Black-Box Control for Linear Dynamical Systems

Xinyi Chen XINYIC@CS.PRINCETON.EDU
Department of Computer Science, Princeton University and Google AI Princeton

Elad Hazan EHAZAN@PRINCETON.EDU
Department of Computer Science, Princeton University and Google AI Princeton

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We consider the problem of black-box control: the task of controlling an unknown linear time-invariant dynamical system from a single trajectory without a stabilizing controller. Under the assumption that the system is controllable, we give the first *efficient* algorithm that attains sublinear regret under the setting of online nonstochastic control. This resolves an open problem since the work of Abbasi-Yadkori and Szepesvári (2011) on the stochastic black-box LQR problem, and in a more general setting that allows for adversarial perturbations and adversarially chosen changing convex loss functions.

We give finite-time regret bounds for our algorithm on the order of $2^{\mathrm{poly}(d)} + \tilde{O}(\mathrm{poly}(d)T^{2/3})$ for general nonstochastic control, and $2^{\mathrm{poly}(d)} + \tilde{O}(\mathrm{poly}(d)\sqrt{T})$ for black-box LQR. To complete the picture, we investigate the complexity of the online black-box control problem and give a matching regret lower bound of $2^{\Omega(d)}$, showing that the exponential cost is inevitable. This lower bound holds even in the noiseless setting, and applies to any, randomized or deterministic, black-box control method.

1. Introduction

A major goal in the field of adaptive control and reinforcement learning is to produce a truly independent learning agent. Such an agent can start in an unknown environment, and follow one continuous chain of experiences until it learns to perform as well as the optimal policy.

We consider this problem for the fundamental setting of controlling an unknown, linear time-invariant (LTI) dynamical system. This problem has received significant attention in the recent ML literature; however, nearly all existing methods assume some knowledge about the environment, usually in the form of a stabilizing controller. [1] Such assumptions can be restrictive because stabilizing controllers depend heavily on the unknown system parameters. On the other hand, without a stabilizing controller, technical challenges arise when the control algorithm needs to stabilize and optimally control a potentially unstable system. Prior to this work, the only exception is the seminal paper of Abbasi-Yadkori and Szepesvári (2011), which gives near-optimal regret bounds for certain variants of this problem using an exponential-time algorithm. [2]

In this work, we present a polynomial-time control algorithm that **only has black-box access** to an LTI system, under which the algorithm has no access to a stabilizing controller. The algorithm

1. Roughly speaking, a stabilizing controller is a policy that ensures the system will not explode, i.e. that the states stay bounded, under bounded perturbations. We formally define this concept in later sections.
2. Identification and stabilization of unstable systems have been studied in Faradonbeh et al. (2019) and Sarkar and Rakhlin (2019) in the stochastic setting, but the guarantees are not stated in terms of control cost. See related works section for a discussion.

Algorithm	Regret Bound	Efficient	Disturbances	Cost Functions
(Abbasi-Yadkori and Szepesvári, 2011)	$2^{\tilde{O}(d)}\sqrt{T}$	No	Stochastic	Quadratic
Algorithm 1	$2^{\text{poly}(d)} + \sqrt{T}$	Yes	Adversarial	Quadratic
Algorithm 1	$2^{\text{poly}(d)} + T^{2/3}$	Yes	Adversarial	General convex

Table 1: Summary of settings and results

is guaranteed to attain sublinear regret, converging on average to the performance of the best controller in hindsight among a set of reference controllers. Furthermore, the guarantees hold under the nonstochastic control setting, where both the perturbations and cost functions can be adversarially chosen. The question of controlling unknown systems under adversarial noise was posed in (Tu, 2019); our results quantify the difficulty of this task and provide a polynomial time solution. As far as we know, these results are the first finite-time sublinear regret bounds known for black-box, single-trajectory control in the nonstochastic setting. Table 1 provides a summary of results, and for clarity we omit polynomial dependence on system parameters and logarithmic dependence on T in the regret bound.

Our regret bounds are accompanied by a novel lower bound on the cost of black-box control. We show that this cost is inherently exponential in the natural parameters of the problem for *any*, deterministic or randomized, control method. As far as we know, this is the first finite-time lower bound for the online control problem that is exponential in the *system dimension*.

1.1. Statement of results

Consider an LTI dynamical system with black-box access. The controller can only interact with the system by sequentially observing states x_t and applying controls u_t. The state evolves according to the dynamics equation

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

where $x_t \in \mathbb{R}^{d_x}$, $u_t \in \mathbb{R}^{d_u}$. The system dynamics A, B are unknown to the controller, and the disturbance w_t can be adversarially chosen at the start of each time step. An adversarially chosen convex cost function $c_t(x, u)$ is revealed after the controller's action, and the controller suffers $c_t(x_t, u_t)$. In this model, a controller $\mathcal{A}$ is simply a mapping from all previous states and costs to a control. The total cost of executing a controller $\mathcal{A}$, whose sequence of controls is denoted as $u_t^{\mathcal{A}}$, is defined as

$$J_T(\mathcal{A}) = \sum_{t=1}^{T} c_t(x_t^{\mathcal{A}}, u_t^{\mathcal{A}}).$$

For a randomized control algorithm, we consider the expected cost. Under the special case of quadratic cost functions, if the disturbances are i.i.d. stochastic, we refer to the setting as **online LQR**; if the disturbances are adversarial, we refer to it as **nonstochastic online LQR**.

In the nonstochastic setting, the optimal controller cannot be determined a priori and depends on the disturbance realization. Consequently, we consider a comparative performance metric that takes into account the disturbance realization in hindsight, namely the regret. The goal of the learning algorithm is to choose a sequence of controls $\{u_t\}_{t=1}^T$ such that the total cost over T iterations is competitive with that of the best controller in a reference controller class Π given the disturbances.

We consider the worst-case regret over all possible disturbance realizations. Thus, the learner **with only black-box access, and in a single trajectory**, seeks to minimize regret defined as

$$\text{Regret}_T(\mathcal{A}) = \max_{w_{1:T}} \left\{ J_T(\mathcal{A}) - \min_{\pi \in \Pi} J_T(\pi) \right\}. \tag{1}$$

For the comparator class, we consider the set of Disturbance Action Controllers (DACs, see Definition 3), whose control is a linear function of past disturbances. This is a general class of controllers known to approximate any stabilizing linear controllers, in particular the H_2 optimal controller (Agarwal et al., 2019).

Let $\mathcal{L}$ denote the upper bound on the system's natural parameters, and κ^* be the controllability parameter of the stabilized system (Section 2.1). Let $\tilde{\kappa}$ denote an upper bound on the stability parameters of the recovered controller (Section 4.3). The following statements summarize our main results in Theorem 6 and Theorem 9:

1. We give an efficient algorithm whose regret with high probability satisfies

$$\text{Regret}_T(\mathcal{A}) \leq 2^{O(\mathcal{L} \log \mathcal{L})} + \tilde{O}(\text{poly}(\mathcal{L}, \kappa^*) T^{2/3}).$$

2. For the nonstochastic online LQR problem, we give an efficient algorithm, whose regret is with high probability at most

$$\text{Regret}_T(\mathcal{A}) \leq 2^{O(\mathcal{L} \log \mathcal{L})} + \tilde{O}(\text{poly}(\mathcal{L}, \tilde{\kappa}) \sqrt{T}).$$

3. We show that **any control algorithm** (randomized or deterministic) must suffer exponential regret in the worst case due to limited information. Formally, we show that for every controller $\mathcal{A}$, there exists an LTI dynamical system where (with high probability if the algorithm is randomized)

$$\text{Regret}_T(\mathcal{A}) \geq 2^{\Omega(\mathcal{L})}.$$

   Interestingly, this lower bound holds even for benign systems where the control-input matrix B is full rank and no disturbances are present. From existing results by Cassel et al. (2020), in general the online LQR problem has regret lower bound $\Omega(\sqrt{T})$. Therefore, any algorithm must incur regret at least $2^{\Omega(\mathcal{L})} + \Omega(\sqrt{T})$.

To the best of our knowledge, we provide the first finite-time regret bounds for control in a single trajectory with black-box access to the system in the nonstochastic setting. In particular, it is the first polynomial-time algorithm with optimal regret for nonstochastic black-box online LQR.

The main challenge of designing an efficient algorithm is obtaining a stabilizing controller from black-box interactions in the presence of adversarial noise. As our lower bound shows, this is a difficult task even when the system is well-conditioned, noiseless, and the cost functions are time invariant. Our method consists of three phases. In the first phase, we identify the dynamics matrices coarsely by injecting large controls into the system. Previous works on system identification under adversarial noise either require stable dynamics, or the knowledge of a strongly stable controller. Our approach is not limited by these requirements.

In the second phase, we use an SDP relaxation for the LQR setting by Cohen et al. (2018) to obtain a strongly stable controller given the system estimates. After we identify a strongly stable

controller, we use the techniques of Hazan et al. (2020) for regret minimization in the third phase for general convex costs, and those of Simchowitz (2020) for the nonstochastic online LQR problem.

For the lower bound, our approach is inspired by lower bounds for gradient-based methods from the optimization literature (Braverman et al., 2020). We give two separate lower bounds: one for deterministic algorithms, and one for randomized algorithms. The deterministic lower bound is less general but has better constants. Given a controller, we show system constructions that force the states, and thus costs, to grow exponentially before enough information about the system is revealed.

1.2. Related work

The focus of our work is adaptive control, where the controller does not have a priori knowledge of the underlying dynamics and has to learn them in addition to controlling the system. This task, under the nonstochastic control setting recently put forth in the machine learning literature, differs substantially from classical control theory that we survey below in the following aspects:

1. The system is unknown to the learner, and no stabilizing controllers are given.
2. The cost functions are unknown to the learner and can be chosen adversarially.
3. The disturbances are not assumed to be stochastic and can be chosen adversarially.

Robust and Optimal Control. When the underlying system and the cost functions are known, the noise is stochastic, one can compute the optimal controller a priori in some settings. For example, in the LQR setting, the system has linear dynamics and the cost functions are quadratic in the state and the control; it follows from the Bellman equations that the infinite horizon optimal policy is linear: $u_t = Kx_t$, where K is the solution to the algebraic Ricatti equation (Stengel, 1994; Zhou et al., 1996; Bertsekas, 2017). Control that is robust to worst-case noise is studied in the framework of H_∞ control, which computes the best linear controller over worst-case noise given system dynamics and cost functions, see e.g. the text by Zhou et al. (1996).

Online Control. Recent literature in the machine learning community considers the online LQR setting (Dean et al., 2018; Mania et al., 2019; Cohen et al., 2018), where the noise remains stochastic but the performance metric is regret instead of cost. Under this setting, polynomial time algorithms in (Mania et al., 2019; Cohen et al., 2019, 2018) attain $\sqrt{T}$ regret which also depends polynomially on relevant problem parameters. Regret bounds for partially observed systems are studied in (Lale et al., 2020a,b,c). However, all the results above assume the learner is given a stabilizing controller or the system is stable.

Black-box control of an unknown LDS was studied in Abbasi-Yadkori and Szepesvári (2011) and $\sqrt{T}$ regret was obtained, though the algorithm is inefficient in the sense that it may take exponential running time in the worst case. In contrast, our algorithm runs in polynomial time, and our setting permits adversarial noise sequences and cost functions.

Regret **lower bounds** for online LQR were studied in Cassel et al. (2020) and Simchowitz and Foster (2020), who show polynomial lower bounds in terms of the parameter T. In comparison, our exponential lower bound is stated in the system dimension rather than time.

Concurrently and independently, recent work by Lale et al. (2020d) considers the black-box online LQR setting and obtains $\tilde{O}(2^{\mathcal{L}} + \mathrm{poly}(d)\sqrt{T})$ regret under the weaker condition of stabilizability. However, their setting is restricted to stochastic noise and quadratic cost functions.

Nonstochastic Control: Moving away from stochastic noise, the nonstochastic control problem for linear dynamical systems was posed in Agarwal et al. (2019) to capture more robust online control (see survey (Hazan, 2020)). In this setting, the controller has no knowledge of the system dynamics or the adversarial noise sequence. The controller generates controls u_t at each iteration to minimize regret over sequentially revealed adversarial convex cost functions, against Disturbance Action Controllers. If a strongly stable controller is known, Hazan et al. (2020) give an algorithm that achieves $\tilde{O}(poly(\mathcal{L}, \kappa^*)T^{2/3})$ regret, where $\mathcal{L}$ is an upper bound on the system's natural parameters and κ^* is the controllability parameter of the stabilized system, as formalized in Section 2.1. This was recently extended in Simchowitz et al. (2020) to partially observed systems, and better bounds for certain families of loss functions with semi-adversarial noise. In Simchowitz (2020), $\tilde{O}(\mathrm{poly}(\mathcal{L}, \tilde{\kappa})\sqrt{T})$ regret was obtained for the nonstochastic LQR problem, where $\tilde{\kappa}$ is an upper bound on the parameters of the strongly stable controller, see Section 4.3. However, all of the above works assume a stabilizing controller is given to the learner, and are **not black-box** as per our definition.

Identification and Stabilization of Linear Systems: If the system is stabilized and has stochastic noise, the least squares method can be used to identify the dynamics in the partially observable and fully observable settings (Oymak and Ozay, 2019; Simchowitz et al., 2018). The algorithm by Simchowitz et al. (2019) tolerates adversarial noise and the guarantees only hold for stable systems; this work also shows that least squares can yield inconsistent estimates if the system is not stable.

However, least squares can still be used to estimate unstable systems if the closed-loop dynamics satisfy regularity conditions (Faradonbeh et al., 2017; Sarkar and Rakhlin, 2019). Using this method as a subroutine, for the setting of stochastic noise, recent work by Faradonbeh et al. (2019) and Shirani Faradonbeh et al. (2019) stabilize general systems in finite time with high probability.

In contrast, our system identification procedure is deterministic and permits adversarial noise. We further provide explicit finite-time bounds for optimally controlling the system. Our results do not assume stability of the system (spectral radius bounded by 1), but the weaker condition of controllability. It remains open to relax this assumption even further, to that of stabilizability in the nonstochastic black-box model.

2. Setting and Background

To enable the analysis of non-asymptotic regret bounds, we consider regret minimization against the class of strongly stable linear controllers. The notion of strong stability was formalized in Cohen et al. (2018) to characterize controllers under which a stochastic system converges to the steady-state distribution exponentially fast. Throughout the paper $\|\cdot\|$ denotes the spectral norm for matrices and the ℓ_2 norm for vectors.

Definition 1 (Strong Stability) *K is a (κ, γ) strongly stable controller for (A, B) if $\|K\| \le \kappa$, and there exist matrices H, L such that $A + BK = HLH^{-1}$, and $\|H\|\|H^{-1}\| \le \kappa$, $\|L\| \le 1 - \gamma$.*

The regret definition in Section 1.1 is meaningful only when the comparator set Π is non-empty. As shown in Cohen et al. (2018), a system (A, B) has a strongly stable controller if it is strongly controllable. This notion is formalized in the next definition.

Definition 2 (Strong Controllability) *Given a system (A, B), let C_k denote*

$$C_k = [B \; AB \; A^2B \; \cdots A^{k-1}B] \in \mathbb{R}^{d_x \times k d_u}.$$

Then (A, B) is (k, κ) strongly controllable if C_k has full row-rank, and $\|(C_k C_k^\top)^{-1}\| \leq \kappa$.

Assumption 1 *The system (A, B) is (k, κ) strongly controllable for $\kappa \geq 1$, and $\|A\|, \|B\| \leq \beta$ for some $\beta \geq 1$.*

Assumption 1 implies the existence of a strongly stable controller, and in Section 2.3 we give an explicit bound on its parameters. As a consequence of the Cayley-Hamilton theorem, a controllable system's controllability index k is at most d_x. Finally we make the following mild assumptions on the noise sequence and the cost functions. Similar assumptions appear in the nonstochastic control literature, see Simchowitz (2020), Hazan et al. (2020), Agarwal et al. (2019), Ghai et al. (2020).

Assumption 2 *The noise sequence is bounded such that $\|w_t\| \leq 1$ for all t.*

Assumption 3 *The cost functions are convex, and for all x, u such that $\|x\|, \|u\| \leq D$, $\|\nabla_{(x,u)} c_t(x, u)\| \leq GD$. Without loss of generality, assume $c_t(0, 0) = 0$.*

2.1. Notations

Inspired by the convention from the theory of Linear Programming (Nemirovski, 1994-1995), we use $\mathcal{L}$ to denote an upper bound on the natural parameters, which we interpret as the complexity of the system, i.e.

$$\mathcal{L} = kd_u + d_x + G + \beta + \kappa, \text{ where}$$

- κ, k are the controllability parameter and controllability index of the true system, respectively.
- d_x, d_u are the dimension of the states $x_t \in \mathbb{R}^{d_x}$ and dimension of the controls $u_t \in \mathbb{R}^{d_u}$.
- G is an upper bound on the Lipschitz constant of the cost functions c_t.
- β is an upper bound on the spectral norm of system dynamics A, B.

Given a $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable controller K, we denote κ^* as the upper bound on the controllability parameter of the stabilized system $(A + BK, B)$, and $\tilde{\kappa}$. We henceforth prove an upper bound on κ^* for the controller we recover, and show in Section 4.3 that $\kappa^* \leq \text{poly}(\kappa, \beta^k, d_x)$. We use $\tilde{O}$ to denote bounds that hold with probability at least $1 - \delta$, and omit the $\log(\delta^{-1})$ and $\log(T)$ factors.

2.2. Disturbance Action Controllers

In the regret formulation in 1, we take the reference policy class Π to be the class of Disturbance Action Controllers (DACs) (Agarwal et al., 2019; Hazan et al., 2020; Simchowitz, 2020), defined below. This class of policies can approximate any strongly stable controller in terms of cost, so we can compete with strongly stable controllers if we can compete with DACs. Moreover, DACs also include some classes of Linear Dynamic Controllers (LDCs) (Simchowitz et al., 2020). LDCs are a generalization of static feedback controllers, and both $\mathcal{H}_2$ and $\mathcal{H}_\infty$ optimal controllers under partial observation can be well-approximated by LDCs.

Definition 3 (Disturbance Action Controllers) *A Disturbance Action Controller with parameters (K, M) where $M = [M^0, M^1, \ldots, M^{H-1}]$ outputs control u_t at state x_t,*

$$u_t = Kx_t + \sum_{i=1}^{H} M^{i-1} w_{t-i}.$$

Definition 4 (Linear Dynamic Controllers) *A linear dynamic controller π is a linear dynamical system $(A_\pi, B_\pi, C_\pi, D_\pi)$ with internal state $s_t \in \mathbb{R}^{d_\pi}$, input $x_t \in \mathbb{R}^{d_x}$ and output $u_t \in \mathbb{R}^{d_u}$ that satisfies*

$$s_{t+1} = A_\pi s_t + B_\pi x_t, \;\; u_t = C_\pi s_t + D_\pi x_t.$$

The class of DACs enables the use of online convex optimization techniques in control. In the canonical parameterization of the nonstochastic control problem, the total cost of a linear controller $J(K)$ is not convex in K. However, as shown in Agarwal et al. (2019), the total cost of DACs is convex with respect to their parameters.

2.3. SDP Relaxation for LQ Control

In Linear Quadratic control the cost functions are known ahead of time and are fixed,

$$c_t(x, u) = x^\top Q x + u^\top R u,$$

and the disturbances are i.i.d., $w_t \sim N(0, W)$. Given an instance of the LQ control problem defined by (A, B, Q, R, W), the learner can obtain a strongly stable controller by solving the SDP relaxation for minimizing steady-state cost, proposed in Cohen et al. (2018). For $\nu > 0$, the SDP is given by

$$\begin{aligned} \text{minimize} \quad & J(\Sigma) = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \bullet \Sigma \\ \text{subject to} \quad & \Sigma_{xx} = \begin{pmatrix} A & B \end{pmatrix} \Sigma \begin{pmatrix} A & B \end{pmatrix}^\top + W, \;\; \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{xu}^\top & \Sigma_{uu} \end{pmatrix}. \\ & \Sigma \succeq 0, \;\; \mathrm{Tr}(\Sigma) \le \nu. \end{aligned}$$

Indeed, a strongly stable controller can be extracted from any feasible solution to the SDP, as guaranteed by the following lemma.

Lemma 5 (Lemma 4.3 in (Cohen et al., 2018)) *Assume that $W \succeq \sigma^2 I$ and let $\kappa = \sqrt{\nu}/\sigma$. Let Σ be any feasible solution for the SDP, then the controller $K = \Sigma_{xu}^\top \Sigma_{xx}^{-1}$ is $(\kappa, 1/2\kappa^2)$ strongly stable.*

Existence of Strongly Stable Controllers Under Assumption 1, the noiseless dynamical system $x_{t+1} = Ax_t + Bu_t$ starting from x_1 can be driven to the zero state in k steps. Furthermore, Lemma B.4 in Cohen et al. (2018) gives an upper bound on the reset cost, defined as $\sum_{t=1}^{k} \|x_t\|^2 + \|u_t\|^2$. Suppose the reset cost is at most $C\|x_1\|^2$, then Thereom B.5 in Cohen et al. (2018) suggests that the SDP for the noisy system $x_{t+1} = Ax_t + Bu_t + w_t$ with $w_t \sim N(0, W)$ and $\nu = C \cdot \mathrm{Tr}(W)$ is feasible. Taking $W = I$, the system (A, B) has a $(\sqrt{Cd_x}, 1/(2Cd_x))$ strongly stable controller. Lemma B.4 in Cohen et al. (2018) shows that under Assumption 1, $C = 3\kappa^2 k^2 \beta^{6k}$.

3. Algorithm and Main Theorem

Now we describe our main algorithm for the black-box control problem, Algorithm 1. Overall we use the explore-then-commit strategy, and split the algorithm into three phases. In phase 1, we identify the underlying system dynamics coarsely with large controls. In phase 2, we extract a strongly stable controller for the estimated system using the SDP in Section 2.3, and show that it is also strongly stable for the true system. We then alleviate the effects of using large controls by

Algorithm 1 Nonstochastic Control with Black-box Access

Input: horizon T, k, κ such that the system (A, B) is (k, κ) strongly controllable, $\beta \geq 1$ such that $\|A\|, \|B\| \leq \beta$.
Set $\kappa' = \sqrt{Cd_x}$, $\gamma' = 1/(2\kappa'^2)$, where $C = 3\kappa^2 k^2 \beta^{6k}$.
Phase 1: **Black-box System Identification**
Set $\varepsilon = \frac{\gamma'^2}{10^5 d_x^2 \kappa'^8}$, $\lambda = 8\beta$.
$(\hat{A}, \hat{B}) \leftarrow \text{AdvSysId}(\varepsilon, \lambda, x_1, k, \kappa)$ for $T_1 = d_u(k+1) + 1$ rounds.
Phase 2: **Stable Controller Recovery**
$\hat{K} \leftarrow \text{ControllerRecovery}(\hat{A}, \hat{B}, \varepsilon, \kappa', \gamma')$, set $\tilde{\kappa} = \frac{2\kappa'^2 d_x^{1/2}}{\gamma'^{1/2}}$, $\tilde{\gamma} = \frac{\gamma'}{16 d_x \kappa'^4}$.
Execute $\hat{K}$ for $T_2 = \max\{\frac{\ln(\tilde{\gamma}\|x_{T_1}\|)}{\tilde{\gamma}}, 0\}$ rounds .
Phase 3: **Nonstochastic Control**
Set $\kappa^* = 4\tilde{\kappa}^2 k^2 \beta^{2k} \kappa$, $W = 2\kappa^*/\tilde{\gamma}$.
General convex costs: call Algorithm 1 in Hazan et al. (2020) with inputs $\hat{K}$, κ^*, $\tilde{\gamma}$, W for $T - T_1 - T_2$ rounds.
Quadratic costs: call Algorithm 3 in Simchowitz (2020) for $T - T_1 - T_2$ rounds.

decaying the system to a state with constant magnitude. Finally in phase 3, we invoke Algorithm 1 in Hazan et al. (2020) or Algorithm 3 in Simchowitz (2020) to achieve sublinear regret with the obtained strongly stable controller.

Our main theorem below is stated using asymptotic notation that hides constants independent of the system parameters, and uses $\mathcal{L}$ for an upper bound on the system parameters as defined in section 2.1. Exact constants appear in the proofs.

Theorem 6 *Under Assumptions 1, 2, 3, with high probability the regret of Algorithm 1 satisfies*

$$\textit{Regret}_T(\mathcal{A}_1) \;\leq\; 2^{O(\mathcal{L}\log\mathcal{L})} + \tilde{O}(\text{poly}(\mathcal{L}, \kappa^*)T^{2/3}).$$

If the loss functions are in addition α-strongly convex, quadratic, and without loss of generality assuming $\tilde{\kappa} \geq \tilde{\gamma}^{-1}$, the regret of Algorithm 1 satisfies

$$\textit{Regret}_T(\mathcal{A}_1) \;\leq\; 2^{O(\mathcal{L}\log\mathcal{L})} + \tilde{O}(\text{poly}(\mathcal{L}, \tilde{\kappa}, \alpha^{-1})\sqrt{T}).$$

This is composed of

1. *Phase 1: after T_1 rounds we have $\|x_{T_1}\|^2 \leq 2^{O(\mathcal{L}\log\mathcal{L})}$. The total cost is at most $2^{O(\mathcal{L}\log\mathcal{L})}$.*
2. *Phase 2: Computing $\hat{K}$ has zero cost. Decaying the system has total cost at most $O\big(G\tilde{\kappa}^4\|x_{T_1}\|^3\tilde{\gamma}^{-3}\big)$, where $\tilde{\kappa}$, $\tilde{\gamma}$ are as defined in the algorithm. This phase has total cost $2^{O(\mathcal{L}\log\mathcal{L})}$.*
3. *Phase 3: Nonstochastic control with a known strongly stable controller for general convex costs incurs regret at most $\tilde{O}(\text{poly}(\mathcal{L}, \kappa^*)(T - T_1 - T_2)^{2/3})$ with high probability. If the cost functions are α-strongly convex and quadratic, with high probability the regret is bounded by $\tilde{O}(\text{poly}(\tilde{\kappa}, \mathcal{L}, \alpha^{-1})\sqrt{T - T_1 - T_2})$.*

4. Analysis Outline

We provide an outline of our analysis in this section, and the formal statements are in the appendix.

4.1. Black-box system identification

In this phase we obtain estimates of the system $\hat{A}, \hat{B}$ without knowing a stabilizing controller. Recall the definition of $C_k = [B, AB, \ldots, A^{k-1}B]$, and let $Y = [AB\ A^2B\ \cdots\ A^kB]$. The procedure AdvSysId (Algorithm 2) consists of two steps. In the first step, we estimate each A^jB for $j = 0 \ldots, k$ (in particular we obtain $\hat{B}$ close to B), and guarantee that $\|C_k - C_0\|_F$, $\|Y - C_1\|_F$ are small. In the second step, we take $\hat{A}$ to be the solution to the system of equations in X: $XC_0 = C_1$.

For the first step, the algorithm estimates matrices A^jB by using controls that are scaled standard basis vectors once every $k+1$ iterations, and using zero controls for the iterations in between. The state evolution satisfies

$$x_{t+1} = A^t x_1 + \sum_{i=1}^{t} (A^{t-i}Bu_i + A^{t-i}w_i).$$

Intuitively, we choose scaling factors ξ_i such that j iterations after a non-zero control $\xi_i \cdot e_i$ is used, the state is dominated by $\xi_i A^{j-1}Be_i$, the scaled i-th column of $A^{j-1}B$. In the algorithm $\hat{M}_j$ is the concatenation of estimates for A^jBe_i, and we concatenate the $\hat{M}_j$'s to obtain C_0, C_1. We show in Lemma 12 that $\|\hat{M}_j - A^jB\|_F \le O(d_u^2 k \lambda^{2k} \varepsilon_0)$, which implies the closeness of C_0, C_1 to C_k, Y, respectively.

Under the assumption that (A, B) is (k, κ) strongly controllable, A is the unique solution to the system of equations in X: $XC_k = Y$. By perturbation analysis of linear systems, the solution to the system of equations $XC_0 = C_1$ is close to A, as long as $\|C_0 - C_k\|_F, \|C_1 - Y\|_F$ are sufficiently small. By our choice of ε_0, we conclude that $\|\hat{A} - A\| \le \varepsilon$, $\|\hat{B} - B\| \le \varepsilon$. Lemma 14 shows that the total cost of this phase is bounded by $2^{O(\mathcal{L} \log \mathcal{L})}$. [3]

4.2. Computing a stabilizing controller

The goal of phase 2 is to recover a strongly stable controller from system estimates obtained in phase 1 by solving the SDP presented in Section 2.3. The key to our task is setting the trace upper bound ν appropriately, so that the SDP is feasible and the recovered controller is strongly stable even for the original system. We justify our choice of ν in Lemma 18, and show that by our choice of ε, $\hat{A}, \hat{B}$ are sufficiently accurate and $\hat{K}$ is $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable for the true system. We remark that Simchowitz and Foster (2020) has an alternative procedure for recovering K given system estimates.

4.2.1. DECAYING THE SYSTEM

In phase 1 the algorithm uses large controls to estimated the system, and after T_1 iterations the state can have an exponentially large magnitude. Equipped with a strongly stable controller, we decay the system so that the state has a constant magnitude before starting phase 3. We show in Lemma 19 that following the policy $u_t = \hat{K}x_t$ for T_2 iterations decays the state to at most $2\tilde{\kappa}/\tilde{\gamma}$ in magnitude.

4.3. Nonstochastic control

Given a $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable controller $\hat{K}$ for the true system, we use existing algorithms for nonstochastic control. Both algorithms in phase 3 follow the recipe of system identification and

3. Different from many existing system identification routines, Algorithm 2 is deterministic and our guarantee does not have a failure probability.

Algorithm 2 AdvSysId

Input: accuracy parameter $\varepsilon < 1/2$, $\|x_1\| \le 1$. Let $\lambda \ge 1$ be such that $\|A\|, \|B\| \le \frac{1}{4}\lambda - 1$, (k, κ) such that the system (A, B) is (k, κ) strongly controllable.
Set $\varepsilon_0 = \frac{\varepsilon}{10^2 d_u^2 k^2 \lambda^{3k} d_x \kappa^{1/2}}$.
for $t = 1, \ldots, (k+1)d_u$ **do**
  observe x_t.
  if $t = 1 \pmod{k+1}$ **then**
    Let $i = (t-1)/(k+1) + 1$.
    control with $u_t = \xi_i \cdot e_i$ for $\xi_i = \lambda^{t-1}\varepsilon_0^{-i}$, where e_i is the i-th standard basis vector.
  else
    control with $u_t = 0$.
  end if
  pay cost $c_t(x_t, u_t)$.
end for
For $0 \le j \le k$, $1 \le i \le d_u$, define $l(i,j) = (i-1)(k+1) + j + 2$. Let $x_i^j = x_{l(i,j)}$. Construct

$$\hat{M}_j = [\frac{x_1^j}{\xi_1} \; \frac{x_2^j}{\xi_2} \; \cdots \; \frac{x_{d_u}^j}{\xi_{d_u}}] \in \mathbb{R}^{d_x \times d_u}.$$

Define $C_0 = [\hat{M}_0 \; \hat{M}_1 \; \cdots \; \hat{M}_{k-1}]$, $C_1 = [\hat{M}_1 \; \hat{M}_2 \; \cdots \; \hat{M}_k] \in \mathbb{R}^{d_x \times d_u k}$.
Output $\hat{A} = C_1 C_0^\top (C_0 C_0^\top)^{-1}$, $\hat{B} = \hat{M}_0$.

Algorithm 3 ControllerRecovery

Input: κ', γ' such that there exists K that is (κ', γ') strongly stable for (A, B); accuracy parameter ε, and $\hat{A}$, $\hat{B}$ such that $\|A - \hat{A}\| \le \varepsilon$, $\|B - \hat{B}\| \le \varepsilon$.
Set $\nu = \frac{2\kappa'^4 d_x}{\gamma' - 2\varepsilon\kappa'^2}$.
Solve the following SDP:

$$\begin{aligned} \text{minimize} \quad & 0 \\ \text{subject to} \quad & \Sigma_{xx} = \begin{pmatrix} \hat{A} & \hat{B} \end{pmatrix} \Sigma \begin{pmatrix} \hat{A} & \hat{B} \end{pmatrix}^\top + I, \text{ where} \\ & \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{xu}^\top & \Sigma_{uu} \end{pmatrix}, \; \Sigma \succeq 0, \; \mathrm{Tr}(\Sigma) \le \nu. \end{aligned}$$

Denote a feasible solution as $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{xx} & \hat{\Sigma}_{xu} \\ \hat{\Sigma}_{xu}^\top & \hat{\Sigma}_{uu} \end{pmatrix}$, return $\hat{K} = \hat{\Sigma}_{xu}^\top \hat{\Sigma}_{xx}^{-1}$.

then policy regret minimization with gradient-based methods. Different from phase 1, the system can be estimated to arbitrary accuracy without prohibitive cost given a stabilizing controller.

If the costs are general convex functions, we run Algorithm 1 in Hazan et al. (2020) (Algorithm 4 in the appendix) which achieves sublinear regret. By Lemma 21, the system $(A + B\hat{K}, B)$ is $(k, 4\tilde{\kappa}^2 k^2 \beta^{2k} \kappa)$ strongly controllable. If we start Algorithm 4 from $t = T_1 + T_2$, the setting is consistent with the nonstochastic control setting where the noise is bounded by $\|x_{T_1+T_2}\|$, and with

total iteration number $T - T_1 - T_2$. By Theorem 12 in Hazan et al. (2020), setting $\kappa^* = 4\tilde{\kappa}^2 k^2 \beta^{2k} \kappa$, $W = 2\kappa^*/\tilde{\gamma}$, and noticing that $\tilde{\gamma}^{-1} = \text{poly}(\kappa^*)$, with high probability, our total regret is at most $\tilde{O}(\text{poly}(\kappa^*, k, d_x, d_u, G)T^{2/3})$.

If the cost functions are α-strongly convex and quadratic, we use Algorithm 3 in Simchowitz (2020). Note that this algorithm does not need controllability assumptions on the system. By Theorem 3.2 in Simchowitz (2020), and without loss of generality assuming $\tilde{\kappa} \geq \tilde{\gamma}^{-1}$, with high probability the total regret of this phase is bounded by $\tilde{O}(\text{poly}(\tilde{\kappa}, \beta, d_x, d_u, G, \alpha^{-1})\sqrt{T})$.

5. Lower Bound on Black-box Control

In this section we prove that with high probability, any randomized black-box control algorithm incurs a loss which is exponential in the system dimension, even for noiseless LTI systems. Our lower bound is partially based on the construction in Braverman et al. (2020) and uses technique from the optimization literature. In addition, we provide a lower bound for deterministic black-box control algorithms in Appendix E with improved constants. We first define the relevant concepts.

Definition 7 (Black-box Control Algorithm) *A randomized black-box control algorithm $\mathcal{A}$ has a random string σ_t and outputs a control u_t at each iteration t, where u_t is a function of past information and the random string, i.e. $u_t = \mathcal{A}(x_1, ..., x_t, c_1, \ldots, c_t, u_1, \ldots, u_{t-1}, \sigma_t)$.*

Definition 8 (Control Problem Instance) *An instance of a control problem is defined by a noiseless system (A, B), an initial state x_1, and a sequence of oblivious convex cost functions $\{c_t\}$.*

Theorem 9 *Let $\mathcal{A}$ be a randomized control algorithm as per Definition 7. Then there exists a control problem instance with system dimension d_x, where the system is stabilizable and $(1, 1)$-strongly controllable, such that with $T = d_x/8$, with probability at least $1 - \exp(-\frac{d_x}{100})$, we have*

$$\textit{Regret}_T(\mathcal{A}) \geq 2^{\Omega(\mathcal{L})}.$$

Proof We first consider deterministic black-box control algorithms. We show that there exists a distribution over control problem instances, such that with high probability, the total cost of any deterministic control algorithm is exponential in the system dimension. Then we treat a randomized algorithm as a distribution over deterministic algorithms, and use a probabilistic argument to show that there exists a hard control problem for every randomized algorithm.

The construction of the hard distribution below follows from the intuition that a matrix with i.i.d. random Gaussian entries is rotation invariant, and therefore for such a matrix of dimension d, a deterministic control algorithm needs to observe at least $O(d)$ matrix-vector products to gain enough information.

The construction. Fix $x_1 = e_1$ and $c_t(x, u) = \|x\|^2 + \|u\|^2$ for all t. Let $N(m, n, \sigma)$ denote a distribution over matrices of dimension $m \times n$, where each coordinate is Gaussian with mean 0 and variance σ. Consider the distribution of control problems specified by $\{(A, I), x_1, \{c_t\}\}$, where $A \sim N(d_x, d_x, \frac{\gamma}{d_x})$ for some $\gamma > 0$. For a realization of A, the system is $x_{t+1} = Ax_t + u_t$. Note that for any A, this system is $(1, 1)$-strongly controllable, and $-A$ is a stabilizing controller that gives constant regret. Moreover, with high probability, the system has bounded size: by Corollary 35 in (Vershynin, 2011), with probability at least $1 - 2\exp(-\frac{d_x}{2})$, $\|A\| \leq 3\sqrt{\gamma}$. Let $\mathcal{L}(A)$ denote the system upper bound of the control problem instance defined by our choice of $x_1, \{c_t\}$, and A.

Under this event, we have $\mathcal{L}(A) \leq 2d_x + 4 + 1 + 3\sqrt{\gamma} \leq 4d_x$ for $\gamma = 40$ and d_x large.

To show that the above distribution is hard for deterministic control algorithms, we first frame the control task under an information model with queries and observations, similar to the setting in Braverman et al. (2020). This framing facilitates our analysis by making the information a controller receives in each time step explicit.

Information model. At every iteration the controller observes x_t, then computes u_t as a deterministic function of $x_1, x_2, \ldots, x_t$, $u_1, u_2, \ldots u_{t-1}$, and then observes $x_{t+1} = Ax_t + u_t$. Without loss of generality, we can assume that the controller also observes $Au_1, \ldots, Au_{t-1}$ before computing u_t, but does not act on this information. Then in the following information model, the controller can be seen as a player making adaptive queries to an unknown matrix A, and receives observations in the form of matrix-vector products: the controller makes deterministic queries defined by vectors $x_1, u_1, x_2, u_2, \ldots, x_{t-1}, u_{t-1}, x_t, u_t$ and observes $Ax_1, Au_1, \ldots, Ax_{t-1}, Au_{t-1}, Ax_t, Au_t$. Each pair of queries x_t, u_t are deterministic functions of previous queries and observations: $x_1, \ldots, x_{t-1}$, $u_1, \ldots, u_{t-1}, Ax_1, \ldots, Ax_{t-1}.Au_1, \ldots, Au_{t-1}$. Note that even though u_t can depend on x_t, we have $x_t = Ax_{t-1} + u_{t-1}$, so without loss of generality we can assume u_t only depends on previous queries and observations. However, u_t cannot depend on Ax_t since this is a future observation.

Under this information model, for every x_t, there exists a subspace $(V_{t-1}^{\perp})^\top$ such that $(V_{t-1}^{\perp})^\top x_t$ has a random component. Importantly, the subspace only depends on the queries and observations so far and not on any future information. Further, we show that with high probability, the magnitude of the random component grows exponentially with time.

Lemma 10 *Let $T = d_x/8$. There exists a sequence of orthonormal matrices $V_1, \ldots, V_T$, such that for $t \in [T]$, V_t only depends on $x_1, \ldots, x_t, u_1, \ldots, u_t$ and they satisfy the following condition: Let r_t denote the rank of span$(x_1, \ldots, x_t, u_1, \ldots, u_t)$, and let $V_t^{\|}$ denote the first r_t columns of V_t, and let $V_t^{\perp}$ denote the last $d - r_t$ columns of V_t. Let $h_t = (V_{t-1}^{\perp})^\top x_t$, then for all $t \in [T]$, conditioned on $x_1, x_2, \ldots, x_t$, $u_1, u_2, \ldots, u_t, Au_1, \ldots, Au_{t-1}$, we have $(V_t^{\perp})^\top x_{t+1} = c_t + z_t$, where the coordinates of z_t are i.i.d. normally distributed, i.e. $z_t(i) \sim N(0, \frac{\gamma \|h_t\|^2}{d})$.*

Lemma 11 *Let $V_1, \ldots, V_T$ be as in Lemma 10, and $T = d_x/8$. Let $h_t = (V_{t-1}^{\perp})^\top x_t$, with probability at least $1 - \exp(-\frac{d_x}{25})$, conditioned on $x_1, x_2, \ldots, x_t, u_1, u_2, \ldots, u_t$, $Au_1, \ldots, Au_{t-1}$, we have $\|(V_t^{\perp})^\top x_{t+1}\|^2 \geq \frac{\gamma \|h_t\|^2}{20}$.*

Consider the construction of matrices $V_1, V_2, \ldots, V_T$ as in Lemma 10. Then conditioned on x_1, u_1, we have $h_1 = (V_0^{\perp})^\top x_1 = x_1$, and $\|h_1\| = 1$. Here u_1 can depend on x_1 because x_1 is independent of A. By Lemma 11, for $t \leq T$, conditioned on $x_1, \ldots, x_t, u_1, \ldots, u_t, Au_1, \ldots, Au_{t-1}$, with probability at least $1 - \exp(-\frac{d_x}{25})$, we have $\|h_{t+1}\|^2 \geq 2\|h_t\|^2$ with our choice of γ. Therefore, with probability at least $(1 - \exp(-\frac{d_x}{25}))^{T-1}$, $\|h_T\|^2 \geq 2^{T-1}$. Note that $\|x_T\|^2 = \|V_{T-1} x_T\|^2 \geq \|V_{T-1}^{\perp} x_T\|^2 = \|h_T\|^2 \geq 2^{d_x/8-1}$. Since for small ε, we have $(1 - \varepsilon)^t \geq 1 - 2t\varepsilon$, we have $(1 - \exp(-\frac{d_x}{25}))^{\frac{d_x}{8}-1} \geq 1 - \frac{d_x}{4}\exp(-\frac{d_x}{25}) \geq 1 - \exp(-\frac{d_x}{50})$ for d_x large. Therefore with high probability, the total cost of any deterministic black-box control algorithm $\mathcal{A}$ over T iterations is at least $2^{d_x/8-1}$ by our choice of cost functions. Note that this result holds with any realization of

$Au_1, Au_2, \ldots, Au_T$. Since there exists a stabilizing controller that incurs constant cost, we conclude that $\text{Regret}_T(\mathcal{A}) \geq 2^{\Omega(d_x)}$ with probability at least $1 - \exp(-\frac{d_x}{50})$, for any deterministic $\mathcal{A}$.

Now we consider randomized control algorithms. For any randomized algorithm $\mathcal{A}_{\text{rand}}$, its randomness is independent of the distribution over the system, and can be considered as a random string whose value is chosen before the start of the algorithm. Let σ_T denote the randomness of $\mathcal{A}_{\text{rand}}$ over T iterations, and for any value b_T of σ_T, let $\mathcal{A}_{\text{rand}}(b_T)$ denote the algorithm which is $\mathcal{A}_{\text{rand}}$ with σ_T fixed to b_T. Then $\mathcal{A}_{\text{rand}}(b_T)$ is a deterministic algorithm. Let $\text{Regret}_T(\mathcal{A}_{\text{rand}}(b_t), A)$ denote the regret of $\mathcal{A}_{\text{rand}}(b_T)$ on the system A. We can write

$$\begin{aligned}\mathbb{P}_{A,\sigma_T}[\text{Regret}_T(\mathcal{A}_{\text{rand}}, A) \geq 2^{\Omega(d_x)}] &= \sum_{b_T} \mathbb{P}[\sigma_T = b_T]\, \mathbb{P}_A[\text{Regret}_T(\mathcal{A}_{\text{rand}}(b_T), A) \geq 2^{\Omega(d_x)}] \\ &\geq \min_{\mathcal{A}_{\text{det}}} \mathbb{P}_A[\text{Regret}_T(\mathcal{A}_{\text{det}}, A) \geq 2^{\Omega(d_x)}] \geq 1 - \exp(-\frac{d_x}{50}).\end{aligned}$$

In addition to having regret exponential in the system dimension, we also need the size of the system to be bounded. Let $\mathcal{D}$ denote the distribution of A conditioned on the event $\mathcal{E} : \mathcal{L}(A) \leq O(d_x)$. Since the event $\mathcal{E}^c$ happens with probability at most $2\exp(-\frac{d_x}{2})$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{A}_{\text{rand}}, A\sim\mathcal{D}}[\text{Regret}_T(\mathcal{A}_{\text{rand}}, A) \geq 2^{\Omega(d_x)}] &\geq \mathbb{P}_{\mathcal{A}_{\text{rand}}, A}[\text{Regret}_T(\mathcal{A}_{\text{rand}}, A) \geq 2^{\Omega(d_x)}] - \mathbb{P}[\mathcal{E}^c] \\ &\geq 1 - 2\exp(-\frac{d_x}{50}) \\ &\geq 1 - \exp(-\frac{d_x}{100}).\end{aligned}$$

It follows that there exist a system $A^\star$ with $\mathcal{L}(A^\star) \leq O(d_x)$, such that over the randomness of $\mathcal{A}_{\text{rand}}$, with probability at least $1 - \exp(-\frac{d_x}{100})$, $\text{Regret}_T(\mathcal{A}_{\text{rand}}, A^\star) \geq 2^{\Omega(\mathcal{L}(A^\star))}$. ■

6. Conclusion

We present the first end-to-end, efficient black-box control algorithm for unknown linear dynamical systems in the nonstochastic control setting. This improves upon previous work in black-box control (Abbasi-Yadkori and Szepesvári, 2011) in several dimensions: computational efficiency (previous methods were exponential time), robustness (tolerating adversarial noise), and generality (our algorithm permits a broader set of cost functions than quadratic functions).

The startup cost of our algorithm is exponential in the system dimension. However we show that this cost is nearly optimal by giving a novel lower bound for any randomized or deterministic black-box control algorithm. Combined with previous results, our algorithm applied to the nonstochastic online LQR setting achieves near optimal regret.

One intriguing open problem in our setting is whether or not it is possible to achieve our regret upper bound with control signals that are not exponential in the system parameters. Large magnitude controls are often impossible to implement, and a more practical algorithm is desirable. As far as we know, our lower bound does not prohibit such a method and this possibility remains open.

Acknowledgments

We thank Blake Woodworth and Max Simchowitz for very helpful discussions.

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, 2011.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119, 2019.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmont, MA, USA, 4th edition, 2017.

Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 627–647. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/braverman20a.html.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently, 2020.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1300–1309, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cohen19b.html.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

M. K. S. Faradonbeh, A. Tewari, and G. Michailidis. Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64(8):3498–3505, 2019.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *CoRR*, abs/1710.01852, 2017. URL http://arxiv.org/abs/1710.01852.

Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1714–1757. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/ghai20a.html.

Elad Hazan. Lecture notes: Computational control theory. https://sites.google.com/view/cos59x-cct/lecture-notes, 2020. [Online; accessed 15-Jan-2021].

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret bound of adaptive control in linear quadratic gaussian (lqg) systems, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems, 2020b.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control, 2020c.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators, 2020d.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Arkadi Nemirovski. Lecture on information-based complexity of convex programming, 1994-1995.

S. Oymak and N. Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pages 5655–5661, 2019.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/sarkar19a.html.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Randomized algorithms for data-driven stabilization of stochastic linear systems. In *2019 IEEE Data Science Workshop (DSW)*, pages 170–174, 2019. doi: 10.1109/DSW.2019.8755578.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic, 2020.

Max Simchowitz and Dylan J. Foster. Naive exploration is optimal for online lqr, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/simchowitz18a.html.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine*

Learning Research, pages 2714–2802, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/simchowitz19a.html.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control, 2020.

Robert F. Stengel. *Optimal Control and Estimation*. 1994.

Stephen Lyle Tu. *Sample Complexity Bounds for the Linear Quadratic Regulator*. PhD thesis, UC Berkeley, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice-Hall, Inc., USA, 1996. ISBN 0134565673.

Appendix A. Proofs for Section 4.1

In this section we present proofs for phase 1 of Algorithm 1. We show that the estimates $\hat{A}, \hat{B}$ satisfy $\|\hat{A} - A\| \le \varepsilon, \|\hat{B} - B\| \le \varepsilon$, and bound the total cost of this phase. We first bound the magnitude of states in each iteration to guide our choice of scaling factors ξ_i. In Algorithm 2, for all $t = 2, \ldots, (k+1)d_u$, let $j = t - 2 \pmod{k+1}$, $i = (t-2-j)/(k+1) + 1$, we have $\|x_t\| \le \lambda^{t-1}\varepsilon_0^{-i}$.

Proof This can be seen by induction. For our base case, consider x_2, where $i = 1, j = 0$. $\|x_2\| \le \|A\| + \|B\|\|u_1\| + 1 \le \frac{1}{4}\lambda(1 + \varepsilon_0^{-1}) + 1 \le \lambda\varepsilon_0^{-1}$. Assume $\|x_t\| \le \lambda^{t-1}\varepsilon_0^{-i}$ for $t = (i-1)(k+1) + j + 2$. If $j = k$, then $t = i(k+1) + 1$, $\|u_t\| = \lambda^{t-1}\varepsilon_0^{-i-1}$, and $t + 1 = i(k+1) + 2$.

$$\|x_{t+1}\| \le \|A\|\|x_t\| + \|B\|\|u_t\| + \|w_t\| \le \frac{1}{4}\lambda(\lambda^{t-1}\varepsilon_0^{-i} + \lambda^{t-1}\varepsilon_0^{-i-1}) + 1 \le \lambda^t\varepsilon_0^{-i-1}.$$

Otherwise, we have $0 \le j \le k-1$, and $u_t = 0$. Moreover, $t + 1 \in \{(i-1)(k+1) + 3, \ldots, (i-1)(k+1) + 2 + k\}$. Therefore

$$\|x_{t+1}\| \le \|A\|\|x_t\| + \|B\|\|u_t\| + \|w_t\| \le \frac{1}{4}\lambda^t\varepsilon_0^{-i} + 1 \le \lambda^t\varepsilon_0^{-i}.$$

■

With appropriate choice of ξ_i, we ensure that $\hat{M}_j$ and A^jB are close in the Frobenius norm.

Lemma 12 *For $j = 0, \ldots, k$, $\hat{M}_j$ satisfies*

$$\|\hat{M}_j - A^jB\|_F \le 3d_u^2 k\lambda^{2k}\varepsilon_0.$$

In particular, $\|\hat{M}_0 - B\| \le 3d_u^2 k\lambda^{2k}\varepsilon_0 \le \varepsilon$.

Proof Observe that by definition,

$$x_{t+1} = A^t x_1 + \sum_{s=1}^{t} A^{t-s}Bu_s + A^{t-s}w_s.$$

We have $\|A^t x_1 + \sum_{s=1}^t A^{t-s}w_s\| \le \lambda^t + \sum_{s=1}^t \lambda^{t-s} \le (t+1)\lambda^t$. Note that the magnitude of this term should be small once we normalize by ξ_i. Let $j = t-2 \pmod{k+1}$, $i = (t-2-j)/(k+1)+1$, then $t = (i-1)(k+1) + j + 2$. We proceed to bound $\|x_t/\xi_i - (A^jB)_i\|$, where $(A^jB)_i$ is the ith column of A^jB. The largest sum in x_t can be analyzed as follows,

$$\begin{aligned}
\sum_{s=1}^{t-1} A^{t-1-s}Bu_s &= \sum_{s=1}^{(i-1)(k+1)+j+1} A^{(i-1)(k+1)+j+1-s}Bu_s \\
&= \sum_{r=0}^{i-1} A^{(i-1-r)(k+1)+j}Bu_{r(k+1)+1} \\
&= \sum_{r=0}^{i-1} A^{(i-1-r)(k+1)+j}B\xi_{r+1}e_{r+1}
\end{aligned}$$

$$= \sum_{r=1}^{i} \varepsilon_0^{-r} \lambda^{(r-1)(k+1)} A^{(i-r)(k+1)+j} B e_r.$$

Normalizing by the scaling factor,

$$\begin{aligned}
\frac{1}{\xi_i} \sum_{s=1}^{t-1} A^{t-1-s} B u_s &= \varepsilon_0^i \lambda^{(1-i)(k+1)} \sum_{r=1}^{i} \varepsilon_0^{-r} \lambda^{(r-1)(k+1)} A^{(i-r)(k+1)+j} B e_r \\
&= \sum_{r=1}^{i} \varepsilon_0^{i-r} \lambda^{(r-i)(k+1)} A^{(i-r)(k+1)+j} B e_r \\
&= A^j B e_i + \sum_{r=1}^{i-1} \varepsilon_0^{i-r} \lambda^{(r-i)(k+1)} A^{(i-r)(k+1)+j} B e_r.
\end{aligned}$$

The second term can be bounded as

$$\begin{aligned}
\| \sum_{r=1}^{i-1} \varepsilon_0^{i-r} \lambda^{(r-i)(k+1)} A^{(i-r)(k+1)+j} B e_r \| &\le \sum_{r=1}^{i-1} \varepsilon_0^{i-r} \lambda^{(r-i)(k+1)} \| A^{(i-r)(k+1)+j} B \| \\
&\le \lambda^{j+1} \sum_{r=1}^{i-1} \varepsilon_0^{i-r} \le (d_u - 1) \lambda^{2k} \varepsilon_0.
\end{aligned}$$

Let $(\hat{M}_j)_i$ denote the i-th column of $\hat{M}_j$, then we have

$$\begin{aligned}
\| (\hat{M}_j)_i - (A^j B)_i \| &= \| \frac{x_i^j}{\xi_i} - (A^j B)_i \| \\
&\le \frac{1}{\xi_i} \| A^{t-1} x_1 + \sum_{s=1}^{t-1} A^{t-1-s} w_s \| + \| \frac{1}{\xi_i} \sum_{s=1}^{t-1} A^{t-1-s} B u_s - (A^j B)_i \| \\
&\le \frac{1}{\xi_i} t \lambda^{t-1} + (d_u - 1) \lambda^{2k} \varepsilon_0 \\
&\le t \varepsilon_0^i \lambda^{2k} + (d_u - 1) \lambda^{2k} \varepsilon_0 \le 3 d_u k \lambda^{2k} \varepsilon_0.
\end{aligned}$$

Thus we can bound the Frobenius norm of $\hat{M}_j - A^j B$ by

$$\| \hat{M}_j - A^j B \|_F^2 = \sum_{i=1}^{d_u} \| (\hat{M}_j)_i - (A^j B)_i \|^2 \le 9 d_u^3 k^2 \lambda^{4k} \varepsilon_0^2.$$

■

We show that by our choice of ε_0, $\| C_0 - C_k \|_F$, $\| C_1 - Y \|_F$ are sufficiently small to guarantee $\hat{A}$ and A are close.

Lemma 13 *Algorithm 2 outputs $\hat{A}$ such that $\| \hat{A} - A \| \le \varepsilon$.*

Proof By Lemma 12, for all j, $\|\hat{M}_j - A^j B\|_F \le 3d_u^2 k\lambda^{2k}\varepsilon_0$. Let $C_k = [B\ AB\ A^2B\ \cdots A^{k-1}B]$, and $Y = [AB\ A^2B\ \ \cdots A^kB]$. We have

$$\|C_0 - C_k\|_F^2 = \sum_{j=0}^{k-1} \|\hat{M}_j - A^j B\|_F^2 \le 9d_u^4 k^3 \lambda^{4k}\varepsilon_0^2.$$

Similarly, $\|C_1 - Y\|_F^2 \le 9d_u^4 k^3\lambda^{4k}\varepsilon_0^2$.

Recall that A is the unique solution to the system of equations in X: $XC_k = Y$. Let A_i denote the i-th row of A, and let $\hat{A}_i$ denote the i-th row of $\hat{A}$. By Lemma 22 in (Hazan et al., 2020), as long as $\|C_0 - C_k\|_F \le \sigma_{min}(C_k)$,

$$\|A_i - \hat{A}_i\| \le \frac{\|C_1 - Y\|_F + \|C_0 - C_k\|_F\|A_i\|}{\sigma_{min}(C_k) - \|C_0 - C_k\|_F}$$

By our assumption, $\|(C_kC_k^\top)^{-1}\| \le \kappa$, so $\sigma_{min}(C_k) \ge \kappa^{-1/2}$. We have

$$\|C_0 - C_k\|_F \le 3d_u^2k^2\lambda^{2k}\varepsilon_0 \le \frac{\varepsilon}{2\lambda^k d_x\sqrt{\kappa}} \le \kappa^{-1/2}/2 \le \sigma_{min}(C_k).$$

Further notice that $\|A\| \le \lambda$ implies $\|A_i\| \le \|A\|_F \le \lambda\sqrt{d_x}$,

$$\|A_i - \hat{A}_i\| \le \frac{3d_u^2k^2\lambda^{2k}\varepsilon_0(1+\lambda\sqrt{d_x})}{\kappa^{-1/2} - 3d_u^2k^2\lambda^{2k}\varepsilon_0} \le \frac{\varepsilon}{\sqrt{d_x}}.$$

Finally, we have

$$\|A - \hat{A}\| \le \|A - \hat{A}\|_F = \sqrt{\sum_{i=1}^{d_x} \|A_i - \hat{A}_i\|^2} \le \varepsilon.$$

■

Lemma 14 *The total cost of estimating A, B starting from $\|x_1\| \le 1$ is bounded by*

$$G(10^5\lambda^{10k}\varepsilon^{-2}\kappa d_x^2k^5d_u^5)^{d_u}.$$

Proof The magnitude of the state and control is bounded by

$$\|x_t\|^2 + \|u_t\|^2 \le 2\lambda^{2t-2}\varepsilon_0^{-2d_u} \le 2\lambda^{4d_uk}\varepsilon_0^{-2d_u} = 2(\lambda^{4k}\varepsilon_0^{-2})^{d_u} = 2(10^4\varepsilon^{-2}d_u^4k^4\lambda^{10k}d_x^2\kappa)^{d_u}$$

By Assumption 3, taking $D^2 = 2(10^4\varepsilon^{-2}d_u^4k^4\lambda^{10k}d_x^2\kappa)^{d_u}$,

$$c_t(x_t,u_t) \le \|\nabla_{(x,u)}c_t(x_t,u_t)\|\|(x_t,u_t)\| \le 2GD^2.$$

Summing over $(k+1)d_u \le 2kd_u$ iterations, the total cost is upper bounded by

$$8Gkd_u(10^4\varepsilon^{-2}d_u^4k^4\lambda^{10k}d_x^2\kappa)^{d_u} \le G(10^5\lambda^{10k}\varepsilon^{-2}\kappa d_x^2k^5d_u^5)^{d_u}.$$

■

Using our choice of ε and λ, the total cost is bounded by

$$\begin{aligned} G(10^5\lambda^{10k}\varepsilon^{-2}\kappa d_x^2 k^5 d_u^5)^{d_u} &\le G(10^{25k}\beta^{10k}\gamma'^{-4}\kappa d_x^6 k^5 d_u^5 \kappa'^{16})^{d_u} \\ &\le G(10^{30k}\beta^{10k}\kappa d_x^6 k^5 d_u^5 \kappa'^{24})^{d_u} \\ &\le G(10^{30k}\beta^{10k}\kappa d_x^{18} k^5 d_u^5 C^{12})^{d_u} \\ &\le G(10^{40k}\beta^{82k} d_x^{18} k^{30} d_u^5 \kappa^{25})^{d_u} \\ &= 2^{O(\mathcal{L}\log\mathcal{L})}. \end{aligned}$$

Appendix B. Proofs for Section 4.2

In this section we prove that a $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable controller can be obtained by solving the SDP in Algorithm 3. We first argue that for two systems close in spectral norm, a strongly stable controller for one system is also strongly stable for ther other system.

Lemma 15 *If K is (κ, γ) strongly stable for a system (A, B) with $\kappa \ge 1$, and if $\hat{A}$, $\hat{B}$ satisfy $\|A - \hat{A}\| \le \varepsilon$, $\|B - \hat{B}\| \le \varepsilon$, then K is $(\kappa, \gamma - 2\varepsilon\kappa^2)$ strongly stable for $(\hat{A}, \hat{B})$.*

Proof By definition, we have

$$\begin{aligned} \hat{A} + \hat{B}K &= A + BK - A - BK + \hat{A} + \hat{B}K \\ &= HLH^{-1} + (\hat{A} - A) + (\hat{B} - B)K \\ &= H(L + H^{-1}(\hat{A} - A + (\hat{B} - B)K)H)H^{-1} \end{aligned}$$

The lemma follows by observing that

$$\|L + H^{-1}(\hat{A} - A + (\hat{B} - B)K)H\| \le 1 - \gamma + \kappa\varepsilon(1 + \kappa) \le 1 - \gamma + 2\varepsilon\kappa^2.$$

■

Now, we use Lemma 15 twice to show that the recovered controller $\hat{K}$ is strongly stable for the original system (A, B). The following lemma computes $\tilde{\kappa}, \tilde{\gamma}$ in terms of ε.

Lemma 16 *Algorithm 3 returns $\hat{K}$ that is $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable for A and B, where*

$$\tilde{\kappa} = \Big(\frac{\kappa'^4 2d_x}{\gamma' - 2\varepsilon\kappa'^2}\Big)^{1/2}, \quad \tilde{\gamma} = \frac{\gamma' - 2\varepsilon\kappa'^2}{4d_x\kappa'^4} - 2\varepsilon\tilde{\kappa}^2.$$

Proof We show in Section 2.3 that a (κ', γ') strongly stable controller exists for (A, B). Let K be a (κ', γ') strongly stable controller. By Lemma 12, 13, and 15, K is $(\bar{\kappa}, \bar{\gamma})$ strongly stable for $\hat{A}, \hat{B}$, where $\bar{\kappa} = \kappa', \bar{\gamma} = \gamma' - 2\varepsilon\kappa'^2$. With knowledge of $\bar{\kappa}, \bar{\gamma}$, we can set the trace upper bound appropriately to extract a strongly stable controller from a feasible solution of the SDP. Specifically, we set

$$\nu = \frac{2\bar{\kappa}^4 d_x}{\bar{\gamma}}$$

as in Lemma 18, and the SDP is feasible. We obtain $\hat{K}$ that is $(\hat{\kappa}, \hat{\gamma})$ strongly stable for the system $\hat{A}, \hat{B}$, where $\hat{\kappa} = \frac{\bar{\kappa}^2\sqrt{2d_x}}{\sqrt{\bar{\gamma}}} = \big(\frac{\kappa'^4 2d_x}{\gamma' - 2\varepsilon\kappa'^2}\big)^{1/2}$, $\hat{\gamma} = \frac{\bar{\gamma}}{4d_x\bar{\kappa}^4} = \frac{\gamma' - 2\varepsilon\kappa'^2}{4d_x\kappa'^4}$. We apply Lemma 15 again and conclude that $\hat{K}$ is $(\hat{\kappa}, \hat{\gamma} - 2\varepsilon\hat{\kappa}^2)$ strongly stable for A, B. ■

With our choice of ε, we compute the final values of $\tilde{\kappa}, \tilde{\gamma}$.

Lemma 17 *Setting* $\varepsilon = \frac{\gamma'^2}{10^5 d_x^2 \kappa'^8}$, $\hat{K}$ *returned by Algorithm 3 is* $(\frac{2\kappa'^2 d_x^{1/2}}{\gamma'^{1/2}}, \frac{\gamma'}{16 d_x \kappa'^4})$ *strongly stable for* (A, B).

Proof With this choice of ε, we have $2\varepsilon\kappa'^2 = \frac{2\gamma'^2}{10^5 d_x^2 \kappa'^6} \leq \frac{\gamma'}{2}$. It follows that

$$\tilde{\kappa} = \left(\frac{\kappa'^4 2 d_x}{\gamma' - 2\varepsilon\kappa'^2}\right)^{1/2} \leq \frac{2\kappa'^2\sqrt{d_x}}{\sqrt{\gamma'}}.$$

Therefore we have $2\varepsilon\tilde{\kappa}^2 \leq \frac{\gamma'}{10^2 d_x \kappa'^4}$. We obtain a lower bound on $\tilde{\gamma}$ as follows

$$\tilde{\gamma} = \frac{\gamma' - 2\varepsilon\kappa'^2}{4 d_x \kappa'^4} - 2\varepsilon\tilde{\kappa}^2 \geq \frac{\gamma' - 2\varepsilon\kappa'^2}{4 d_x \kappa'^4} - \frac{\gamma'}{10^2 d_x \kappa'^4} \geq \frac{\gamma'}{8 d_x \kappa'^4} - \frac{\gamma'}{10^2 d_x \kappa'^4} \geq \frac{\gamma'}{16 d_x \kappa'^4}.$$

■

The following lemma details how we set the trace upper bound ν in the SDP, and our application of results from Cohen et al. (2018) to extract $\hat{K}$.

Lemma 18 *For any system* A, B *with a* (κ, γ) *strongly stable controller, the SDP in Algorithm 3 defined by* (A, B) *with trace constraint* $\nu = \frac{2\kappa^4 d_x}{\gamma}$ *is feasible. Moreover, a policy* K *such that* K *is* $(\frac{\kappa^2\sqrt{2d_x}}{\sqrt{\gamma}}, \frac{\gamma}{4 d_x \kappa^4})$ *strongly stable for* A, B *can be extracted from any feasible solution of the SDP.*

Proof We first show that the SDP is feasible. Let K be the (κ, γ) strongly stable controller for (A, B), and consider the system with Gaussian noise $x_{t+1} = Ax_t + Bu_t + w_t$, $w_t \sim N(0, I)$. This system will converge to a steady state where the state covariance $X = \mathbb{E}[xx^\top]$ satisfies

$$X = (A + BK)X(A + BK)^\top + I.$$

Let $KXK^\top$ be the steady-state covariance of u when following K. By Lemma 3.3 in (Cohen et al., 2018), $\mathrm{Tr}(X) \leq \frac{\kappa^2 d_x}{\gamma}$, $\mathrm{Tr}(KXK^\top) \leq \frac{\kappa^4 d_x}{\gamma}$.

Consider the matrix

$$\Sigma = \begin{pmatrix} X & XK^\top \\ KX & KXK^\top \end{pmatrix}.$$

By Lemma 4.1 in (Cohen et al., 2018), Σ is feasible for the SDP if $\nu \geq \mathrm{Tr}(X) + \mathrm{Tr}(KXK^\top)$; since $\nu = \frac{2\kappa^4 d_x}{\gamma}$, Σ is feasible for the SDP. Now let $\hat{\Sigma}$ be any feasible solution of the SDP, and write

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{xx} & \hat{\Sigma}_{xu} \\ \hat{\Sigma}_{xu}^\top & \hat{\Sigma}_{uu} \end{pmatrix}.$$

Consider $\hat{K} = \hat{\Sigma}_{xu}^\top \hat{\Sigma}_{xx}^{-1}$, which is well-defined because $\hat{\Sigma}_{xx} \succeq I$ by the steady-state constraint. As shown in Lemma 4.3 in (Cohen et al., 2018), $\hat{K}$ is $(\sqrt{\nu}, 1/(2\nu))$ strongly stable for A, B. Under our choice of ν, $\hat{K}$ is $(\frac{\kappa^2\sqrt{2d_x}}{\sqrt{\gamma}}, \frac{\gamma}{4 d_x \kappa^4})$ strongly stable for A, B. ■

B.1. Decaying the System

Lemma 19 *Let K be a $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable controller for the system. and x_1 be any starting state. Suppose $\tilde{\kappa} \geq 1$. After following K for $T_2 = \max\{\frac{\ln(\tilde{\gamma}\|x_1\|)}{\tilde{\gamma}}, 0\}$ iterations , the final state x_{T_2+1} satisfies $\|x_{T_2+1}\| \leq 2\tilde{\kappa}/\tilde{\gamma}$, and the total cost is bounded by*

$$O\big(G\tilde{\kappa}^4\|x_1\|^3\tilde{\gamma}^{-3}\big).$$

Proof Under the controller K, the state evolution satisfies

$$x_{t+1} = (A+BK)^t x_1 + \sum_{i=1}^{t}(A+BK)^{t-i}w_i.$$

By definition of strong stability, $\|(A+BK)^t\| \leq \|H\|\|H^{-1}\|\|L\|^t \leq \tilde{\kappa}(1-\tilde{\gamma})^t$. It follows that

$$\|x_{t+1}\| \leq \tilde{\kappa}(1-\tilde{\gamma})^t\|x_1\| + \tilde{\kappa}\sum_{i=1}^{t}(1-\tilde{\gamma})^{t-i} \leq \tilde{\kappa}(1-\tilde{\gamma})^t\|x_1\| + \frac{\tilde{\kappa}}{\tilde{\gamma}}.$$

Let $T_2 = \max\{\frac{\ln(\tilde{\gamma}\|x_1\|)}{\tilde{\gamma}}, 0\}$. If $\ln(\tilde{\gamma}\|x_1\|) \geq 0$, we have $T_2 \geq -\frac{\ln(\tilde{\gamma}\|x_1\|)}{\ln(1-\tilde{\gamma})}$, hence $(1-\tilde{\gamma})^{T_2} \leq 1/(\tilde{\gamma}\|x_1\|)$ and $\|x_{T_2+1}\| \leq 2\tilde{\kappa}/\tilde{\gamma}$. Otherwise $T_2 = 0$ and $\|x_1\| \leq 1/\tilde{\gamma} < 2\tilde{\kappa}/\tilde{\gamma}$. Notice that $\|x_t\| \leq \tilde{\kappa}\|x_1\| + \tilde{\kappa}/\tilde{\gamma}$, $\|u_t\| \leq \tilde{\kappa}^2\|x_1\| + \tilde{\kappa}^2/\tilde{\gamma}$ for all $t \in [T_2+1]$. Taking $D = \tilde{\kappa}^2\|x_1\| + \tilde{\kappa}^2/\tilde{\gamma}$ and assuming $\ln(\tilde{\gamma}\|x_1\|) \geq 0$, the total cost of decaying the system is bounded by

$$\begin{aligned}
2(T_2+1)GD^2 &= 2G(\frac{\ln(\tilde{\gamma}\|x_1\|)}{\tilde{\gamma}} + 1)(\tilde{\kappa}^2\|x_1\| + \tilde{\kappa}^2/\tilde{\gamma})^2 \\
&\leq 4G(\frac{\ln(\tilde{\gamma}\|x_1\|)}{\tilde{\gamma}} + 1)\tilde{\kappa}^4(\|x_1\|^2 + \frac{1}{\tilde{\gamma}^2}) \\
&\leq 8G(\frac{\ln(\|x_1\|)}{\tilde{\gamma}} + 1)\tilde{\kappa}^4\|x_1\|^2\tilde{\gamma}^{-2} \\
&\leq 8G(\ln(\|x_1\|) + 1)\tilde{\kappa}^4\|x_1\|^2\tilde{\gamma}^{-3} \\
&\leq 16G\tilde{\kappa}^4\|x_1\|^3\tilde{\gamma}^{-3}.
\end{aligned}$$

The same upper bound holds for $T_2 = 0$. ■

Appendix C. Proofs for Section 4.3

In this section we give an upper bound on quantities related to the controllability of the stabilized system $(A + B\hat{K}, B)$, and include the main results in Hazan et al. (2020) for completeness. The following lemma is an equivalent characterization of strong controllability.

Lemma 20 *A system defined by $x_{t+1} = Ax_t + Bu_t$ is (k, κ)-strongly controllable if and only if it can drive $x_1 = 0$ to any state x_f where $\|x_f\| = 1$ in k steps with control cost at most κ. I.e., there exists $u_1, \ldots, u_k, x_2, \ldots, x_{k+1}$ such that $x_{k+1} = x_f$, $x_{t+1} = Ax_t + Bu_t$, and*

$$\sum_{t=1}^{k}\|u_t\|^2 \leq \kappa.$$

Proof Consider the quadratic program:

$$\begin{aligned} \min_{(u_t)_{t=1}^k} \quad & \sum_{t=1}^{k} \|u_t\|^2 \\ \text{s.t.} \quad & x_{t+1} = Ax_t + Bu_t \\ & x_{k+1} = x_f,\ x_1 = 0 \end{aligned} \tag{2}$$

Recall $C_k = [B\ AB\ \cdots\ A^{k-1}B]$, and let $(v_1, v_2, \ldots, v_k) \in \mathbb{R}^{kn}$ denote the concatenation of k n-dimensional vectors. Then this is equivalent to

$$\begin{aligned} \min_{(u_t)_{t=1}^k} \quad & \sum_{t=1}^{k} \|u_t\|^2 \\ \text{s.t.} \quad & C_k(u_k, u_{k-1}, \ldots, u_1) = x_f \end{aligned} \tag{3}$$

Suppose the system is (k, κ) strongly controllable, then C_k has full row-rank, and $C_k C_k^\top$ is invertible with $\|(C_k C_k^\top)^{-1}\| \le \kappa$. Therefore (3) is feasible for all unit vectors x_f. By Lemma B.6 in Cohen et al. (2018), an optimal solution to (3) is given by $C_k^\top (C_k C_k^\top)^{-1} x_f$, and its value is at most

$$\sum_{t=1}^{k} \|u_t\|^2 = \|C_k^\top (C_k C_k^\top)^{-1} x_f\|^2 = x_f^\top (C_k C_k^\top)^{-1} x_f \le \|(C_k C_k^\top)^{-1}\| = \kappa.$$

Now suppose for any unit vector x_f, there exists $u_1, \ldots, u_k, x_2, \ldots, x_{k+1}$ such that $x_{k+1} = x_f$, $x_{t+1} = Ax_t + Bu_t$, and $\sum_{t=1}^k \|u_t\|^2 \le \kappa$. Then (3) is feasible for any unit vector x_f, implying that C_k has full row-rank and $(C_k C_k^\top)$ is invertible. Moreover, the optimal value is at most κ. Let x_f be the eigenvector corresponding to the largest eigenvalue of $(C_k C_k^\top)^{-1}$. Then an optimal solution to (3) is $C_k^\top (C_k C_k^\top)^{-1} x_f$, and the value satisfies $\|C_k^\top (C_k C_k^\top)^{-1} x_f\|^2 = x_f^\top (C_k C_k^\top)^{-1} x_f \le \kappa$. We conclude that $\|(C_k C_k^\top)^{-1}\| \le \kappa$, and the system is (k, κ) strongly controllable. ■

Using our characterization, we show an upper bound on the controllability parameter of $(A + BK, B)$ where K is any linear controller with a bounded spectral norm.

Lemma 21 *Suppose (A, B) is (k, κ) strongly controllable and $\|A\|, \|B\| \le \beta$. Let K be a linear controller with $\|K\| \le \kappa'$, then the system $(A + BK, B)$ is (k, κ_0) strongly controllable, with $\kappa_0 = 4\kappa'^2 k^2 \beta^{2k} \kappa$.*

Proof Let $C_k = [B\ AB\ \cdots\ A^{k-1}B]$. By the definition of strong controllability, C_k has full row-rank, and under the noiseless system $x_{t+1} = Ax_t + Bu_t$, any state is reachable by time $k+1$ starting from $x_1 = 0$. We will show that any state is reachable at time $t+1$ for the system $(A + BK, B)$ as well. Let $v \in \mathbb{R}^m$ be an arbitrary state, and the sequence of controls $(u_1, u_2, \ldots, u_k) = C_k^\top (C_k C_k^\top)^{-1} v$ can be used to reach v from initial state $x_1 = 0$, i.e.

$$x_{k+1} = \sum_{i=1}^{k} A^{k-i} B u_i = C_k(u_1, u_2, \ldots, u_k) = v.$$

Let $\{x_t\}$ denote the state trajectory under controls $\{u_t\}$, where $x_{k+1} = v$. Consider the system $y_{t+1} = (A + BK)y_t + Bz_t = Ay_t + B(z_t + Ky_t)$, where y_t's are states and z_t's are controls. We

claim that the sequence of controls $z_t = u_t - Ky_t$ can be used to drive the system to v in $k+1$ steps from initial state $y_1 = 0$. Let $\{y_t\}$ denote the system's trajectory under controls $\{z_t\}$. For our base case, we have $y_2 = B(z_1 + Ky_1) = Bu_1 = x_2$, since $y_1 = x_1 = 0, z_1 = u_1 - Ky_1$. Assume $x_t = y_t$ for some $t \le k$. For $t+1$, $y_{t+1} = Ay_t + B(z_t + Ky_t) = Ay_t + Bu_t = Ax_t + Bu_t = x_{t+1}$. We conclude that the trajectories $\{x_t\}$ and $\{y_t\}$ are the same and $v = y_{k+1}$. Since we can write $y_{k+1} = \sum_{i=1}^{k}(A+BK)^{k-i}Bz_i$, y_{k+1} is in the range of the matrix $C_k' = [B\ (A+BK)B\ \cdots\ (A+BK)^{k-1}B]$; therefore C_k' has full row-rank.

Now we show the controls $\{z_t\}$ satisfy $\sum_{t=1}^{k}\|z_t\|^2 \le 4\kappa'^2k^2\beta^{2k}\kappa\|v\|^2$. By our choice of z_t, we have $z_t = u_t - Ky_t = u_t - Kx_t$; therefore $\sum_{t=1}^{k}\|z_t\|^2 \le 2\sum_{t=1}^{k}(\|u_t\|^2 + \|K\|^2\|x_t\|^2)$. By our choice of u_t, we have

$$\sum_{t=1}^{k}\|u_t\|^2 = \|C_k^\top(C_kC_k^\top)^{-1}v\|^2 = v^\top(C_kC_k^\top)^{-1}v \le \kappa\|v\|^2.$$

Further, the trajectory $\{x_t\}_{t=1}^{k}$ satisfies

$$\|x_t\|^2 = \|\sum_{i=1}^{t-1}A^{t-1-i}Bu_i\|^2 \le k\sum_{i=1}^{t-1}\|A^{t-1-i}B\|^2\|u_i\|^2 \le k\beta^{2k}\kappa\|v\|^2$$

Hence we have

$$\sum_{t=1}^{k}\|z_t\|^2 \le 2\kappa\|v\|^2 + 2\kappa'^2k^2\beta^{2k}\kappa\|v\|^2 \le 4\kappa'^2k^2\beta^{2k}\kappa\|v\|^2.$$

By Lemma 20, $(A+BK, B)$ is $(k, 4\kappa'^2k^2\beta^{2k}\kappa)$ strongly controllable. ■

Algorithm 4 is the main algorithm (Algorithm 1) in Hazan et al. (2020), where T_0, η, H are internal parameters that can be set by the learner. In line 10, let $\Pi_{\mathcal{M}}$ denote projection onto the set $\mathcal{M}$, and let f_t denote the surrogate cost at time t as in Definition 11 of Hazan et al. (2020). Theorem 22 gives the regret bound for the algorithm when the internal parameters are set appropriately.

Theorem 22 *[Theorem 12 in Hazan et al. (2020)] Suppose $\hat{K}$ is $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable for (A, B), and the system $(A + B\hat{K}, B)$ is (k, κ^*) strongly controllable. In addition, assume that the noise sequence w_t satisfies $\|w_t\| \le W$ for all t. Then Algorithm 4 with $H = \Theta(\tilde{\gamma}^{-1}\log((\kappa^*)^2T)), \eta = \Theta(GW\sqrt{T})^{-1}, T_0 = \Theta(T^{2/3}\log(1/\delta))$, incurs regret upper bounded by*

$$\textit{Regret} = O(\mathrm{poly}(\kappa^*, \tilde{\gamma}^{-1}, k, d_x, d_u, G, W)T^{2/3}\log(1/\delta)).$$

with probability at least $1-\delta$ for controlling an unknown LDS.

Appendix D. Proofs for Lower Bound for Randomized Black-box Control Algorithms

Lemma 23 *Let $T = d_x/8$. There exists a sequence of orthonormal matrices $V_1, \ldots, V_T$, such that for $t \in [T]$, V_t only depends on $x_1, \ldots, x_t, u_1, \ldots, u_t$ and they satisfy the following condition: Let r_t denote the rank of span$(x_1, \ldots, x_t, u_1, \ldots, u_t)$, and denote the first r_t columns of V_t as $V_t^{\|}$, and the last $d - r_t$ columns of V_t as $V_t^{\perp}$. Let $h_t = (V_{t-1}^{\perp})^\top x_t$, then for all $t \in [T]$, conditioned on $x_1, x_2, \ldots, x_t, u_1, u_2, \ldots, u_t, Au_1, \ldots, Au_{t-1}$, we have $(V_t^{\perp})^\top x_{t+1} = c_t + z_t$, where the coordinates of z_t are iid normally distributed, i.e. $z_t(i) \sim N(0, \frac{\gamma\|h_t\|^2}{d})$.*

Algorithm 4 Adversarial Control via System Identification

Input: Number of iterations T, $\tilde{\gamma}, \hat{K}$ such that $\hat{K}$ is $(\tilde{\kappa}, \tilde{\gamma})$ strongly stable, κ^*, k such that $(A + B\hat{K}, B)$ is (k, κ^*) strongly controllable, and $\kappa^* \geq \tilde{\kappa}$.

Phase 1: **System Identification.**

Call Algorithm 2 in (Hazan et al., 2020) with a budget of T_0 rounds to obtain system estimates $\tilde{A}, \tilde{B}$.

Phase 2: **Robust Control.**

Define the constraint set $\mathcal{M} = \{M = \{M^0 \dots M^{H-1}\} : \|M^{i-1}\| \leq \kappa^4(1-\gamma)^i\}$.

Initialize $\hat{w}_{T_0} = x_{T_0+1}$ and $\hat{w}_t = 0$ for $t < T_0$.

for $t = T_0 + 1, \dots, T$ **do**

Choose the action:

$$u_t = \hat{K}x_t + \sum_{i=1}^{H} M_t^{i-1}\hat{w}_{t-i}.$$

Observe the new state x_{t+1} and cost $c_t(x_t, u_t)$.

Record estimate $\hat{w}_t = x_{t+1} - \tilde{A}x_t - \tilde{B}u_t$.

Update:

$$M_{t+1} = \Pi_{\mathcal{M}}(M_t - \eta \nabla f_t(M_t | \tilde{A}, \tilde{B}, \{\hat{w}\}))$$

end for

Proof Fix $t \leq T$, and condition on $x_1, \dots, x_t$, $u_1, \dots, u_{t-1}, Au_1, \dots, Au_t$. Let $V_1, \dots, V_{t-1}$ be constructed as in Corollary 26. By construction, the first r_{t-1} columns of V_{t-1}, denoted as $V_{t-1}^{\|}$, form a basis for span$(x_1, \dots, x_{t-1}, u_1, \dots, u_{t-1})$. Recall the last $d - r_{t-1}$ columns of V_{t-1} is denoted as $V_{t-1}^{\perp}$. We have

$$\begin{aligned}
(V_t^{\perp})^\top x_{t+1} &= (V_t^{\perp})^\top A x_t + (V_t^{\perp})^\top u_t \\
&= (V_t^{\perp})^\top A V_{t-1} V_{t-1}^\top x_t && ((V_t^{\perp})^\top u_t = 0) \\
&= (V_t^{\perp})^\top \left[\begin{array}{c|c} AV_{t-1}^{\|} & AV_{t-1}^{\perp} \end{array} \right] V_{t-1}^\top x_t \\
&= \left[\begin{array}{c|c} (V_t^{\perp})^\top AV_{t-1}^{\|} & (V_t^{\perp})^\top AV_{t-1}^{\perp} \end{array} \right] V_{t-1}^\top x_t \\
&= \left[\begin{array}{c|c} (V_t^{\perp})^\top AV_{t-1}^{\|} & (V_t^{\perp})^\top AV_{t-1}^{\perp} \end{array} \right] \left[\begin{array}{c} (V_{t-1}^{\|})^\top \\ (V_{t-1}^{\perp})^\top \end{array} \right] x_t \\
&= (V_t^{\perp})^\top AV_{t-1}^{\|}(V_{t-1}^{\|})^\top x_t + (V_t^{\perp})^\top AV_{t-1}^{\perp}(V_{t-1}^{\perp})^\top x_t
\end{aligned}$$

Denote $(V_t^{\perp})^\top AV_{t-1}^{\|}(V_{t-1}^{\|})^\top x_t$ as $c_t \in \mathbb{R}^{d-r_t}$, and recall $h_t = (V_{t-1}^{\perp})^\top x_t$. We have

$$(V_t^{\perp})^\top x_{t+1} = c_t + G_t h_t,$$

where $G_t = (V_t^{\perp})^\top AV_{t-1}^{\perp}$. By Corollary 26, conditioned on $x_1, \dots, x_t, u_1, \dots, u_t, Au_1, \dots, Au_{t-1}$, $G_t \sim N(d-r_t, d-r_{t-1}, \frac{\gamma}{d_x})$, and therefore the coordinates of $z_t = G_t h_t$ are iid normally distributed with zero mean and $\frac{\gamma\|h_t\|^2}{d_x}$ variance. ■

Lemma 24 *Let $V_1, \ldots, V_T$ be as in Lemma 10, and $T \leq d_x/8$. Let $h_t = (V_{t-1}^{\perp})^\top x_t$, with probability at least $1 - \exp(-\frac{d_x}{25})$, conditioned on $x_1, x_2, \ldots, x_t, u_1, u_2, \ldots, u_t$, $Au_1, \ldots, Au_{t-1}$, we have $\|(V_t^{\perp})^\top x_{t+1}\|^2 \geq \frac{\gamma\|h_t\|^2}{20}$.*

Proof By Lemma 10, conditioned on $x_1, \ldots, x_t, u_1, \ldots, u_t, Au_1, \ldots, Au_t$, we have $h_{t+1} = (V_t^{\perp})^\top x_{t+1} \sim N(c_t, \frac{\gamma\|h_t\|^2}{d_x} I)$. There exists a rotation R of h_{t+1}, such that $Rh_{t+1} \sim N(\|c_t\|e_1, \frac{\gamma\|h_t\|^2}{d_x} I)$. Let $Rh_{t+1}(i)$ denote the i-th coordinate of the vector Rh_{t+1}, then we have $\sum_{i=2}^{d_x - r_t} Rh_{t+1}(i)^2 \sim \frac{\gamma\|h_t\|^2}{d_x}\chi_{d_x - r_t - 1}$ follows a chi-square distribution. By Lemma 1 in Laurent and Massart (2000), for a random variable $Y \sim \chi_k$, $\mathbb{P}[Y \leq k - 2\sqrt{kx}] \leq \exp(-x)$. Therefore for $t \leq T < \frac{d_x}{8}$, $r_t \leq 2t \leq \frac{d_x}{4}$, we have

$$\begin{aligned}
\mathbb{P}[\sum_{i=2}^{d_x - r_t} Rh_{t+1}(i)^2 \leq \frac{\gamma\|h_t\|^2}{20}] &= \mathbb{P}[\frac{d_x}{\gamma\|h_t\|^2}\sum_{i=2}^{d_x - r_t} Rh_{t+1}(i)^2 \leq \frac{d_x}{20}] \\
&\leq \mathbb{P}[\frac{d_x}{\gamma\|h_t\|^2}\sum_{i=2}^{d_x - r_t} Rh_{t+1}(i)^2 \leq d_x - r_t - 1 - 2\sqrt{(d_x - r_t - 1)d_x/25}] \\
&\leq \exp(-\frac{d_x}{25})
\end{aligned}$$

We conclude that

$$\begin{aligned}
\mathbb{P}[\|(V_t^{\perp})^\top x_{t+1}\|^2 \geq \frac{\gamma\|h_t\|^2}{20}] = \mathbb{P}[\|Rh_{t+1}\|^2 \geq \frac{\gamma\|h_t\|^2}{20}] &\geq \mathbb{P}[\sum_{i=2}^{d_x - r_t} Rh_{t+1}(i)^2 \geq \frac{\gamma\|h_t\|^2}{20}] \\
&\geq 1 - \exp(-\frac{d_x}{25}).
\end{aligned}$$

■

Lemma 25 *Consider the observation model, where $A \sim N(d, d, \gamma)$, and a player can make queries defined by vectors $q_1, q_2, \ldots, q_T$, $T \leq d$. In turn, the player observes $w_1 = Aq_1, w_2 = Aq_2, \ldots, w_T = Aq_T$. For any $t \leq T$, the player is allowed to choose q_t as a deterministic function of the previous queries and observations. Let r_t denote the rank of span$(q_1, \ldots, q_t)$. Then for all $t \leq T$, there exists an orthonormal matrix V_t that can be constructed only as a function of $q_1, \ldots, q_t$, such that with $V_t^{\perp}$ denoting the last $d - r_t$ columns of V_t, the following hold:*

1. *Conditioned on $q_1, q_2, \ldots, q_t, w_1, \ldots, w_{t-1}$, $(V_t^{\perp})^\top A V_{t-1}^{\perp} \sim N(d - r_t, d - r_{t-1}, \gamma)$.*
2. *Conditioned on $q_1, q_2, \ldots, q_t, w_1, \ldots, w_t$, $AV_t^{\perp} \sim N(d, d - r_t, \gamma)$.*

Proof We first construct $V_1, \ldots, V_T$. For $t \leq T$, let r_t be the rank of span$(q_1, q_2, \ldots, q_t)$, $r_t \leq t$, and denote the normalized component of q_t that lies outside of span$(q_1, q_2, \ldots, q_{t-1})$ as $\tilde{q}_t$. Let $W_1, \ldots, W_T$ be orthonormal matrices, such that if $\tilde{q}_t = 0$, $W_t = I$; otherwise, the first r_{t-1} columns of W_t are standard basis vectors $e_1, \ldots, e_{r_{t-1}}$, and the r_t-th column of W_t, denoted as z_t, is such that $W_1 W_2 \cdots W_{t-1} z_t = \tilde{q}_t$. Such a W_t exists because the product $W_1 \cdots W_{t-1}$ is an orthonormal matrix and thus is full rank. Moreover, z_t is orthogonal to $e_1, \ldots e_{r_t - 1}$ by the construction of $\tilde{q}_t$.

Let $V_t = W_1 W_2 \cdots W_t$. Then by definition, V_t is orthonormal, and the first r_t columns of V_t form a basis for span$(q_1, q_2, \ldots, q_t)$. Moreover, V_t only depends on $q_1, \ldots, q_t$. Denote the first r_t columns of V_t by $V_t^{\|}$, and recall that the last $d - r_t$ columns of V_t is denoted by $V_t^{\perp}$. Now we prove the lemma by induction.

Base case. Define $V_0 = 0$. Since q_1 is chosen without any observations, it is independent of A, and hence V_1 is independent of A. Therefore conditioned on q_1, $AV_1^{\|}$ is independent of $AV_1^{\perp}$, and $(V_1^{\perp})^{\top} AV_0^{\perp} = (V_1^{\perp})^{\top} AI \sim N(d - r_1, d, \gamma)$, $AV_1^{\perp} \sim N(d, d - r_1, \gamma)$. Since w_1 only depends on $AV_1^{\|}$, it is independent of $AV_1^{\perp}$. We conclude that conditioned on q_1 and w_1, $AV_1^{\perp} \sim N(d, d - r_1, \gamma)$.

Inductive step. Suppose for all $s < t$, the two conditions in the lemma hold. By definition, q_t is a deterministic function of $q_1, \ldots, q_{t-1}, w_1, \ldots, w_{t-1}$, so by the inductive hypothesis, conditioned on $q_1, \ldots, q_{t-1}, q_t, w_1, \ldots, w_{t-1}$, $AV_{t-1}^{\perp} \sim N(d, d - r_{t-1}, \gamma)$, and we can obtain W_t and V_t. Since V_t is only a function of $q_1, \ldots, q_t$, we have $(V_t^{\perp})^{\top} AV_{t-1}^{\perp} \sim N(d - r_t, d - r_{t-1}, \gamma)$. Denote the last $d - r_t$ columns of W_t as Z_t. Now observe

$$V_t^{\perp} = V_{t-1} W_t \begin{bmatrix} 0_{r_t \times r_t} \\ I_{d-r_t} \end{bmatrix} = \left[\begin{array}{c|c} V_{t-1}^{\|} & V_{t-1}^{\perp} \end{array} \right] Z_t$$

By construction the columns of Z_t are orthogonal to $e_1, \ldots, e_{r_{t-1}}$, therefore their first r_{t-1} coordinates are all zero, and we can write $Z_t = \begin{bmatrix} 0_{(r_{t-1}) \times (d-r_t)} \\ \tilde{Z}_t \end{bmatrix}$, where $\tilde{Z}_t \in \mathbb{R}^{(d-r_{t-1}) \times (d-r_t)}$ have orthonormal columns. Therefore we have

$$V_t^{\perp} = V_{t-1}^{\perp} \tilde{Z}_t$$

Since $\tilde{Z}_t$ is independent of $AV_{t-1}^{\perp}$, we have $AV_t^{\perp} = AV_{t-1}^{\perp} \tilde{Z}_t \sim N(d, d - r_t, \gamma)$. Now we need to show that this distribution doesn't change conditioned on w_t. If $q_t \in$ span$(q_1, \ldots, q_{t-1})$, then $w_t = Aq_t$ can be determined by $w_1, \ldots, w_{t-1}$, so the distribution of $AV_t^{\perp}$ remains the same conditioned on w_t. Now assume $q_t \notin$ span$(q_1, \ldots, q_{t-1})$, and $r_t = r_{t-1} + 1$. Since w_t is determined by $AV_t^{\|}$, it suffices to show that $AV_t^{\|}$ is independent of $AV_t^{\perp}$ conditioned on $q_1, \ldots, q_t, w_1, \ldots, w_{t-1}$. Consider the following decomposition

$$AV_t^{\|} = A \left[\begin{array}{c|c} V_{t-1}^{\|} & V_t e_{r_t} \end{array} \right] = \left[\begin{array}{c|c} AV_{t-1}^{\|} & AV_t e_{r_t} \end{array} \right].$$

By the construction of $V_{t-1}^{\|}$, $AV_{t-1}^{\|}$ can be determined by $w_1, \ldots, w_{t-1}$. Therefore, by the inductive hypothesis, $AV_{t-1}^{\|}$ is independent of $AV_{t-1}^{\perp} \tilde{Z}_t = AV_t^{\perp}$. Now we expand $V_t e_{r_t} = V_{t-1} W_t e_{r_t}$, and as before, let $z_t = W_t e_{r_t}$. Since z_t is orthogonal to $e_1, \ldots, e_{r_{t-1}}$, the first r_{t-1} coordinates of z_t are zero. Let the last $d - r_{t-1}$ coordinates of z_t be y_t, then we have $V_t e_{r_t} = V_{t-1} z_t = V_{t-1}^{\perp} y_t$, and y_t is orthogonal to the columns of $\tilde{Z}_t$. By the inductive hypothesis, $AV_{t-1}^{\perp} \sim N(d, d - r_{t-1}, \gamma)$, so $AV_{t-1}^{\perp} y_t$ is independent of $AV_{t-1}^{\perp} \tilde{Z}_t$. We conclude that $AV_t e_{r_t}$ is independent of $AV_t^{\perp}$, so $AV_t^{\|}$ is independent of $AV_t^{\perp}$ and conditioned on w_t, $AV_t^{\perp} \sim N(d, d - r_t, \gamma)$. ■

Corollary 26 *Consider an alternative observation model, where $A \sim N(d, d, \gamma)$, and a player can make two queries at a time: $p_1, q_1, p_2, q_2, \ldots, p_T, q_T, T < d/2$. The player observes $v_t =$*

$Ap_t, w_t = Aq_t$ for $t \in [T]$, and the player can choose p_t, q_t as deterministic functions of $\{p_s\}_{s=1}^{t-1}$, $\{q_s\}_{s=1}^{t-1}$, $\{v_s\}_{s=1}^{t-1}$, $\{w_s\}_{s=1}^{t-1}$. Let r_t denote the rank of span$(\{p_s\}_{s=1}^{t}, \{q_s\}_{s=1}^{t})$. Then for all $t \le T$, there exists an orthonormal matrix V_t that can be constructed only as a function of $\{p_s\}_{s=1}^{t}, \{q_s\}_{s=1}^{t}$, such that with $V_t^{\perp}$ denoting the last $d - r_t$ columns of V_t, the following hold:

1. *Conditioned on $\{p_s\}_{s=1}^{t}, \{q_s\}_{s=1}^{t}, \{v_s\}_{s=1}^{t-1}, \{w_s\}_{s=1}^{t-1}$, $(V_t^{\perp})^\top A V_{t-1}^{\perp} \sim N(d-r_t, d-r_{t-1}, \gamma)$.*
2. *Conditioned on $\{p_s\}_{s=1}^{t}, \{q_s\}_{s=1}^{t}, \{v_s\}_{s=1}^{t}, \{w_s\}_{s=1}^{t}$, $AV_t^{\perp} \sim N(d, d - r_t, \gamma)$.*

Proof The proof is very similar to the proof of Lemma 25. ■

Appendix E. Lower Bound for Deterministic Black-box Control Algorithms

Theorem 27 *Let $\mathcal{A}$ be a deterministic black-box control algorithm. Then there exists a stabilizable system that is also $(1,1)$-strongly controllable, and a sequence of oblivious perturbations and costs, such that with $x_1 = e_1$, and $T = d_x$, we have*

$$\textit{Regret}_T(\mathcal{A}) = 2^{\Omega(\mathcal{L})}.$$

Let $c_t(x,u) = \|x\|^2 + \|u\|^2$ for all t. Consider the noiseless system $x_{t+1} = Q^\top V x_t + u_t$ for some Q and orthogonal V. Under this system $w_t = 0$ for all t, and a stabilizing controller is $K = -Q^\top V$. Observe that $J(K)$ is constant. The system is also $(1,1)$ strongly controllable because $B = I$. Let $V_i, Q_i \in \mathbb{R}^{1 \times d_x}$ denote the rows of V and Q, respectively. Fix a deterministic algorithm $\mathcal{A}$, and let $u_t = \mathcal{A}(x_1, x_2, \ldots, x_t, c_1, \ldots, c_t)$ be the control produced by $\mathcal{A}$ at time t. There exists Q, V such that under this system, $\mathcal{A}$ outputs controls such that $\|x_{d_x}\| \ge 2^{d_x - 1}$.

The construction. Set $x_1 = e_1$. We construct Q and V as follows: let $y_0 = e_1$, set $V_1 = y_0^\top = e_1^\top$; for $i = 1, \ldots, d_x - 1$, define

$$z_i = \begin{cases} u_i & \text{if } u_i \notin \text{span}(V_1^\top, \ldots, V_i^\top) \\ v \text{ s.t. } v \in \text{span}(V_1^\top, \ldots, V_i^\top)^\perp, \|v\| = 1 & \text{otherwise} \end{cases}$$

Let y_i be the component of z_i that is independent of $V_1^\top, \ldots, V_i^\top$,

$$y_i = \frac{z_i - \sum_{j=1}^{i} \Pi_{V_j^\top}(z_i) V_j^\top}{\|z_i - \sum_{j=1}^{i} \Pi_{V_j^\top}(z_i) V_j^\top\|},$$

where $\Pi_v(z)$ denotes the projection of z onto vector v. Set $Q_i = d_i y_i^\top$ for some $d_i \neq 0$ to be specified later, and set $V_{i+1} = y_i^\top$.

The next lemma justifies this iterative construction of V by showing that the trajectory $x_1, \ldots, x_t$ is not affected by the choice of V_i, Q_i for $i \ge t$. As a result, without loss of generality we can set V_t after obtaining x_t, and set Q_t after receiving u_t.

Lemma 28 *As long as V is orthogonal, the states satisfy $x_t = \sum_{i=1}^{t-1} c_i^t V_i^\top + c_t^t y_{t-1}$ for some constants c_i^t that only depend on $\mathcal{A}$ and $\{Q_i\}_{i=1}^{t-1}$.*

Proof We prove the lemma by induction. For our base case, x_1 is trivially $c_1^1 e_1$ and it is fixed for all choices of Q, V. Set $V_1 = e_1^\top$. Assume the lemma is true for x_t, and we have specified V_i for $i \le t$, Q_i for $i < t$. The specified rows of V are orthonormal by construction. Note that by our construction, x_t is obtained first, and then we set $V_t = y_{t-1}^\top$. Since u_t only depends on the current trajectory up to x_t, it is well-defined, and we can obtain z_t. By definition of y_t, we can write $u_t = \sum_{i=1}^t a_i^t V_i^\top + a_{t+1}^t y_t$ for some coefficients a_i^t. Set $Q_t = d_t y_t^\top$ as in the lemma. The next state is then

$$\begin{aligned}
x_{t+1} = Q^\top V x_t + u_t &= Q^\top V \sum_{i=1}^t c_i^t V_i^\top + \sum_{i=1}^t a_i^t V_i^\top + a_{t+1}^t y_t && V_t = y_{t-1}^\top \\
&= \sum_{i=1}^t c_i^t Q^\top e_i + \sum_{i=1}^t a_i^t V_i^\top + a_{t+1}^t y_t && V \text{ is orthogonal} \\
&= \sum_{i=1}^t c_i^t Q_i^\top + \sum_{i=1}^t a_i^t V_i^\top + a_{t+1}^t y_t \\
&= \sum_{i=1}^{t-1} c_i^t d_i V_{i+1}^\top + c_t^t d_t y_t + \sum_{i=1}^t a_i^t V_i^\top + a_{t+1}^t y_t && Q_i = d_i y_i^\top = d_i V_{i+1} \\
&= \sum_{i=1}^t c_i^{t+1} V_i^\top + c_t^t d_t y_t + a_{t+1}^t y_t && (4)
\end{aligned}$$

We have shown in the inductive step that x_{t+1} does not depend on the choice of V_{t+1} as long as V is orthogonal, hence we can set $V_{t+1} = y_t^\top$. Moreover, x_{t+1} is not affected by Q_i for $i \ge t+1$ by inspection. ■

The magnitude of the state. In this section we specify the constants d_i in the construction to ensure that the state has an exponentially increasing magnitude. Let $u_i = \sum_{j=1}^i a_j^i V_j^\top + a_{i+1}^i y_i$, $x_i = \sum_{j=1}^{i-1} c_j^i V_j^\top + c_i^i y_{i-1}$. Set $d_i = \operatorname{sign}(c_i^i)\operatorname{sign}(a_{i+1}^i) \cdot 2$. The quantities c_i^i and a_{i+1}^i are well-defined when we set Q_i after obtaining x_i and u_i. Intuitively, $Q^\top V$ applied to x_i aligns y_{i-1} to y_i, and we grow the magnitude of the y_i component in x_{t+1} multiplicatively.

Lemma 29 *The states satisfy* $x_t = \sum_{i=1}^t c_i^t V_i^\top$, *and* $|c_t^t| \ge 2|c_{t-1}^{t-1}|$.

Proof By equation 4 in Lemma 28, we can express $x_{t+1} = \sum_{i=1}^t c_i^{t+1} V_i^\top + c_t^t d_t y_t + a_{t+1}^t y_t$. As we claimed before, since x_{t+1} does not depend on the choice of V_{t+1}, we set $V_{t+1} = y_t$, and write $x_{t+1} = \sum_{i=1}^{t+1} c_i^{t+1} V_i^\top$. By our choice of d_t, we have

$$c_{t+1}^{t+1} = c_t^t d_t + a_{t+1}^t = \operatorname{sign}(c_t^t)\operatorname{sign}(a_{t+1}^t) \cdot 2c_t^t + a_{t+1}^t = \operatorname{sign}(a_{t+1}^t)(2|c_t^t| + |a_{t+1}^t|).$$

We conclude that $|c_{t+1}^{t+1}| = 2|c_t^t| + |a_{t+1}^t| \ge 2|c_t^t|$. ■

Observe that $x_1 = c_1^1 e_1$ where $|c_1^1| = 1$; therefore we have $\|x_{d_x}\| \ge |c_{d_x}^{d_x}| \ge 2^{d_x - 1}$.

Size of the system. Our construction only requires $Q_1, \ldots, Q_{d_x - 1}$ to be specified, and without loss of generality we take $Q_{d_x} = d_{d_x} V_1 = 2V_1$. By inspection, Q can be written as $Q = DPV$, where $D = Diag(d_1, d_2, \ldots, d_{d_x})$ and P is a permutation matrix that satisfies $(PV)_i = V_{i+1 \pmod{d_x}}$. Therefore the spectral norm of $Q^\top V$ is at most $\|Q\|\|V\| \leq 2$. We conclude that for this system, $\mathcal{L} = d_u + d_x + 7$, and the total cost is at least $2^{\Omega(\mathcal{L})}$.