Learning Controllers for Unstable Linear Quadratic Regulators from a Single Trajectory

Lenart Treven
ETH Zürich
trevenl@ethz.ch

Sebasitan Curi
ETH Zürich
scuri@inf.ethz.ch

Mojmir Mutny
ETH Zürich
mmutny@inf.ethz.ch

Andreas Krause
ETH Zürich
krausea@ethz.ch

Abstract

We present the first approach for learning – from a single trajectory – a linear quadratic regulator (LQR), even for unstable systems, without knowledge of the system dynamics and without requiring an initial stabilizing controller. Our central contribution is an efficient algorithm – *eXploration* – that quickly identifies a stabilizing controller. Our approach utilizes robust System Level Synthesis (SLS), and we prove that it succeeds in a constant number of iterations. Our approach can be used to initialize existing algorithms that require a stabilizing controller as input. When used in this way, it yields a method for learning LQRs from a single trajectory and even for unstable systems, while suffering at most $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

1 Introduction

Dynamical systems are ubiquitous in real world applications, ranging from autonomous robots (Ribeiro et al., 2017), energy systems (Haddad et al., 2005) to manufacturing (Singh, 2010). Control theory (Trentelman et al., 2001) seeks to find an optimal input to the system to ensure a desired behavior while suffering low cost. In particular, *linear* dynamical systems with quadratic costs can model a variety of practical problems (Tornambè et al., 1998), and enjoy an elegant solution referred to as *Linear Quadratic Regulator (LQR)*, whose history goes back to Kalman (1960).

Despite the long and rich history of the LQR problem, *learning* dynamical systems and their optimal controllers is still an actively studied problem. On one hand, there are systems that can be reset to an initial condition. For such systems, the multiple-trajectory (episodic) setting is natural and exploration costs can be controlled by resetting the system. This setting is well studied and efficient algorithms rely on *certainty equivalent control* (CEC) (Mania et al., 2019). On the other hands, other systems cannot be reset and must thus be learnt online from a *single* trajectory. In this setting, the OSLO algorithm (Cohen et al., 2019) is provably efficient. It is based on the optimism-in-the-face-of-uncertainty (OFU) principle, whereas epsilon-greedy (certainty equivalent with additive random noise) is also provably efficient (Simchowitz and Foster, 2020). Crucially, both algorithms require prior knowledge in form of an initial stabilizing controller. This privileged information is essential to ensure that unstable systems do not "explode". However, such prior knowledge is not always available.

Contributions In this work, we develop an approach for provably efficiently learning an LQR for potentially unstable systems from a *single trajectory* and *without* the knowledge of an initial stabilizing controller. Our central contribution is an initial *eXploration* phase that *does not* require a stabilizing controller. During this phase, we estimate the parameters of the system and stop once we have identified a controller that *provably* stabilizes the underlying system. Crucially, we prove that this phase ends after *constant* time, essentially adding no regret to the algorithm. Our algorithm can be used to initialize existing approaches (such as OSLO and CEC) that require a stabilizing controller, *without* introducing (more than constant) additional regret. Together, this yields the X-OSLO and X-CEC algorithms with provable $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. Our basic *eXploration* approach uses zero-mean Gaussian exploration which can cause instability (albeit only for provably constant rounds). We

Preprint. Under review.

also introduce heuristics for finding a stabilizing controller even during the unstable eXploration phase, which empirically reduces the length of the phase and *substantially* lowers the total regret. We demonstrate the practicality of our method on common benchmark problems from Dean et al. (2019).

1.1 Related Work

Linear dynamical systems have been extensively studied in control theory (cf., Zhou et al. (1996)). Here, we focus on the most closely related recent work on learning LQR controllers. Simchowitz and Foster (2020) establish fundamental limits that show that that the minimal regret *any* algorithm can attain is at least of order $\Omega(\sqrt{T})$. We seek algorithms that match this lower bound.

Multi-trajectory (Episodic) Setting Dean et al. (2019) present the Coarse-ID control algorithm that is provably efficient. Coarse-ID control explores episodically until it outputs a single controller, by solving a robust control synthesis problem, which provably stabilizes the underlying system. Unfortunately, Coarse-ID control has sub-optimal $\widetilde{\mathcal{O}}(T^{2/3})$ regret (Dean et al., 2018). The sub-optimality arises from the fact that the output controller is being robust (or pessimistic). Mania et al. (2019) prove that using Certainty Equivalent Control (CEC) with epsilon-greedy exploration during each episode is enough to prove an optimal $\widetilde{\mathcal{O}}(T^{1/2})$ regret bound.

Single-trajectory (Online) Setting Some systems cannot be reset and the episodic setting is inappropriate to model them. For this more challenging setting, Abbasi-Yadkori and Szepesvari (2011) propose an algorithm based on the OFU principle that has provable $\widetilde{\mathcal{O}}(T^{1/2})$ regret, but it requires to solve at every time-step a *non-convex* optimization problem. Cohen et al. (2019) overcome this drawback by relaxing the non-convex optimization problem into an efficient semi-definite program (SDP), yielding the OSLO method. Perhaps surprisingly, Simchowitz and Foster (2020) prove that CEC with epsilon-greedy exploration is also optimal in this setting. Crucially, all works in this setting require the knowledge of an *initial stabilizing controller*, which might not always be available.

Online Learning of Unstable Systems Recent results suggest that detecting unstable modes of the system is *easier* than learning stable ones (Simchowitz et al., 2018), hence suggest that learning general systems with possibly unstable modes is possible. In particular, Sarkar and Rakhlin (2019) propose to use an ordinary least squares (OLS) estimator for the parameters of general linear dynamical systems and present a non-asymptotic analysis for the estimation error. In this work, we use regularized least squares (RLS) estimates and refine their bounds to build data-dependent confidence intervals on the parameter errors. We use these bounds together with robust control synthesis in a similar spirit of Coarse-ID (Dean et al., 2019). Our main contribution is an algorithm that outputs a provably robust controller for general systems in constant time, adding no regret to the CEC and OSLO algorithms described previously.

2 Problem Statement and Background

We consider a system evolving with the following linear dynamics

$$x_{i+1} = A_* x_i + B_* u_i + w_{i+1}, \quad x_0 = 0, \tag{1}$$

where $x_i \in \mathbb{R}^d$ are states, $u_i \in \mathbb{R}^k$ actions, $(w_i)_{i \geq 1} \overset{i.i.d.}{\sim} \mathrm{subG}_d(\sigma^2)$ unobserved non degenerated sub-Gaussian noise in $\mathbb{R}^d$ with unknown variance proxy σ^2. The matrices $A_* \in \mathbb{R}^{d \times d}, B_* \in \mathbb{R}^{d \times k}$ are unknown transition matrices. We further assume that the system is *stabilizable*, which means that there exists a matrix $K \in \mathbb{R}^{k \times d}$ for which it holds that the spectral radius $\rho(A_* + B_* K) < 1$.

At every step i, the system incurs a cost c_i given by,

$$c_i = x_i^\top Q x_i + u_i^\top R u_i, \tag{2}$$

where $Q \in \mathbb{R}^{d \times d}, R \in \mathbb{R}^{k \times k}$ are known positive definite matrices. The objective of the learner is to minimize the *expected infinite horizon cost* $\lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}[c_i]$.

When the system matrices A_*, B_* are known, the optimal solution is given by the fixed map $u_i = K_* x_i$ and the optimal cost is J_* (Bertsekas, 2000). Hereby, $K_* = -(R + B_*^\top P B_*)^{-1} B_*^\top P A_*$, where P is the solution to *discrete algebraic Ricatti equation* of the system, $P = DARE(A_*, B_*, Q, R)$. When the matrices A_*, B_* are *unknown*, the learner can use all the information up to time i to play an

action u_i. We call a map π from all the observed past states $(x_j)_{j=0}^{i}$ and actions $(u_j)_{j=0}^{i-1}$ to an action u_i a *policy*. For such a policy, we define its *regret* $R_\pi(T)$, where T is the number of steps taken, as

$$R_\pi(T) = \sum_{i=1}^{T} (c_i - J^*), \tag{3}$$

where J_* is the optimal cost, and c_i is the cost incurred by π at time i. We seek a policy with *sublinear regret* with high probability.

2.1 Background on System Level Synthesis (SLS)

In this section, we summarize the SLS framework (Wang et al., 2019). We use SLS to find a linear controller K, i.e. $u_i = Kx_i$, that *provably* stabilizes the true underlying system.

Assume that we have estimates $(\widehat{A}, \widehat{B})$. Dean et al. (2019) show that a controller K stabilizes *all* systems (A, B) with $\|\widehat{A} - A\| \le \varepsilon_A$ and $\|\widehat{B} - B\| \le \varepsilon_B$ if :

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K \end{pmatrix} \left(zI - \widehat{A} - \widehat{B}K \right)^{-1} \right\|_{\mathcal{H}_\infty} < 1. \tag{4}$$

The $\mathcal{H}_\infty$-norm for a function $f : \mathbb{C} \to \mathbb{C}^{d \times d}$ is defined as $\|f\|_{\mathcal{H}_\infty} = \sup_{z \in \partial\mathbb{D}} \|f(z)\|$, where $\mathbb{D} = \{z \in \mathbb{C} | \, \|z\| < 1\}$ is a unit disk in the complex plane. Dean et al. (2019) further prove that condition (4) is equivalent to the feasibility of the following SDP:

$$\begin{aligned} \min_{s \in [0,1), X \succ 0, Z} \; & s \\ \text{s.t.} \; & \begin{pmatrix} X - I & \widehat{A}X + \widehat{B}Z & 0 & 0 \\ X\widehat{A}^\top + Z^\top \widehat{B}^\top & X & \epsilon_A X & \epsilon_B Z^\top \\ 0 & \epsilon_A X & \frac{1}{2}sI & 0 \\ 0 & \epsilon_B Z & 0 & \frac{1}{2}sI \end{pmatrix} \succcurlyeq 0. \end{aligned} \tag{5}$$

The *robust* policy is then extracted from optimal solution of SDP (5) as $K = ZX^{-1}$. The convex SDP (5) can be solved efficiently (e.g., using MOSEK (ApS, 2020)).

2.2 Background on online LQR Algorithms

In this section, we review the two optimal algorithms for online LQR, namely OSLO (Cohen et al., 2019) and CEC (Algorithm 1 of Simchowitz and Foster (2020)). Both algorithms require as input an initial stabilizing controller K_0. Then they split their algorithms into two phases: a warm-up (*safe exploration*), and an *exploitation* phase.

The safe exploration phase is equivalent in both algorithms. The learner explores with actions $u_i \sim \mathcal{N}(K_0 x_i, \Sigma)$, where Σ is an appropriately chosen covariance matrix as shown in appendices C and D. It then uses RLS to refine the estimates $(\widehat{A}, \widehat{B})$ of the true system. This phase is crucial to bound the closed-loop dynamics that arise during exploitation.

The main difference between both algorithm comes in the final phase. OSLO solves an *optimistic* SDP to carefully balance exploration and exploitation. CEC, instead, greedily exploits the current estimates and injects random noise that decays as $O(1/\sqrt{i})$, to balance exploration and exploitation. As CEC does not need to solve an optimistic problem, it attains better regret bounds in terms of dimensions and the optimization problem enjoys a closed form solution. For general systems, such naive greedy exploitation is inefficient, but for the particular structure of LQR, it is optimal.

3 Efficiently Learning a Stabilizing Controller with Unstable Exploration

We now show how to use the SLS approach (c.f., section 2.1) to *provably* find a robust controller. In section 3.2, we show how to combine this initial eXploration phase with the OSLO (Cohen et al., 2019) or the CEC algorithms (Simchowitz and Foster, 2020), transforming them into the X-OSLO, X-CEC algorithm respectively.

3.1 Efficient eXploration for identifying a stabilizing controller

We now present an algorithm that finds a controller which provably stabilizes the true system in constant time. The learner explores using independent Gaussian noise $u_i \sim \mathcal{N}(0, \sigma_u^2 I)$ and uses a RLS estimator to estimate the (A_*, B_*) parameters of the dynamical system, together with high-probability confidence bounds on these parameters. Algorithm 1 ends when we can find a feasible solution to a robust controller synthesis problem and returns a controller K_0. Using section 2.1, we can prove, with high probability, that the controller K_0 stabilizes the *true* underlying system, i.e., $\rho(A_* + B_* K_0) < 1$. Next, we prove that *eXploration*, shown in Algorithm 1, finishes in $\widetilde{O}(1)$ time.

Algorithm 1 eXploration

1: **Input:** $x_0 = 0, \vartheta$, s.t. $\|(A_* \; B_*)\| \leq \vartheta$
2: **for** $s = 1, \ldots$ **do**
3: Play $u_{s-1} \sim \mathcal{N}(0, \sigma_u^2 I)$ and observe state x_s
4: Let (A_s, B_s) be a minimizer of $\operatorname{argmin}_{A,B} \sum_{i<s} \|x_{i+1} - (A \; B) \begin{pmatrix} x_i \\ u_i \end{pmatrix} \|^2 + \lambda \|(A \; B)\|_F^2$
5: Solve SDP (5) with $(\widehat{A}, \widehat{B}) = (A_s, B_s)$ and $\varepsilon_A, \varepsilon_B$ from Corollary 2 (use ϑ in upper bound).
6: **if** SDP is feasible **return** associated controller K_0
7: **end for**

The first step towards proving that Algorithm 1 finishes in constant time is to show that the estimators A_s, B_s in Algorithm 1 converge towards A_*, B_*. To identify the system matrices A_*, B_* we consider the RLS estimator, defined as:

$$A_s, B_s = \operatorname*{argmin}_{A,B} \sum_{i=0}^{s-1} \left\| x_{i+1} - (A \; B) \begin{pmatrix} x_i \\ u_i \end{pmatrix} \right\|^2 + \lambda \|(A \; B)\|_F^2 . \tag{6}$$

Sarkar and Rakhlin (2019) show that as long as matrix A_* is *regular* (i.e., all its eigenvalues with absolute value larger than 1 have geometric multiplicity 1) and the actions taken are independent and non degenerated $u_i \sim \mathrm{subG}_k(\sigma_u^2)$, the OLS estimator is consistent. In appendix A, we adapt the analysis of Sarkar and Rakhlin (2019) for OLS to the RLS estimator that we use, yielding:

Corollary 1 (based on Theorem 2 of Sarkar and Rakhlin (2019))**.** *Let us be in the setting of* (1) *and suppose that* $(u_i)_{i \geq 0} \overset{i.i.d.}{\sim} \mathrm{subG}_k(\sigma_u^2)$ *are non-degenerate independent of* $(w_i)_{i \geq 1}$*. Further assume that matrix* A_* *is regular. Then, with probability at least* $1 - \delta$ *for the RLS estimators* (6) *it holds:*

$$\max\left(\|A_s - A_*\|, \|B_s - B_*\|\right) \leq \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}},$$

whenever $s \geq poly(\log \frac{1}{\delta})$.

Improved upper bound The upper bound of Corollary 1 on the estimation error is more of theoretical interest, and rather loose in practice. While running Algorithm 1, we prefer an upper bound that is as tight as possible so that we can stop early. Here, we present an improved, data-dependent upper bound on the estimation errors $\|A_s - A_*\|, \|B_s - B_*\|$ that we can use in Algorithm 1. The key insight is that after running the algorithm for s steps, we – in addition to the knowledge of actions $(u_i)_{i \leq s}$ – observe the states $(x_i)_{i \leq s}$ and can also leverage them. For the RLS estimator given in (6), we derive in appendix A that:

$$((A_s \; B_s) - (A_* \; B_*))^\top = (V_s + \lambda I)^{-1} S_s - \lambda (V_s + \lambda I)^{-1} (A_* \; B_*)^\top, \tag{7}$$

where $V_s = \sum_{i=0}^{s-1} z_i z_i^\top$ and $S_s = \sum_{i=0}^{s-1} z_i w_{i+1}^\top$, with $z_i = (x_i^\top u_i^\top)^\top$. Since we know an upper bound ϑ for $\|(A_* \; B_*)\| \leq \vartheta$ and we observe V_s, the only term that we still have to deal with, in order to upper bound the estimation error, is S_s. In the following, we will show an upper bound for $\left\|(V_s + \lambda I)^{-\frac{1}{2}} S_s\right\|$. The result builds on ideas from Abbasi-Yadkori et al. (2011) and Sarkar and Rakhlin (2019).

Proposition 1. *In the aforementioned setting let* $\varepsilon \in (0, 1)$ *arbitrary. Then it holds w.p. at least* $1 - \delta$*:*

$$\forall s \geq 0 : \left\|(V_s + \lambda I)^{-\frac{1}{2}} S_s\right\|^2 \leq \frac{2R^2}{(1-\epsilon)^2} \log\left(\frac{\det(V_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\epsilon}\right)^d}{\delta}\right)$$

The proof can be found in appendix A. Collecting the results together, we arrive at an upper bound that we apply in the algorithm.

Corollary 2. *For the RLS estimates from Algorithm 1 it holds w.p. at least* $1-\delta$ *for every* $s \geq 0$*:*

$$\begin{aligned}\|A_s - A_*\| &\leq \frac{R}{1-\epsilon}\sqrt{2\log\left(\frac{\det(V_s+\lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}}\frac{\left(1+\frac{2}{\epsilon}\right)^d}{\delta}\right)}\left\|(I_d\ 0)(V_s+\lambda I)^{-1/2}\right\| \\ &\quad + \lambda\left\|(I_d\ 0)(V_s+\lambda I)^{-1}\right\|\vartheta \\ \|B_s - B_*\| &\leq \frac{R}{1-\epsilon}\sqrt{2\log\left(\frac{\det(V_s+\lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}}\frac{\left(1+\frac{2}{\epsilon}\right)^d}{\delta}\right)}\left\|(0\ I_k)(V_s+\lambda I)^{-1/2}\right\| \\ &\quad + \lambda\left\|(0\ I_k)(V_s+\lambda I)^{-1}\right\|\vartheta\end{aligned}$$

Since the upper bound in Corollary 2 holds for every $\varepsilon \in (0,1)$, we optimize over ε to obtain the best possible bound while running the algorithm.

Feasibility of SLS synthesis The next building block is to find a condition under which the SDP (5) becomes feasible. For large $\varepsilon_A, \varepsilon_B$ the problem (5) is usually infeasible. Nevertheless, for small ε_A and ε_B, we can guarantee that the SLS synthesis (5) is feasible.

Lemma 1. *Let* K *be* any *controller for which it holds that* $\rho(A_* + B_*K) \leq \gamma_0 < 1$*. If*

$$\max(\varepsilon_A, \varepsilon_B) \leq \frac{1}{3(1+\|K\|)C(\gamma_0, A_*, B_*, K)},$$

then Problem (5) *is feasible.*

Here, $C(\gamma_0, A_*, B_*, K)$ is a constant that depends only on γ_0 and norms of matrices A_*, B_* and K. The explicit relation together with the proof of the Lemma is given in appendix B.

Main Result We combine our results on SLS synthesis in lemma 1 with the tighter upper bounds of corollary 1 to obtain the bound on the maximum number of steps taken by algorithm 1.

Theorem 1. *Let* $\delta, \varepsilon \in (0,1)$ *and* K any *controller which stabilizes the underlying system* A_*, B_* *with* $\rho(A_* + B_*K) \leq \gamma_0 < 1$*. If we play actions* $u_i \sim \mathcal{N}(0, \sigma_u^2 I)$*), Algorithm 1 terminates with probability at least* $1-\delta$ *in at most* $\mathcal{O}((poly(\log\frac{1}{\delta})(1+\|K\|)C(\gamma_0, A_*, B_*, K))^{2+\epsilon})$ *steps.*

3.2 X-OSLO and X-CEC

In this section, we extend both OSLO and CEC to the case where an initial stabilizing controller is not available, by initializing them with our novel eXploration algorithm 1. We call the resulting algorithms X-OSLO, X-CEC, respectively. We prove that both algorithms still achieve regret $\widetilde{\mathcal{O}}(\sqrt{T})$.

Instead of a stabilizing controller, X-OSLO and X-CEC only require two scalar parameters ϑ, with $\|(A_*, B_*)\| \leq \vartheta$, and γ_0, such that there exists a controller K with $\rho(A_* + B_*K) \leq \gamma_0 < 1$. We assume that the system is stabilizable, hence γ_0 exists. The parameter ϑ can be a very loose bound at the beginning, since we can update it while we run the algorithm via $\vartheta_{i+1} = \min(\vartheta_i, \|\widehat{A}_i\| + a_i + \vartheta_i/b_i)$, where usually a_i is of order $\frac{1}{\sqrt{i}}$ and b_i grows linearly. Both algorithms further require time horizon T and cost matrices Q, R which the control designer provides.

The general structure of X-OSLO and X-CEC is the following. *Phase I*: run eXploration until we find a controller K_0 which provably stabilizes the true system. *Phase II*: play $u_i \sim \mathcal{N}(K_0 x_i, \Sigma)$ to tighten the bounds until we can prove that exploiting yields $\mathcal{O}(\sqrt{T})$ regret. *Phase III*: exploit the knowledge of the tight estimates and play either optimistically (X-OSLO) or greedily (X-CEC).

X-CEC Algorithm: We do not change the implementation of Phase II and III of the original CEC algorithm. Namely, in Phase II we use the policy $u_i \sim \mathcal{N}(K_0 x_i, I)$ and in Phase III the certainty equivalent controller together with epsilon greedy noise scaled as $\mathcal{O}(1/\sqrt{i})$.

X-OSLO Algorithm: We modify the analysis of the warm-up phase of OSLO (Phase II). Cohen et al. (2019) require the *strong stability* parameters of the *unknown* closed loop matrix $A_* + B_*K_0$ to prove that this phase terminates in $\mathcal{O}(\sqrt{T})$ time. Instead of the required *strong stability* parameters,

we leverage the knowledge of $\rho(A_* + B_* K_0) < 1$. In appendix C we prove that this is sufficient to *provably* terminate this phase in $\mathcal{O}(\sqrt{T})$. Practically, we can also terminate Phase II before by evaluating the upper bounds from corollary 2. Once we establish that $\|\widehat{A} - A_*\|^2, \|\widehat{B} - B_*\|^2 \leq \mathcal{O}(1/\sqrt{T})$ we move on to OSLO's Phase III using an optimistic strategy.

We now state the theorem which bounds the total regret of the X-OSLO algorithm. In appendix D we provide an analogous theorem for X-CEC.

Theorem 2. *Suppose A_* is regular and that there exists K with $\rho(A_* + B_* K) \leq \gamma_0 < 1$. Suppose that for known matrices Q, R there exists α_0 such that $0 \preceq \alpha_0 I \preceq Q, R$. Further assume that $T \geq poly\left(n, \vartheta, \alpha_0^{-1}, \sigma^{-1}, (1-\gamma_0)^{-1}, \log \frac{1}{\delta}, \|A_*\|, \|B_*\|, \|K\|\right)$. Then the total regret of the* X-OSLO *algorithm is, with high probability, bounded by*

$$R_T = \mathcal{O}\left(\sqrt{T} \log^2 T\right).$$

Proof sketch. By theorem 1, we know that after a constant (in T) number of steps the eXploration algorithm will yield a stabilizing controller. Hence the regret of the first phase is constant. Next, while running the warm-up phase for $\mathcal{O}(\sqrt{T} \log T)$, the squares of states norm are bounded by $\mathcal{O}(\log T)$, X-OSLO incurs regret of $\mathcal{O}(\sqrt{T} \log^2 T)$ and we obtain $\widehat{A}, \widehat{B}$ with $\max(\|\widehat{A} - A_*\|^2, \|\widehat{B} - B_*\|^2) \leq \mathcal{O}(1/\sqrt{T})$. With the obtained estimates, in the final exploitation phase we run the OSLO algorithm until the end. This last phase again suffers $\mathcal{O}(\sqrt{T} \log^2 T)$ regret. Hence the total regret is of order $\mathcal{O}(\sqrt{T} \log^2 T)$. □

Stable Systems If the *true* underlying system is stable, then the eXploration phase still terminates in constant time and the results in this section still hold. Practically, this means that X-OSLO and X-CEC will have two (safe) warm-up phases with possibly different stabilizing controllers.

4 Improved eXploration policies

The basic eXploration approach (Phase I of X-OSLO and X-CEC) takes random actions $u_i \sim \mathcal{N}(0, \sigma_u^2 I)$. For this choice we can guarantee that Phase I terminates after constant time, depending solely on the system parameters. However, as we demonstrate in our experiments (c.f., section 5), the states can grow *exponentially* during this phase, which can be highly problematic for certain applications. We now propose improved, *data-dependent* policies to counteract this blow-up.

In particular, we consider playing $u_i \sim \mathcal{N}(K_i x_i, \sigma_u^2 I)$, where K_i is a controller picked at time i. With such a controller, we generally lose the theoretical guarantee that the Phase I will end. However, the upper bounds on estimation errors from Corollary 2 (and thus the validity of the stopping condition) still hold and we can run Algorithm 1. Here, we discuss different choices for controller K_i that we study in our experiments.

As first possibility, we could act as if the estimators A_i, B_i are the true system matrices and we compute the controller K_i as the optimal controller: $K_i = -(R + B_i^\top P B_i)^{-1} B_i^\top P A_i$, where $P = DARE(A_i, B_i, Q, R)$, i.e., we act using *certainty equivalent control*.

As second alternative, we could use *robust control*. In particular, we start with error estimates $\varepsilon_A, \varepsilon_B$ given by Corollary 2 and successively half them until the SDP (5) becomes feasible. For controller K_i, we then take the resulting controller. Since $\varepsilon_A, \varepsilon_B$ are upper bounds on estimation errors, we expect that the robust controller will stabilize the system much earlier than we have a theoretical guarantee for that. The latter expectation is indeed supported by our experiments.

For third option, we could again act as if the estimators A_i, B_i are the true system matrices and we compute the controller K_i as the controller: $K_i = (R + B_i^\top P B_i)^{-1} B_i^\top P A_i$, where $P = DARE(A_i, B_i, Q, R)$, i.e., we act using *negative certainty equivalent control* (NegCEC). Sarkar and Rakhlin (2019) show that identification of purely explosive systems happens with exponential speed, hence we expect that with this controller Phase I ends fast. However, since we destabilize the system, we also expect that the norm of the state will grow fast.

Lastly, the last two controllers are a mixture of the NegCEC and either CEC or robust controller. For the *mixed* controller we use the following heuristic: we choose for K_i NegCEC whenever $\|x_i\| \leq M$,

where M is a predefined margin, and switch to using for K_i either CEC or *robust controller* when $\|x_i\| > M$.

In order to obtain a guarantee that the Phase I ends when we use a non-trivial K_i, we restrict ourselves to the case when matrix B_* is known and has a full row rank. In this case, we can without loss of generality assume that B_* is the identity, and the learner only needs to learn matrix A_*. This setting is, in the one dimensional case, discussed by Rantzer (2018). They show that as long as actions u_i are measurable functions of the past (any controller K_i satisfies this) it holds for OLS estimator A_s that $\|A_s - A_*\| \leq \mathcal{O}(1/\sqrt{s})$. A natural question that arises then is whether one can obtain estimation error of $\mathcal{O}(1/\sqrt{s})$ for arbitrary measurable actions of the past also for the case of state dimension d with $d \geq 2$. As we show in appendix E, when $d \geq 2$, perhaps surprisingly, there exists a controller for which the OLS estimator is *not consistent*. The intuition behind this lies in the fact that in the one-dimensional case the smallest singular value of the empirical covariance matrix $\sum_i x_i x_i^\top$ is equal to the largest one, while in case $d \geq 2$ this does not hold, and the estimation procedure might not be consistent anymore. However, we can still prove convergence under some additional assumptions, as stated in the next theorem.

Theorem 3. *Let $x_{i+1} = A_* x_i + u_i + w_{i+1}$, $x_0 = 0$, where $x_i \in \mathbb{R}^d$, $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$ and actions $u_i = K_i x_i$ are chosen in such a way that for every time i it holds: $\|x_i\| \leq M_1$ and $\|A_* + K_i\| \leq M_2$ for some constants M_1, M_2, which do not depend on time. Then at every time s it holds for the RLS estimator A_s of the matrix A_* with probability at least $1 - \delta$:*

$$\|A_s - A_*\| \leq \frac{\mathcal{O}(1)(d \log s + \log \frac{1}{\delta})}{\sqrt{s}}$$

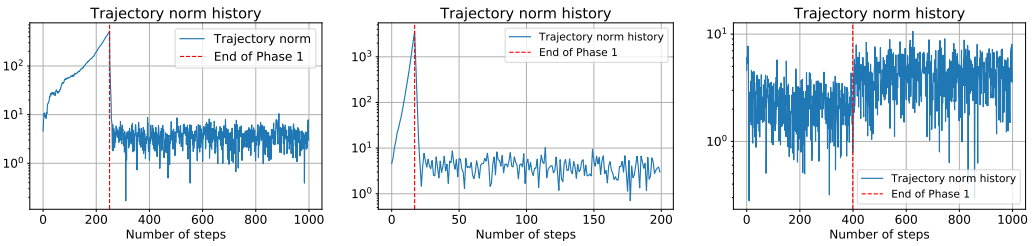
The proof of the theorem and the constants that hide in $\mathcal{O}(1)$ are given in appendix E.

5 Experiments

To evaluate our eXploration approach, we run the X-OSLO for the time it needs to finish the Phase I (note that this first phase is identical for X-CEC as well). We try to understand how fast we can obtain a stabilizing controller, and how much regret we suffer until that happens. Due to very similar performance of CEC and robust controller we show the result here only for robust controller, the results for CEC we present in appendix F. In appendix F we test X-OSLO also on different dynamical systems. Here we run X-OSLO on the following dynamical system:

$$A_* = \begin{pmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{pmatrix}, \quad B_* = I, \quad Q = I, \quad R = I, \tag{8}$$

introduced by Dean et al. (2019) and assume both A_* and B_* are unknown. The noise follows $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ and actions $(u_i)_{i\geq 0} \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$.



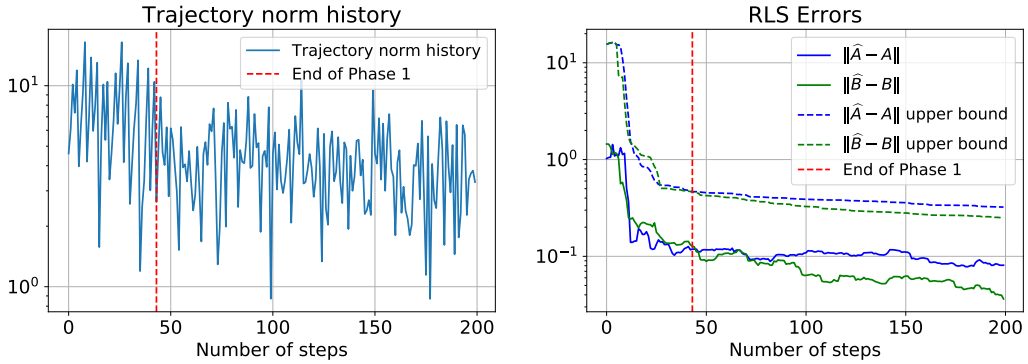
(a) Zero mean Gaussian actions (b) NegCEC (c) Robust control

Figure 1: Comparison of state magnitude for different controllers used in Phase I (eXploration). NegCEC leads to state blow-up, but quickly ends Phase I. In contrast, the robust controller avoids blow-up at the cost of delayed convergence.

Since two eigenvalues of the matrix A are larger than 1, the norm of the state $\|x_i\|$ is exponentially increasing, which is expected, since the input actions are zero mean Gaussian. However by Theorem 1 we know that after constantly many steps, we will find a stabilizing controller as we see in Figure 1a.

Improved action selection during Phase I As discussed in section 4, instead of playing standard Gaussian actions, at each step, we can select an action u_i as $u_i \sim \mathcal{N}(K_i x_i, I)$, where K_i is a controller picked at step i. In fig. 1b and fig. 1c we present the case where we choose for K_i the *NegCEC* and *robust* controller respectively.

We empirically observe that when we use the *robust controller*, the states do not blow up. However it usually takes more time to find a provably stabilizing controller. At the same time, we observe that using the *NegCEC*, the norm of the state increases faster, but the stabilizing controller is also found much faster. In fig. 2, we test the behavior of the *mixed* controller between NegCEC and robust controller from section 4 with margin $M = 10$ on system (8).



(a) Trajectory norm

(b) Estimation errors

Figure 2: (a) Our mixed controller achieves the best of both worlds: fast termination of Phase I, while successfully avoiding state blowup. (b) Our data-dependent bounds overestimate the true estimation errors, but qualitatively well capture their decay.

As shown in fig. 2 and further experiments in appendix F, our *mixed* controller bounds the state norms similarly to the *robust* controller, and the time it takes to find a stabilizing controller is comparable to that of *NegCEC*.

Lastly, table 1 compares the number of steps taken (length of Phase I) and cost suffered to find a stabilizing controller, between multi- (episodic) and single trajectory (online) exploration. In the multi-trajectory setting, we reset the system after every *rollout length* number of steps and use all collected data to find the RLS estimates. Beyond allowing resets, we additionally provide as bounds on the estimation error the *true* errors (not available in practice). We ran every approach 20 times and present the average number of steps taken and incurred cost. We see that except for the NegCEC, the incurred costs are comparable between the multi- and single trajectory setting. We expect that looser bounds on the estimation errors would result in a larger difference between the cost with zero mean actions ($K_i = 0$) and multi-trajectory setting. This is because the states would grow exponentially for longer time in the single trajectory setting compared to the multi-trajectory setting, where we ensure the boundedness of the states by resetting the system. We also observe that with looser estimation bounds, the *robust* controller needs longer to finish Phase I, while such looser bounds do not strongly influence neither the number of steps taken nor the size of the states for the *mixed* controller.

Table 1: Comparison of incurred cost and steps taken until we find a stabilizing controller

	Rollout length	Steps taken	Cost
Multi traj.	6	60	1474
	10	62	2231
	15	64	3451
	20	78	5874

	Controller	Steps taken	Cost
Single traj.	$K_i = 0$	31	4558
	NegCEC	13	800435
	Robust	20	2471
	Mixed, $M = 10$	13	1979

6 Conclusions

We presented the first approach for the fundamental problem of learning an LQR controller, even for unstable systems and from a single trajectory. In particular, we analyzed an algorithm – *eXploration* – that provably learns a stabilizing controller from a single trajectory, and can be used to initialize OSLO and CEC, yielding an algorithm with optimal regret. Beyond theoretically analyzing Gaussian eXploration, we tested different controllers in order to speed up the system identification and bound the state blow up.

Broader Impact

Currently, there is a lack of provably efficient *and* practical algorithms for real-world RL problems, limiting their applicability to promising applications such as personal robotics, efficient transportation, control of renewable energy systems etc. While restricted to linear quadratic regulators, our paper makes important contributions in this regard.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2011). Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems. *arXiv e-prints*, page arXiv:1102.2670.

Abbasi-Yadkori, Y. and Szepesvari, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *COLT*.

ApS, M. (2020). *MOSEK Optimizer API for Python 9.2.4*.

Bertsekas, D. P. (2000). *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition.

Cohen, A., Hassidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. (2018). Online Linear Quadratic Control. *arXiv e-prints*, page arXiv:1806.07104.

Cohen, A., Koren, T., and Mansour, Y. (2019). Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In Chaudhuri, K. and Salakhutdinov, R., 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. PMLR.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 4188–4197. Curran Associates, Inc.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2019). On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*.

Dowler, D. A. (2013). Bounding the Norm of Matrix Powers. *Theses and Dissertations. 3692.*

Gil', M. (2014). A new identity for resolvents of matrices. *Linear and Multilinear Algebra*, 62(6):715–720.

Haddad, W. M., Chellaboina, V., and Nersesov, S. G. (2005). *Thermodynamics: A Dynamical Systems Approach*. Princeton University Press.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:6 pp.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Mania, H., Tu, S., and Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 10154–10164. Curran Associates, Inc.

Nielsen, B. (2008). Singular vector autoregressions with deterministic terms: Strong consistency and lag order determination.

Phillips, P. C. and Magdalinos, T. (2013). Inconsistent var regression with common explosive roots. *Econometric Theory*, 29(4):808–837.

Rantzer, A. (2018). Concentration bounds for single parameter adaptive control. *2018 Annual American Control Conference (ACC)*, pages 1862–1866.

Ribeiro, F., Lopes, G., Maia, T., Ribeiro, H., Osório, P., Roriz, R., and Ferreira, N. (2017). Motion control of mobile autonomous robots using non-linear dynamical systems approach. In Garrido, P., Soares, F., and Moreira, A. P., editors, *CONTROLO 2016*, pages 409–421, Cham. Springer International Publishing.

Sarkar, T. and Rakhlin, A. (2019). Near optimal finite time identification of arbitrary linear dynamical systems. In Chaudhuri, K. and Salakhutdinov, R., 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. PMLR.

Simchowitz, M. and Foster, D. J. (2020). Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR.

Singh, T. (2010). *Optimal reference shaping for dynamical systems: theory and applications*. CRC Press, Boca Raton.

Tornambè, A., Conte, G., and Perdon, A. (1998). *Theory and Practice of Control and Systems: Proceedings of the 6th IEEE Mediterranean Conference, Alghero, Sardinia, Italy, 9-11 June 1998*. World Scientific.

Trentelman, H., Stoorvogel, A., and Hautus, M. (2001). *Control Theory for Linear Systems*. Communications and Control Engineering. Springer London.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv e-prints*, page arXiv:1011.3027.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, Y.-S., Matni, N., and Doyle, J. C. (2019). A system-level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 64(10):4079–4093.

Zhou, K., Doyle, J. C., and Glover, K. (1996). *Robust and Optimal Control*. Prentice-Hall, Inc., USA.

Contents of Appendix

A System identification

In this section we derive the difference between RLS estimator and the true system given in eq. (7), show how we can use the analysis of Sarkar and Rakhlin (2019) to prove the RLS convergence and derive the data-dependent upper bounds, which we use while running algorithm eXploration.

A.1 RLS estimator

Let us first derive the eq. (7). From the definition of RLS in eq. (6) follows that A_s, B_s minimize the expression

$$\|(x_1 \quad \dots \quad x_s) - (A\ B)\,(z_0 \quad \dots \quad z_{s-1})\|_F^2 + \lambda\,\|(A\ B)\|_F^2 \tag{9}$$

in variables A, B. Deriving eq. (9) with respect to $(A\ B)$ and setting the derivative to zero we obtain:

$$(A_s\ B_s)^\top = \left(\sum_{i=0}^{s-1} z_i z_i^\top + \lambda I\right)^{-1} \left(\sum_{i=0}^{s-1} z_i x_{i+1}^\top\right).$$

Using the relation $x_{i+1} = (A_*\ B_*)z_i + w_{i+1}$ we obtain:

$$\begin{aligned}
(A_s\ B_s)^\top &= \left(\sum_{i=0}^{s-1} z_i z_i^\top + \lambda I\right)^{-1} \left(\sum_{i=0}^{s-1} z_i z_i^\top (A_*\ B_*)^\top + z_i w_{i+1}^\top\right) \\
&= \left(\sum_{i=0}^{s-1} z_i z_i^\top + \lambda I\right)^{-1} \left(\sum_{i=0}^{s-1} \left(z_i z_i^\top + \lambda I\right)(A_*\ B_*)^\top - \lambda (A_*\ B_*)^\top + z_i w_{i+1}^\top\right) \\
&= (A_*\ B_*)^\top + (V_s + \lambda I)^{-1} S_s - \lambda\,(V_s + \lambda I)^{-1}\,(A_*\ B_*)^\top,
\end{aligned}$$

which yields eq. (7).

A.2 Proof of Corollary 1

Proof of Corollary 1. Rewrite eq. (1) as:

$$\begin{pmatrix} x_{i+1} \\ u_{i+1} \end{pmatrix} = \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_i \\ u_i \end{pmatrix} + \begin{pmatrix} w_{i+1} \\ u_{i+1} \end{pmatrix}. \tag{10}$$

Since $v_{i+1} := \begin{pmatrix} w_{i+1} \\ u_{i+1} \end{pmatrix}$ are $\mathrm{subG}_{d+k}(\sigma_*^2)$, where $\sigma_* = \max(\sigma, \sigma_u)$, and since by further denoting $A = \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}$ the eq. (10) can be rewritten as $z_{i+1} = Az_i + v_{i+1}$, we are in the setting analyzed by Sarkar and Rakhlin (2019). Since matrix A_* is regular, also matrix $\begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}$ is regular since the eigenvalues of matrix A are $\{\text{eigenvalues of matrix } A_*\} \cup \{k \text{ zero eigenvalues}\}$. Further denote by $S'_s = \sum_{i=0}^{s-1} z_i v_{i+i}^\top$. From the proof of Theorem 2 in their paper follows that $\left\|V_s^{-1/2} S'_s\right\| \le poly(\log s, \log \frac{1}{\delta})$ and $V_s \succeq \Omega(s) I$ (the latter can be observed from eq. (116) in (Sarkar and Rakhlin, 2019)). Since $\left\|(V_s + \lambda I)^{-1/2} S'_s\right\| \le \left\|V_s^{-1/2} S'_s\right\|$ and $V_s + \lambda I \succeq V_s \succeq \Omega(s) I$ we obtain:

$$\begin{aligned}
\|(A_s\ B_s) - (A_*\ B_*)\| &= \left\|(I_d \quad 0)\left(\begin{pmatrix} A_s & B_s \\ * & * \end{pmatrix} - \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}\right)\right\| \le \left\|\begin{pmatrix} A_s & B_s \\ * & * \end{pmatrix} - \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}\right\| \\
&\le \left\|(V_s + \lambda I)^{-1/2}\right\| \left\|(V_s + \lambda I)^{-1/2} S'_s\right\| + \lambda \left\|(V_s + \lambda I)^{-1}\right\| \left\|\begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}\right\| \\
&\le \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}} + \lambda \vartheta \mathcal{O}(\frac{1}{\sqrt{s}}) = \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}}.
\end{aligned}$$

We finish the proof by observation $\max\left(\|A_s - A_*\|, \|B_s - B_*\|\right) \le \|(A_s\ B_s) - (A_*\ B_*)\|$. □

A.3 Data-dependent estimation error upper bound

In this section we prove the results which lead to the data-dependent upper bounds which we use in the eXploration algorithm. First we prove Proposition 1. For that we use *ε-net covering* arguments. For the sake of completeness let us first define it.

Definition 1. *Let (X, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ is called an ϵ-net if $\forall x \in X$ $\exists y \in \mathcal{N}_\epsilon$ such that $d(x, y) \leq \epsilon$. The minimal cardinality (finite) of an ϵ-net is denoted by $\mathcal{N}(X, \epsilon)$ and is called covering number.*

For the sake of completeness we further state next lemma which was shown in (Vershynin, 2018).

Lemma 2. *For $\mathbb{R}^n$ equipped with euclidean metric it holds: $\mathcal{N}(S^{n-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n$.*

Using ϵ-net argument the next proposition was proven. Proposition 2 we will use later in order to prove an estimation error upper bound.

Proposition 2 (Proposition 8.1 in Sarkar and Rakhlin (2019))**.** *Let $M \in \mathbb{R}^{n\times d}$ be a random matrix. Then for any $\epsilon \in (0, 1)$ there exist $u \in S^{d-1}$ such that it holds:*

$$\mathbb{P}\left(\|M\| > z\right) \leq \left(1 + \frac{2}{\epsilon}\right)^d \mathbb{P}\left(\|Mu\| > (1 - \epsilon)z\right).$$

With the use of the Proposition 2 let us now state and prove Lemma 3 which will help us with data-dependent bounds. Before stating Lemma 3 let us introduce some notation.

Let $\mathcal{F} = (\mathcal{F}_i)_{i\geq 0}$ be a filtration, $(x_i)_{i\geq 0}$ stochastic process in $\mathbb{R}^d$ adopted to $\mathcal{F}$ and $(w_i)_{i\geq 1}$ zero mean, conditionally $\mathrm{subG}_l(\sigma^2)$, meaning that it holds for every $\|u\| = 1$, $\gamma \geq 0$ and $i \geq 1$:

$$\mathbb{E}\left[e^{\gamma(u^\top w_i)}|\mathcal{F}_{i-1}\right] \leq e^{\frac{\gamma^2\sigma^2}{2}}.$$

Further denote $\mathcal{V}_s = \sum_{i=0}^{s-1} x_i x_i^\top$ and $\mathcal{S}_s = \sum_{i=0}^{s-1} x_i w_{i+1}^\top$.

Lemma 3. *Let us be in the aforementioned setting. It holds w.p. at least $1 - \delta$:*

$$\forall s \geq 0 : \|\mathcal{S}_s\|^2_{(\mathcal{V}_s+\lambda I)^{-1}} \leq \frac{2\sigma^2}{(1-\epsilon)^2} \log\left(\frac{\det(\mathcal{V}_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\epsilon}\right)^l}{\delta}\right)$$

Proof. For $\epsilon \in (0, 1)$ we obtain from Proposition 2:

$$\begin{aligned}\mathbb{P}\left(\|\mathcal{S}_s\|_{(\mathcal{V}_s+\lambda I)^{-1}} > y\right) &\leq \left(1 + \frac{2}{\epsilon}\right)^l \mathbb{P}\left(\|\mathcal{S}_s u\|_{(\mathcal{V}_s+\lambda I)^{-1}} > (1 - \epsilon)y\right) \\ &= \left(1 + \frac{2}{\epsilon}\right)^l \mathbb{P}\left(\|\mathcal{S}_s u\|^2_{(\mathcal{V}_s+\lambda I)^{-1}} > (1 - \epsilon)^2 y^2\right),\end{aligned}$$

where $u \in \mathbb{R}^l$ is an appropriate unit vector. Since $\mathcal{S}_s u = \sum_{i=1}^{s} z_{i-1}(w_i^\top u)$ and w_i are independent $\mathrm{subG}_l(\sigma^2)$ random variables, $w_s^\top u$ are independent $\mathrm{subG}(\sigma^2)$ random variables. Hence we can apply Theorem 3 of Abbasi-Yadkori et al. (2011). Setting

$$y^2 = \frac{2\sigma^2}{(1-\epsilon)^2} \log\left(\frac{\det(\mathcal{V}_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\epsilon}\right)^l}{\delta}\right)$$

we obtain that with probability at least $1 - \delta$ it holds for every $s \geq 0$:

$$\|\mathcal{S}_s\|^2_{(\mathcal{V}_s+\lambda I)^{-1}} \leq \frac{2\sigma^2}{(1-\epsilon)^2} \log\left(\frac{\det(\mathcal{V}_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\epsilon}\right)^l}{\delta}\right).$$

□

To prove the Proposition 1 we can now simply use Lemma 3.

Proof of Proposition 1. Denote by $\mathcal{F}_i = \sigma((u_j)_{j\leq i}, (w_j)_{j\leq i})$. We then apply Lemma 3 with $\mathcal{S}_s = S_s$ and $\mathcal{V}_s = V_s$ and the result follows. □

Proof of Corollary 2. Since the analysis for matrix A_s is very much the same as for the matrix B_s we will do it just for A_s. From eq. (7) we obtain:

$$(A_s - A)^\top = (I_d\ 0)\,(V_k + \lambda I)^{-1}\,S_k - \lambda (I_d\ 0)\,(V_k + \lambda I)^{-1}\,(A\ B)^\top.$$

Next using triangle inequality we obtain:

$$\|A_s - A\| \leq I_1 + I_2,$$

where $I_1 = \left\|(I_d\ 0)\,(V_k + \lambda I)^{-1}\,S_k\right\|$ and $I_2 = \left\|\lambda (I_d\ 0)\,(V_k + \lambda I)^{-1}\,(A\ B)\right\|$. The first term is by Lemma 3 bounded w.p. at least $1-\delta$:

$$\begin{aligned} I_1 &\leq \left\|(I_d\ 0)\,(V_k + \lambda I)^{-\frac{1}{2}}\right\| \left\|(V_k + \lambda I)^{-\frac{1}{2}}\,S_k\right\| \\ &\leq \left\|(I_d\ 0)\,(V_k + \lambda I)^{-\frac{1}{2}}\right\| \frac{R}{1-\epsilon}\sqrt{2\log\left(\frac{\det(V_t + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}}\frac{\left(1+\frac{2}{\epsilon}\right)^d}{\delta}\right)}. \end{aligned}$$

With the bound on I_2 term:

$$I_2 = \left\|\lambda (I_d\ 0)\,(V_k + \lambda I)^{-1}\,(A\ B)\right\| \leq \lambda \left\|(I_d\ 0)\,(V_k + \lambda I)^{-1}\right\| \vartheta,$$

we conclude the proof. □

B System level synthesis

In this section we prove Lemma 1 and Theorem 1. First, for a matrix A, define its *resolvent* as $\mathfrak{R}_A(z) = (zI - A)^{-1}$ for $z \in \mathbb{C} \setminus \sigma(A)$, where $\sigma(A)$ is specter of matrix A. Next we, for the sake of completeness, state Theorem 2.3 of Gil' (2014) and show its corollary which will come handy.

Theorem 4 (Theorem 2.3 in (Gil', 2014)). *Let $A \in \mathbb{R}^{d\times d}$ and $z \notin \sigma(A)$. Denote $\rho(A, z) = \min_{k=1}^{d} |\lambda_k(A) - z|$, then it holds:*

$$\|\mathfrak{R}_A(z)\|_2 \leq \frac{1}{\rho(A,z)}\left(1 + \frac{1}{d-1}\left(1 + \frac{\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|}{\rho(A,z)^2}\right)\right)^{\frac{d-1}{2}}.$$

Corollary 3. *Let $A \in \mathbb{R}^{d\times d}$ and $\rho(A) < 1$, then it holds:*

$$\|\mathfrak{R}_A\|_{\mathcal{H}_\infty} \leq \frac{1}{1-\rho(A)}\left(1 + \frac{1}{d-1}\left(1 + \frac{\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}}.$$

To prove Lemma 1 we also need the following result of Dean et al. (2019).

Lemma 4 (Lemma 4.2 in Dean et al. (2019)). *Let K be a controller such that $\rho(A_* + B_*K) < 1$ If $(\epsilon_A + \epsilon_B \|K\|)\,\|\mathfrak{R}_{A_*+B_*K}\|_{\mathcal{H}_\infty} \leq \frac{1}{1+\sqrt{2}}$, then the SDP (5) is feasible.*

To further ease the notation we will by $A_{K,*}$ denote the true closed loop matrix if we choose controller K i.e. $A_{K,*} = A_* + B_*K$. We have now all the machinery in order to prove Lemma 1.

Proof of Lemma 1. Since $\rho(A_{K,*}) \leq \gamma_0 < 1$ we have by Corollary 3:

$$\begin{aligned}\left\|\mathfrak{R}_{A_{K,*}}\right\|_{\mathcal{H}_\infty} &\leq \frac{1}{1-\gamma_0}\left(1 + \frac{1}{d-1}\left(1 + \frac{\|A_{K,*}\|_F^2 - \left|\mathrm{Tr}(A_{K,*}^2)\right|}{(1-\gamma_0)^2}\right)\right)^{\frac{d-1}{2}} \\ &\leq \frac{1}{1-\gamma_0}\left(1 + \frac{1}{d-1}\left(1 + \left(\frac{\|A_*\|_F + \|B_*\|_F \|K\|)_F}{(1-\gamma_0)^2}\right)^2\right)\right)^{\frac{d-1}{2}}.\end{aligned}$$

With

$$(\epsilon_A + \epsilon_B \|K\|) \leq \max(\varepsilon_A, \varepsilon_B)(1 + \|K\|),$$

we obtain that the condition $(\epsilon_A + \epsilon_B \|K\|)\,\|\mathfrak{R}_{A_*+B_*K}\|_{\mathcal{H}_\infty} \leq \frac{1}{1+\sqrt{2}}$ from Lemma 4 is satisfied if:

$$\max(\varepsilon_A, \varepsilon_B) \leq \frac{1}{3(1+\|K\|)C(\gamma_0, A_*, B_*, K))},$$

where we denote:

$$C(\gamma_0, A_*, B_*, K) = \frac{1}{1-\gamma_0}\left(1 + \frac{1}{d-1}\left(1 + \left(\frac{\|A_*\|_F + \|B_*\|_F \|K\|)_F}{(1-\gamma_0)^2}\right)^2\right)\right)^{\frac{d-1}{2}}.$$

□

The proof of Theorem 1 then follows.

Proof of Theorem 1. From Lemma 1 we know that as soon as $\max(\varepsilon_A, \varepsilon_B) \leq \frac{1}{3(1+\|K\|)C(\gamma_0, A_*, B_*, K)}$ the SDP (5) in Algorithm 1 will be feasible and Algorithm 1 will terminate. At the same time from Corollary 1 we know that if $s \geq poly(\log \frac{1}{\delta})$ with probability at least $1 - \delta$ it holds:

$$\max(\varepsilon_A, \varepsilon_B) \leq \frac{poly\left(\log s, \log \frac{1}{\delta}\right)}{\sqrt{s}} \leq \frac{poly\left(\log \frac{1}{\delta}\right)}{s^{1/2-\varepsilon/6}}.$$

Since $\frac{1}{2+\varepsilon} \leq \frac{1}{2} - \frac{\varepsilon}{6}$, for $\varepsilon \in (0,1)$ we obtain that as soon as

$$s \geq \left(poly\left(\log \frac{1}{\delta}\right)(1+\|K\|)C(\gamma_0, A_*, B_*, K)) \right)^{2+\epsilon}$$

Algorithm 1 has, with probability at least $1-\delta$, found a controller which stabilizes A_*, B_*. □

Remark 1. *With the result of Theorem 1 we can upper bound the largest norm of states in eXploration phase while playing zero Gaussian actions with* $\mathcal{O}\left(\|A_*\|^{\left(poly\left(\log \frac{1}{\delta}\right)(1+\|K\|)C(\gamma_0, A_*, B_*, K))\right)^{2+\epsilon}} \right)$.

C OSLO

In this section we will first find $\gamma < 1$ for which it holds $\rho(A_* + B_* K_0) \leq \gamma$ show how to compute γ and then use it in order to provide refined analysis of warm-up (Phase II) of X-OSLO. To run the Phase III of X-OSLO we further require an upper bound on optimal expected infinite horizon cost J_*. We show how we can compute this using the SDP (5) which terminates eXploration phase. For the ease of presentation we assume in this section that $(w_i)_{i \geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$.

C.1 Compute γ with $\rho(A_* + B_* K_0) \leq \gamma < 1$

Let us first state a result which comes as a consequence of transformation of eq. (4) to SDP (5).

Lemma 5. *Let s^* be the minimal value, $\widehat{A}, \widehat{B}$ the estimates, $\varepsilon_A, \varepsilon_B$ upper bounds and K_0 the obtained controller of the SDP with which we stop the Algorithm 1. Then it holds:*

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K_0 \end{pmatrix} \left(zI - \widehat{A} - \widehat{B} K_0 \right)^{-1} \right\|_{\mathcal{H}_\infty} \leq \sqrt{s^*}.$$

Next we show a result which show an upper bound on the norm of resolvent of perturbed matrix.

Lemma 6. *Let $D \in \mathbb{R}^{d \times d}$ with $\rho(D) < 1$. Then for $\varepsilon \leq \frac{1-\rho(D)}{2\rho(D)}$ it holds:*

$$\left\| (zI - (1+\varepsilon)D)^{-1} \right\|_{\mathcal{H}_\infty} \leq \frac{2}{1-\rho(D)} \left(1 + \frac{1}{d-1} \left(1 + \frac{4(\|D\|_F^2 - \left|\mathrm{Tr}(D^2)\right|)}{\rho(D)^2(1-\rho(D)^2)} \right) \right)^{\frac{d-1}{2}}$$

Proof. By Corollary 3 we have:

$$\left\| (zI - (1+\varepsilon)D)^{-1} \right\|_{\mathcal{H}_\infty} \leq \frac{1}{1-(1+\varepsilon)\rho(D)} \left(1 + \frac{1}{d-1} \left(1 + \frac{4(1+\varepsilon)^2(\|D\|_F^2 - \left|\mathrm{Tr}(D^2)\right|)}{(1-(1+\varepsilon)\rho(D))^2} \right) \right)^{\frac{d-1}{2}}$$

Plugging in the bound $\varepsilon \leq \frac{1-\rho(D)}{2\rho(D)}$ we obtain the result. □

In what comes next we will denote

$$f(D) := \frac{2}{1-\rho(D)} \left(1 + \frac{1}{d-1} \left(1 + \frac{4(\|D\|_F^2 - \left|\mathrm{Tr}(D^2)\right|)}{\rho(D)^2(1-\rho(D)^2)} \right) \right)^{\frac{d-1}{2}}.$$

Proposition 3. *Let s^* be the minimal value, $\widehat{A}, \widehat{B}$ the estimates, $\varepsilon_A, \varepsilon_B$ upper bounds and K_0 the obtained controller of the SDP with which we stop the Algorithm 1 Denote $D = \widehat{A} + \widehat{B} K_0$. Then it holds:*

$$\rho(A_* + B_* K_0) < \frac{1}{1+\varepsilon},$$

where $\varepsilon = \min\left(\frac{1-\rho(D)}{2\rho(D)}, \frac{\sqrt{1+(1/s^*-1)\|D\|f(D)}-1}{2\|D\|f(D)} \right)$.

Proof. Observe that the condition $\rho(A_* + B_* K_0) < \frac{1}{1+\varepsilon}$ is equivalent to $\rho(A'_* + B'_* K_0) < 1$, where we denote by $A'_* = (1+\varepsilon)A_*, B'_* = (1+\varepsilon)B_*$. Let us further denote by $\widehat{A}' = (1+\varepsilon)\widehat{A}, \widehat{B}' = (1+\varepsilon)\widehat{B}$ and $\varepsilon'_A = (1+\varepsilon)\varepsilon_A, \varepsilon'_B = (1+\varepsilon)\varepsilon_B$.

From the discussion in section 2.1 we obtain that the sufficient condition for $\rho(A'_* + B'_* K_0) < 1$ is:

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon'_A I \\ \sqrt{2}\varepsilon'_B K_0 \end{pmatrix} \left(zI - \widehat{A}' - \widehat{B}' K_0 \right)^{-1} \right\|_{\mathcal{H}_\infty} < 1,$$

which is equivalent to:

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K_0 \end{pmatrix} (zI - (1+\varepsilon)D)^{-1} \right\|_{\mathcal{H}_\infty} < \frac{1}{1+\varepsilon},$$

Next denote by $C = \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K_0 \end{pmatrix}$ and bound:

$$\begin{aligned} \left\| C\left(zI - (1+\varepsilon)D\right)^{-1} \right\|_{\mathcal{H}_\infty} &= \left\| C\left((zI - D)^{-1} + \varepsilon \left(zI - D\right)^{-1} D \left(zI - (1+\varepsilon)D\right)^{-1} \right) \right\|_{\mathcal{H}_\infty} \\ &\leq \left\| C\left(zI - D\right)^{-1} \right\|_{\mathcal{H}_\infty} \left(1 + \varepsilon \left\| D \right\| \left\| (zI - (1+\varepsilon)D)^{-1} \right\|_{\mathcal{H}_\infty} \right), \end{aligned}$$

where we used the equality $(X+Y)^{-1} = X^{-1} + X^{-1}Y(X+Y)^{-1}$. Then by Lemmas 5 and 6 follows:

$$\left\| C\left(zI - (1+\varepsilon)D\right)^{-1} \right\|_{\mathcal{H}_\infty} \leq s_* \left(1 + \varepsilon \left\| D \right\| f(D)\right).$$

The right hand side is smaller than $1/(1+\varepsilon)$ by setting $\varepsilon = \min\left(\frac{1-\rho(D)}{2\rho(D)}, \frac{\sqrt{1+(1/s^*-1)\|D\|f(D)}-1}{2\|D\|f(D)} \right)$.
For such a choise of ε then follows:

$$\rho(A_* + B_* K_0) < \frac{1}{1+\varepsilon}.$$

□

Since matrix $\widehat{A} + \widehat{B}K_0$ and s^* are known while running the algorithm, we can compute ε given in Proposition 3 and hence we found γ, defined as $\gamma = \frac{1}{1+\varepsilon}$, which we can compute, and for which it holds $\rho(A_* + B_* K_0) \leq \gamma < 1$.

C.2 Refined analysis of OSLO's warm-up phase

Next we present the pseudo code of the warm-up (Phase II) of X-OSLO and provide a refined analysis given in Cohen et al. (2019) in which they show that running Phase II for $\mathcal{O}(\sqrt{T})$ rounds tightens the estimates of A_*, B_* enough that we can start with optimistic exploitation and at the same time yields the regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$.

Algorithm 2 Phase II: Tighten the bounds

1: **Input:** K_0 from Phase I, T
2: Denote $\kappa_0 = \max(1, \|K_0\|)$
3: **for** $i = 1, \dots \mathcal{O}(\sqrt{T})$ **do**
4: **observe** state x_i
5: **play** $u_i \sim \mathcal{N}(K_0 x_i, 2\sigma^2 \kappa_0^2 I)$
6: **end for**

Before starting with the analysis let us state some useful results which will come handy. First we state a result which with high probability bounds the norm of zero mean Gaussians.

Theorem 5 (Hanson-Wright (Proposition 1.1 in (Hsu et al., 2012))). *Let $x \sim \mathcal{N}(0, I_n)$ and let $A \in \mathbb{R}^{m \times n}$. Denote by $\Sigma = A^\top A$. Then for all $z > 0$ it holds:*

$$\mathbb{P}\left(\|Ax\|^2 > \operatorname{Tr}(\Sigma) + 2\sqrt{\operatorname{Tr}(\Sigma^2) z} + 2\|\Sigma\| z \right) \leq e^{-z}$$

Corollary 4. *Let $x \sim \mathcal{N}(0, \Sigma)$. Then for any $\delta \in (0, \frac{1}{e})$ it holds with probability at least $1 - \delta$:*

$$\|x\|^2 \leq 5 \operatorname{Tr}(\Sigma) \log \frac{1}{\delta}$$

Proof. Since $\delta \in (0, \frac{1}{e})$ it holds $\log \frac{1}{\delta} > 1$. Hence if we set $z = \log \frac{1}{\delta}$ we have $\sqrt{z} \leq z$. Inserting $z = \log \frac{1}{\delta}$ to Theorem 5 we have that it holds w.p at least $1 - \delta$:

$$\|x\|^2 \leq \mathrm{Tr}(\Sigma) + (2\|\Sigma\|_F + 2\|\Sigma\|) \log \frac{1}{\delta}.$$

Hence it is enough to show that $\|\Sigma\|, \|\Sigma\|_F \leq \mathrm{Tr}(\Sigma)$. Since Σ is symmetric positive semi-definite matrix its eigenvalues are equal to singular values. Hence it is enough to show:

$$\sqrt{\sum_i \lambda_i(\Sigma)^2} \leq \sum_i \lambda_i(\Sigma).$$

Since

$$\left(\sum_i \lambda_i(\Sigma)\right)^2 - \sum_i \lambda_i(\Sigma)^2 = \sum_{i \neq j} \lambda_i(\Sigma)\lambda_j(\Sigma) \geq 0,$$

we have $\|\Sigma\|_F \leq \mathrm{Tr}(\Sigma)$. For every matrix it also holds $\|\Sigma\| \leq \|\Sigma\|_F$, hence we showed that $\|\Sigma\|, \|\Sigma\|_F \leq \mathrm{Tr}(\Sigma)$. □

To analyze the case when we would like to utilize $\rho(A_* + B_* K_0) < 1$ we further need to know how to bound the norm of power of closed loop matrix. The following theorem and its corollary will come handy.

Theorem 6 (Theorem 2.16 in (Dowler, 2013))**.** *Let $A \in \mathbb{R}^{d \times d}$ be a square matrix and let Γ be a positively oriented Jordan curve in complex plane which contains ball $B(\rho(A))$ in its interior. Then it holds:*

$$A^k = \frac{1}{2\pi i} \int_\Gamma z^k \mathfrak{R}_A(z) dz.$$

Lemma 7 (Matrix power norm bound)**.** *Let $A \in \mathbb{R}^{d \times d}$ with $\rho(A) < 1$. Then it holds:*

$$\left\|A^k\right\| \leq \left(\frac{1+\rho(A)}{2}\right)^{k+1} \frac{2}{1-\rho(A)} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}}$$

Proof. Since $\rho(A) < 1$ the curve which parametrizes the circle $\partial B\left(\frac{1+\rho(A)}{2}\right)$ in the positive way contains in its interior all the eigenvalues of A. Hence we can use Theorem 6 and compute

$$\begin{aligned}
\left\|A^k\right\| &\leq \frac{1}{2\pi} \int_{\partial B\left(\frac{1+\rho(A)}{2}\right)} |z|^k \left\|\mathfrak{R}_A(z)\right\| dz \\
&\leq \frac{1}{\pi(1-\rho(A))} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}} \int_{\partial B\left(\frac{1+\rho(A)}{2}\right)} |z|^k dz \\
&= \left(\frac{1+\rho(A)}{2}\right)^{k+1} \frac{2}{1-\rho(A)} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}},
\end{aligned}$$

where we used Theorem 4 in the second inequality. □

Now we present the refined analysis of Cohen et al. (2019) where we leverage the knowledge of $\rho(A_* + B_* K_0) < 1$. From Proposition 3 we obtain that there exist γ, which we can compute after Phase I, with $\rho(A_* + B_* K_0) \leq \gamma < 1$. In the rest of this section we denote by

$$C_0 = \frac{2}{1-\gamma} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A_* + B_* K_0\|_F^2 - \left|\mathrm{Tr}((A_* + B_* K_0)^2)\right|)}{\gamma^2(1-\gamma)^2}\right)\right)^{\frac{d-1}{2}}.$$

Since we know an upper bound ϑ with $\|(A\ B)\| \leq \vartheta$ we can compute an upper bound for C_0 after we finish with Phase I.

Lemma 8. *Let $x_0, x_1, \ldots$ be a sequence of states starting from state x_0 and generated by dynamics (1) following a policy K_0, obtained from Algorithm 1. Then we have:*

$$\|x_i\| \leq C_0 \left(\frac{1+\gamma}{2}\right)^{i+1} \|x_0\| + \frac{2C_0}{1-\gamma} \max_{j=0}^{i-1} \|w_j\|$$

Proof. Since we stick to the policy K_0, we have $x_{i+1} = (A + BK_0)x_i + B\eta_i + w_{i+1}$, where $\eta_i \sim \mathcal{N}(0, 2\kappa_0^2\sigma^2 I)$. From there it follows:

$$x_i = (A + BK_0)^i x_0 + \sum_{j=0}^{i-1} (A + BK_0)^{i-j-1}(Bv_j + w_{j+1}).$$

Using first triangle inequality and then Lemma 7 we obtain:

$$\begin{aligned}
\|x_i\| &\leq \left\|(A + BK_0)^i\right\| \|x_0\| + \sum_{j=0}^{i-1} \left\|(A + BK_0)^{i-j-1}\right\| \|Bv_j + w_{j+1}\| \\
&\leq C_0 \left(\frac{1+\gamma}{2}\right)^{i+1} \|x_0\| + C_0 \max_{i=1}^{k} \|w_i\| \sum_{i=0}^{\infty} \left(\frac{1+\gamma}{2}\right)^i \\
&= C_0 \left(\frac{1+\gamma}{2}\right)^{i+1} \|x_0\| + \frac{2C_0}{1-\gamma} \max_{j=0}^{i-1} \|Bv_j + w_{j+1}\|
\end{aligned}$$

□

In the next lemma we will apply a corollary 4 of Hanson-Wright inequality and bound the maximal norm of the noise.

Lemma 9. *Let $\delta \in (0, \frac{1}{e})$. With probability at least $1 - \delta$ for all $i = 1, \ldots T_0$ holds:*

$$\|x_i\| \leq C_0 \left(\frac{1+\gamma}{2}\right)^{i+1} \|x_0\| + \frac{2\sqrt{5}C_0\sigma}{1-\gamma} \sqrt{(d + 2k\kappa_0^2\vartheta^2) \log \frac{T_0}{\delta}}$$

Proof. In order to use Lemma 8 we need to bound $\max_{j=0}^{T_0-1} \|B\eta_j + w_{j+1}\|$. Since $B\eta_j + w_{j+1} \sim \mathcal{N}(0, 2\sigma^2\kappa_0^2 BB^\top + \sigma^2 I)$ we can use Corollary 4. For every $0 \leq j \leq T_0 - 1$ it holds w.p. at least $1 - \frac{\delta}{T_0}$:

$$\|B\eta_j + w_{j+1}\|^2 \leq 5\sigma^2(d + 2\kappa_0^2 \|B\|_F^2) \log \frac{T_0}{\delta}.$$

Using union bound and Lemma 8 we obtain that it holds w.p. at least $1 - \delta$:

$$\|x_i\| \leq C_0 \left(\frac{1+\gamma}{2}\right)^{i+1} \|x_0\| + \frac{2\sqrt{5}C_0\sigma}{1-\gamma} \sqrt{(d + 2k\kappa_0^2\vartheta^2) \log \frac{T_0}{\delta}},$$

which finishes the proof. □

The rest of the analysis which shows that running Phase II for $\widetilde{\mathcal{O}}(\sqrt{T})$ rounds yields a controller with tight enough estimates to start Phase III is very similar to the analysis presented in the proof of Theorem 20 in (Cohen et al., 2019) and hence we omit it here.

C.3 Optimal infinite horizon cost upper bound

We can start with Phase III of X-OSLO when we have estimates $\widehat{A}, \widehat{B}$ with $\|(\widehat{A}\ \widehat{B}) - (A_*\ B_*)\|_F^2 \leq c\frac{\alpha_0^5\sigma^{10}}{\nu^5\vartheta\sqrt{T}}$. Here $\alpha_0 = \min(\lambda_{min}(Q), \lambda_{min}(R))$, c universal constant, σ, ϑ, T as defined above and ν an upper bound for the optimal expected infinite horizon cost J_*. We need upper bound ν in order to know when we can start with Phase III. We now show how we can compute ν from the optimal solution of SDP with which we finish Phase I of X-OSLO.

If we choose action $u_i = Kx_i$, where $K \in \mathbb{R}^{k\times d}$ is fixed matrix for which it holds $\rho(A_* + B_*K) < 1$, then the infinite horizon cost associated with this policy is equal to the solution of the minimization problem (c.f. (Cohen et al., 2018)):

$$\begin{aligned} \min_{X\succeq 0} \ & \operatorname{Tr}((Q + K^\top RK)X), \\ & \text{s.t. } X = (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I. \end{aligned} \tag{11}$$

We denote the expected infinite horizon cost for such a policy with $J(A_*, B_*, K)$. We will now show that we obtain the same result if we replace equality constraint in (11) with PSD inequality one.

Lemma 10. *Let ν^* be minimal value of* (11) *and ν' minimal value of*

$$\begin{aligned} \min_{X\succeq 0} \ & \operatorname{Tr}((Q + K^\top RK)X), \\ & \textit{s.t. } X \succeq (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I. \end{aligned} \tag{12}$$

Then it holds $\nu^* = \nu'$.

Proof. Since constraint in (12) is weaker than the constraint in (11) we have $\nu^* \geq \nu'$. To prove the equality we will show that in the optimal solution of minimization problem (12) it holds: $X = (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I$. Suppose that for X it holds $X \succeq (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I$ but not $X = (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I$. Then there exist $E \succeq 0$, $E \neq 0$ such that: $X - E = (A_* + B_*K)X(A_* + B_*K)^\top + \sigma^2 I$. We will show that then also $X - E$ satisfy the constraints of problem (12). Since:

$$\begin{aligned} & X - E \succeq (A_* + B_*K)(X - E)(A_* + B_*K)^\top + \sigma^2 I \\ \iff & X - E \succeq (A_* + B_*K)X(A_* + BK_*)^\top - (A_* + BK_*)E(A_* + BK_*)^\top + \sigma^2 I \\ \iff & X - E \succeq X - E - (A_* + B_*K)E(A_* + B_*K)^\top \\ \iff & 0 \succeq -(A_* + B_*K)E(A_* + B_*K)^\top, \end{aligned}$$

$X - E$ indeed satisfy the constraints of (12). Since $\operatorname{Tr}((Q + K^\top RK)X) > \operatorname{Tr}((Q + K^\top RK)(X - E))$ we obtain that in the optimal solution of (12) we have the equality in the constraint and it holds $\nu_* = \nu'$. □

The next lemma will show how can we remove the σ^2 term from constraint to the minimization term.

Lemma 11. *The minimal value of the optimization problem* (11) *is equal to the optimal value of:*

$$\begin{aligned} \min_{X\succeq 0} \ & \sigma^2 \operatorname{Tr}((Q + K^\top RK)X), \\ & \textit{s.t. } X = (A_* + B_*K)X(A_* + B_*K)^\top + I. \end{aligned} \tag{13}$$

Proof. First notice that the optimal value of (11) is equal to:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \mathbb{E}\left(x_i^\top Q x_i + u_i^\top R u_i\right).$$

Since $u_i = Kx_i$ we obtain:

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \mathbb{E}\left(x_i^\top Q x_i + u_i^\top R u_i\right) &= \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \mathbb{E}\left(x_i^\top Q x_i + x_i^\top K^\top RK x_i\right) \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \operatorname{Tr}\left(\left(Q + K^\top RK\right)\mathbb{E}[x_i x_i^\top]\right) \\ &= \operatorname{Tr}\left(\left(Q + K^\top RK\right) \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \mathbb{E}[x_i x_i^\top]\right) \end{aligned}$$

Let us look at the term $\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}[x_i x_i^\top]$. Since $x_i=\sum_{j=1}^{i}(A_*+B_*K)^{i-j}w_j$ we obtain:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}[x_i x_i^\top]=\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}\left[\sum_{j=1}^{i}(A_*+B_*K)^{i-j}w_j w_l^\top((A_*+B_*K)^\top)^{i-j}\right]$$
$$=\sigma^2\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}\left[\sum_{j=1}^{i}(A_*+B_*K)^{i-j}\frac{w_j}{\sigma}\frac{w_l}{\sigma}^\top((A_*+B_*K)^\top)^{i-j}\right].$$

Hence (due to the linearity of trace operator) if we calculate as if that the process noise has covariance matrix equal to I we need to multiply the infinite horizon cost with σ^2 to obtain the true infinite horizon cost. □

To finish we will first show a result presented by Dean et al. (2019) and then use it to to arrive at an upper bound for J_*.

Lemma 12. *Let $X, K=ZX^{-1}, s$ be a feasible solution for SDP* (5)*. Then it holds:*

$$J(A_*,B_*,K)\leq\frac{1}{1-\sqrt{s}}J(\widehat{A},\widehat{B},K).$$

Lemma 13. *Let s^*, X, K be parameters of the optimal solution of SPD problem which are returned by Algorithm 1. Then it holds:*

$$J^*\leq\frac{\sigma^2}{1-\sqrt{s^*}}\operatorname{Tr}((Q+K^\top RK)X)$$

Proof. First we will use the fact that if a matrix is positive semi definite then all its minors are also positive semi definite. Since s^*, X, K are optimal solution to SDP (5) they are also feasible solution and hence we have:

$$\begin{pmatrix}X-I & (\widehat{A}+\widehat{B}K)X\\ X(\widehat{A}+\widehat{B}K)^\top & X\end{pmatrix}\succeq 0.$$

Since $X\succ 0$, the latter is by Schur's complement lemma equivalent to:

$$X-I-(\widehat{A}+\widehat{B}K)XX^{-1}X(\widehat{A}+\widehat{B}K)^\top\succeq 0.$$

Reordering the terms we obtain:

$$X\succeq(\widehat{A}+\widehat{B}K)X(\widehat{A}+\widehat{B}K)^\top+I$$

Since $X\succeq(\widehat{A}+\widehat{B}K)X(\widehat{A}+\widehat{B}K)^\top+I$, we obtain:

$$J(\widehat{A},\widehat{B},K)\leq\sigma^2\operatorname{Tr}\left(\left(Q+K^\top RK\right)X\right).$$

To finish the proof let us use lemma 12:

$$J_*\leq J(A_*,B_*,K)\leq\frac{1}{1-\sqrt{s^*}}J(\widehat{A},\widehat{B},K)\leq\frac{\sigma^2}{1-\sqrt{s^*}}\operatorname{Tr}\left(\left(Q+K^\top RK\right)X\right).$$

□

D Certainty equivalent control

In this section we first show a theorem which show that the regret of X-CEC is bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$. We then further provide pseudo code of algorithm CEC, introduced by Simchowitz and Foster (2020), which we use as Phase II and III of algorithm X-CEC.

Theorem 7. *Let $\delta \in (0, \frac{1}{T})$. Then the Regret of the X-CEC algorithm is bounded by:*

$$R_T = \mathcal{O}\left(\sqrt{k^2 dT \log \frac{1}{\delta}}\right).$$

Proof Sketch. By Theorem 2 of Simchowitz and Foster (2020) the regret of Phase II and III is bounded by $\mathcal{O}\left(\sqrt{k^2 dT \log \frac{1}{\delta}}\right)$. Since running eXploration as Phase I adds $\widetilde{\mathcal{O}}(1)$ to the total regret the result follows. □

Note that since $\delta \in (0, \frac{1}{T})$ the factor $\log \frac{1}{\delta}$ is of order $\log T$. Hence the regret scale asymptotically as $\mathcal{O}(\sqrt{T \log T})$. We included the state and action dimension in the $\mathcal{O}(\cdot)$, due to the lower bound on regret by Simchowitz and Foster (2020) which states that any algorithm suffers at least $\Omega(\sqrt{k^2 dT})$.

Let us know show the pseudo code of CEC. In the pseudo code we use the notation $P(A_i, B_i) = DARE(A_i, B_i, Q, R)$ and $K(A_i, B_i) = -(R + B_i^\top P(A_i, B_i) B_i)^{-1} B_i^\top P(A_i, B_i) A_i$.

As long as the variable safe in Algorithm 3 is set on $False$ we are in the Phase II and play $u_i \sim \mathcal{N}(K_0 x_i, I)$. Once we can with high probability ensure that the obtained estimates are close to the true underlying system we set variable safe to $True$ and start with Phase III, where we greedily exploit the tightness of the estimates, while we additionally explore by injecting Gaussian noise with variance scaling with $\mathcal{O}(1/\sqrt{i})$. Note that the end of Phase II is again as the end of Phase I completely data dependent.

Algorithm 3 Phase II and III of X-CEC: Tighten the bounds and exploit greedily

Input: K_0 with $\rho(A_* + B_* K_0) < 1$, δ
safe $= False$, $K = K_0$, $\sigma^2 = 1$
for $i = 1, \ldots$ **do**
  observe state x_i
  if $i = 2^j$ **then**
    Let (A_i, B_i, V_i) be OLS estimators and covariance matrix from samples $2^{j-1}, \ldots, 2^j - 1$
    if safe $= False$ **then**
      Conf$_i$ = $6\lambda_{min}(V_i)^{-1}(n \log 5 + \log(4i^2 \det(3V_i)/\delta))$ (∞, if $\det(V_i) = 0$)
      if $V_i \succeq I$ and $1/\text{Conf}_i \geq 54 \left\|P(A_i, B_i)\right\|^5$ **then**
        safe $= True$
        $B_{\text{safe}} = \{(A, B) | \left\|A - A_i\right\| \leq \text{Conf}_i, \left\|B - B_i\right\| \leq \text{Conf}_i\}$
        $\sigma_{in}^2 = \sqrt{d} \left\|P(A_i, B_i)\right\|^{9/2} \max(1, \left\|B_i\right\|) \sqrt{\log(\left\|P(A_i, B_i)\right\| / \delta)}$
      end if
    else if safe $= True$ **then**
      Let $(\widetilde{A}_i, \widetilde{B}_i)$ be projection of (A_i, B_i) on B_{safe}
      $K = K(\widetilde{A}_i, \widetilde{B}_i)$
      $\sigma^2 = i^{-1/2} \min(1, \sigma_{in}^2)$
    end if
  end if
  play $u_i \sim \mathcal{N}(Kx_i, \sigma^2 I)$
end for

E Controller analysis

The section consists of two parts. In first part we prove that with the right choice of controller K_i the OLS estimator is inconsistent and in the second part we prove Theorem 3. The setting in this section is the following. The system evolves as $x_{i+1} = A_* x_i + u_i + w_{i+1}, x_0 = 0$, where $x_i \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}, u_i = K_i x_i$ and $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$. We define the RLS estimator of A_* as

$$A_s = \operatorname*{argmin}_A \sum_{i=0}^{s-1} \|x_{i+1} - u_i - Ax_i\| + \lambda \|A\|_F^2 .$$

A similar derivation as in appendix A would show that we have:

$$(A_s - A_*)^\top = (V_s + \lambda I)^{-1} S_s - \lambda (V_s + \lambda I)^{-1} A^\top, \tag{14}$$

where $V_s = \sum_{i=0}^{s-1} x_i x_i^\top$ and $S_s = \sum_{i=0}^{s-1} x_i w_{i+1}^\top$. We obtain the OLS estimator A_s^o of A_* by setting $\lambda = 0$ in RLS.

E.1 Inconsistency of OLS in case $d > 1$

The construction will be based on the inconsistency of OLS estimator. Nielsen (2008) and Phillips and Magdalinos (2013) shows that in the case when A_* is irregular and the system evolves as $x_{i+1} = A_* x_i + w_{i+1}$, the OLS estimator is inconsistent. Their result shows that

$$(A_s^o - A_*)^\top = \left(\sum_{i=0}^{s-1} x_i x_i^\top \right)^{-1} \sum_{i=0}^{s-1} x_i w_{i+1}^\top$$

does not converge in probability towards zero. To show that we can take such actions u_i, which will lead to inconsistent OLS estimator A_s^o of matrix A_* we will assume that we know matrix A_*, however we would still like to compute its OLS estimator. For that select u_i as $u_i = (2I_d - A_*)x_i$. Since u_i is a measurable function of x_i it is also a measurable function of $(x_j)_{j\leq i}$. With such a control the system evolves as:

$$x_{i+1} = A_* x_i + u_i + w_{i+1} = A_* x_i + (2I_d - A_*)x_i + w_{i+1} = 2I_d x_i + w_{i+1}.$$

At the same time for OLS estimator A_s^o it holds:

$$(A_s^o - A_*)^\top = \left(\sum_{i=0}^{s-1} x_i x_i^\top \right)^{-1} \sum_{i=0}^{s-1} x_i w_{i+1}^\top \tag{15}$$

Since $2I_d$ is irregular matrix, the right hand side of the eq. (15) by result of Nielsen (2008) does not converge towards zero. Hence we have shown that there exist a sequence of measurable actions for which the OLS estimator does not converge.

E.2 Proof of Theorem 3

The strategy will be the following. We first use eq. (14) to bound:

$$\|A_s - A_*\| \leq \left\| (V_s + \lambda I)^{-\frac{1}{2}} \right\| \left\| (V_s + \lambda I)^{-\frac{1}{2}} S_s \right\| + \lambda \|A\| \left\| (V_s + \lambda I)^{-1} \right\|,$$

and then show that $\left\| (V_s + \lambda I)^{-\frac{1}{2}} \right\| = \mathcal{O}(1/\sqrt{s})$ and $\left\| (V_s + \lambda I)^{-\frac{1}{2}} S_s \right\| = \mathcal{O}(1) \left(d \log s + \log \frac{1}{\delta} \right)$.

The toughest part is to show that $\left\| (V_s + \lambda I)^{-\frac{1}{2}} \right\| = \mathcal{O}(1/\sqrt{s})$, which is equivalent to show that $V_s \succeq \Omega(s) I$. Let us begin with a simple lemma which was proven in (Sarkar and Rakhlin, 2019).

Lemma 14. *Let $P, Q \in \mathbb{R}^{d\times d}$ such that $P \succ 0$. Assume $\|Q\|_{P^{-1}} \leq \gamma$. Then for every vector v for which it holds $v^\top P v = \alpha$ we have: $\left\| v^\top Q \right\| \leq \sqrt{\alpha}\gamma$*

Next we show a decomposition of V_s to three parts. We will later show that the sum of the first two terms contribute at least $-\Theta(\log s)$ and the last term at least $\Omega(s)$ to the smallest eigenvalue of V_s with high probability.

Lemma 15. *Let $y_i = (A_* + K_i)x_i$. Then for every $s \geq 1$ it holds:*

$$V_s = \sum_{i=0}^{s-2} y_i y_i^\top + \sum_{i=0}^{s-2} \left(y_i w_{i+1}^\top + w_{i+1} y_i^\top\right) + \sum_{i=0}^{s-2} w_{i+1} w_{i+1}^\top$$

Proof. By inserting $V_s = \sum_{i=0}^{s-1} x_i x_i^\top$ and use the initial condition $x_0 = 0$ we obtain:

$$\begin{aligned}
V_s &= \sum_{i=0}^{s-1} x_i x_i^\top = \sum_{i=1}^{s-1} x_i x_i^\top = \sum_{i=0}^{s-2} x_{i+1} x_{i+1}^\top \\
&= \sum_{i=0}^{s-2} ((A_* + K_i)x_i + w_{i+1})((A_* + K_i)x_i + w_{i+1})^\top = \sum_{i=0}^{s-2} (y_i + w_{i+1})(y_i + w_{i+1})^\top \\
&= \sum_{i=0}^{s-2} \left(y_i y_i^\top + y_i w_{i+1}^\top + w_{i+1} y_i^\top + w_{i+1} w_{i+1}^\top\right) \\
&= \sum_{i=0}^{s-2} y_i y_i^\top + \sum_{i=0}^{s-2} \left(y_i w_{i+1}^\top + w_{i+1} y_i^\top\right) + \sum_{i=0}^{s-2} w_{i+1} w_{i+1}^\top
\end{aligned}$$

□

We now show an upper bound on the norm of middle term, normalized with the regularized first term of the V_s decomposition.

Lemma 16. *Let us be in setting of this section. Then it holds w.p. at least $1 - \delta$:*

$$\forall s \geq 1: \quad \left\| \sum_{i=0}^{s-2} y_i w_{i+1}^\top \right\|^2_{\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)^{-1}} \leq 8\sigma^2 d \left(\log s + \log \frac{5 M_1^2 M_2^2}{\delta^{1/d}}\right)$$

Proof. Denote by $\mathcal{F}_s = \sigma\left((w_i)_{i \leq s}\right)$ and $\mathcal{F} = (\mathcal{F}_s)_{s \geq 0}$. With this notation $(y_i)_{i \geq 0}$ is stochastic process in $\mathbb{R}^d$ adopted to filtration $\mathcal{F}$. Further denote by $V = I$. Now we apply Lemma 3 with $\varepsilon = \frac{1}{2}$ and obtain:

$$\forall s \geq 1: \quad \left\| \sum_{i=0}^{s-2} y_i w_{i+1}^\top \right\|^2_{\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)^{-1}} \leq 8\sigma^2 \log \left(\frac{\det\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)}{\det(I)} \frac{5^d}{\delta} \right). \tag{16}$$

Since

$$\sum_{i=0}^{s-2} y_i y_i^\top \preceq \sum_{i=0}^{s-2} \|y_i\|^2 I \preceq \sum_{i=0}^{s-2} \|A_* + K_i\|^2 \|x_i\|^2 I \preceq M_2^2 M_1^2 (s-1) I,$$

it holds

$$\det\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right) \leq \det\left(M_1^2 M_2^2 s I\right) = \left(M_1^2 M_2^2 s\right)^d.$$

Therefore the upper bound from eq. (16) is upper bounded by

$$\begin{aligned}
8\sigma^2 \log \left(\frac{\det\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)}{\det(I)} \frac{5^d}{\delta} \right) &\leq 8\sigma^2 \log \left(\frac{(5 M_1^2 M_2^2 s)^d}{\delta} \right) \\
&= 8\sigma^2 d \left(\log s + \log \frac{5 M_1^2 M_2^2}{\delta^{1/d}}\right),
\end{aligned}$$

which concludes the proof. □

Next we show that the sum of first two terms contributes at least $-\Theta(\log s)$ towards the smallest eigenvalue of V_s.

Lemma 17. *For any $u \in S^{d-1}$ it holds w.p at least $1-\delta$ for every $s \geq 1$:*

$$u^\top \sum_{i=0}^{s-1} y_i y_i^\top u + u^\top \sum_{i=0}^{s-2} \left(y_i w_{i+1}^\top + w_{i+1} y_i^\top\right) u \geq -8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right) - 1$$

Proof. First observe that the LHS can be rewritten as:

$$u^\top \sum_{i=0}^{s-1} y_i y_i^\top u + 2u^\top \sum_{i=0}^{s-2} y_i w_{i+1}^\top u = u^\top P u + 2u^\top Q u,$$

where $P = \sum_{i=0}^{s-1} y_i y_i^\top$ and $Q = \sum_{i=0}^{s-2} y_i w_{i+1}^\top$ By Lemma 16 we have:

$$\|Q\|_{(P+I)^{-1}} \leq \sqrt{8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right)}$$

Denote by $u^\top (P+I) u = \alpha^2$. Then we have by Lemma 14:

$$\left\|u^\top Q\right\| \leq \alpha \sqrt{8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right)}$$

Hence:

$$\begin{aligned}
u^\top P u + 2u^\top Q u &= u^\top (P+I) u + 2u^\top Q u - 1 \\
&\geq \alpha^2 - 2\left\|u^\top Q\right\| \|u\| - 1 \\
&= \alpha^2 - 2\left\|u^\top Q\right\| - 1 \\
&\geq \alpha^2 - 2\alpha \sqrt{8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right)} - 1
\end{aligned}$$

The last expression is quadratic function in α which attains its minimum at

$$\alpha = \sqrt{8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right)}.$$

Plugging this expression for α we arrive at:

$$u^\top P u + 2u^\top Q u \geq -8\sigma^2 d \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d}}\right) - 1.$$

□

Next theorem tells us how to bound the smallest singular value of a matrix which rows are independent Gaussian vectors. We will use this theorem to first show that that the last term of V_s decomposition is lower bounded by $\Omega(s)I$ in Corollary 5. We then join this result with result from Lemma 17 to obtain $V_s \succeq \Omega(s)I$ in Proposition 4.

Theorem 8 (Corollary 5.35 in Vershynin (2010))**.** *Let W be a $s \times d$ matrix, whose rows are independent $\mathcal{N}(0, I)$ random vectors in $\mathbb{R}^d$. Then for every $t \geq 0$ with probability at least $1 - e^{-\frac{t^2}{2}}$ it holds:*

$$\sqrt{s} - \sqrt{d} - t \leq \sigma_d(W)$$

Corollary 5. *Let $(w_i)_{i \geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$. Then for every $s \geq 1$ it holds w.p. at least $1-\delta$:*

$$\sum_{i=1}^{s-1} w_i w_i^\top \succeq \sigma^2 \left(\sqrt{s-1} - \sqrt{d} - \sqrt{2\log \frac{1}{\delta}}\right)^2 I$$

Proof. First observe $\sum_{i=1}^{k-1} w_i w_i^\top = \sigma^2 W^\top W$, where

$$W^\top = \begin{pmatrix} \frac{1}{\sigma} w_1 & \frac{1}{\sigma} w_2 & \cdots & \frac{1}{\sigma} w_{s-1} \end{pmatrix} \in \mathbb{R}^{d \times (k-1)}.$$

From Theorem 8 it follows $\sigma_d(W) \geq \sqrt{s-1} - \sqrt{d} - \sqrt{2 \log \frac{1}{\delta}}$, which implies $\sigma_d(W^\top W) \geq \left(\sqrt{s-1} - \sqrt{d} - \sqrt{2 \log \frac{1}{\delta}}\right)^2$. □

Proposition 4. *Let us be in the setting of this section. Then it holds for all $s \geq 1$ w.p. at least $1 - \delta$:*

$$V_s \succeq \sigma^2 \left(\left(\sqrt{s} - \sqrt{d} - \sqrt{2 \log \frac{2}{\delta}} \right)^2 - 8d \left(\log s + \log \frac{5 M_1^2 M_2^2}{\delta^{1/d}} \right) - 1 \right) I$$

Proof. By Lemma 15 we have:

$$V_s = \sum_{i=0}^{s-2} y_i y_i^\top + \sum_{i=0}^{s-2} \left(y_i w_{i+1}^\top + w_{i+1} y_i^\top \right) + \sum_{i=0}^{s-2} w_{i+1} w_{i+1}^\top$$

Let $u \in S^{d-1}$ be arbitrary. We will now lower bound $u^\top V_k u$:

$$u^\top V_s u = \underbrace{u^\top \sum_{i=0}^{s-2} y_i y_i^\top u + 2 u^\top \sum_{i=0}^{s-2} y_i w_{i+1}^\top u}_{\text{Part 1}} + \underbrace{u^\top \sum_{i=0}^{s-2} w_i w_i^\top u}_{\text{Part 2}}.$$

By Lemma 17 part 1 is lower bounded by $-8\sigma^2 d \left(\log s + \log \frac{5 M_1^2 M_2^2 2^{1/d}}{\delta^{1/d}} \right) - 1$ w.p. at least $1 - \frac{\delta}{2}$. By Corollary 5 the part 2 term is lower bounded w.p. at least $1 - \frac{\delta}{2}$ by $\sigma^2 \left(\sqrt{s-1} - \sqrt{d} - \sqrt{2 \log \frac{2}{\delta}} \right)^2$. Using union bound we obtain that w.p. at least $1 - \delta$ it holds:

$$u^\top V_s u \geq \sigma^2 \left(\sqrt{s-1} - \sqrt{d} - \sqrt{2 \log \frac{2}{\delta}} \right)^2 - 8\sigma^2 d \left(\log s + \log \frac{5 M_1^2 M_2^2 2^{1/d}}{\delta^{1/d}} \right) - 1.$$

Since $u \in S^{d-1}$ was arbitrary we obtain that w.p. at least $1 - \delta$ it holds:

$$V_s \succeq \sigma^2 \left(\left(\sqrt{s-1} - \sqrt{d} - \sqrt{2 \log \frac{2}{\delta}} \right)^2 - 8d \left(\log s + \log \frac{5 M_1^2 M_2^2 2^{1/d}}{\delta^{1/d}} \right) - 1 \right) I,$$

which concludes the proof. □

Since by Proposition 4 we have $\sigma_d(V_s) \geq \mathcal{O}(s)$ it also holds: $\sigma_d(V_s + \lambda I) \geq \sigma_d(V_s) \geq \mathcal{O}(s)$. Now the proof of Theorem 3 easily follows. By application of Lemma 3 with $\epsilon = \frac{1}{2}$ we further obtain that it holds w.p. at least $1 - \delta$:

$$\begin{aligned} \left\| (V_s + \lambda I)^{-\frac{1}{2}} S_s \right\|^2 &\leq 8\sigma^2 \log \left(\frac{\det \left(\sum_{i=0}^{s-1} x_i x_i^\top + \lambda I \right)}{\det(\lambda I)} \frac{5^d}{\delta} \right) \\ &\leq 8\sigma^2 \log \left(\frac{((s-1)M_1 + \lambda)^d}{\lambda^d} \frac{5^d}{\delta} \right) \\ &= 8\sigma^2 d \left(\log \frac{(s-1)M_1 + \lambda}{\lambda} + \log \frac{5}{\delta^{1/d}} \right) \end{aligned}$$

Using union bound we obtain that it holds w.p. at least $1 - 2\delta$:

$$\|A_s - A_*\| \leq \frac{8d\left(\log\frac{(s-1)M_1+\lambda}{\lambda} + \log\frac{5}{\delta^{1/d}}\right)}{\sqrt{\left(\left(\sqrt{s-1} - \sqrt{d} - \sqrt{2\log\frac{2}{\delta}}\right)^2 - 8d\left(\log s + \log\frac{5M_1^2M_2^2 2^{1/d}}{\delta^{1/d}}\right) - 1\right)}} + \frac{\lambda\|A_*\|}{\sigma^2\left(\left(\sqrt{s-1} - \sqrt{d} - \sqrt{2\log\frac{2}{\delta}}\right)^2 - 8d\left(\log s + \log\frac{5M_1^2M_2^2 2^{1/d}}{\delta^{1/d}}\right) - 1\right)}$$

Hence we have established that $\|A_s - A_*\| \leq \frac{\mathcal{O}(1)(d\log s + \log\frac{1}{\delta})}{\sqrt{s}}$. The same analysis as in the proof of Theorem 1 then shows that in this setting eXploration finishes in constant (in T) time.

F Additional experiments

We will first run additional experiments on system (8) and then present the behavior of some controllers also on larger and more explosive system. Let us first show the similarity in performance of CEC and robust controller.

We empirically observe that if we use CEC, Phase I usually ends a bit faster and the norm of states are comparable to the case when we use robust controller. However sometimes, as is also the case in one among 20 runs in the experiment shown in tables below, the states magnitude at the beginning increase much more compared to the case of robust control.

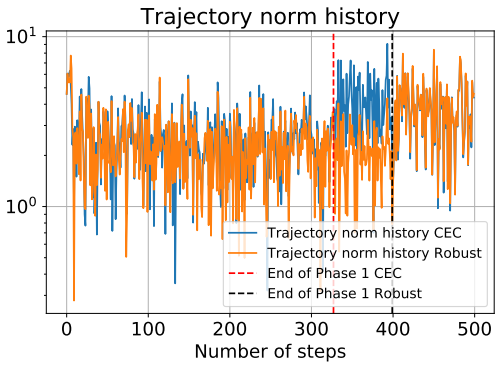


Figure 3: Comparison of CEC and robust controller

In the following tables we present the results of Table 1 more in depth. We analyze 3 variables - steps taken (S), total cost (C) and per step cost (PsC) until we end Phase I in different settings. We present average, standard deviation and median of variables S and C and average of PsC. The *Steps Taken* and *Cost* from Table 1 are here S_{avg} and C_{avg}. Every setting we run for 20 times with the same collection of random seeds. We present two cases with either true estimation error upper bounds or data-dependent ones. For the data-dependent upper bounds we use Corollary 2 which also holds for the multi episodic setting. To compute the estimators in multi episodic setting we use whole trajectories from the episodes and not just the last ones, for which Dean et al. (2019) derived upper bounds.

	Rollout length	S_{avg}	S_{std}	S_{med}	C_{avg}	C_{std}	C_{med}	PsC_{avg}
Multi traj.	6	60	22	54	$1.47 \cdot 10^3$	$7.57 \cdot 10^2$	$1.30 \cdot 10^3$	$2.39 \cdot 10^1$
	10	62	30	55	$2.23 \cdot 10^3$	$1.49 \cdot 10^3$	$1.78 \cdot 10^3$	$3.39 \cdot 10^1$
	15	64	25	67	$3.45 \cdot 10^3$	$1.62 \cdot 10^3$	$3.24 \cdot 10^3$	$5.53 \cdot 10^1$
	20	78	39	65	$5.87 \cdot 10^3$	$3.70 \cdot 10^3$	$5.06 \cdot 10^3$	$6.98 \cdot 10^1$

Table 2: Phase I statistics in Multi episodic setting with true estimation errors

	Rollout length	S_{avg}	S_{std}	S_{med}	C_{avg}	C_{std}	C_{med}	PsC_{avg}
Multi traj.	6	161	25	158	$3.81 \cdot 10^3$	$7.81 \cdot 10^2$	$3.57 \cdot 10^3$	$2.38 \cdot 10^1$
	10	171	23	159	$6.47 \cdot 10^3$	$8.01 \cdot 10^2$	$6.57 \cdot 10^3$	$3.84 \cdot 10^1$
	15	202	42	210	$1.11 \cdot 10^4$	$3.03 \cdot 10^3$	$1.03 \cdot 10^4$	$5.56 \cdot 10^1$
	20	224	39	221	$1.66 \cdot 10^4$	$3.26 \cdot 10^3$	$1.57 \cdot 10^4$	$7.55 \cdot 10^1$

Table 3: Phase I statistics in Multi episodic setting with data-dependent estimation error bounds

As expected the per step cost PsC_{avg} is not much influenced in multi episodic setting whether we use true estimation (not available in practice) errors or looser data-dependent ones. At the same time for $K_i = 0$ in single trajectory setting using data-dependent upper bounds and running Phase I a bit longer significantly increases PcS_{avg} since we let the system grow exponentially for a bit longer time. At the same time we observe that using any among controllers CEC, robust controller or mixed controller with either CEC (MCEC) or robust one (MRobust) the PsC_{avg} is not much influenced. We further observe that using controller MCEC or MRobust significantly reduces the number of steps taken in Phase I compared to the use of pure CEC or robust controller.

	Controller	S_{avg}	S_{std}	S_{med}	C_{avg}	C_{std}	C_{med}	PsC_{avg}
Single traj.	$K_i = 0$	31	10	28	$4.56 \cdot 10^3$	$5.06 \cdot 10^3$	$2.94 \cdot 10^3$	$1.31 \cdot 10^2$
	NegCEC	13	2	12	$8.00 \cdot 10^5$	$1.02 \cdot 10^6$	$3.21 \cdot 10^5$	$6.30 \cdot 10^4$
	CEC	21	11	18	$7.55 \cdot 10^4$	$3.23 \cdot 10^5$	$5.35 \cdot 10^2$	$8.36 \cdot 10^3$
	Robust	20	19	14	$2.47 \cdot 10^3$	$4.83 \cdot 10^3$	$7.71 \cdot 10^2$	$2.78 \cdot 10^2$
	MCEC, $M = 10$	16	5	15	$2.04 \cdot 10^3$	$7.30 \cdot 10^2$	$1.76 \cdot 10^3$	$1.28 \cdot 10^2$
	MRobust, $M = 10$	13	2	14	$1.98 \cdot 10^3$	$8.19 \cdot 10^2$	$1.65 \cdot 10^3$	$1.58 \cdot 10^2$

Table 4: Phase I statistics in Single-Trajectory setting with true estimation errors

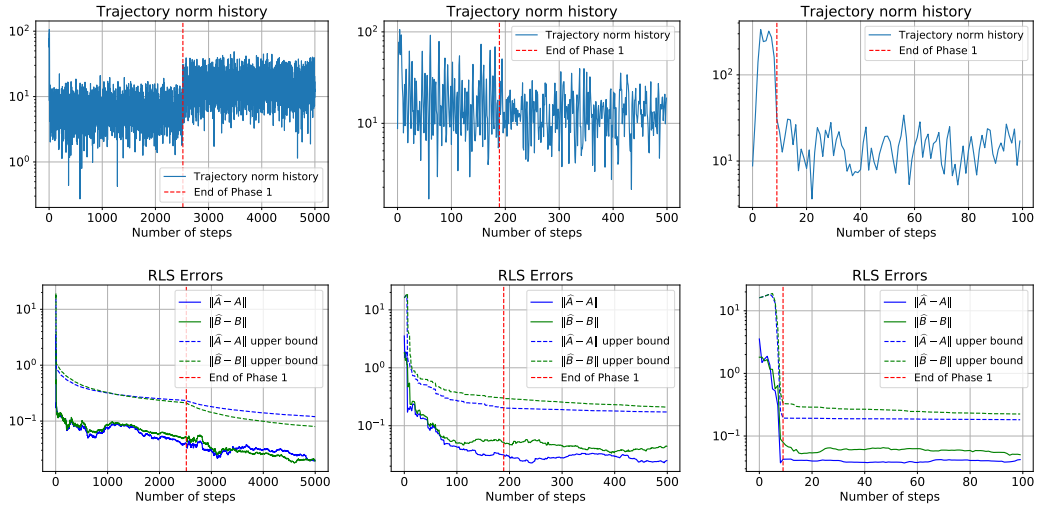
	Controller	S_{avg}	S_{std}	S_{med}	C_{avg}	C_{std}	C_{med}	PsC_{avg}
Single traj.	$K_i = 0$	268	39	262	$2.90 \cdot 10^9$	$9.47 \cdot 10^9$	$2.21 \cdot 10^8$	$8.65 \cdot 10^6$
	NegCEC	18	2	18	$2.51 \cdot 10^8$	$5.47 \cdot 10^8$	$4.73 \cdot 10^7$	$1.36 \cdot 10^7$
	CEC	288	75	313	$7.96 \cdot 10^4$	$3.26 \cdot 10^5$	$4.34 \cdot 10^3$	$6.26 \cdot 10^3$
	Robust	375	39	396	$7.38 \cdot 10^3$	$4.09 \cdot 10^3$	$6.27 \cdot 10^3$	$2.05 \cdot 10^1$
	MCEC, $M = 10$	65	18	60	$8.61 \cdot 10^3$	$2.22 \cdot 10^3$	$8.32 \cdot 10^3$	$1.33 \cdot 10^2$
	MRobust, $M = 10$	44	9	43	$6.12 \cdot 10^3$	$1.08 \cdot 10^3$	$5.75 \cdot 10^3$	$1.42 \cdot 10^2$

Table 5: Phase I statistics in Single-Trajectory setting with data-dependent estimation error bounds

In the following we test the influence of margin M on the controller MCEC on a bit larger and more explosive system:

$$A_* = \begin{pmatrix} 1.5 & 1.0 & 0.4 & 2.3 \\ 0.0 & 1.3 & 1.3 & 1.1 \\ 0.0 & 0.0 & 1.0 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.8 \end{pmatrix}, \quad B_* = \begin{pmatrix} 0.6 & 0.7 & 0.3 \\ 0.8 & 1.1 & 1.1 \\ 1.2 & 0.2 & 2.3 \\ 2.1 & 0.4 & 0.4 \end{pmatrix}, \quad R = I, \quad Q = I, \tag{17}$$

with $\sigma^2 = \sigma_u^2 = 1$.



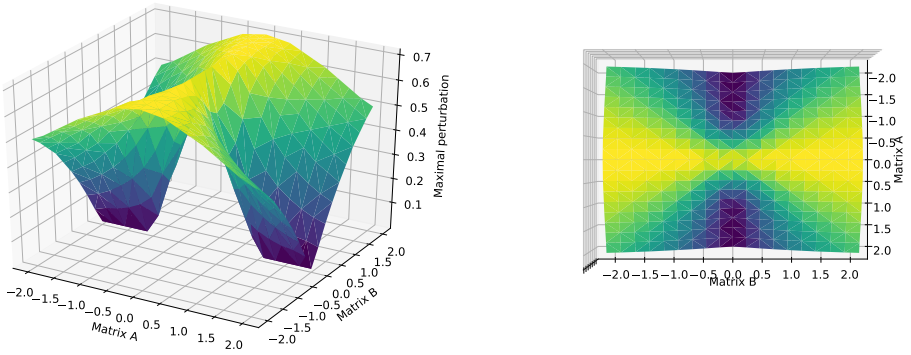
(a) $M = 0$ (b) $M = 10$ (c) $M = 100$

Figure 4: Comparison of different choices for M. With larger M Phase I finishes faster, however the norm of the states can grow more during that time.

In Figure 4 we compare MCEC controllers with different margins M. In case $M = 0$, MCEC is in fact the same as CEC. We observe that with larger M we find controller faster at the cost of larger

state sizes. It is then up to the controller designer to decide how large M can he tolerate. Note that for many systems the size of the norm state is hard constraint and we do not want to breach it. Is such a case it is probably the best to use either pure robust or CEC controller during the eXploration phase.

To prove that controller stabilizes the underlying system we need to solve SDP (5). With the following experiment we would like to give some intuition on how small the estimation error needs to be in order for SDP (5) to be feasible. For that consider the case when state and action are one dimensional. For every $\widehat{A}, \widehat{B}$ we we search for the largest ε such that the SDP (5) is feasible with $\varepsilon_A = \varepsilon_B = \varepsilon$. In other words we search for the largest perturbation ε for which SDP (5) finds a controller K which stabilizes all the systems in the set $\{(A, B) | \|\widehat{A} - A\| \leq \varepsilon, \|\widehat{B} - B\| \leq \varepsilon\}$. In Figure 5 we show the plot $\varepsilon(\widehat{A}, \widehat{B})$.



(a) Perturbation analysis from side

(b) Perturbation analysis from top

Figure 5: The size of the perturbation ε for which SDP (5) still finds a stabilizing controller.