Domain Randomization is Sample Efficient for Linear Quadratic Control

Tesshu Fujinami FTESSHU@SEAS.UPENN.EDU
Bruce D. Lee BRUCELE@SEAS.UPENN.EDU
Nikolai Matni NMATNI@SEAS.UPENN.EDU
George J. Pappas PAPPASG@SEAS.UPENN.EDU
All authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

We study the sample efficiency of domain randomization and robust control for the benchmark problem of learning the linear quadratic regulator (LQR). Domain randomization, which synthesizes controllers by minimizing average performance over a distribution of model parameters, has achieved empirical success in robotics, but its theoretical properties remain poorly understood. We establish that with an appropriately chosen sampling distribution, domain randomization achieves the optimal asymptotic rate of decay in the excess cost, matching certainty equivalence. We further demonstrate that robust control, while potentially overly conservative, exhibits superior performance in the low-data regime due to its ability to stabilize uncertain systems with coarse parameter estimates. We propose a gradient-based algorithm for domain randomization that performs well in numerical experiments, which enables us to validate the trends predicted by our analysis. These results provide insights into the use of domain randomization in learning-enabled control, and highlight several open questions about its application to broader classes of systems.

Keywords: Learning-Enabled Control, Domain Randomization, Reinforcement Learning

1. Introduction

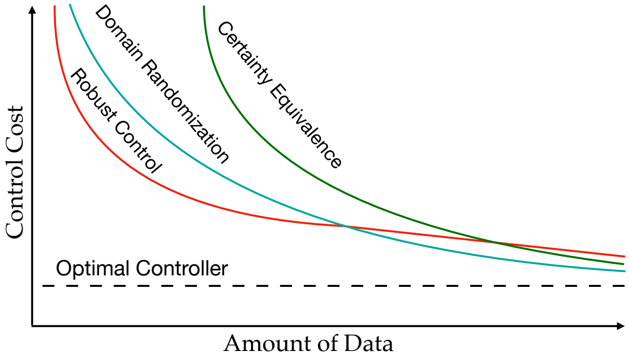

Figure 1: Illustration of the sample efficinecy of various synthesis methods.

The use of learned world models to synthesize controllers via policy optimization is becoming increasingly prevalent in reinforcement learning (Wu et al., 2023; Matsuo et al., 2022). The performance of the resulting controller depends heavily upon the synthesis procedure. Simple approaches that do not account for uncertainty, known as certainty equivalence, can overfit to errors in the learned model. Robust control approaches can tolerate some error in the learned model, but may be overly conservative and computationally demanding. Consequently, *domain randomization* has emerged as a dominant paradigm in robotics for enabling transfer of policies optimized in simulation on a learned or physics-based simulator to the real world by randomizing system parameters during policy optimization (Tobin et al., 2017; Akkaya et al., 2019). Despite the empirical success of domain randomization, it remains poorly understood how to select the randomization distribution, and how much of a discrepancy between the learned model and the real world can be tolerated.

We seek to address these issues by restricting attention to the benchmark problem of learning the linear quadratic regulator (LQR). This problem consists of collecting experimental interaction data from a linear dynamical system, and using this data to synthesize a controller that optimizes a control objective. The linear dynamical system is described by

$$X_{t+1} = A(\theta^\star)X_t + B(\theta^\star)U_t + W_t \text{ for } t = 1, \dots, T, \tag{1}$$

where $X_t \in \mathbb{R}^{d_\mathsf{X}}$ is the system state, $U_t \in \mathbb{R}^{d_\mathsf{U}}$ is the control input, $W_t \in \mathbb{R}^{d_\mathsf{X}}$ is i.i.d. mean zero Gaussian noise, and $\theta^\star \in \mathbb{R}^{d_\theta}$ is an unknown parameter. The goal is to design a controller K to minimize the objective $C(K, \theta^\star)$, defined as

$$C(K,\theta) \triangleq \limsup_{T\to\infty} \mathbf{E}_\theta^K \left[\frac{1}{T} \sum_{t=1}^{T} \left(\|X_t\|_Q^2 + \|U_t\|_R^2 \right) \right], \tag{2}$$

for Q and R positive definite weight matrices. The subscript on the expectation denotes that the states evolves according to $X_{t+1} = A(\theta)X_t + B(\theta)U_t + W_t$, and the superscript denotes that the inputs are selected according to the linear feedback $U_t = KX_t$. This benchmark problem has been used to study the sample efficiency of certainty equivalence (Mania et al., 2019) and robust control (Dean et al., 2020) by quantifying how many experiments from the system (1) are sufficient to achieve some level of control performance. In this work, we study the sample efficiency of domain randomization, which chooses a control policy as

$$K_{\mathsf{DR}}(\mathcal{D}) = \operatorname*{argmin}_K \mathbf{E}_{\theta\sim\mathcal{D}}[C(K,\theta)], \tag{3}$$

for a sampling distribution $\mathcal{D}$ determined using the dataset of experiments collected from (1).

1.1. Contributions

Our contribution is to study the sample efficiency of domain randomization and robust control to establish the relationships visualized as a conceptual diagram in Figure 1. We achieve the following:

- **Sample Effiency of Domain Randomization:** We prove that with an appropriately chosen sampling distribution $\mathcal{D}$, domain randomization (3) achieves the optimal asymptotic rate of decay with the number of samples, thereby matching the performance of certainty equivalence in the large sample regime. We further conjecture that the burn-in time for domain randomization lies between that of robust control and certainty equivalence, and verify this numerically.
- **Sample Effiency of Robust Control:** We prove the tightest known bound on robust control, improving the asymptotic rate of decay with the number of samples N from $1/\sqrt{N}$ to $1/N$. The upper bounds indicate a gap between the asymptotic rate of decay for robust control, and the rate of decay for domain randomization and certainty equivalence in terms of system-theoretic quantities. We conjecture that this gap is fundamental, due to the conservative nature of robust control. However, we establish that robust control can achieve a smaller burn-in time relative to certainty equivalence, due to its ability to stabilize the system with a coarse estimate.

- **Algorithm for Domain Randomization:** We propose a policy gradient algorithm to solve (3) by finite sample approximation of objective (2), which proves effective in numerical experiments. This enables verification of the trends predicted in the aforementioned results, and aligns with the conceptual diagram in Figure 1. While the focus of this work is restricted to linear systems, the proposed algorithm can in principle extend to general nonlinear systems, whereas extensions for robust control face computational challenges. We test domain randomization on a pendulum example in the extended manuscript (Fujinami et al., 2025),

By providing this characterization, our work demonstrates the potential of domain randomization in learning-enabled control, and partially explains the empirical success that it has achieved in robotics applications. We therefore conclude by highlighting several interesting open questions regarding the use of domain randomization for learning-enabled control.

1.2. Related Work

Domain Randomization Domain randomization, introduced by Tobin et al. (2017), is widely used for *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. This approach has been applied in areas like autonomous racing (Loquercio et al., 2019) and robotic control (Peng et al., 2018; Akkaya et al., 2019). However, its success depends heavily on selecting an effective sampling strategy (Mehta et al., 2020), which is often challenging. While previous work has explored generalization of domain randomization in discrete Markov Decision Processes (Chen et al., 2021; Zhong et al., 2019; Jiang, 2018), formalizing generalization for continuous control remains an open problem, which we address in this work.

Identification and Control The linear quadratic regulator problem has become a key benchmark for evaluating reinforcement learning in continuous control (Abbasi-Yadkori and Szepesvári, 2011; Recht, 2019). The offline setting has been extensively studied: Dean et al. (2020) analyzed the sample efficiency of robust control, while Mania et al. (2019); Wagenmaker et al. (2021); Lee et al. (2023) showed that certainty equivalence is asymptotically instance-optimal, achieving the best possible sample efficiency with respect to system-theoretic quantities. Extensions to smooth nonlinear systems were made by Wagenmaker et al. (2024); Lee et al. (2024). However, certainty equivalence can perform poorly with limited data. Alternative Bayesian approaches (von Rohr et al., 2022; Chiuso et al., 2023) can mitigate such limitations. We therefore show that such uncertainty-aware synthesis methods can match the asymptotic efficiency of certainty equivalence while achieving better performance in low-data regimes.

Robust Control: The control community has traditionally addressed policy synthesis with imperfect models using methods like $\mathcal{H}_\infty$ control, which focuses on worst-case uncertainty (Zhou et al., 1996; Başar and Bernhard, 2008; Doyle, 1982; Fan et al., 1991). Randomized approaches to robust control emphasizing high-probability guarantees have also been explored (Calafiore and Campi, 2006; Stengel and Ray, 1991; Ray and Stengel, 1993). Vidyasagar (2001) proposed an average performance metric similar to domain randomization but focused on a fixed distribution rather than one informed by data. Early data-driven synthesis efforts combined classical system identification (Ljung, 1998) with worst-case robust control (Gevers, 2005), while recent work has developed robust synthesis methods that bypass explicit models (Berberich et al., 2020). To the best of our knowledge, existing analyses of statistical efficiency in robust control yield suboptimal rates, with excess control cost decreasing at $1/\sqrt{N}$ (Dean et al., 2020), compared to the faster $1/N$ rate achieved by certainty equivalence. This work refines robust synthesis analysis, demonstrating the $1/N$ rate and a short burn-in period, highlighting its advantages with limited data.

Notation: The operator $\mathsf{vec}(A)$ stacks the columns of A into a vector, and its inverse $\mathsf{vec}^{-1}(\mathsf{vec}(A), n)$ returns a matrix with n rows by successively stacking chunks of n elements into columns. The Kronecker product of A and B is $A \otimes B$. $x \vee y$ denotes the max of x and y.

2. Problem Formulation

Consider the linear dynamical system (1). We assume that$(A(\theta^\star), B(\theta^\star))$ is stabilizable. For ease of exposition, we restrict attention to the case where $\begin{bmatrix} A(\theta) & B(\theta) \end{bmatrix} = \mathsf{vec}^{-1}(\theta, d_{\mathsf{X}})$, i.e. all entries of the state and input matrices are unknown. We additionally assume that $\Sigma_w = I$, and the cost matrices $Q \succeq I$ and $R = I$ are known.[1]

When θ is known, the optimal controller that minimizes $C(K, \theta)$ is given by

$$P(\theta) \triangleq A(\theta)^\top P(\theta) A(\theta) - A(\theta)^\top P(\theta) B(\theta)(B(\theta)^\top P(\theta) B(\theta) + R)^{-1} B(\theta)^\top P(\theta) A(\theta) + Q$$
$$K(\theta) \triangleq -(B(\theta)^\top P(\theta) B(\theta) + R)^{-1} B(\theta)^\top P(\theta) A(\theta),$$

where $P(\theta)$ is the positive definite solution to the discrete algebraic Ricatti equation, and $K(\theta)$ is the LQR solution corresponding to a system with parameters θ.

To design a controller for the unknown system (1), we suppose that we have run N experiments with control input $U_t \sim \mathcal{N}(0, \Sigma_u)$ for $\Sigma_u \succ 0$.[2] From these experiments, we collect a dataset

$$\{(X_t^n, U_t^n, X_{t+1}^n)\}_{t=1,n=1}^{T,N} \tag{4}$$

consisting of N trajectories of length T from (1). We use this dataset to design a controller $\hat{K}$ such that the cost, $C(\hat{K}, \theta^\star)$, is small.[3] In our analysis, we consider several approaches to achieve this objective. The approaches will be contrasted by examining the rate at which this cost decays to the optimal cost as the number of experiments in the dataset increases.

2.1. Controller Synthesis Approaches

All the synthesis approaches under consideration are model-based approaches which determine $\hat{\theta}$ from the dataset (4) via the following least squares problem:

$$\hat{\theta} \triangleq \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{t=1}^{T} \left\| X_{t+1}^n - \begin{bmatrix} A(\theta) & B(\theta) \end{bmatrix} \begin{bmatrix} X_t^n \\ U_t^n \end{bmatrix} \right\|^2. \tag{5}$$

The estimation error for $\hat{\theta}$ can be characterized using the Fisher information matrix:

$$\mathsf{FI}(\theta^\star) \triangleq \mathbf{E}_{\theta^\star}^{U_t \sim \mathcal{N}(0, \Sigma_u)} \left[\sum_{t=1}^{T} \begin{bmatrix} X_t \\ U_t \end{bmatrix} \begin{bmatrix} X_t \\ U_t \end{bmatrix}^\top \right] \otimes \Sigma_W^{-1}, \tag{6}$$

[1] Extension to $\Sigma_w \succ 0$, $Q \succ 0$, and $R \succ 0$ is possible by scaling the cost and changing the state and input basis.

[2] Our results extend to general exploration policies by modifying the identification bounds to display dependence on the Fisher Information of the exploration policy, as in (Lee et al., 2024).

[3] Note that we collect finite horizon trajectories, but evaluate infinite horizon cost.

see, e.g. Lee et al. (2024). Since this depends on the unknown parameter, we define the estimate

$$\hat{\mathsf{FI}} \triangleq \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\begin{bmatrix} X_t^n \\ U_t^n \end{bmatrix}\begin{bmatrix} X_t^n \\ U_t^n \end{bmatrix}^\top \otimes \Sigma_w^{-1}, \tag{7}$$

which quantifies the uncertainty of the estimation procedure.

We consider three approaches to control synthesis using the estimates $\hat{\theta}$ and $\hat{\mathsf{FI}}$:

- **Certainty Equivalence (CE)** uses the estimate $\hat{\theta}$ to minimize the control objective (2) by treating the estimate as though it were ground truth: $K_{\mathsf{CE}}(\hat{\theta}) = \operatorname{argmin}_K C(K, \hat{\theta})$.
- **Robust Control (RC)** constructs a high confidence ellipsoid around the nominal estimate $\hat{\theta}$ using the estimated fisher information matrix $\hat{\mathsf{FI}}$ as

$$G = \left\{\theta : (\theta - \hat{\theta})^\top (N\hat{\mathsf{FI}})(\theta - \hat{\theta}) \le 16(d_\theta + \log(2/\delta))\right\}. \tag{8}$$

  Such a set can be shown to contain the true parameter $\theta^\star$ with probability at least $1-\delta$ as long as the number of experiments, N, is sufficiently large (see Fujinami et al. (2025)). RC then uses the confidence ellipsoid to determine a controller that minimizes the worst case value of the control objective over all members of the confidence set as

$$K_{\mathsf{RC}}(G) = \operatorname*{argmin}_K \sup_{\theta \in G}(C(K,\theta) - C(K(\theta),\theta)).$$

  This formulation is nonstandard, as the controller minimizes the worst-case suboptimality gap, rather than the worst case cost (Gevers, 2005). However, it simplifies the analysis.
- **Domain Randomization (DR)** constructs a sampling distribution $\mathcal{D}$ using the least squares estimate $\hat{\theta}$ and the estimated Fisher Information $\hat{\mathsf{FI}}$. It then synthesizes a controller by minimizing the average control cost as in (3).[4] By ensuring good performance on average over a distribution, DR serves as a middle ground between CE and RC. In particular, it can be interpreted as enforcing a high probability robust stability constraint over the sampling region.[5] Therefore, careful choice of the distribution is critical for downstream performance, and our analysis informs this choice.

The goal of this paper is to study the sample efficiency of these three approaches. In particular, we consider upper bounds on the gap $C(\hat{K}, \theta^\star) - C(K(\theta^\star), \theta^\star)$, where $\hat{K}$ is a controller synthesized with CE, RC, or DR. We express these bounds in terms of system-theoretic quantities, and the number of experiments collected from the system. Doing so provides an indication of the types of systems on which these methods perform well. We focus our attention on two key quantities: the burn-in time required to ensure finite bounds, and the asymptotic rate of decay in these bounds.

3. Sample Efficiency Bounds for Controller Synthesis Approaches

Our sample efficiency bounds are summarized in Table 1, where they are compared with existing bounds for CE. The leading term in the bound characterizes the asymptotic rate of decay. For all

[4] Note that DR directly minimizes the average cost rather than the average sub-optimality gap, unlike RC. However, doing so for DR leads to the same optimal solution by linearity of expectation.

[5] If the controller is not stabilizing for a subset of systems with nonzero mass, the cost will be infinite.

[6] Due to slight discrepancies in the setting (e.g. we consider multiple trajectories for identification), the exact version of the certainty equivalent bound considered is presented in Fujinami et al. (2025).

Method	Leading Term	Burn-in Time	Scalable Alg.	Source of Bounds
CE	$\frac{1}{N}\operatorname{tr}\big(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\big)$	$\lVert P(\theta^\star)\rVert^{10}$	Yes	Wagenmaker et al. (2021)[6]
RC	$\frac{1}{N}d_\theta \left\lVert H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right\rVert$	$\lVert P(\theta^\star)\rVert^{4}\vee\frac{1}{r^2}$	No	This paper
DR	$\frac{1}{N}\operatorname{tr}\big(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\big)$	$\lVert P(\theta^\star)\rVert^{11}\tau_{B(\theta^\star)}^{16}$	Yes	This paper

Table 1: Comparison of sample efficiency bounds for CE, RC, and DR. The leading term is stated up to universal constants. The burn-in time reports the components which are different between the three approaches. We use the shorthand $\tau_{B(\theta^\star)} = \lVert B(\theta^\star)\rVert \vee 1$. We classify algorithms as scalable if they are possible to implement via first order gradient-based approaches.

three synthesis approaches described in Section 2.1, the leading term depends on four quantities: the parameter dimension d_θ, the number of experiments N, the Fisher Information matrix $\mathsf{FI}(\theta^\star)$, and a matrix $H(\theta^\star)$ capturing the sensitivity of synthesis to the identification error:

$$H(\theta^\star) = \nabla_\theta^2 C(K(\theta), \theta^\star)|_{\theta=\theta^\star}.$$

Wagenmaker et al. (2021) show that $\frac{1}{N}\operatorname{tr}\big(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\big)$ is the optimal asymptotic rate achievable by any algorithm mapping a dataset (4) to a controller.[7] Accordingly, both CE and DR can be classified as sample-efficient, as they achieve this optimal rate of decay with respect to system-theoretic quantities. The bound on RC instead has a leading term of $d_\theta \left\lVert H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right\rVert \geq \operatorname{tr}\big(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\big)$, and therefore cannot be classified as efficient.[8]

Table 1 also highlights the burn-in time, the number of samples that suffice for the sample efficiency bounds to hold. The reported values omit terms common to all three methods. Among the differing quantities, the burn-in for CE scales with $\lVert P(\theta^\star)\rVert^{10}$, for DR it scales with $\lVert P(\theta^\star)\rVert^{11}\tau_{B(\theta^\star)}^{16}$, and for RC it scales with the max of $\lVert P(\theta^\star)\rVert^{4}$ and $\frac{1}{r^2}$, a term quantifying the robust stabilizability of $\theta^\star$. Although $\frac{1}{r^2}$ can be as large as $\lVert P(\theta^\star)\rVert^{10}$, it is often much smaller (see Section 3.2), suggesting that RC can achieve a much lower burn-in than the alternative approaches for many system instances.[9] Experiments further suggest that DR's burn-in can fall between that of CE and DR.

Finally, Table 1 highlights that CE and DR give rise to scalable gradient-based policy optimization algorithms, which can easily be extended to nonlinear and high dimensional systems. In contrast, solving the RC problem requires computationally challenging LMI-based approaches, even for fully observed linear systems.

3.1. Sample Efficiency of Domain Randomization

We now establish the characterization of DR in the final row of Table 1. To this end, we first define a burn-in time which enables a bound on the the least squares error:

$$N_{\mathsf{ID}} \triangleq \mathsf{poly}\left(\sum_{t=0}^{T-1}\left\lVert A(\theta^\star)^t \begin{bmatrix} I & B(\theta^\star)\end{bmatrix}\right\rVert, \lVert\Sigma_u\rVert, \lVert\Sigma_w\rVert, \lVert\mathsf{FI}(\theta^\star)\rVert, \frac{1}{\lambda_{\min}(\mathsf{FI}(\theta^\star))}\right). \tag{9}$$

A bound on least squares using this quantity and proofs for the remainder of the results in this section may be found in Fujinami et al. (2025). We first state a bound on the performance of DR that holds for general sampling distributions centered at $\hat{\theta}$.

[7] This lower bound is specified to our setting in Fujinami et al. (2025).

[8] While we lack a formal lower bound proving the inefficiency of RC, experiments support this conclusion (see Fig. 2).

[9] We emphasize that these burn-in conditions are sufficient but not necessary; refining them is left to future work.

Lemma 1 *Suppose the dataset $\left\{(X_t^n, U_t^n, X_{t+1}^n)\right\}_{t=1,n=1}^{T,N}$ is collected from N trajectories of the system (1) via a random control input $U_t \sim \mathcal{N}(0, \Sigma_u)$. Let $\hat{\theta}$ be the least square estimate computed by (5). Let $\mathcal{D}$ be any distribution with mean $\hat{\theta}$, and which is supported on a set with diameter bounded by $\frac{1}{256}\left\|P(\theta^\star)\right\|^{-5}$. It holds with probability at least $1-\delta$ that*

$$C(K_{\mathsf{DR}}(\mathcal{D}), \theta^\star) - C(K(\theta^\star), \theta^\star) \leq \frac{8\mathrm{tr}(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1})}{N} + 2\,\mathrm{tr}(\mathbf{V}(\mathcal{D})H(\theta^\star))$$
$$+ 16\frac{\left\|H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right\|}{N}\log\frac{2}{\delta} + L_{\mathsf{DR}}(\theta^\star)\frac{\left\|\mathsf{FI}(\theta^\star)^{-1}\right\|^{3/2}}{N^{3/2}}, \tag{10}$$

where $L_{\mathsf{DR}}(\theta^\star, \delta) = \mathsf{poly}(d_\theta, \max\{1, \|B(\theta^\star\|\}, \left\|P(\theta^\star)\right\|, \log\frac{1}{\delta})$, as long as the number of experiments N satisfies $N \geq \max\left\{N_{\mathsf{ID}},\ \frac{c\|P(\theta^\star)\|^{11}(\|B(\theta^\star)\|\vee 1)^{16}\left(d_\theta+\log\frac{2}{\delta}\right)}{\lambda_{\min}(\mathsf{FI}(\theta^\star))}\right\}$ for a universal constant c.

To minimize the above upper bound, choosing a sampling distribution with a variance of zero would be best and would lead to completely canceling the term with $\mathbf{V}(\mathcal{D})$. However, this would eliminate any benefits that we hope to see in the low data regime. Instead, to maximize the potential robustness benefits, we should choose the distribution with the most spread, which does not significantly degrade the performance achieved for large N (the regime where the above bound is valid). We therefore propose choosing $\mathcal{D}$ as a uniform distribution over the confidence ellipsoid constructed for RC (8). Computing the variance of this quantity demonstrates that the term $\mathrm{tr}(\mathbf{V}(\mathcal{D})H(\theta^\star))$ scales as $\frac{1}{N}\,\mathrm{tr}\left(\hat{\mathsf{FI}}^{-1}H(\theta^\star)\right)\left(1+\frac{1}{d_\theta}\log\frac{2}{\delta}\right)$, leading to the following bound.[10]

Theorem 1 *Under the setting of Lemma 1, let $\mathcal{D}$ be a uniform distribution over the confidence ellipsoid G defined in Equation (8). Then it holds with probability at least $1-\delta$ that*

$$C(K_{\mathsf{DR}}(\mathcal{D}), \theta^\star) - C(K(\theta^\star), \theta^\star) \leq \frac{40\left(1+\log\left(\frac{2}{\delta}\right)/d_\theta\right)\mathrm{tr}(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1})}{N}$$
$$+ 16\frac{\left\|H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right\|}{N}\log\frac{2}{\delta} + L_{\mathsf{DR}}(\theta^\star, \delta)\frac{\left\|\mathsf{FI}(\theta^\star)^{-1}\right\|^{3/2}}{N^{3/2}}, \tag{11}$$

as long as N satisfies the burn-in time of Lemma 1.

The burn-in time from Lemma 1 provides the powers of $\|P(\theta^\star)\|$ and $\tau_{B(\theta^\star)}$ listed in Table 1. The leading term follows from the fact that $L_{\mathsf{DR}}(\theta^\star)$ is multiplied by $N^{-3/2}$, thus decays faster. The deviation terms (quantities multiplied by $\log\frac{2}{\delta}$) are not considered leading terms, because they are dominated by the term $\frac{1}{N}\,\mathrm{tr}\left(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right)$ when the bounds are converted from high probability bounds to bounds in expectation. This leaves a universal constant multiplied by $\mathrm{tr}\left(H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right)$, leading to the characterization of DR with the chosen distribution as sample efficient.

The bound above fails to demonstrate a clear advantage of DR over CE in the low-data regime, despite empirical evidence suggesting such benefits (Figure 2). By designing the sampling distribution to have support on the confidence ellipsoid (8), we ensure that the true system $\theta^\star$ has positive density in the sampling distribution with high probability. This design raises the hope that DR could reduce the burn-in time. However, we have not been able to prove this property, as we cannot exclude the possibility that for distributions with large support, the DR controller (3) might incur very high costs near $\theta^\star$ while performing well elsewhere. In contrast, we show in the sequel that RC can provide such benefits, albeit at the expense of sacrificing asymptotic efficiency.

[10] Note that any bounded distribution with a variance decaying faster than $\frac{1}{N}\mathsf{FI}(\theta^\star)$ achieves the same leading term.

3.2. Sample Efficiency of Robust Control

We now analyze the efficiency of RC. Our goal is to demonstrate how RC addresses the limitations of CE with limited data. To achieve this, we introduce a formal definition of robust stabilizability. In this definition, we denote the state covariance of system θ under controller K by $\Sigma^K(\theta)$.

Definition 2 (Robust Stabilizability) *Let $M \in \mathbb{R}$ and G be a set. G is M-robustly stabilizable by CE if $\exists \theta \in G$ such that for all $\theta' \in G$, $\left\|\Sigma^{K(\theta)}(\theta')\right\| \leq M$. Let also $r \in \mathbb{R}$. θ is (M, r)-robustly stabilizable by CE if $\forall A \subseteq \mathcal{B}(\theta, r)$, A is M-robustly stablizable by CE.*

The above definition captures the idea that a CE controller synthesized for a some parameter in a set can stabilize every member of that set. This is stronger than merely assuming the existence of an arbitrary controller that stabilizes all members, and plays a key role in our analysis of RC. Relaxing this condition is left for future work. Nevertheless, for any stabilizable system, we can choose r sufficiently small to satisfy the condition. Specifically, by Theorem 3 of Simchowitz and Foster (2020), if $r \leq \frac{1}{256}\left\|P(\theta)\right\|^{-5}$, then θ is (M, r)-robustly stabilizable by CE with $M = 2\left\|P(\theta)\right\|$. However, the given condition is system-specific and can often be satisfied with larger r.

Example 1 *Consider a scalar linear dynamical system: $X_{t+1} = a^\star X_t + b^\star U_t + W_t \quad \forall t \geq 0$, where $a^\star = 1.05$ and $b^\star = 1$. Suppose that only $a^\star$ is unknown. Consider the LQR problem defined by $Q = 1$, and $R = 1000$. Computing $\|P(\theta^\star)\|$, it holds by Simchowitz and Foster (2020) that the instance is (M, r) stabilizable with $M = 225$ and $r = 2 \times 10^{-13}$. However, we can achieve a better characterization for this instance by noting that for any subset G of the interval $[0.3, 1.8]$, synthesizing an LQR controller k using the largest value of the parameter in G ensures that $a + b^\star k < 0.97$ for all $a \in G$, thereby ensuring that $\left\|\Sigma^k(a)\right\| \leq \frac{1}{1-.97^2} \leq 20$ for all $a \in G$. Then the instance is $(20, 0.75)$-robustly stabilizable by CE.*

To illustrate the impact on control performance, suppose we have an estimate $\hat{a} = 1.01$. The CE controller derived from this estimate is $k = -0.0424$, which fails to stabilize the true system. In contrast, if we apply RC over any uncertainty set within the interval $[0.3, 1.8]$ that includes $a^\star$, we could synthesize a controller that stabilizes the system, thereby avoiding infinite cost.

With the definition of robust stabilizability by CE in hand, we proceed to state an upper bound on the excess cost incurred by the robust controller.

Theorem 3 *Suppose the dataset $\left\{(X_t^n, U_t^n, X_{t+1}^n)\right\}_{t=1,n=1}^{T,N}$ is collected from N trajectories of the system (1) via a random control input $U_t \sim \mathcal{N}(0, \Sigma_u)$. Let $\hat{\theta}$ be the least square estimate computed by (5), and G be the confidence ellipsoid of (8). Choose $r > 0$. Let M be the smallest real number such that $\theta^\star$ is (M, r)-robustly stabilizable. It holds that with probability at least $1 - \delta$*

$$C(K_{RC}(G), \theta^\star) - C(K(\theta^\star), \theta^\star) \leq \frac{64\left(d_\theta + \log \frac{2}{\delta}\right) \left\|H(\theta^\star)\mathsf{FI}(\theta^\star)^{-1}\right\|}{N} + L_{\mathsf{RC}}(\theta^\star, \delta, M) \frac{\left\|\mathsf{FI}(\theta^\star)^{-1}\right\|^{3/2}}{N^{3/2}},$$

where $L_{\mathsf{RC}}(\theta^\star, \delta, M) = \mathsf{poly}(d_\theta, \max\{1, \|B(\theta^\star)\|\}, \|P(\theta^\star)\|, \log \frac{1}{\delta}, M)$, as long as $N \geq \max\left\{N_{\mathsf{ID}}, \frac{c\|P(\theta^\star)\|^4\left(d_\theta + \log \frac{2}{\delta}\right)}{\lambda_{\min}(\mathsf{FI}(\theta^\star))}, \frac{c\left(d_\theta + \log \frac{2}{\delta}\right)}{r^2 \lambda_{\min}(\mathsf{FI}(\theta^\star))}\right\}$ for a universal constant c.

A proof of this result is provided in Fujinami et al. (2025). Similar to Theorem 1, the leading term in Table 1 is obtained by omitting the deviation term ($\log \frac{2}{\delta}$) and the lower-order term scaling with

$N^{-3/2}$. While we lack a lower bound for RC, we conjecture that the leading term is tight. The reasoning behind this is that RC selects the worst-case perturbation of the parameter within the confidence set, which naturally leads to dependence on the operator norm, as opposed to the trace observed in alternative approaches. The burn-in time from the theorem demonstrates how the robust stabilizability condition combined with the RC approach leads to the value reported in Table 1.

4. Numerical Experiments

We validate the trends predicted in Table 1 through a case study on the linear system

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix},$$
$$B = I, Q = 10^{-3} I, R = I. \tag{12}$$

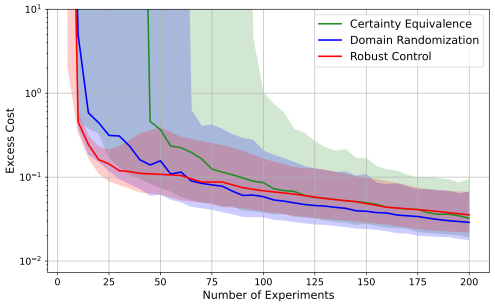

Figure 2: Excess cost of controllers found via three methods using models fit with various amounts of data.

We first estimate A and B using least squares identification, then synthesize CE, DR, and RC controllers based on the identified models using the confidence parameter $\delta = 0.5$. Further details are provided in Fujinami et al. (2025).

In Figure 2, we plot the median and shade 25% to 75% quantile over 500 random seeds. This result reveals two key observations. First, both DR and RC stabilize the system with fewer experiments than CE. While the improved performance of DR in the low-data regime lacks concrete theoretical justification, Theorem 3 supports this trend for RC. Second, after an initial period, DR converges faster than RC, eventually matching the convergence rate of CE. This aligns with the conclusion of Theorem 1 and is consistent with the conceptual illustration in Figure 1.

5. Discussion

Algorithmic considerations: Our numerical experiments use a gradient-based algorithm for DR, detailed in Algorithm 1. This algorithm builds on the scenario approach of Vidyasagar (2001) and the policy gradient method introduced for the LQR by Fazel et al. (2018). It initially samples a number of scenarios from the distribution. At each iteration, the gradient update is performed on the cost summed over all scenarios. Since the control cost gradient becomes infinite if the current iterate does not stabilize a system, we incorporate only the gradients for system which the current iterate stabilizes. Such rejection sampling only kicks in when the uncertainty set is large (i.e. when data is scarce). Thus it does not affect the asymptotic behavior of the method. We lack formal convergence guarantees for this algorithm; however, numerical experiments indicate that it converges with a sufficiently small step size. This raises questions which may be fruitful directions for future work.

- **Convergence analysis of Algorithm 1**: Extending the convergence analysis of policy gradient methods for LQR by Fazel et al. (2018); Hu et al. (2023) to the proposed algorithm could provide theoretical guarantees for convergence to the solution of (3). Such an analysis would offer strong evidence of the algorithm's applicability beyond the toy numerical example presented here.
- **Alternative choices of distribution:** We proposed sampling from a uniform distribution over the confidence ellipsoid (8) to maximize the spread of the sampling distribution without sacrificing asymptotic efficiency (Theorem 1). Exploring alternative distributions, such as truncated

Algorithm 1 Domain Randomized Policy-Gradient for the Linear Quadratic Regulator

```
Input: Randomization distribution $\mathcal{D}$, estimate $\hat{\theta}$, stepsize $\eta$, # iterations $M$, # scenarios $N$
Initialize: $\hat{K}_1 \leftarrow K_{CE}(\hat{\theta})$,
Sample: $N$ scenarios $\theta_1, \dots, \theta_N \sim \mathcal{D}$
for $i = 1, 2, \dots, M$ do \\ Gradient descent on stable scenarios
    $\hat{K}_{i+1} = \hat{K}_i - \eta \sum_{j=1}^{N} \nabla C(\hat{K}_i, \theta_j) \mathbf{1}(\rho(A(\theta_j) + B(\theta_j)\hat{K}_i) < 1)$
Return: $\hat{K}_M$.
```

normal distributions, is an interesting direction. Such alternatives could yield improved empirical performance, and also refine the analysis of Theorem 1, particularly with respect to burn-in time.

- **Extension to nonlinear systems:** Algorithm 1 can, in principle, be extended to nonlinear systems, provided that gradients of the control objective can be obtained via Monte Carlo sampling. The least-squares analysis for CE in Table 1 also extends nonlinear systems (Lee et al., 2024), suggesting that the proposed DR approach could also be effective for such systems. This may yield sample efficiency guarantees analogous to those studied here.

Theoretical extensions: There are several ways to tighten the analysis and generalize the setting.

- **Improved burn-in time for domain randomization:** While this work empirically demonstrates that DR can stabilize the system even in the low-data regime, the burn-in time derived in Lemma 1 does not reflect this advantage, as it is larger than that of CE. The only known way to improve the burn-in time is by adopting a robustly stabilizable condition, which considers the worst case but results in a conservative upper bound on the excess cost, as shown in Theorem 3. A promising direction for future work is to tighten the analysis for burn-in requirements and higher order terms of DR while maintaining its asymptotic efficiency.
- **Robust control lower bound:** Our upper bounds feature a gap between the asymptotic convergence rate of RC and that of CE and DR. We conjecture that this gap is fundamental; however, a lower bound on the sample efficiency of RC would be required to formalize this.
- **Misspecification:** This work has focused on the use of DR to address model uncertainty arising from variance in model fitting. Specifically, the assumption of the dynamics in (1) imposes a realizability condition, ensuring that a suitably constructed distribution $\mathcal{D}$ contains $\theta^\star$ in its support with high probability. However, a key explanation for the empirical success of DR in many robotics applications is its robustness to model misspecification (Tobin et al., 2017). Investigating this theoretically represents a promising direction for future work.

6. Conclusion

By analyzing the sample efficiency of learning the linear quadratic regulator via domain randomization and robust control, our work provides insights into the tradeoffs present for approaches to incorporate uncertainty quantification into learning-enabled control. Our analysis demonstrates that if one is strategic about the design of the sampling distribution, then the benefits of domain randomization over robust control may extend beyond computational considerations, and to the sample efficiency. This is particularly exciting due to the prominence of domain randomization in practice for robot learning. We believe that this line of analysis exposes a wide spread of interesting questions regarding the use of domain randomization for learning-enabled control.

Acknowledgments

We thank Manfred Morari, Anastasios Tsiamis, Ingvar Ziemann, and Thomas Zhang for several 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

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Julian Berberich, Anne Koch, Carsten W. Scherer, and Frank Allgöwer. Robust data-driven state-feedback design. In *2020 American Control Conference (ACC)*, pages 1532–1538, 2020. doi: 10.23919/ACC45564.2020.9147320.

Giuseppe Carlo Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753, 2006.

Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*, 2021.

Alessandro Chiuso, Marco Fabris, Valentina Breschi, and Simone Formentin. Harnessing uncertainty for a separation principle in direct data-driven predictive control. *arXiv preprint arXiv:2312.14788*, 2023.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

John Doyle. Analysis of feedback systems with structured uncertainties. In *IEE Proceedings D (Control Theory and Applications)*, volume 129, pages 242–250. IET Digital Library, 1982.

M.K.H. Fan, A.L. Tits, and J.C. Doyle. Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics. *IEEE Transactions on Automatic Control*, 36(1):25–38, 1991. doi: 10.1109/9.62265.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

Tesshu Fujinami, Bruce D Lee, Nikolai Matni, and George J Pappas. Domain randomization is sample efficient for linear quadratic control. *arXiv preprint arXiv:2502.12310*, 2025.

Michel Gevers. Identification for control: From the early achievements to the revival of experiment design. *European journal of control*, 11(4-5):335–352, 2005.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

Nan Jiang. Pac reinforcement learning with an imperfect model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bruce D Lee, Ingvar Ziemann, Anastasios Tsiamis, Henrik Sandberg, and Nikolai Matni. The fundamental limitations of learning linear-quadratic regulators. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4053–4060. IEEE, 2023.

Bruce D Lee, Ingvar Ziemann, George J Pappas, and Nikolai Matni. Active learning for control-oriented identification of nonlinear systems. *arXiv preprint arXiv:2404.09030*, 2024.

Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

Antonio Loquercio, Elia Kaufmann, René Ranftl, Alexey Dosovitskiy, Vladlen Koltun, and Davide Scaramuzza. Deep drone racing: From simulation to reality with domain randomization. *IEEE Transactions on Robotics*, 36(1):1–14, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J Pal, and Liam Paull. Active domain randomization. In *Conference on Robot Learning*, pages 1162–1176. PMLR, 2020.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528.

Laura Ryan Ray and Robert F Stengel. A monte carlo approach to the analysis of control system robustness. *Automatica*, 29(1):229–236, 1993.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, 2019.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Robert F Stengel and Laura 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.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter 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)*, pages 23–30, 2017.

Mathukumalli Vidyasagar. Randomized algorithms for robust controller synthesis using statistical learning theory. *Automatica*, 37(10):1515–1528, 2001.

Alexander von Rohr, Friedrich Solowjow, and Sebastian Trimpe. Improving the performance of robust control through event-triggered learning. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 3424–3430. IEEE, 2022.

Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652. PMLR, 2021.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

Yuren Zhong, Aniket Anand Deshmukh, and Clayton Scott. Pac reinforcement learning without real-world feedback. *arXiv preprint arXiv:1909.10449*, 2019.

K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Feher/Prentice Hall Digital and. Prentice Hall, 1996. ISBN 9780134565675.