\|_F^2, \end{aligned} \tag{5}$$

for $K, K' \in \mathcal{K}^i_{\mathbf{s}\zeta}$. Then $J_{SA}(K)$ is also twice continuously differentiable, and by summing (5) over $i = 1, \cdots, M$, it holds that $J_{SA}(K)$ is L-smooth on $\mathcal{K}_\zeta$ with $L = \frac{\sum_{i=1}^M L_i}{M}$. □

The smoothness constant L_i can be explicitly written in terms of system parameters [14, Theorem 7]. Instantiating these bounds, we find that the smoothness constant L of $J_{SA}(K)$ over the sublevel set of value $J_{SA}(K_0)$ satisfies

$$\begin{aligned} L \le \mathsf{poly}\Big(&\max_{i\in\{1,\cdots,M\}} J(K_0, \theta_i), \max_{i\in\{1,\cdots,M\}} \|\nabla J(K_0, \theta_i)\|, \\ &\max_{i\in\{1,\cdots,M\}} \|A(\theta_i)\|, \max_{i\in\{1,\cdots,M\}} \|B(\theta_i)\|\Big). \end{aligned}$$

A consequence of the above result is that the gradient descent procedure converges to a fixed point.

Lemma III.2 (Convergence to a Fixed Point)**.** *Suppose that the collection $(\theta_1, \ldots, \theta_M)$ is $(B, \mathbf{s})$-bounded. Let $\zeta_0 = J_{SA}(K_0)$ and L be the smoothness constant of $J_{SA}(K)$ on $\mathcal{K}_{\zeta_0}$. Consider the policy gradient method (3). Then, for any $0 < \alpha < \frac{2}{L}$, we have $K_n \in \mathcal{K}_{\zeta_0}$ and $J_{SA}(K_{n+1}) \le J_{SA}(K_n)$ for all n. Furthermore, we have the convergence of the gradient $\nabla_K J_{SA}(K) \to 0$ with the rate*

$$\min_{0\le l\le k} \|\nabla_K J_{SA}(K)\|_F^2 \le \frac{\zeta_0}{C(k+1)}, \tag{6}$$

with $C = \alpha - \frac{L\alpha^2}{2} > 0$.

This is the direct consequence of Lemma III.1 and Theorem 1 of [18]. Lemma III.2 ensures convergence of the PG procedure to a fixed point; however, it does not imply that this fixed point is the global minimizer of (2). In the case of a single dynamical system, Fazel et al. [14] show that a property known as gradient domination is satisfied:

$$J(K, \theta) - J(K^\star, \theta) \le \|\Sigma_{K^\star}\| \, \|\nabla_K J(K, \theta)\|_F^2.$$

It is not clear if such a condition holds for the sum of LQR objectives defined by multiple systems. In particular,

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

extending the analysis in Lemma 11 of [14] leads to the following weaker result.

Lemma III.3 (Approximate Gradient Domination)**.** *For any K that simultaneously stabilizes $\theta_1, \dots, \theta_M$, it holds that*

$$J_{SA}(K) - J_{SA}(K^\star_{SA}) \le \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla_K J(K, \theta_i) (\Sigma^i_K)^{-1} \Sigma^i_{K^\star_{SA}} \right\|_F^2.$$

Proof. Let $\widetilde{E} = \sum_{i=1}^M \left(\Sigma^i_{K^\star_{SA}} \otimes I \right) \mathsf{vec}\, E^i_K$ and $\widetilde{G} = \sum_{i=1}^M \Sigma^i_{K^\star_{SA}} \otimes \left(B(\theta_i)^T P^i_K B(\theta_i) + R \right)$. By Lemma 12 of [14] and the identity $\mathrm{tr}(A^\top B C D^\top) = \mathsf{vec}(A)^\top (D \otimes B)\, \mathsf{vec}(C)$, we have that

$$\begin{aligned} J_{SA}(K) - J_{SA}(K^\star_{SA}) &= \frac{1}{M} \sum_{i=1}^{M} J(K, \theta_i) - J(K^\star_{SA}, \theta_i) \\ &= -\frac{1}{M} \mathsf{vec}(K - K^\star_{SA})^\top (\widetilde{G}\, \mathsf{vec}(K - K^\star_{SA}) - 2\widetilde{E}) \\ &\le \frac{1}{M} \widetilde{E} \widetilde{G}^{-1} \widetilde{E} \le \frac{1}{M^2} \left\| \widetilde{E} \right\|^2 = \left\| \frac{1}{M} \sum_{i=1}^{M} E^i_K \Sigma^i_{K^\star_{SA}} \right\|_F^2, \end{aligned}$$

where the first inequality results from completing the square and the second inequality follows from the fact that $\widetilde{G} \succeq MI$. The last equality holds by the identity $\mathsf{vec}(XYZ) = (Z^\top \otimes X)\, \mathsf{vec}(Y)$. Conclude by applying (4). □

Critically, we observe that convergence of the gradient $\frac{1}{M} \sum_{i=1}^M \nabla_K J(K, \theta_i)$ to zero does not immediately suffice to ensure that the sum of gradients weighted by $(\Sigma^i_K)^{-1} \Sigma^i_{K^\star_{SA}}$ appearing in the approximate gradient domination converges to zero. However, we may write the upper bound as a gradient term, plus a remainder term as

$$\begin{aligned} &J_{SA}(K) - J_{SA}(K^\star_{SA}) \\ &\le 2 \left(\|\nabla_K J_{SA}(K)\|_F^2 + \|\mathsf{R}(K, K^\star_{SA})\|_F^2 \right), \end{aligned} \tag{7}$$

where

$$\mathsf{R}(K, K^\star_{SA}) \triangleq \frac{1}{M} \sum_{i=1}^{M} E^i_K \left(\Sigma^i_{K^\star_{SA}} - \Sigma^i_K \right).$$

The remainder term above becomes small as $K \to K^\star_{SA}$. In particular, if $\mathsf{R}(K, K^\star_{SA})$ is bounded by a small constant multiplied by the cost gap, we achieve gradient domination.

Lemma III.4. *Let $\mu \in [0, \frac{1}{2})$ and $\zeta_0 = J_{SA}(K_0)$. Suppose*

$$\|\mathsf{R}(K, K^\star_{SA})\|_F^2 \le \mu (J_{SA}(K) - J_{SA}(K^\star_{SA})) \tag{8}$$

for all $K \in S \subseteq \mathcal{K}_{\zeta_0}$. Then for all $K \in S$,

$$J_{SA}(K) - J_{SA}(K^\star_{SA}) \le \frac{2}{1 - 2\mu} \|\nabla_K J_{SA}(K)\|_F^2.$$

Proof. The result follows by substitution of (8) into (7). □

The above result in combination with Lemma III.2 establishes convergence of algorithm (3) to the solution of objective (2) from an initial iterate satisfying Assumption III.1 as long as the collection $\theta_1, \dots, \theta_M$ is $(B, \mathbf{s})$-bounded, and the iterates of the gradient updates (3) converge to a set S for which the condition (8) is satisfied. The convergence rate is determined by the stepsize α, the initial cost ζ_0, and the smoothness constant L of the sample average control cost.

The heterogeneity bound of Assumption II.1 ensures that $(B, \mathbf{s})$-boundedness is satisfied, the proof of which can be found in the appendix of [22].

Lemma III.5. *Under Assumption II.1, it holds that any collection of samples $\theta_1, \dots, \theta_M$ drawn from p_Θ is $(B, \mathbf{s})$-bounded with $B = 8$ and $\mathbf{s} = 2$.*

If additionally the cost of the initial iterate satisfies the same condition as Lemma III.5, then condition (8) holds.

Lemma III.6. *Suppose Assumption II.1 holds and that*

$$\zeta_0 = J_{SA}(K_0) \le 8 J_{SA}(K^\star_{SA}).$$

Then condition (8) is satisfied with $\mu = \frac{1}{4}$ and

$$S = \left\{ K \in \mathcal{K}_{\zeta_0} : \|\nabla_K J_{SA}(K)\|_F \le \frac{1}{256 \tau_B J_{SA}(K)^{5/2}} \right\}.$$

Before proving this fact, we state an intermediate result, the proof of which is deferred to the appendix of [22].

Lemma III.7. *Suppose that Assumption II.1 holds and that $J_{SA}(K_0) \le 8 J_{SA}(K^\star_{SA})$. Then*

$$\begin{aligned} \|\mathsf{R}(K, K^\star_{SA})\|_F &\le \frac{1}{4} \|K - K^\star_{SA}\|_F \\ \|\mathsf{R}(K^\star_{SA}, K)\|_F &\le \frac{1}{4} \|K - K^\star_{SA}\|_F \end{aligned}$$

for all $K \in S$ with S defined by Lemma III.6.

We now prove Lemma III.6.

Proof. To bound the result by $J_{SA}(K) - J_{SA}(K^\star_{SA})$, note that by [14, Lemma 12],

$$\begin{aligned} &J_{SA}(K) - J_{SA}(K^\star_{SA}) \\ &\ge \frac{2}{M} \mathrm{tr} \left((K - K^\star_{SA})^T \left(\sum_{i=1}^{M} E^i_{K^\star_{SA}} \Sigma^i_K \right) \right) + \|K - K^\star_{SA}\|_F^2 \\ &\ge \|K - K^\star_{SA}\|_F^2 - 2 \|K - K^\star_{SA}\|_F \|\mathsf{R}(K^\star_{SA}, K)\|_F. \end{aligned}$$

Thus by Lemma III.7, it follows that $J_{SA}(K) - J_{SA}(K^\star_{SA}) \ge \frac{1}{2} \|K - K^\star_{SA}\|_F^2$, which can in turn be substituted to show $\|\mathsf{R}(K, K^\star_{SA})\|_F \le \frac{1}{4} \|K - K^\star\|_F \le \frac{1}{2} \sqrt{J_{SA}(K) - J_{SA}(K^\star_{SA})}$. Squaring completes the proof. □

Ultimately, synthesizing these results provides the following convergence bound.

Theorem III.1. *Suppose Assumption II.1 holds. Let $K^\star_{SA}$ be an optimal policy for the sample average objective (2), and let K_0 be a controller which simultaneously stabilizes $\theta_1, \dots, \theta_M$ achieving average cost $J_{SA}(K_0) \le 8 J_{SA}(K^\star_{SA})$. Let L be the smoothness constant of $J_{SA}(K)$ over K_{ζ_0}. Consider running algorithm (3) with stepsize $\alpha = 1/L$. Then for $n \ge N := 2^9 L \tau_B \zeta_0^{7/2}$, we have the linear convergence*

$$J_{SA}(K_n) - J_{SA}(K^\star_{SA}) \le \left(1 - \frac{4}{L} \right)^{(n-N)} \left(\frac{1}{64 \tau_B \zeta_0^{5/2}} \right),$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

and it holds that $J_{SA}(K_n) - J_{SA}(K^\star_{SA}) \leq \epsilon$ *after* $n \geq N + \frac{L}{4}\log\frac{\zeta_0}{\varepsilon}$ *iterations.*

Proof. From Lemma III.2 and Lemma III.6, the gradient domination holds when $n \geq N$. Then, from Theorem 1 of [18], $J_{SA}(K)$ converges at the linear rate. □

The above result demonstrates that under a sufficiently small randomization distribution, the approximate gradient domination condition suffices to ensure convergence to the global optimal solution for the sample average LQR objective. The heterogeneity requirements of Assumption II.1 are stronger than desired, as they exclude particular cases in which it is possible to achieve a controller that simultaneously stabilizes multiple systems, but for which the systems are not close in the sense of Assumption II.1. We leave relaxing this condition or determining if it is fundamental to future work. The condition $J_{SA}(K_0) \leq 8J_{SA}(K^\star_{SA})$ means K_0 is not too far in cost from the optimal – this can be guaranteed by the discounting scheme in the sequel.

IV. Stabilization

The previous section analyzed the convergence of the PG method on a collection of sampled instances starting from an initial stabilizing controller. Here, we remove the requirement that we have access to an initial controller which simultaneously stabilizes all systems and incurs a small cost. We do so by incorporating a discount annealing scheme. In particular, consider scaling the matrices $A(\theta_i)$ and $B(\theta_i)$ by a factor $\sqrt{\gamma}$. For any controller K, if $\gamma < \min_{i\in\{1,\dots,M\}} \rho(A(\theta_i)+B(\theta_i)K)^{-2}$, then K stabilizes $\sqrt{\gamma}A(\theta_i)$ and $\sqrt{\gamma}B(\theta_i)$ for $i = 1,\dots,M$.[4] Therefore, if we start from a small discount factor, run several iterations of (3), and then increase the discount factor, we will be able to ensure convergence to a controller that simultaneously stabilizes all systems. Algorithm 1 extends a similar scheme proposed in Perdomo et al. [21] for the single-system setting to the multi-system setting we consider. We let $J_{SA}(\cdot \mid \gamma)$ denote the sample average cost (2) evaluated on the discounted systems $(\sqrt{\gamma}A(\theta_i), \sqrt{\gamma}B(\theta_i))$.

Subproblems (9) and (11) can be solved via bisection, while subproblem (10) can be solved with algorithm (3) starting from $K_0 \leftarrow K$ by choosing an appropriate stepsize, and running sufficiently many iterations, as in Theorem III.1. Extending the analysis of [21, Theorem 1] to the setting of multiple systems leads to the following result.

Theorem IV.1 (Stabilization by Discount Annealing)**.** *Suppose that Assumption II.1 holds. Let* γ_0 *be the largest* γ *satisfying* (9)*. Then Algorithm 1 is guaranteed to cease after at most* $1024J_{SA}(K^\star)^4\log(1/\gamma_0)$ *iterations. At the start of any iteration, it holds that* $J_{SA}(K|\gamma) \leq 8\inf_{\tilde{K}} J_{SA}(K|\gamma)$*. The final controller* K' *jointly stabilizes all* M *systems, and satisfies* $J_{SA}(K') - J_{SA}(K^\star_{SA}) \leq \varepsilon$*.*

[4]Scaling A and B by $\sqrt{\gamma} < \rho(A+BK)^{-1}$ is equivalent to introducing a discount factor in the LQR cost which ensures the cost is finite even for a controller which does not stabilize A, B.

Algorithm 1 Discount Annealing

Input: collection of systems, $\theta_1,\dots,\theta_M$, optimization tolerance ε, state dimension d_{X}

$K \leftarrow 0$

Find the largest $\gamma \in (0,1]$ satisfying

$$J_{SA}(K|\gamma) \leq 8d_{\mathsf{X}}. \tag{9}$$

while $\gamma < 1$ **do**

  Set $K \leftarrow K'$, where K' is such that

$$J_{SA}(K' \mid \gamma) - \inf_{\tilde{K}} J_{SA}(\tilde{K} \mid \gamma) \leq d_{\mathsf{X}}. \tag{10}$$

  Find a discount factor $\gamma' \in [\gamma, 1]$ such that

$$2.5J_{SA}(K \mid \gamma) \leq J_{SA}(K \mid \gamma') \leq 4J_{SA}(K \mid \gamma) \tag{11}$$

  $\gamma \leftarrow \gamma'$

Run (3) starting from K to find K' such that

$$J_{SA}(K') - \inf_{\tilde{K}} J_{SA}(\tilde{K}) \leq \varepsilon.$$

Return K'

Notably, the bound on the discounted cost of a controller K at the start of every iteration suffices to satisfy the requisite condition on the cost of the K_0 in Theorem III.1.

Proof. We first show that at the start of every iteration $J_{SA}(K|\gamma) \leq 8\inf_{\tilde{K}} J_{SA}(\tilde{K}|\gamma)$. This holds for the first iteration by the fact that γ_0 satisfies the bound (9), and $d_{\mathsf{X}} \leq \inf_{\tilde{K}} J_{SA}(\tilde{K}|\gamma)$. It holds for subsequent iterations by the fact that update (10) achieves a controller within a factor of 2 of the optimal cost, and the choice of γ in (11) inflates the cost by at most a factor of 4. This implies that $J_{SA}(K|\gamma) \leq 8\inf_{\tilde{K}} J_{SA}(K|\gamma)$ by the fact that the cost is nondecreasing with γ.

To bound the number of iterations, let $P^i_{K,\gamma}$ denote the Ricatti equation solution corresponding to the discounted system $(\sqrt{\gamma}A(\theta_i), \sqrt{\gamma}B(\theta_i))$. By Perdomo et al. [21, Theorem 1], if $\tilde{\gamma} \leq \min_{i\in\{1,\dots,M\}} \left(1/(8\left\|P^i_{K,\gamma}\right\|^4)+1\right)^2\gamma$, then $\operatorname{tr}(P^i_{K,\gamma}) \leq 2\operatorname{tr}(P^i_{K,\gamma})$ for $i \in \{1,\dots,M\}$. Therefore $J_{SA}(K|\tilde{\gamma}) \leq 2J_{SA}(K|\gamma)$. Thus the value γ' from (11) satisfies

$$\frac{\gamma'}{\gamma} \geq \min_{i\in\{1,\dots,M\}} \left(\frac{1}{8\left\|P^i_{K,\gamma}\right\|^4}+1\right)^2 \geq \left(\frac{1}{2^{11}J_{SA}(K^\star_{SA})^4}+1\right)^2.$$

This follows from the bound $\left\|P^i_{K,\gamma}\right\| \leq \operatorname{tr}(P^i_{K,\gamma}) \leq 2J_{SA}(K|\gamma) \leq 4J_{SA}(K^\star_{SA}|\gamma)$. Here, the second inequality follows from Lemma III.5 paired with the fact that $J_{SA}(K|\gamma) \leq 8\inf_{\tilde{K}} J_{SA}(\tilde{K}|\gamma)$. The last inequality follows from the fact that K satisfies the error bound of (10), and $d_{\mathsf{X}} \leq J_{SA}(K^\star_{SA}|\gamma)$. Then using the inequality $\log(1+x) \leq x$, it holds that $\gamma = 1$ after the given number of iterations.

The bound on the sample average cost of the final iterate follows immediately from the last step of the algorithm. □

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

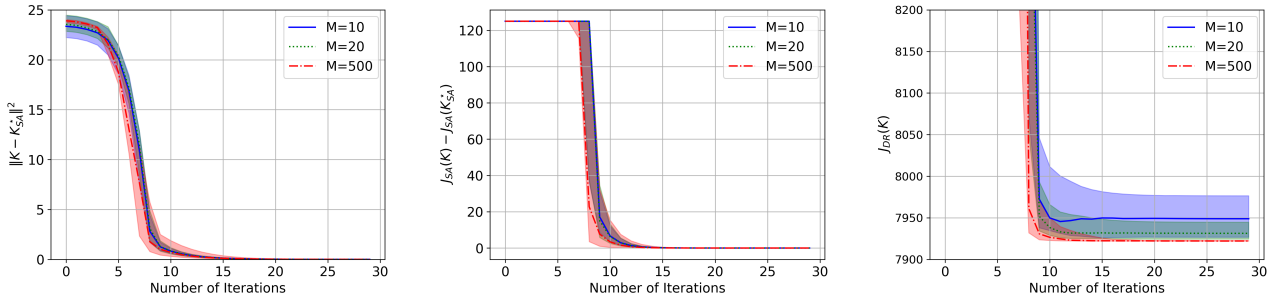


Fig. 1. Convergence of policy gradient with domain randomization. A controller initialized at $K = 0$ converges to the optimal controller via progressive discounting and gradient descent (left), resulting in the convergence of the excess sample average cost (center). With more samples, the minimizer for the sample average cost and the domain randomization objective start to match (right).

V. Sampling

The previous sections provided guarantees ensuring that PG algorithms converge to the minimizer of objective (2). In this section, we bound the suboptimality incurred by optimizing the sample average cost (2) rather than the desired DR objective (1). In particular, let $\tilde{K}$ be the iterate of running the aforementioned gradient procedure. It holds that

$$J_{DR}(\tilde{K}) - J_{DR}(K^\star_{DR}) = J_{DR}(\tilde{K}) - J_{SA}(\tilde{K}) + J_{SA}(\tilde{K}) - J_{SA}(K^\star_{DR}) + J_{SA}(K^\star_{DR}) - J_{DR}(K^\star_{DR}).$$

The results of the previous section ensure that

$$J_{SA}(\tilde{K}) - J_{SA}(K^\star_{DR}) \leq J_{SA}(\tilde{K}) - J_{SA}(K^\star_{SA}) \leq \varepsilon.$$

Then to characterize the gap $J_{DR}(\tilde{K}) - J_{DR}(K^\star_{DR})$, it suffices to achieve upper bounds on $\left|J_{DR}(\tilde{K}) - J_{SA}(\tilde{K})\right|$ and $|J_{DR}(K^\star_{DR}) - J_{SA}(K^\star_{DR})|$. Note that if $J(K,\theta) \leq \mathcal{J}$, then by Hoeffding's inequality [23], it holds that with probability at least $1-\delta$,

$$|J_{DR}(K) - J_{SA}(K)| \leq \sqrt{\frac{2\mathcal{J}^2}{M} \log \frac{2}{\delta}}.$$

To establish the bound $\mathcal{J}$, consider $\tilde{\theta}, \theta \in \mathcal{S}$. We show in Lemma IX.9 that under Assumption II.1, $J(K^\star_{DR}, \tilde{\theta}) \leq 6J(K(\theta),\theta)$ and $J(\tilde{K},\tilde{\theta}) \leq 48J(K(\theta),\theta)$. This leads to the following characterization of the excess cost.

Theorem V.1 (Population Excess Cost Bound)**.** *Suppose Assumption II.1 holds. Consider sampling M systems from Θ. Let $\tilde{K}$ be the iterate achieved by running Algorithm 1. Consider an arbitrary $\theta \in \mathcal{S}$. With probability at least $1-\delta$,*

$$J_{DR}(\tilde{K}) - J_{DR}(K^\star_{DR}) \leq \varepsilon + 200\sqrt{\frac{J(K(\theta),\theta)}{M} \log \frac{2}{\delta}}.$$

VI. Numerical Experiments

We conduct numerical experiments on the discretized and linearized inverted pendulum defined by

$$A = \begin{bmatrix} 1 & dt \\ \frac{g}{\ell} dt & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ \frac{dt}{m\ell^2} \end{bmatrix}. \quad (12)$$

We suppose that the parameters $dt = 0.01$ and $g = 10$ are known. The unknown parameters m and ℓ are modeled with a uniform distribution over the interval $[0.75, 1.25]$. For policy update in Algorithm 1, we run 20 iterations of (3) with a stepsize $\alpha = 1 \times 10^{-3}$ instead of explicitly evaluating the inequality (10). Refer to the code for more details.[5]

a) Verification of Proposed Approach: Figure 1 demonstrates the application of Algorithm 1 to a collection of $M = 10, 20, 500$ samples from the distribution of systems (12). We run the algorithm over 100 random seeds. We plot the median and shade 25% to 75% quantile for the distance of the iterate from the optimal controller, the excess sample average control cost, and a Monte Carlo approximation of the domain randomized objective (1) formed with 10^5 samples. We see that the discount annealing algorithm succeeds in synthesizing a controller that simultaneously stabilizes all systems, and that the policy update (3) converges to the minimizer of sample average cost. This is predicted by the analysis of Theorem III.1 and Theorem IV.1. The rightmost figure demonstrates the impact of optimizing the sample average objective defined in terms of M samples. As suggested in Theorem V.1, sampling more systems results in a controller achieving lower DR objective.

b) Alternative Risk Metrics: We can consider risk metrics other than the expectation for defining the objective (1). For example, we could replace the expectation the entropic risk measure of temperature t:

$$J^t_{ER}(K) \triangleq \frac{1}{t} \log \mathbf{E}_\Theta[\exp(tJ(K,\Theta))]. \quad (13)$$

The gradient-based optimization of a Monte Carlo approximation proposed in the previous sections can still be applied in this setting by replacing the average gradient of (3) with the gradient of (13):

$$K \leftarrow K - \alpha \sum_{i=1}^{M} \frac{\exp(tJ(K,\theta_i))}{\sum_{m=1}^{M} \exp(tJ(K,\theta_i))} \nabla_K J(K,\theta_i).$$

We lack theoretical convergence guarantees for this approach; however, Figure 2 validates it numerically. We see that the proposed algorithm still converges with this alternative risk metric and achieves the optimal controller for the risk-sensitive objective. This result opens the future direction of risk-sensitive domain randomized control.

[5]Code available at this GitHub repository

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

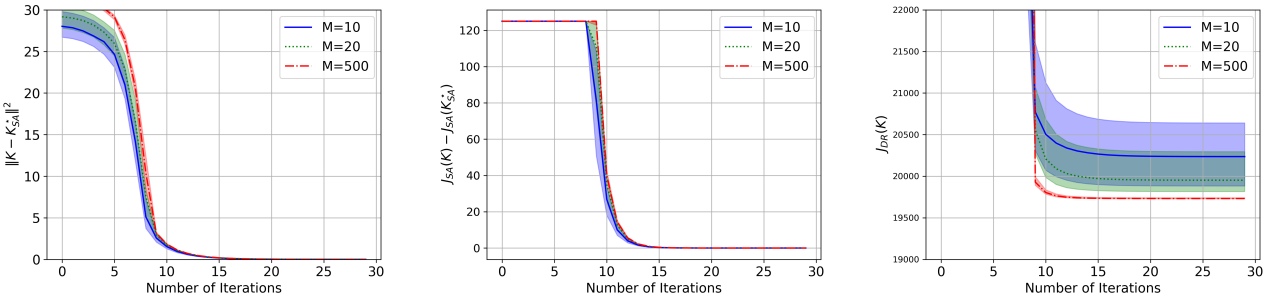


Fig. 2. Convergence of policy gradient with the entropic risk measure using $t = 1.0$

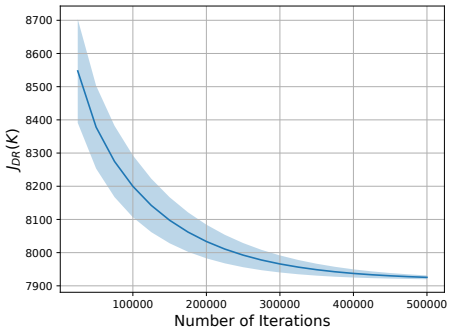


Fig. 3. Stochastic Gradient Descent (Algorithm 2) applied to the linearized inverted pendulum of (12).

c) Stochastic Gradient Descent: Prior experiments verified that the approach analyzed in sections Section III-V is valid for the numerical example of stabilizing a pendulum linearized about the upright position. In Figure 3, we demonstrate a stochastic gradient descent approach which samples a new system at every iteration and takes a gradient step with that system, detailed in Algorithm 2.[6] This also converges, although we lack theoretical guarantees theoretical convergence guarantees. The figure plots the mean and standard deviation over 5 random seeds. The cost $J_{DR}(K)$ is approximated via Monte Carlo, as in the previous experiments. The initial stepsize is chosen as $\alpha = 2 \times 10^{-4}$ and the discount coefficient is $\gamma_0 = 0.99$. We see that the controller converges to achieve approximately the same cost on the domain randomization objective as the batch algorithm with $M = 500$ samples from Figure 1.

Algorithm 2 Stochastic Gradient Descent

Input: distribution p_Θ, iterations N, stepsize α, discount coefficient γ_0
$K \leftarrow 0$,
for $n = 1, \ldots, N$ **do**
  Sample $\theta \sim p_\Theta$
  $\gamma \leftarrow \min\{\gamma_0 \rho(A(\theta) + B(\theta)K)^{-2}, 1\}$
  $K \leftarrow K - \alpha \nabla_K J(K, \theta)$
Return: K

d) Rotary Inverted Pendulum Experiments: We also verify the Algorithm 1 on a Quanser Qube Servo 2 rotary inverted pendulum [24] to study the impact on the sim-to-real gap. We consider the scenario where we have coarse estimates of several physical parameters, and compare the DR controller against the LQR controller which takes estimates as ground truth. As expected, the former achieves a higher success rate by overcoming the sim-to-real gap.[7] Details are deferred to the appendix of [22].

VII. Discussion

There are several exciting possibilities for theoretical and empirical extensions of the results presented in this paper.

- **Relaxed heterogeneity assumptions:** Our theory required strong requirements on the distance between systems. These requirements can almost certainly be made weaker; however, it is unclear the extent to which they are fundamental. Our numerical experiments failed to discover any instances where PG did not converge due to a large distance between systems, unless it was impossible to simultaneously stabilize these systems.
- **Extensions of the convergence analysis:** Our numerical experiments tested a stochastic variation of the algorithm and an alternative risk metric for DR for which we do not have theoretical convergence guarantees. It would be interesting to provide convergence proofs for these approaches. In particular, stochastic gradient descent closer resembles the way DR is implemented in practice. Meanwhile, the entropic risk (or alternative metrics) may better represent the goal of DR to provide robustness to uncertainty. Indeed, entropic risk interpolates between DR as $t \to 0$, and robust control as $t \to \infty$.
- **Practical implications:** The study conducted in this paper provides a better understanding of the application of gradient-based approaches for RL with DR. It may be possible to use this perspective to design alternative hyperparameter schedules for DR as applied in practice [3].

VIII. Conclusion

Our work proves the convergence of policy gradient methods applied to linear quadratic control with domain randomization. The analysis relies upon a generalization of the gradient domination condition from prior work. This generalization results in guarantees that are sensitive to the initialization; however, a curriculum learning approach with an appropriate schedule can bypass this sensitivity. We believe that this line of analysis has potential to demystify

[6]The update is performed with the gradient of the discounted cost for the sampled system. The discounting ensures stabilization by the current iterate.

[7]The experiment video can be found at this YouTube link.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.

and improve heuristic approaches for taming the optimization landscape of DR from the robot learning literature [3].

Acknowledgements

We thank Leonardo Toso for instructive conversations. TF is supported by JASSO Exchange Support program and UTokyo-TOYOTA Study Abroad Scholarship. TF and GP are supported in part by NSF Award SLES 2331880 and NSF TRIPODS EnCORE 2217033. BL and NM are supported by NSF Award SLES-2331880, NSF CAREER award ECCS-2045834 and AFOSR Award FA9550-24-1-0102.

References

[1] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.

[2] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3803–3810.

[3] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[4] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull, "Active domain randomization," in *Conference on Robot Learning*. PMLR, 2020, pp. 1162–1176.

[5] T. Fujinami, B. D. Lee, N. Matni, and G. J. Pappas, "Domain randomization is sample efficient for linear quadratic control," 2025. [Online]. Available: https://arxiv.org/abs/2502.12310

[6] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, "Deep drone racing: From simulation to reality with domain randomization," *IEEE Transactions on Robotics*, vol. 36, no. 1, pp. 1–14, 2019.

[7] G. C. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on automatic control*, vol. 51, no. 5, pp. 742–753, 2006.

[8] R. F. Stengel and L. R. Ray, "Technical notes and correspondence: Stochastic robustness of linear time-invariant control systems," *NASA. Langley Research Center, Joint University Program for Air Transportation Research, 1990-1991*, 1991.

[9] L. R. Ray and R. F. Stengel, "A monte carlo approach to the analysis of control system robustness," *Automatica*, vol. 29, no. 1, pp. 229–236, 1993.

[10] M. Vidyasagar, "Randomized algorithms for robust controller synthesis using statistical learning theory," *Automatica*, vol. 37, no. 10, pp. 1515–1528, 2001.

[11] X. Chen, J. Hu, C. Jin, L. Li, and L. Wang, "Understanding domain randomization for sim-to-real transfer," *arXiv preprint arXiv:2110.03239*, 2021.

[12] Y. Zhong, A. A. Deshmukh, and C. Scott, "Pac reinforcement learning without real-world feedback," 2019. [Online]. Available: https://arxiv.org/abs/1909.10449

[13] F. Muratore, F. Ramos, G. Turk, W. Yu, M. Gienger, and J. Peters, "Robot learning from randomized simulations: A review," *Frontiers in Robotics and AI*, vol. 9, p. 799893, 2022.

[14] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International conference on machine learning*. PMLR, 2018, pp. 1467–1476.

[15] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 179–190.

[16] S. Rathod, M. Bhadu, and A. De, "Global convergence using policy gradient methods for model-free markovian jump linear quadratic control," *arXiv preprint arXiv:2111.15228*, 2021.

[17] J. Umenberger, M. Simchowitz, J. C. Perdomo, K. Zhang, and R. Tedrake, "Globally convergent policy search over dynamic filters for output estimation," *arXiv preprint arXiv:2202.11659*, 2022.

[18] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, no. 1, pp. 123–158, 2023.

[19] B. Gravell, P. M. Esfahani, and T. Summers, "Learning robust control for lqr systems with multiplicative noise via policy gradient," *arXiv preprint arXiv:1905.13547*, 2019.

[20] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free learning with heterogeneous dynamical systems: A federated lqr approach," *arXiv preprint arXiv:2308.11743*, 2023.

[21] J. Perdomo, J. Umenberger, and M. Simchowitz, "Stabilizing dynamical systems via policy gradient methods," *Advances in neural information processing systems*, vol. 34, pp. 29 274–29 286, 2021.

[22] T. Fujinami, B. D. Lee, N. Matni, and G. J. Pappas, "Policy gradient for lqr with domain randomization," 2025. [Online]. Available: https://arxiv.org/abs/2503.24371

[23] R. Vershynin, "High-dimensional probability," 2009.

[24] K. Polzounov, R. Sundar, and L. Redden, "Blue River Controls: A toolkit for Reinforcement Learning Control Systems on Hardware," Accepted at the Workshop on Deep Reinforcement Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada., 2019.

[25] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.