2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

How are policy gradient methods affected by the limits of control?

Ingvar Ziemann[1], Anastasios Tsiamis[2], Henrik Sandberg[1], and Nikolai Matni[3]

[1]Division of Decision and Control Systems, KTH Royal Institute of Technology
[2]Automatic Control Laboratory, ETH Zurich
[3]Department of Electrical and Systems Engineering, University of Pennsylvania

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992612

***Abstract*— We study stochastic policy gradient methods from the perspective of control-theoretic limitations. Our main result is that ill-conditioned linear systems in the sense of Doyle inevitably lead to noisy gradient estimates. We also give an example of a class of stable systems in which policy gradient methods suffer from the curse of dimensionality. Finally, we show how our results extend to partially observed systems.**

I. Introduction

Reinforcement learning (RL) methods have shown great empirical success in controlling complex dynamical systems [1]. While these methods are promising, we have only begun to understand performance guarantees and fundamental limitations in continuous state and action problems. Providing such guarantees and understanding such limitations is crucial to deploying these methods in safety-critical systems. In this paper, we focus on a particular class of such methods; namely, we seek to understand fundamental limitations for policy gradient methods.

Policy gradient methods are a relatively simple class of algorithms that have been recently analyzed in the context of the linear quadratic regulator (LQR), [2], [3]. The motivation for studying policy gradients in the context of LQR stems from that it serves as an analytically tractable benchmark for RL in continuous state and action spaces. For instance, by direct arguments on can show that control-theoretic parameters affect the hardness of both offline and online learning in LQR [4]–[7]. Here, we extend this line of work and show that the popular policy gradient methods degrade similarly for systems with poor controllability and observability. To be precise, we show that ill-conditioned systems lead to arbitrarily noisy stochastic gradients.

a) Problem Formulation: We are interested in studying how policy gradient methods applied to the linear system

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad x_0 = 0 \quad t = 0, 1, \dots \tag{1}$$

are affected by the fundamental limits of control. Above, $x_t \in \mathbb{R}^{d_x}, A \in \mathbb{R}^{d_x \times d_x}$, $u_t \in \mathbb{R}^{d_u}$ and $w_t \in \mathbb{R}^{d_x}$ is an i.i.d. mean zero sequence of Gaussian noise with covariance matrix $\Sigma_W \in \mathbb{R}^{d_x \times d_x}$.

The learning task is to minimize

$$J_S(K) \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{K,S} \left[x_t^\top Q x_t + u_t^\top R u_t \right] \tag{2}$$

subject to the dynamics (1) without access to the model parameter $S = (A, B)$. In equation (2), $\mathbf{E}_{K,S}$ denotes expectation under the control law K with dynamics S.

In this work, we relate the efficiency of stochastic policy gradient methods to certain control-theoretic parameters. Namely, we analyze algorithms of the form

$$\widehat{K} \leftarrow \widehat{K} - \alpha \widehat{\nabla_K J(K; S)}$$

for some learning rate $\alpha \in \mathbb{R}_+$, and where at each iteration $\widehat{\nabla_K J(K; S)}$ is estimated using data from the system (1). Such algorithms have been shown to converge for LQR by [2]. The purpose of this work is to demonstrate that any estimate $\widehat{\nabla_K J(K; S)}$ from an (arbitrarily) ill-conditioned system (1) is (arbitrarily) noisy.

To make this statement rigorous we need to model the statistical information available to the learner. Here, we model this as follows: the learner is given access to $N \in \mathbb{N}$ experiments $(x_{0,n}, \dots, x_{T,n}), n \in [N]$ of length $T \in \mathbb{N}$ and a total input budget of βNT with $\beta \in \mathbb{R}_+$. More precisely, the learner is allowed to freely choose $u_{t,n}$ as a function of past observations $(x_{0,n}, \dots, x_{t,n})$ and past trajectories $(x_{0,m}, \dots, x_{T,m}), m < n$ and possible auxiliary randomization, while being constrained to a total budget

$$\sum_{n=1}^{N} \sum_{t=0}^{T-1} \mathbf{E} u_{t,n}^\top u_{t,n} \le \beta NT. \tag{3}$$

This formulation allows both open- and closed-loop experiments but normalizes the average input energy to β.

A. Related Work

The first proof that policy gradient methods converge for LQR is given in [2], which provides nonasymptotic guarantees that are polynomial in relevant problem parameters. Convergence guarantees for more general MDPs and other versions of LQR are given in [8]–[12]. Extensions to partially observed systems are considered in [13]–[15]. A popular alternative approach to policy gradients for LQR is based on certainty equivalence [16].

Most closely related to our work are [3], [17], [18]. While we give lower bounds valid for any estimator in this work, [17] analyzes the variance of a particular gradient estimator known as REINFORCE. Similarly, [3] provides algorithm specific lower bounds which demonstrate, among other things, that if $R = 0$ and B is invertible, then learning

fails as $\|B\| \to 0$. Further, [3] gives a generic performance lower bound for offline methods which however does not scale with relevant system-theoretic quantities.

Our work also relates to [4]–[7], which also study fundamental limits in learning-enabled control. From a broader perspective, the present work fits into a line of work that strives to ascertain the interplay between control-theoretic performance, stability and robustness notions, and learning [19]–[23]. While we are mostly interested in offline methods in this work, analyses of online LQR can be found in the literature, see [6], [24]–[26] and the references therein.

B. Contribution

We show that policy gradient methods are very much affected by the limits of control. Our main result (Theorem 1) demonstrates that state feedback systems operating near marginal stability suffer from noisy gradients. This happens for instance if the system has poorly controllable unstable modes. We also provide an analogue of this result for partially observed systems (Theorem 2), which we use to show that systems with bad (small) Markov parameters also lead to noisy gradients. Compared to previous literature on this topic [3], [17], [18], our results provide a more fine-grained theoretical understanding of when and why gradient methods applied to dynamical systems fail.

C. Preliminaries

A matrix A is stable if $\rho(A) < 1$. A matrix K is stabilizing for the system $S = (A, B)$ if $A + BK$ is a stable matrix. If K is stabilizing for the system $S = (A, B)$, the closed-loop controllability gramian

$$\Gamma_{K,S} \triangleq \sum_{t=0}^{\infty} (A+BK)^t \Sigma_W (A+BK)^{t,\top} \tag{4}$$

is well-defined. The set of all systems S for which there exists a stabilizing K is denoted by $\mathcal{S} = \mathcal{S}_{d_x,d_u}$, which is an open subset of $\mathbb{R}^{d_x \times d_x + d_x \times d_u}$ in norm topology. Further, we define $K_\star(S)$ as (any element of) $K_\star(S) \in \operatorname{argmin}_K J_S(K)$. Moreover, we denote the matrix operator norm (induced $l^2(\mathbb{R}^d) \to l^2(\mathbb{R}^d)$) by $\|\cdot\|_{\mathsf{op}}$.

We also require the following information-theoretic quantities. We define the Kullback-Leibler divergence between two probability measures $\mathbf{P}$ and $\mathbf{Q}$ as $d_{\mathsf{KL}}(\mathbf{P}, \mathbf{Q}) \triangleq \int \frac{d\mathbf{P}}{d\mathbf{Q}} d\mathbf{P}$ and the total variation distance as $d_{\mathsf{TV}}(\mathbf{P}, \mathbf{Q}) \triangleq \int |d\mathbf{P} - d\mathbf{Q}|$. When $\mathbf{P}$ and $\mathbf{Q}$ correspond to induced probability measures from two systems S_1 and S_2 we abuse notation and write $d_{\mathsf{KL}}(S_1, S_2) = d_{\mathsf{KL}}(\mathbf{P}, \mathbf{Q})$ and $d_{\mathsf{TV}}(S_1, S_2) = d_{\mathsf{TV}}(\mathbf{P}, \mathbf{Q})$ for divergences between the corresponding parametric families.

It will be convenient to introduce the shorthand $a_t \lesssim b_t$ if there exists a universal constant C such that $a_t \leq C b_t$ for every $t \geq t_0$ and some $t_0 \in \mathbb{N}$. If $a_t \lesssim b_t$ and $b_t \lesssim a_t$ we write $a_t \asymp b_t$. For an integer N, we also define $[N] \triangleq \{1, \dots, N\}$.

a) Policy Gradients: We begin by recalling a standard characterization of the LQR cost (2) for a linear controller K. A version of the following lemma can also be found in for instance [2].

Lemma 1: If K is stabilizing for $S = (A, B)$, the LQR cost can be written as

$$J(K; S) = \operatorname{tr} P_K \Sigma_W$$

where P_K satisfies the Lyapunov equation

$$P_K = Q + K^\top R K + (A+BK)^\top P_K (A+BK). \tag{5}$$

Lemma 1 allows us to conveniently characterize the policy gradient $\nabla_K J(K; S)$.

Lemma 2: Let K be stabilizing for $S = (A, B)$. The policy gradient $\nabla J(K; S)$ can be written as

$$\nabla_K J(K; S) = 2\left((R + B^\top P_K B)K + B^\top P_K A\right) \Gamma_{K,S} \tag{6}$$

where P_K satisfies the Lyapunov equation (5) and where $\Gamma_{K,S}$ is given by definition (4).

Combining Lemmas 1 and 2 we see that we are almost in the same setting as studied in [2]. The difference is mainly in how information is acquired, since each sample from the system (1) is noisy and conditionally Gaussian. Compared to the noise-free random initial condition setting considered in [2], this simplifies the analysis of the variance of the gradient estimates since we later rely on the closed form of the KL divergence for Gaussians of different means.

II. Policy Gradient Estimation Lower Bounds

Let us begin our study of stochastic gradient methods by the observation that $\nabla_K J(K; S)$ diverges if K does not stabilize the system (1). Consider for instance the following two systems

$$\begin{cases} S_1 : x_{t+1} = a x_t + b u_t + w_t, \\ S_2 : x_{t+1} = a x_t - b u_t + w_t. \end{cases} \tag{7}$$

If $|a| > 1$ there exists no linear feedback controller which stabilizes both S_1 and S_2 of equation (7). Hence, any policy gradient which is finite for the first system will be infinite for the second system and vice versa. Combining this observation with the two point method (Lemma 7) leads to the following conclusion.

Proposition 1: For any $k \in \mathbb{R}$ the global minimax complexity of estimating the policy gradient at k is infinite:

$$\inf_{\widehat{\nabla J}} \sup_{(a,b) \in \mathbb{R}^2} \mathbf{E}_{(a,b)} \left| \frac{d}{dk} J(k; (a,b)) - \widehat{\nabla J} \right| = \infty \tag{8}$$

where the infimum is taken over all measurable functions of the data $(x_{0,n}, u_{0,n}, \dots, u_{T-1,n}, x_{T,n}), n \in [N]$.

Proposition 1 shows that the global minimax complexity of estimating gradients is infinite. While this shows that estimating gradients can be hard, it does not reveal how this hardness depends on control theoretic parameters.

A. Local Minimax Complexities

In order to understand what properties of a particular system makes learning to control hard, we need to consider local complexity measures. Here, we investigate the (d,ε)-local minimax complexity of estimating gradients. We define this as

$$\mathfrak{M}_d(\varepsilon;S,K) \triangleq \inf_{\widehat{\nabla J}} \sup_{S':d(S,S')\le\varepsilon} \mathbf{E}_{S'} \left\| \nabla_K J(K;S') - \widehat{\nabla J} \right\|_{\mathsf{op}} \tag{9}$$

for some metric d on the set of stabilizable systems $\mathcal{S}$ and where the infimum is taken over all measurable functions of the data $(x_{0,n}, u_{0,n}, \dots, u_{T-1,n}, x_{T,n}), n \in [N]$. This captures a more instance-specific notion of how hard it is to estimate gradients. Roughly, this complexity measure corresponds to requiring algorithms to performing well not just on a nominal system S, but also on small ε-perturbations of that system.

Note further that the definition (9) still leaves open the question at which K to measure the complexity of estimating gradients. By equation (6) and Proposition 1 we know that $\nabla_K J(K)$ can be arbitrarily large when evaluated far from a stationary point. As this rather reflects poor initialization than fundamental control-theoretic hardness, we instead seek to lower bound $\mathfrak{M}_d(\varepsilon;S,K)$ near $K_\star(S)$. Arguably, any successful policy gradient algorithm should eventually find itself near $K_\star(S)$. Thus, we provide lower bounds on the gradient estimation error in the vicinity of the stationary point $K_\star(S)$. We denote the associated local complexity by $\mathfrak{M}_d(\varepsilon;S) = \mathfrak{M}_d(\varepsilon;S,K_\star(S))$.

a) Constructing Hard Instances: To simplify the evaluation of the local minimax complexity (9), we mainly consider the construction below. Fix a nominal instance $S_1 = (A,B)$ of system (1) with optimal control law $K_\star = K_\star(S_1)$; then the perturbation

$$S_2: \quad x_{t+1} = A' x_t + B' u_t + w_t \tag{10}$$

is tractable to evaluate. Here, $A' = A - \Delta K_\star$ and $B' = B + \Delta$ for some $\Delta \in \mathbb{R}^{d_x \times d_u}$. This perturbation is convenient since $A' + B'K_\star = A + BK_\star$ for any Δ and has previously been used in [6], [25]. In particular, the system quantities $P_{K_\star,S}$ and $\Gamma_{K_\star,S}$ are invariant as we vary Δ. Combining this observation with the optimality of $K_\star = K_\star(S_1)$ for system S_1, yields the following simple expression for the gradient of system (10).

Lemma 3: The policy gradient for $S_2 = (A',B')$ given by system (10) at $K_\star = K_\star(S_1)$ is given by

$$\nabla_K J(K_\star;S_2) = 2\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}$$

Proof of Lemma 3. By Lemma 2 the policy gradient is given by

$$\begin{aligned}&\nabla_K J(K_\star;S_2)\\ &= 2\left(RK_\star + (B+\Delta)^\top P_{K_\star,S_1}(A+BK_\star)\right)\Gamma_{K_\star,S_1}\end{aligned}$$

where we used that $A + BK_\star = A' + B'K_\star$. On the other hand

$$2\left(RK_\star + B^\top P_{K_\star,S_1}(A+BK_\star)\right)\Gamma_{K_\star,S_1} = 0$$

by optimality of $K_\star$ to S_1. The result follows. ∎

By combining Lemma 3 with Le Cam's two point method [27] (provided in the appendix as Lemma 7) we obtain a generic estimation lower bound for policy gradients evaluated in the vicinity of the optimum $K_\star$.

Theorem 1: Consider two systems $S_1 = (A,B)$ and $S_2(\Delta) = (A',B')$ with $A' = A - \Delta K_\star$ and $B' = B + \Delta$. Let $\mathfrak{M}_d(\varepsilon;S_1) = \mathfrak{M}_d(\varepsilon;S_1,K_\star(S_1))$ and $K_\star = K_\star(S_1)$, then

$$\begin{aligned}&\mathfrak{M}_d(\varepsilon;S_1)\\ &\ge \sup_{d(S_1,S_2(\Delta))\le\varepsilon} \left\|\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}}\\ &\times \left(1 - \sqrt{\frac{1}{2} d_{\mathsf{KL}}(S_1,S_2(\Delta))}\right).\end{aligned} \tag{11}$$

In other words, the local complexity of estimating gradients can be lower bounded by the maximum of $\left\|\Delta^\top P_{K_\star,S_1}(A+BK)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}}$, optimized over Δ and subject to this leading to small differences in the output of systems S_1 and S_2.

Proof of Theorem 1. Define the loss function $L(\mathsf{dec},S) \triangleq \|\nabla_K J(K;S) - \mathsf{dec}\|_{\mathsf{op}}$, where the decision dec is a placeholder variable for the gradient estimate. For any two systems S_1 and S_2 we have that

$$\begin{aligned}&L(\mathsf{dec},S_1) + L(\mathsf{dec},S_2)\\ &= \|\nabla_K J(K_\star;S_1) - \mathsf{dec}\|_{\mathsf{op}} + \|\nabla_K J(K_\star;S_2) - \mathsf{dec}\|_{\mathsf{op}}\\ &\ge \|\nabla_K J(K_\star;S_1) - \nabla_K J(K_\star;S_2)\|_{\mathsf{op}}\end{aligned}$$

by the triangle inequality. Invoking Lemma 3 we thus see that for the choice $S_1 = (A,B)$ and $S_2 = (A',B')$ with $A' = A - \Delta K_\star$ and $B' = B + \Delta$ we have

$$L(\mathsf{dec},S_1) + L(\mathsf{dec},S_2) \tag{12}$$
$$\ge \left\|2\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}} \tag{13}$$

for any $\Delta \in \mathbb{R}^{d_x \times d_u}$. Combining equation (12) with Lemma 7 it follows that

$$\begin{aligned}\mathfrak{M}(\varepsilon,S_1) &\ge \left\|\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}}\\ &\times (1 - d_{\mathsf{TV}}(S_1,S_2))\\ &\ge \left\|\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}}\\ &\times \left(1 - \sqrt{\frac{1}{2} d_{\mathsf{KL}}(S_1,S_2)}\right)\end{aligned}$$

where the second inequality is an application of Pinsker's inequality. ∎

At this point, we note that the right hand side of inequality (11) is large for systems operating near marginal stability. When $A + BK_\star \to 1$ both $P_{K_\star,S}$ and $\Gamma_{K_\star,S}$ tend to infinity. To better understand the practical implications of this, we now turn to interpreting Theorem 1 by instantiating it for three special cases: scalar systems, over-actuated systems and integrator-like systems.

B. Consequences of Theorem 1

a) Scalar Systems: The bound in Theorem 1 is agnostic to the experiment used to generate the dataset, which is

simply reflected in the quantity $\left(1-\sqrt{\frac{1}{2}d_{\mathsf{KL}}(S_1,S_2(\Delta))}\right)$. Let us interpret Theorem 1 by a simple scalar example. To this end, consider the system

$$s_1 : x_{t+1} = ax_t + bu_t + w_t \tag{14}$$

with $a, b \in \mathbb{R}$, which is open-loop unstable $|a| > 1$. Let s_2 be given by the perturbation $s_2 = (a', b') = (a - \Delta k_\star, b + \Delta)$ with $\Delta \in \mathbb{R}$. The divergence $d_{\mathsf{KL}}(s_1, s_2)$ satisfies

$$\begin{aligned} d_{\mathsf{KL}}(s_1,s_2) &= \sum_{n=1}^{N}\sum_{t=0}^{T-1} \mathbf{E}_{s_1}\frac{1}{2}(k_\star \Delta x_t + \Delta u_t)^2 \quad \text{(Lemma 8)} \\ &\le \Delta^2 \sum_{n=1}^{N}\sum_{t=0}^{T-1} \mathbf{E}_{s_1}\left(u_t^2 + k_\star^2 x_t^2\right) \\ &\le \Delta^2 NT(k_\star^2 \Gamma_{k_\star,s_1} + \beta) \quad \text{(by (3))} \\ &= \frac{1}{2} \end{aligned} \tag{15}$$

if $\Delta^2 = \frac{1}{2NT(k_\star^2\Gamma_{k_\star,s_1}+\beta)}$. Plugging inequality (15) into inequality (11), we conclude that

$$\begin{aligned} &\mathfrak{M}_{d_\infty}(\varepsilon_{N,T}, s_1) \\ &\gtrsim \frac{1}{\sqrt{NT(\beta + k_\star^2\Gamma_{k_\star,s_1})}} \left|P_{k_\star,s_1}(a+bk_\star)\Gamma_{k_\star,s_1}\right| \end{aligned} \tag{16}$$

with $\varepsilon_{N,T} \asymp \frac{\max(k_\star,1)}{\sqrt{NT(\beta+k_\star^2\Gamma_{k_\star,s_1})}}$ and where $d_\infty(s_1,s_2) = \max(|a-a'|, |b-b'|)$. In particular as $|b| \to 0$, one may verify that the right hand side of the expression (16) tends to infinity. In other words, as controllability (of unstable modes) is lost, policy gradients become arbitrarily noisy. This is verified via simulations in the appendix (Figure 1) using both a least squares certainty equivalent approach and a 0-th order method (see [2, Algorithm 1]).

b) Multivariate Systems: If we assume that K has a left nullspace, the bound in Theorem 1 becomes tractable to evaluate since we are free to select Δ such that $\Delta K = 0$, which simplifies some calculations. Intuitively, these instances are hard to distinguish between because controllers with left nullspaces lead to identifiability issues regarding the B-matrix [5].

Corollary 1: For any Δ such that $\Delta K_\star = 0$ and $\|\Delta\|_{\mathsf{op}} \le 1$ we have that

$$\begin{aligned} &\mathfrak{M}_{d_\infty}(\varepsilon_{N,T}, S_1) \\ &\gtrsim \frac{1}{\sqrt{\beta NT}} \left\|\Delta^\top P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}\right\|_{\mathsf{op}} \end{aligned}$$

for any $\varepsilon_{N,T} \gtrsim 1/\sqrt{NT}$ and where $d_\infty(S_1,S_2) = \max(\|A-A'\|_{\mathsf{op}}, \|B-B'\|_{\mathsf{op}})$.

Proof of Corollary 1. Fix $\varepsilon > 0$. By Lemma 8 the two systems $S_1 = (A,B)$ and $S_2(N,T) = (A'_{N,T}, B'_{N,T})$ with $A' = A - \frac{\varepsilon}{\sqrt{\beta NT}}\Delta K_\star = A$ and $B' = B + \frac{\varepsilon}{\sqrt{\beta NT}}\Delta$ satisfy $d_{\mathsf{KL}}(S_1, S_2(N,T)) = O(1)$. The result follows by Theorem 1. ∎

In other words, the complexity of estimating gradients can be asymptotically lower bounded at the central limit theorem scale $\sqrt{NT}$ by the part of $P_{K_\star,S_1}(A+BK_\star)\Gamma_{K_\star,S_1}$ that cannot be identified by closed loop experiments using $K_\star$. It so happens that this complexity measure is very similar to that dictating regret lower bounds in adaptive LQR [5], [6]. In the sequel, we exploit this to show that the gradient variance can grow exponentially with the system dimension in the worse case by leveraging certain Riccati calculations due to [7].

c) Policy Gradients and the Curse of Dimensionality: Let us now show that variance of policy gradient estimates can suffer from exponential complexity in the dimension. The proof of this fact relies on a construction due to [7]. Namely, we consider a system consisting of two decoupled subsystems $S_1 = (A,B)$ of the form:

$$x_{t+1} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \rho & 2 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ & & & \ddots & & 0 \\ 0 & 0 & 0 & & \rho & 2 \\ 0 & 0 & 0 & \dots & 0 & \rho \end{bmatrix}}_{=A} x_t + \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \end{bmatrix}}_{=B} u_t + w_t \tag{17}$$

with $\rho \in (0,1)$, $Q = I_{d_x}$ and $R = I_2$. We also define the subsystem

$$A_0 = \begin{bmatrix} \rho & 2 & 0 & \dots & 0 & 0 \\ 0 & \rho & 2 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ & & & \ddots & & 0 \\ 0 & 0 & 0 & & \rho & 2 \\ 0 & 0 & 0 & \dots & 0 & \rho \end{bmatrix}, \; B_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{18}$$

with $Q_0 = I_{d_x-1}$ and $R_0 = 1$ and where $A_0 \in \mathbb{R}^{(d_x-1)\times(d_x-1)}$ and $B_0 \in \mathbb{R}^{d_x-1}$. Note that A_0 is a stable matrix since $|\rho| < 1$. In the notation of Theorem 1, we let $\Delta = \begin{bmatrix} 0 & 0 \\ \Delta_1 & 0 \end{bmatrix}$, so that S_2 consists of two weakly coupled subsystems, with coupling induced by Δ_1 (recall $S_2 = (A', B') = (A - \Delta K, B + \Delta)$).

Denote further by $P_{0,\star}$ the solution to the Lyapunov equation (5) for the subsystem (18) with $K_0 = K_{\star,0}$. Note also that $P_{0,\star}$ satisfies the discrete algebraic Riccati equation for the tuple (A_0, B_0, Q_0, R_0). With these preliminaries established, we now recall the following two lemmas from [7, Appendix E].

Lemma 4: We have:

$$\|\Delta_1^\top P_0(A_0 + B_0K_{\star,0})\|_{\mathsf{op}} \ge \left(\frac{1}{2} + o(1)\right)(B_0'P_0B_0 + R_0),$$

where the term $o(1)$ tends to 0 as d_x tends to infinity.

Lemma 5 (Riccati matrix can grow exponentially): For system (18) we have:

$$B_0^\top P_{0,\star} B_0 + R_0 \ge 2^{2d_x-4} + 1.$$

Combining Corollary 1 with Lemmas 4 and 5 we arrive at the following conclusion:

Proposition 2: For the system S given in equation (17) we have that

$$\mathfrak{M}_{d_\infty}(\varepsilon_T, S) \gtrsim \frac{4^{d_x}}{\sqrt{\beta NT}} \tag{19}$$

for d_x and NT sufficiently large.
In other words, there are classes of stable systems for which the policy gradient suffers from exponential complexity in the state dimension.

III. Extension to Partially Observed Systems

We now demonstrate that our lower bound approach extends to partially observed systems of the form

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t, \quad x_0 = 0 \quad t = 0, 1, \dots \\ y_t &= Cx_t + v_t \end{aligned} \tag{20}$$

in which A and B are as in system (1), $C \in \mathbb{R}^{d_y \times d_x}$ and both w_t and v_t are i.i.d. normal with mean zero and covariance Σ_W, Σ_V. We denote partially observed systems of the form (20) by $G = (A, B, C)$. For system (20) one typically seeks to learn dynamic controllers of the form (see e.g. [13])

$$\begin{aligned} \xi_{t+1} &= A_{\mathsf{dyn}}\xi_t + B_{\mathsf{dyn}}y_t, \quad \xi_0 = 0 \quad t = 0, 1, \dots \\ u_t &= K\xi_t \end{aligned} \tag{21}$$

parametrized by the linear system $K_{\mathsf{dyn}} = (A_{\mathsf{dyn}}, B_{\mathsf{dyn}}, K)$. The objective, as before is to minimize the cost

$$J_G(K_{\mathsf{dyn}}) \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{K_{\mathsf{dyn}},G} \left[x_t^\top Q x_t + u_t^\top R u_t\right] \tag{22}$$

but this time subject to process and controller dynamics (20)-(21).

a) Fully Observed Reformulation: To establish a hardness result, it suffices to focus on the difficulty of estimating gradients with respect to the output matrix of the controller (21), K. We will exploit this by reducing the system (20)-(21) when evaluated near the optimum of J_S (22) to a fully observed system. Namely, at the optimum $K_{\mathsf{dyn},\star} = \operatorname{argmin} J_S(K_{\mathsf{dyn}})$ the dynamics of ξ_t in equation (21) are given by the Kalman filter $\xi_t = \hat{x}_t$, allowing us to write

$$\begin{aligned} \hat{x}_{t+1} &= A\hat{x}_t + Bu_t + \nu_t \\ \nu_t &\sim N(0, \Sigma_t) \end{aligned} \tag{23}$$

which has the same input-output behavior as system (20) and where the sequence of innovations $\{\nu_t\}$ is independent. More precisely, the covariance Σ_t of ν_t is given by

$$\Sigma_t = L_t(CF_{t|t-1}C^\top + \Sigma_V)L_t^\top \tag{24}$$

where $F_{t|t-1}$ satisfies the filter Riccati recursion (see e.g. [28])

$$\begin{aligned} F_{t+1|t} = \Sigma_W + AF_{t|t-1}A^\top \\ - F_{t|t-1}C^\top(CF_{t|t-1}C^\top + \Sigma_V)^{-1}CF_{t|t-1} \end{aligned} \tag{25}$$

and the filter gain L_t is given by

$$L_t = F_{t|t-1}C^\top(CF_{t|t-1}C^\top + \Sigma_V)^{-1}. \tag{26}$$

We now consider the cost $J(K_{\mathsf{dyn}})$ (22) evaluated at the optimal filter (23) and with variable K. With some abuse of notation, we denote this quantity $J(K; G)$ where u_t is given by $u_t = K\hat{x}_t$, and $\hat{x}_t$ is defined by the Kalman filter (23). We shall call the quantity $J(K; G)$ the restricted cost function, and note that it has almost the exact same form as the fully observed cost (2). With these preliminaries established, the following lemma is straightforward to verify using Lemmas 1 and 2 (and justifies the abuse of notation $J(K; G)$ and $K_\star(G)$).

Lemma 6: Consider a partially observed system $G = (A, B, C)$ of the form (20). Then the restricted cost function satisfies

$$J(K; G) = \operatorname{tr} P_K \Sigma_{\nu,G} + \operatorname{tr} Q(I - L_G C)$$

where L_G and $\Sigma_{\nu,G}$ are the steady state quantities corresponding to recursions (25) and (26) respectively and where P_K as before is given by the Lyapunov equation (5). Moreover, the policy gradient is given by

$$\nabla_K J(K; G) = 2\left((R + B^\top P_K B)K + B^\top P_K A\right)\Gamma_{K,\nu,G} \tag{27}$$

where

$$\Gamma_{K,\nu,G} \triangleq \sum_{t=0}^{\infty} (A + BK)^t \Sigma_{\nu,G} (A + BK)^{t,\top}. \tag{28}$$

In other words, near the optimal controller $K_{\mathsf{dyn},\star}$ the gradient with respect to the filter gain K has the same form as in the state-feedback setting (1). However, we stress at this point that neither the realization of the system (20) nor the realization of the controller (21) is unique. To remedy this, we will later verify in a scalar setting that the perfomance limitations outlined here are invariant under similarity transformation (see equation (33)).

b) Recovering Theorem 1: In the partially observed setting, we keep the exploration budget constraint (3) but the observation model is necessarily different. Namely, we assume that the learner instead has access to input-output data of the form $(y_{0,n}, u_{0,n}, \dots, u_{T-1,n}, y_{T,n}), n \in [N]$.

In the partially observed setting, we thus define the analogous local minimax complexity as

$$\begin{aligned} &\mathfrak{M}_d(\varepsilon; G, K) \\ &\triangleq \inf_{\widehat{\nabla J}} \sup_{G' : d(G,G') \le \varepsilon} \mathbf{E}_{G'} \left\| \nabla_K J(K; G') - \widehat{\nabla J} \right\|_{\mathsf{op}} \end{aligned} \tag{29}$$

where the infimum is taken over all measurable functions of the data $(y_{0,n}, u_{0,n}, \dots, u_{T-1,n}, y_{T,n}), n \in [N]$, $\nabla_K J(K; G)$ is given by equation (27) and d again is a metric on system parameters $G = (A, B, C)$. We further set $\mathfrak{M}_d(\varepsilon; G) \triangleq \mathfrak{M}_d(\varepsilon; G, K_\star(G))$.

Equipped with the definition (29) and Lemma 6 the proof of the following result follows similarly to that of Theorem 1.

Theorem 2: Consider two systems $G_1 = (A, B, C)$ and $G_2(\Delta) = (A', B', C')$ with $A' = A - \Delta K_\star$, $B' = B + \Delta$ and $C' = C$. Then the local minimax complexity of estimating

gradients (29) is lower bounded as

$$\begin{aligned}\mathfrak{M}_d(\varepsilon; G_1) \\ \geq \sup_{d(G_1, G_2(\Delta)) \leq \varepsilon} & \left\| \Delta^\top P_{K_\star, G_1}(A + BK)\Gamma_{K_\star, \nu, G_1} \right\|_{\mathsf{op}} \\ & \times \left(1 - \sqrt{\frac{1}{2} d_{\mathsf{KL}}(G_1, G_2(\Delta))}\right). \end{aligned} \tag{30}$$

Above $d_{\mathsf{KL}}(G_1, G_2(\Delta))$ is the divergence between the probability measures over input-output data $(y_{0,n}, u_{0,n}, \ldots, u_{T-1,n}, y_{T,n}), n \in [N]$ when the true model is G_1 or G_2 respectively.

By the data-processing inequality, the lower bound (30) can be brought onto the exact same form as the lower bound (30). Namely, we observe that[1]

$$d_{\mathsf{KL}}(G_1, G_2(\Delta)) \leq d_{\mathsf{KL}}(S_1, S_2(\Delta))$$

where $d_{\mathsf{KL}}(S_1, S_2(\Delta))$ is a slight overload of notation for the divergence between state-input data $(x_{0,n}, u_{0,n}, \ldots, u_{T-1,n}, x_{T,n}), n \in [N]$ between models G_1 and G_2. In other words, all the results of Section II apply with $\Gamma_{K_\star, S_1}$ defined by equation (4) exchanged for $\Gamma_{K_\star, \nu, G_1}$ defined in equation (28) and Σ_w exchanged for Σ_t given by equation (24). While this is true for a fixed parametrization $G = (A, B, C)$, one may wonder whether the lower-bound relies on fundamental system-theoretic quantities or is simply a consequence of poor parametric choice for computing gradients. In the next example we show that the lower bound (30) captures control-theoretic limitations that are independent of the state-space representation (20).

c) Bad Markov Parameters Imply Noisy Gradients: Consider the almost scalar system $g_1 = (a, B, c)$ given by

$$\begin{aligned} x_{t+1} &= ax_t + \begin{bmatrix} b & 0 \end{bmatrix} u_t + w_t \\ y_t &= cx_t + v_t \end{aligned} \tag{31}$$

defined consistently with system (20), but specifically $a, b, c \in \mathbb{R}$ and $Q = \Sigma_V = \Sigma_W = 1$ and $R = I_2$. Note that the maximum singular values of the first Markov parameter of g_1 is equal to the product $m = cb$ and that this is m is invariant under similarity transformation. We consider the two systems $g_1 = (a, B, c)$ and $g_2 = (a, B(\Delta), c)$ and where $B(\Delta) = \begin{bmatrix} b & \Delta \end{bmatrix}$. Observe that the optimal policy to system (31) is of the form $K_\star = \begin{bmatrix} k_\star & 0 \end{bmatrix}$ and that the gramians $P_{K_\star, \nu, g_1}$ and $\Gamma_{K_\star, g_1}$ are scalar and equal to $P_{K_\star, \nu, g_1} = P_{k_\star, \nu, g_1}$ and $\Gamma_{K_\star, g_1} = \Gamma_{k_\star, g_1}$ respectively. In other words, the second input has no effect on the system (31), but as well shall see, g_2 is very sensitive to perturbations Δ whenever the largest singular value of the Markov parameter $m = |cb|$ is small.

If we denote by $d_{\mathsf{KL}}(s_1, s_2)$ the KL divergence between scalar input-state trajectories drawn from g_1 and g_2, we have by Lemma 8 that

$$\begin{aligned} d_{\mathsf{KL}}(s_1, s_2) &= \sum_{n=1}^{N} \sum_{t=0}^{T-1} \mathbf{E}_{g_1} \frac{1}{2} (\Delta u_t)^2 && \text{(Lemma 8)} \\ &\leq \frac{1}{2} \Delta^2 \sum_{n=1}^{N} \sum_{t=0}^{T-1} \mathbf{E}_{g_1} u_t^2 \\ &\leq \frac{1}{2} \Delta^2 NT\beta && \text{(by (3))} \\ &= \frac{1}{2} && \left(\text{if } \Delta^2 = \frac{1}{NT\beta} \right). \end{aligned}$$

Invoking Theorem 2 this implies the local minimax lower bound

$$\mathfrak{M}_{d_\infty}(\varepsilon_{N,T}; g_1) \geq \frac{1}{2\sqrt{NT\beta}} P_{k_\star, \nu, g_1}(a + bk_\star)\Gamma_{k_\star, \nu, g_1} \tag{32}$$

where $\varepsilon_{N,T} \asymp 1/\sqrt{NT}$. Inequality (32) in itself is an instance specific lower bound for scalar partially observed systems of the form (31). Further, the inequality implies that if the Markov parameter $|m| = |cb|$ is small, estimating gradients is always hard. Namely, we make the following observations[2]:

- $P_{k_\star, \nu, g_1}$ tends to infinity at rate $1/b^2$ as b tends to 0. Moreover, $P_{k_\star, g1}$ is always lower-bounded by 1.
- The large and small c asymptotics of $\Sigma_{k_\star, \nu, g_1}$ are proportional to $1/c^2$
- The factor $|a + bk_\star|$ tends to 0 no faster than $1/b^2$ and tends to $\min(1, |a|)$ as $b \to 0$ (and $|a|$ is invariant under similarity transform).

Combining these observations, we see that as the system invariant $|m| = |cb|$ tends to zero, gradients become arbitrarily noisy; the lower bound (32) tends to infinity. In other words, we have established that

$$\liminf_{|cb| \to 0} \mathfrak{M}_d(\varepsilon_{N,T}; g_1) = \infty. \tag{33}$$

Thus, we obtain an RL analogue to the well-known fact that reparametrization cannot help controlling a partially observed system as any gain in observability is offset by a proportional loss in controllability and vice versa.

IV. Discussion

In this work we showed that estimating policy gradients can become arbitrarily hard due to known control-theoretic fundamental limitations [29] by leveraging the classic two point method due to Le Cam [27]. For instance, we showed with system (31) that a partially observed system with small Markov parameters necessarily has noisy policy gradients and that this holds independently of the parametrization. Our bounds also show that learning controllers that are close to marginal stability can be hard. This is similar to what has already been observed for adaptive LQR/LQG in [6]. Leveraging results from [7] we further show that estimating policy gradients can suffer from exponential complexity in

[1] To see this, simply observe that $(y_0, \ldots, y_T)$ is a stochastic function of $(x_{0,n}, u_{0,n}, \ldots, u_{T-1,n}, x_{T,n})$.

[2] To verify these claims, observe that the scalar quantities $P_{k_\star, \nu, g_1}$, $k_\star$ and $\Sigma_{k_\star, \nu, g_1}$ have closed form solutions.

the system dimension. From a broader perspective, these results work toward elucidating when learning to control is feasible and when it is not.

Acknowledgements: Ingvar Ziemann and Henrik Sandberg are supported by the Swedish Research Council (grant 2016-00861) and the Swedish Foundation for Strategic Research through the CLAS project (grant RIT17-0046). Nikolai Matni is supported in part by NSF awards CPS-2038873 and CAREER award ECCS-2045834, and a Google Research Scholar award.

REFERENCES

[1] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[2] 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.

[3] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.

[4] Anastasios Tsiamis and George J Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

[5] Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive lqr with unknown b-matrix. In *Learning for Dynamics and Control*, pages 213–226. PMLR, 2021.

[6] Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *arXiv preprint arXiv:2201.01680*, 2022.

[7] Anastasios Tsiamis, Ingvar Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. *arXiv preprint arXiv:2205.14035*, 2022.

[8] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

[9] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning robust control for lqr systems with multiplicative noise via policy gradient. *arXiv preprint arXiv:1905.13547*, 2019.

[10] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

[11] Kaiqing Zhang, Bin Hu, and Tamer 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*, pages 179–190. PMLR, 2020.

[12] Farnaz Adib Yaghmaie, Fredrik Gustafsson, and Lennart Ljung. Linear quadratic control using model-free reinforcement learning. *IEEE Transactions on Automatic Control*, 2022.

[13] Yujie Tang, Yang Zheng, and Na Li. Analysis of the optimization landscape of linear quadratic gaussian (lqg) control. In *Learning for Dynamics and Control*, pages 599–610. PMLR, 2021.

[14] Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the lack of gradient domination for linear quadratic gaussian problems with incomplete state information. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1120–1124. IEEE, 2021.

[15] Yang Zheng, Luca Furieri, Maryam Kamgarpour, and Na Li. Sample complexity of linear quadratic gaussian (lqg) control for output feedback systems. In *Learning for Dynamics and Control*, pages 559–570. PMLR, 2021.

[16] 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.

[17] James A Preiss, Sébastien MR Arnold, Chen-Yu Wei, and Marius Kloft. Analyzing the variance of policy gradient estimators for the linear-quadratic regulator. *arXiv preprint arXiv:1910.01249*, 2019.

[18] Harish K Venkataraman and Peter J Seiler. Recovering robustness in model-free reinforcement learning. In *2019 American Control Conference (ACC)*, pages 4210–4216. IEEE, 2019.

[19] Natalie Bernat, Jiexin Chen, Nikolai Matni, and John Doyle. The driver and the engineer: Reinforcement learning and robust control. In *2020 American Control Conference (ACC)*, pages 3932–3939. IEEE, 2020.

[20] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

[21] Juan Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the sample complexity of stability constrained imitation learning. *arXiv preprint arXiv:2102.09161*, 2021.

[23] Ingvar Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. *arXiv preprint arXiv:2202.08311*, 2022.

[24] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117. PMLR, 2019.

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

[26] Asaf B Cassel and Tomer Koren. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1304–1313. PMLR, 2021.

[27] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[28] Torsten Söderström. *Discrete-time stochastic systems: estimation and control*. Springer Science & Business Media, 2002.

[29] John C Doyle. Guaranteed margins for lqg regulators. *IEEE Transactions on automatic Control*, 23(4):756–757, 1978.

[30] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

APPENDIX

a) Proofs for the Preliminary Results: Proof of Lemma 1. For $T \in \mathbb{N}$ and law $u_t = Kx_t$ we have

$$\begin{aligned}\sum_{t=0}^{T-1} x_t^\top Q x_t + u_t^\top R u_t &= \sum_{t=0}^{T-1} x_t^\top (Q + K^\top R K) x_t \\ &= \sum_{t=0}^{T-1} \operatorname{tr}\left[(Q + K^\top R K) x_t x_t^\top\right].\end{aligned} \tag{34}$$

Recall $\Gamma_K = \sum_{t=0}^{\infty} (A+BK)^t \Sigma_W (A+BK)^{t,\top}$ so that $\mathbf{E} x_t x_t^\top = \Gamma_K + o(1)$, where $o(1)$ tends to 0 as t tends to infinity. Hence, by averaging and taking limits we see that

$$\begin{aligned}&J(K) \\ &= \operatorname{tr}\left[(Q + K^\top R K)\Gamma_K\right] \\ &= \operatorname{tr}\left[(Q + K^\top R K)\sum_{t=0}^{\infty}(A+BK)^t \Sigma_W (A+BK)^{t,\top}\right] \\ &= \operatorname{tr}\left(\sum_{t=0}^{\infty}\left[(A+BK)^{t,\top}(Q + K^\top R K)(A+BK)^t\right]\Sigma_W\right).\end{aligned} \tag{35}$$

The result follows since $(A+BK)$ is stable by hypothesis. ■

Proof of Lemma 2. Fix a controller K. By Lemma 1 the average cost of the system (1), $J(K)$, is the same as the total cost of the deterministic system $\tilde{x}_{t+1} = A\tilde{x}_t + B\tilde{u}_t$

with $\tilde{x}_0 \sim N(0, \Sigma_W)$:

$$J_S(K) = \mathbf{E}_{K,S} \sum_{t=0}^{\infty} \tilde{x}_t^\top Q \tilde{x}_t + \tilde{u}_t R u_t.$$

The result now follows by Lemma 1 of [2]. ■

b) Information-Theoretic Lower Bounds: Intuitively, estimating a function $f(A, B)$ while only having access to samples from the unknown system (1) becomes hard if there are parameter variations A' and B' such that the behavior of $x'_{t+1} = A'x'_t + B'u'_t + w_t$ is very close to that of system (1) while the difference between $f(A, B)$ and $f(A', B')$ is large. This can be formalized by the Le Cam's two-point method:

Lemma 7 (Le Cam's Two Point Method): Fix two sets $\mathcal{M}$ and $\mathcal{D}$. Let $L : \mathcal{M} \times \mathcal{D} \to \mathbb{R}_+$ be any loss function and suppose that $S_1, S_2 \in \mathcal{M}$ satisfy $L(S_1, \mathsf{dec}) + L(S_2, \mathsf{dec}) \geq \delta$, $\forall \mathsf{dec} \in \mathcal{D}$. Then

$$\inf_K \sup_{S \in \mathcal{S}} \mathbf{E}_S L(S, K) \geq \frac{\delta}{2}(1 - d_{\mathsf{TV}}(P_{S_1}, P_{S_2})).$$

In other words, if for any decision the average loss is large, then a decision-maker that cannot distinguish between these two instances will suffer large loss on average, and therefore also in the worst case.

Proof of Lemma 7. We lower-bound the supremum over $\mathcal{M}$ by an expectation over the two-point mixture distribution supported on S_1 and S_2 as follows:

$$\begin{aligned}
\inf_{\mathsf{dec}} \sup_{S \in \mathcal{S}} \mathbf{E}_S L(S, \mathsf{dec}) &\geq \inf_{\mathsf{dec}} \frac{\mathbf{E}_S L(S_1, \mathsf{dec}) + \mathbf{E}_{S_2} L(S_2, \mathsf{dec})}{2} \\
&= \frac{1}{2} \inf_{\mathsf{dec}} \left(\int L(S_1, \mathsf{dec}(x)) P_{S_1}(dx) \right. \\
&\quad \left. + \int L(S_2, \mathsf{dec}(x)) P_{S_2}(dx) \right) \\
&\geq \frac{\delta}{2} \left(\int \min(P_{S_1}(dx), P_{S_2}(dx)) \right) \\
&= \frac{\delta}{2}(1 - d_{\mathsf{TV}}(P_{S_1}, P_{S_2}))
\end{aligned}$$

as per requirement. ■

Since $d_{\mathsf{TV}} \leq \sqrt{\frac{1}{2} d_{\mathsf{KL}}}$ by Pinsker's inequality, the following result is convenient to state.

Lemma 8: Let $S_0 = (A_0, B_0)$ and $S_1 = (A_1, B_1)$ and denote by P_1 and P_2 the induced probability measures over samples $(x_0, \ldots, x_T)$ satisfying the recursion (1) with $A = A_i, B = B_i, i = 0, 1$. Then

$$d_{\mathsf{KL}}(P_1, P_2) = \sum_{t=0}^{T-1} \frac{1}{2} \mathbf{E}_1 \|(A_0 - A_1)x_t + (B_0 - B_1)u_t\|^2_{\Sigma_W^{-1}}$$

where $\mathbf{E}_1$ denotes integration with respect to $\mathbf{P}_1$ and the norm $\|\cdot\|_{\Sigma_W^{-1}}$ is the Mahalanobis norm with kernel Σ_W^{-1}.

Proof of Lemma 8. Each random variable x_t is conditionally Gaussian given $(x_0, \ldots, x_{t-1})$ so that the KL divergence is given by half the expected square difference in conditional mean. The result follows by straighforward computation and the chain rule. See for instance [30, Chapter 8]. ■

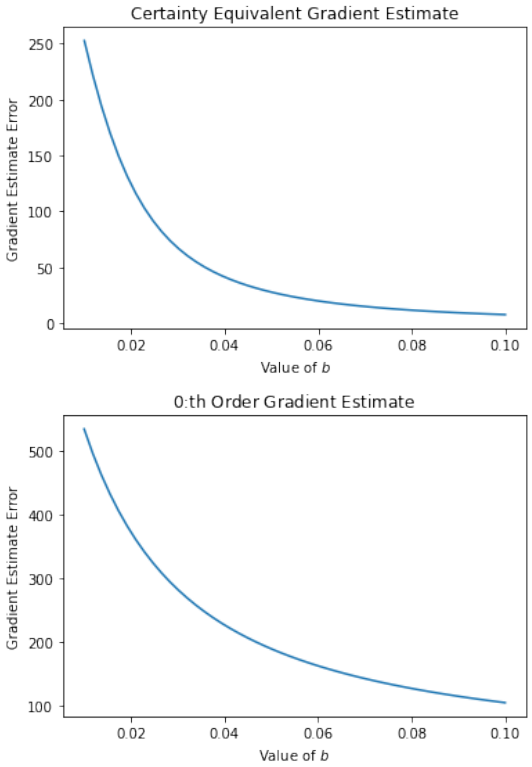

Fig. 1: Gradient estimate spread as a function of b for the scalar system (14). Notice that poor controllability (small b), leads to noisy gradients. The vertical axes show the standard deviation of $\left\| \nabla_K J(K; S) - \widehat{\nabla_K J} \right\|_{\mathsf{op}}$ across multiple trajectories.

c) Simulation: In Figure 1 we numerically verify the claims made for scalar systems in Section II-B with $a = 1$, $\sigma_w = 1$ and variable b. For the first plot, we use trajectories of length $T = 100$ and compute the least squares certainty equivalent (plug-in) gradient estimate using a single trajectory. The error is then averaged over $N = 100$ trajectories. For the second plot, we also use trajectories of length $T = 100$. However, we use $N = 10000$ many trajectories divided into batches, with each batch containing 100 trajectories. Each batch is then used to compute a 0-th order gradient estimate (see [2, Algorithm 1]). The second plot shows the estimator error averaged over these batches. Notice that for either estimator, the performance diverges in the low-controllability regime $b \approx 0$.