Proceedings of Machine Learning Research vol 144:1–38, 2021

Learning Stabilizing Controllers for Unstable Linear Quadratic Regulators from a Single Trajectory

Lenart Treven TREVENL@ETHZ.CH
Sebastian Curi SCURI@INF.ETHZ.CH
Mojmír Mutný MMUTNY@INF.ETHZ.CH
Andreas Krause KRAUSEA@ETHZ.CH
ETH Zürich

Abstract

The principal task to control dynamical systems is to ensure their stability. When the system is unknown, robust approaches are promising since they aim to stabilize a large set of plausible systems simultaneously. We study linear controllers under quadratic costs model also known as linear quadratic regulators (LQR). We present two different semi-definite programs (SDP) which results in a controller that stabilizes all systems within an ellipsoid uncertainty set. We further show that the feasibility conditions of the proposed SDPs are *equivalent*. Using the derived robust controller syntheses, we propose an efficient data dependent algorithm – eXPLORATION – that with high probability quickly identifies a stabilizing controller. Our approach can be used to initialize existing algorithms that require a stabilizing controller as an input while adding constant to the regret. We further propose different heuristics which empirically reduce the number of steps taken by eXPLORATION and reduce the suffered cost while searching for a stabilizing controller.

Keywords: LQR, stabilizing controller, ellipsoid credibility region

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 finding a stabilizing or optimal controller 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 the exploration costs in unstable systems 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 hand, *unstable* systems that cannot be reset must be stabilized online from a *single* trajectory. After stabilization, there are different efficient algorithms that find an optimal controller (Simchowitz and Foster, 2020; Cohen et al., 2019; Abeille and Lazaric, 2020). Crucially, the algorithms that find an optimal controller require an initial stabilizing controller. This

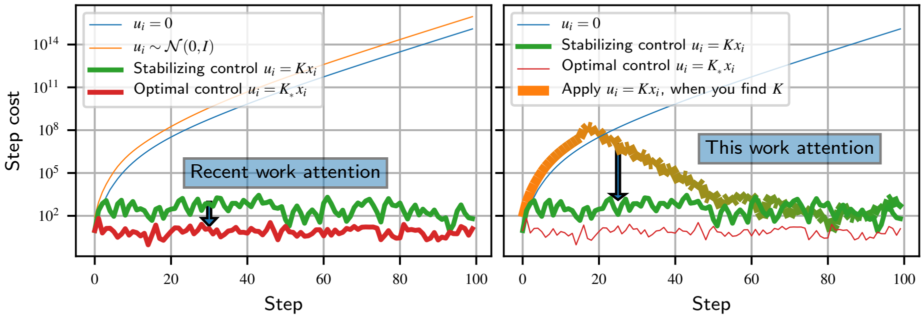

Figure 1: While recent work attention was mostly focused on how to adaptively progress from a given stabilizing controller to an optimal one, this work attention is how to find a stabilizing controller in the single-trajectory setting.

privileged information is essential to ensure that unstable systems do not "explode". However, such prior knowledge is not always available.

In this work, we address the problem of finding a stabilizing controller for a linear dynamical system in a single online trajectory. On the left plot of Figure 1, we show the difference in cost between finding a stabilizing controller and an optimal one. When the true system is unstable, and no knowledge of a stabilizing controller is available, the system costs grow exponentially fast.

Contributions We extend the robust formulation of Dean et al. (2019) from stabilizing all systems within some spectral norm around estimates, to the more general case when the synthesized controller stabilizes all systems within an ellipsoidal uncertainty set. We extend the obtained result to our main contribution where we prove equivalence between common robust controller synthesis algorithms. In particular we show that the controller synthesis of Umenberger et al. (2019), which also tries to stabilize systems inside ellipsoid around estimates, and the derived SLS synthesis for ellipsoids share the same feasibility region: when one algorithm finds a stabilizing controller, so does the other one. Using the proposed robust controller syntheses applied to the ellipsoidal regions obtained from the Bayesian setting we propose an algorithm EXPLORATION. The vanilla version synthesizes a stabilizing controller for the true underlying system in finite time with high probability. The vanilla EXPLORATION approach probes the unknown system with zero-mean Gaussian actions. Additionally, we empirically show that with the robustly motivated choices of system probing which are different from zero-mean Gaussian, we reduce the length of the EXPLORATION and *substantially* lower the total suffered cost. We demonstrate the practicality of our method on the standard common benchmark problems.

1.1. Related Work

Linear dynamical systems have been extensively studied in control theory (cf., Zhou et al. (1996)), nevertheless the interest reemerged in the Machine Learning community after the seminal work of Abbasi-Yadkori and Szepesvári (2011). Cohen et al. (2019); Mania et al. (2019); Simchowitz and Foster (2020); Abeille and Lazaric (2020); Lale et al. (2020); Faradonbeh et al. (2020) and many others show that approaches relying on the optimism in the face of uncertainty (OFU) principle or certainty equivalence principle achieve $\mathcal{O}(\sqrt{T})$ regret in the online single trajectory setting with stochastic disturbances. However, most of the algorithms in this setting require the knowledge of an *initial stabilizing controller*, which might not always be available.

System Identification The first step in EXPLORATION is to *learn* the true system matrices A_*, B_* using System Identification tools. Simchowitz et al. (2018) show that the ordinary least squares (OLS) estimator attains a near optimal error rate $\mathcal{O}(1/\sqrt{i})$, where i is the number of steps in the single trajectory setting for the case when $\rho(A_*) \leq 1$[1]. They further argue that more unstable systems are easier to estimate and prove exponential error decay for one dimensional unstable systems. Faradonbeh et al. (2018a) tackle more challenging general systems with eigenvalues everywhere but on the unit circle. They show that in this case for *regular*[2] systems the ordinary least squares (OLS) estimator is consistent. Further, Sarkar and Rakhlin (2019) extend the OLS consistency to general regular systems. They show that the estimation error scales as $\mathcal{O}(1/\sqrt{i})$. Umenberger et al. (2019) show that the OLS is the same as the maximum (Gaussian) likelihood estimator. They further introduce an ellipsoid region around the estimates where the system lies with high confidence. We extend their idea to the Bayesian setting, where we assume a Gaussian prior on the system parameters, and show that, in this case, the maximum a-posteriori estimator is equivalent to the regularized least squares (RLS) estimator. Applying the analysis of Sarkar and Rakhlin (2019) we show that the associated data dependent high probability credibility regions are consistent.

Controller Synthesis The second step in EXPLORATION is to synthesize a stabilizing controller for all systems in the credibility region that the System Identification step outputs. Dean et al. (2019) derive a robust semi-definite program based on system level synthesis (SLS) whose solution results in a stabilizing controller. They use a multi-trajectory setting to build 2-ball confidence regions around the estimates. We extend their algorithm to a tighter ellipsoidal region around the estimates. It turns out that the SLS synthesis with an ellipsoidal region finds a stabilizing controller as the robust LQR synthesis proposed by Umenberger et al. (2019). Faradonbeh et al. (2018b) propose a non-robust strategy and rely on well-known stability bounds of LQRs (Safonov and Athans, 1977). The main contribution is that they identify the system in closed-loop by sampling different controllers from a Gaussian distribution so that they avoid irregularity of the closed loop matrix a.s. The main practical limitation is that one needs to specify the running time of the algorithm a-priori using unknown system parameters. On the other hand EXPLORATION provably terminates in finite time solving a convex SDP without specifying an a-priori termination time.

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 $x_i \in \mathbb{R}^{d_x}$ are states, $u_i \in \mathbb{R}^{d_u}$ actions and $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$ unobserved Gaussian noise in $\mathbb{R}^{d_x}$. The matrices $A_* \in \mathbb{R}^{d_x \times d_x}, B_* \in \mathbb{R}^{d_x \times d_u}$ are unknown transition matrices. We sample actions u_i from a policy π, which at every step i maps the current history $((x_j)_{j\leq i}, (u_j)_{j<i})$ to a distribution over actions. We assume that the system is *stabilizable*, which means that there exists a matrix $K \in \mathbb{R}^{d_u \times d_x}$ such that $\rho(A_* + B_* K) < 1$. At step i, we incur a cost c_i given by

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

where $Q \in \mathbb{R}^{d_x \times d_x}, R \in \mathbb{R}^{d_u \times d_u}$ are known positive definite matrices.

[1] Spectral radius of matrix A is defined as $\rho(A) = \max\{|\lambda| \,|\exists v : Av = \lambda v\}$

[2] System is regular if every eigenvalue μ of A_* with $|\mu| > 1$, has geometric multiplicity equal to 1.

When the system matrices A_*, B_* are known, the optimal solution in the infinite horizon setting 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 the *discrete algebraic Ricatti equation* of the system, $P = DARE(A_*, B_*, Q, R)$.

While most of the recent work focuses on finding the optimal controller K_* suffering the least possible cummulative cost, they require the knowledge of an initial *stabilizing* controller K_0. In this work, we focus on **finding a stabilizing controller in the single-trajectory setting**. If the system is unstable, the difference between using a stabilizing controller or not using a stabilizing controller results in an *exponential* difference in the suffered cost due to blow-up of the system. On the other hand, the difference between using a stabilizing controller and the optimal one results in a *linear* difference in the suffered cost as summarized in Figure 1. Consequently, it is very desirable that a stabilizing controller is found quickly.

3. Identifying A Stabilizing Controller

As we are trying to stabilize the dynamical system without knowing anything non-trivial about it, certain blow-up of the state from zero is inevitable. Since the state magnitude increases exponentially fast during the blow-up, it is essential that this period is kept very short. There are conceptually two variables we can influence as algorithm designers, a) what control signal we input to the system what we refer to as probing and b) stopping rule, which determines when we should stop probing and sufficient information about the system has been gathered to construct a stabilizing controller.

In this work we focus on the latter and derive a *data-dependent stopping rule* based on feasibility of a semi-definite program. Our method is versatile and can be combined with any control inputs and we show that it terminates in finite time under zero-mean Gaussian control inputs. Our formalism is derived under a general assumption that we can construct estimates of the system matrices and ellipsoidal sets which with high probability contain the true system matrices or they serve as a surrogate for this task as is the case with the Bayesian approach. More specifically, we derive our results under the assumption that after playing a policy π for i steps we have estimates $(\widehat{A}_i, \widehat{B}_i)$ of the system (A_*, B_*) and ellipsoid

$$\Theta_i = \{(A, B) | \Delta^\top D_i \Delta \preceq I, \Delta^\top = (A, B) - (\widehat{A}_i, \widehat{B}_i)\} \tag{3}$$

around estimates for which we believe that $(A_*, B_*) \in \Theta_i$. Here D_i is a data-dependent positive definite matrix. We present an example of how to construct such ellipsoidal region in Section 3.4.

In the following two subsections, we derive two different robust synthesis algorithms for uncertainty sets in the ellipsoidal region (3). In Section 3.3, prove that these two seemingly different approaches are actually equivalent in terms of robust stability.

3.1. Robust System Level Synthesis (SLS)

Our first stopping rule is based on a relaxation stemming from the SLS framework (Wang et al., 2019). In particular, we extend the work of Dean et al. (2019) to ellipsoidal regions (3). Dean et al. (2019) show that a controller K stabilizes *all* systems $(A, B) \in \Theta_i$ if for every $(A\ B) \in \Theta_i$ we have:

$$\left\| \Delta^\top \begin{pmatrix} I \\ K \end{pmatrix} \left(zI - \widehat{A}_i - \widehat{B}_i K \right)^{-1} \right\|_{\mathcal{H}_\infty} < 1, \text{ and } \rho\left(\widehat{A}_i + \widehat{B}_i K \right) < 1, \tag{4}$$

where $\Delta^\top = (A\ B) - (\widehat{A}_i\ \widehat{B}_i)$. The $\mathcal{H}_\infty$-norm for a function $f : \mathbb{C} \to \mathbb{C}^{d\times d}$ is defined as $\|f\|_{\mathcal{H}_\infty} = \sup_{\|z\|=1} \|f(z)\|_2$. With the current formulation, we need to ensure that $\mathcal{H}_\infty$-norm constraint in (4) holds for every $(A, B) \in \Theta_i$. The main difference with Dean et al. (2019) is that we apply the S-Lemma of Luo et al. (2004) to obtain an equivalent formulation with a single $\mathcal{H}_\infty$-norm constraint using the ellipsoidal region instead of a 2-ball. Next, we transform the $\mathcal{H}_\infty$-norm constraint to a convex semi-definite constraint applying the KYP-Lemma (Bart et al., 2018). The equivalent feasibility problem reads:

$$\min_{X\succ 0, S, t\in(0,1)} 0, \qquad \text{s.t.} \quad \begin{pmatrix} X - I & \widehat{A}_i X + \widehat{B}_i S & 0 \\ (\widehat{A}_i X + \widehat{B}_i S)^\top & X & \begin{pmatrix} X \\ S \end{pmatrix}^\top \\ 0 & \begin{pmatrix} X \\ S \end{pmatrix} & tD_i \end{pmatrix} \succeq 0. \tag{5}$$

The stabilizing controller is extracted from the solution of the Robust SLS (5) as $K = SX^{-1}$. We show the derivation details in the Appendix B of the extended paper (Treven et al., 2020).

3.2. Robust Linear Quadratic Regulator (LQR)

The derivation of our second robust controller synthesis is based on the reformulation of the LQR problem, which finds the optimal infinite horizon controller. This reformulations lends itself to an efficient SDP relaxation (Boyd et al., 1994). We follow the exposition from Cohen et al. (2018), assuming we know matrices A_*, B_* we can obtain the optimal infinite controller K_* by first solving

$$\begin{aligned} &\min_{\Sigma\succeq 0} \operatorname{Tr}\left(\begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \Sigma \right) \\ &\text{s.t.} \quad \Sigma_{xx} \succeq (A_*\ B_*)\Sigma(A_*\ B_*)^\top + \sigma_w^2 I, \end{aligned} \tag{6}$$

and then extracting the optimal controller as $K_* = \Sigma_{ux}\Sigma_{xx}^{-1}$. Here $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix}$, where $\Sigma_{xx} \in \mathbb{R}^{d_x\times d_x}$ and $\Sigma_{uu} \in \mathbb{R}^{d_u\times d_u}$, represents the joint covariance matrix of the state and action. The derivation with the motivation behind the SDP (6) is given in Appendix C of the extended paper (Treven et al., 2020), where we also show in Theorem 5 that the semi-definite constraint in the SDP (6) ensures that the controller synthesized as $K = \Sigma_{ux}\Sigma_{xx}^{-1}$ stabilizes the system A_*, B_*. Inspired by Umenberger et al. (2019), the robust formulation of the SDP problem (6) is:

$$\begin{aligned} &\min_{\Sigma\succeq 0} \operatorname{Tr}\left(\begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \Sigma \right) \\ &\text{s.t.}\ \forall (A, B) \in \Theta_i :\ \Sigma_{xx} \succeq (A\ B)\Sigma(A\ B)^\top + \sigma_w^2 I. \end{aligned} \tag{7}$$

As in Section 3.1 we have to ensure that one condition has to hold for every system in the ellipsoid Θ_i. Applying the S-Lemma we reformulate problem given by eq. (7) to an equivalent convex SDP:

$$\begin{aligned} &\min_{\Sigma\succeq 0, t\geq 0} \operatorname{Tr}\left(\begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \Sigma \right) \\ &\text{s.t.} \begin{pmatrix} \Sigma_{xx} - (\widehat{A}_i\ \widehat{B}_i)\Sigma(\widehat{A}_i\ \widehat{B}_i)^\top - (t + \sigma_w^2)I & (\widehat{A}_i\ \widehat{B}_i)\Sigma \\ \Sigma(\widehat{A}_i\ \widehat{B}_i)^\top & tD_i - \Sigma \end{pmatrix} \succeq 0. \end{aligned} \tag{8}$$

The stabilizing controller is extracted from the optimal solution as $K = \Sigma_{ux}\Sigma_{xx}^{-1}$.

3.3. Equivalence

At first glance one could think that we have derived two completely different control synthesis procedures, however as we will see this is not the case and the two optimization problems have a feasible solution at the same time.

Theorem 1 *The Robust SLS given by Equation* (5) *has a nonempty solution if and only if the Robust LQR given by Equation* (8) *has a nonempty solution.*

The proof of the theorem is provided in Appendix D of the extended paper (Treven et al., 2020). Despite the fact that the two SDPs (5) and (8) have a feasible solution at the same time, they differ in the objective. The robust SLS (5) is minimizing a constant, whereas the Robust LQR (8) is minimizing the upper bound of the maximal infinite horizon cost of the systems in Θ_i. The specific nature of the objective is not interesting for the stopping rule, however can have practically dramatic impact in downstream tasks. For different possible objective for the SLS synthesis (5) please refer to Section 4.1.

3.4. Example: Bayesian credible sets

In this section, we show a particular design choice of estimators $\widehat{A}_i, \widehat{B}_i$ and region Θ_i which results from the Bayesian setting. Inspired by the work of Umenberger et al. (2019), we place a Gaussian prior[3] on the system matrices A_*, B_* i.e. $\mathrm{vec}(A_*, B_*) \sim \mathcal{N}\left(0, \sigma_w^2/\lambda I\right)$. In this case we have explicit formulas for the posterior distribution of $\mathrm{vec}(A_*, B_*)|(x_j)_{j\leq i}, (u_j)_{j<i}$. As derived in the Appendix A of the extended paper (Treven et al., 2020) it turns out the posterior is also Gaussian and the MAP estimator $\mathrm{vec}(\widehat{A}_i, \widehat{B}_i)$ is exactly the RLS estimator:

$$\widehat{A}_i, \widehat{B}_i = \underset{A,B}{\mathrm{argmin}} \sum_{j=0}^{i-1} \|x_{j+1} - Ax_j - Bu_j\|_2^2 + \lambda \left\|(A\ B)\right\|_F^2 . \tag{9}$$

Moreover, the Bayesian credibility region is $\Theta_i = \{(A\ B)|\Delta^\top D_i \Delta \preceq I, \Delta^\top = (A\ B) - (\widehat{A}_i, \widehat{B}_i)\}$. Here D_i represents the scaled inverse covariance matrix of the posterior distribution and is explicitly given as $D_i = \frac{1}{c_\delta \sigma_w^2}\left(\sum_{j=1}^{i} z_j z_j^\top + \lambda I\right)$, where $z_j^\top = (x_j^\top\, u_j^\top)$ and c_δ is the $(1-\delta)$-quantile of the χ^2 distribution with $d_x(d_x + d_u)$ degrees of freedom. With this definition of Θ_i we have $(A_*\ B_*) \in \Theta_i$ w.p. $1 - \delta$.

4. EXPLORATION Algorithm

We now show how to use the derived results (c.f., Section 3) to *provably* find a robust controller in the Bayesian setting. In the Appendix E of the extended paper (Treven et al., 2020) we show how to initialize the algorithms, specifically OSLO (Cohen et al., 2019) and CEC (Simchowitz and Foster, 2020), which need a stabilizing controller as an input, with the proposed EXPLORATION algorithm.

[3] Regarding the sense of this assumption and cases when this assumption fails look at Section 5.1

Algorithm 1 EXPLORATION

```
Input: $x_0 = 0, \lambda, \delta$
for $i = 1, \dots$ do
    /* Probing Signal /*
    Play $u_i \sim \pi(\cdot|x_{1:i}, u_{1:i-1})$ and observe state $x_{i+1}$.
    /* Stopping Rule /*
    Build a confidence region $\Theta_i$, such that $(A_*\ B_*) \in \Theta_i$ w.p. $1 - \delta$. (c.f. Section 3.4)
    Solve a robust controller synthesis $\forall (A, B) \in \Theta_i$. (c.f. Section 3.1 or Section 3.2)
    if a controller is found return stabilizing controller $K_0$
end for
```

The learner explores using a policy π that only depends on the past states and inputs. Using the collected data, it builds an empirical estimate and a confidence region around it. Finally, it attempts to solve a robust controller synthesis problem. If it fails, the algorithm continues. If it succeeds, the algorithm terminates and returns a provably stabilizing controller for the *true* underlying system.

The credibility regions must contain the true system with probability $1 - \delta$ only at the time i^* in which a controller is found and not uniformly over all time steps. This is crucial as it allows us to use tight credibility regions. Then, with probability δ, the algorithm might fail to return a stabilizing controller, and it will return a stabilizing controller with probability $1 - \delta$.

In this section, we analyze the VANILLA EXPLORATION variant, in which the policy is to choose independent zero-mean Gaussian action, i.e., $u_i \sim \pi(\cdot|x_{1:i}, u_{1:i-1}) = \mathcal{N}(0, \sigma_u^2 I)$. Next, we prove that VANILLA EXPLORATION finishes in $\widetilde{O}(1)$ time. In Section 4.1, we discuss different probing heuristics that perform well in practice but where we lose the finte time termination guarantee. Nevertheless, the algorithm still remains valid: if it terminates, the resulting controller provably stabilizes the system.

Theorem 2 *Assuming the aforementioned setting, then with probability* $1 - \delta$ VANILLA EXPLORATION *returns a stabilizing controller for* (A_*, B_*) *in time:*

$$\widetilde{\mathcal{O}}\left(\text{polylog}(\delta)(1 + \|K\|_2)^2 \left\|(zI - A_* - B_*K)^{-1}\right\|_{\mathcal{H}_\infty}^2\right), \tag{10}$$

where K *is any stabilizing static controller.*

Proof Sketch: First, note that since we assume that entries of A_*, B_* are sampled from independent Gaussian, system (A_*, B_*) is stabilizable a.s. If K is a stabilizing controller we derive in the Appendix F of the extended paper (Treven et al., 2020), extending the results of Dean et al. (2019), that the SLS synthesis (5) is feasible if

$$\mathcal{O}\left((1 + \|K\|_2)^2 \left\|(zI - A_* - B_*K)^{-1}\right\|_{\mathcal{H}_\infty}^2\right) \leq \lambda_{min}(D). \tag{11}$$

To obtain the best bound we choose a stabilizing controller K such that the left hand side of Equation (11) is minimized. From Equation (11) follows that as soon as the smallest eigenvalue of the matrix D is large enough, the robust synthesis will be feasible. At the same time from the analysis of Sarkar and Rakhlin (2019) follows that $\widetilde{\Omega}(i)I \preceq D_i$ for every regular system. Again, since we assume a Gaussian prior on (A_*, B_*), the system is regular a.s. Assembling the pieces together we arrive at the result, for which we provide more detailed proof in the Appendix F of the extended paper (Treven et al., 2020).

4.1. Different probing signals with EXPLORATION

The VANILLA EXPLORATION approach takes random actions $u_i \sim \mathcal{N}(0, \sigma_u^2 I)$. For such a choice we can guarantee that Algorithm 1 terminates after constant time, depending only on the system parameters. However, as we demonstrate in our experiments (c.f., Appendix H of the extended paper (Treven et al., 2020)), the states 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 Algorithm 1 will terminate. However, the data dependent credibility region on estimation errors from section 3.4 (and thus the validity of the stopping condition) is still valid and we can run Algorithm 1. With data dependent inputs, we cannot guarantee that the minimum eigenvalue of D_i grows as $\tilde{\Omega}(i)$. Next, we discuss different choices for controller K_i that we study in our experiments.

CEC As first possibility, we act as if the estimators $\widehat{A}_i, \widehat{B}_i$ are the true system matrices and we compute the controller K_i as the optimal controller:

$$K_i = -(R + \widehat{B}_i^\top P \widehat{B}_i)^{-1} \widehat{B}_i^\top P \widehat{A}_i, \tag{12}$$

where $P_i = \mathrm{DARE}(\widehat{A}_i, \widehat{B}_i, Q, R)$, i.e., we act using Certainty Equivalent Control (CEC).

MINMAX For the second K_i we consider controller which minimizes the maximal closed loop norm of the systems in Θ_i. At every time step we synthesize the controller K_i as

$$K_i = \operatorname*{argmin}_K \max_{(A,B) \in \Theta_i} \|A + BK\|_2 \tag{13}$$

The controller defined in Equation (13) can be efficiently computed via a convex SDP. We derive the convex SDP formulation of the min max problem given by Equation (13) in Appendix G of the extended paper (Treven et al., 2020).

RELAXEDSLS As a third alternative we relax the constraint $t \in (0, 1)$ to $t \geq 0$ in the SDP feasibility problem (5), and minimize the value of t, i.e.:

$$\min_{X \succ 0, S, t \geq 0} t \qquad \text{s.t. semi-definite constraint (5)} \tag{14}$$

The controller is then synthesized as $K_i = SX^{-1}$. With such relaxation, the SDP is always feasible. The interpretation of this relaxation is that when $t \geq 1$ we find a controller that stabilizes all systems (A, B) in a *smaller* confidence region around the estimates $(\widehat{A}_i, \widehat{B}_i)$. Furthermore, this algorithms returns a provably stabilizing controller when $t < 1$. Although in principle we could also increase D_i in the Robust LQR synthesis in (8), this requires a tedious exponential line search, whereas the RELAXEDSLS synthesis does this automatically.

5. Experiments

In this section, we critically evaluate the different components of EXPLORATION empirically. In Section 5.1, we investigate when the credibility regions are correct and when do they fail on a fixed system A_*, B_*. In particular, the algorithm fails when the prior parameter λ is too large. To overcome this issue, we suggest a way of selecting the prior parameter λ, given some mild privileged information. In Section 5.2, we compare the time it takes to find a stabilizing controller and the total cost suffered using different probing signals. Although VANILLA EXPLORATION provably terminates, the heuristic variants perform better in practice. In all the considered examples the cost matrices Q and R are equal to the identity matrix of the appropriate dimensions. The scales of unobserved and played noise covariance matrices are $\sigma_w^2 = \sigma_u^2 = 1$. We set the probability of failure to $\delta = 0.1$.

5.1. Data Dependent Credibility Region

To illustrate how EXPLORATION builds the credibility regions, we consider a one dimensional system $A_* = 1.5, B_* = 1.8$. We select $\lambda = \frac{1}{4}$ and $\lambda = 3$ and show consecutive credibility regions in Figure 2. As we can see, the credibility region Θ_i shrinks as we see more data. For both choices of λ, Robust SLS and Robust LQR become feasible after 4 iterations and the algorithm terminates. Crucially, when λ is too large (i.e. too small variance of the prior), it may happen that the true parameters A_*, B_* are not inside the credibility region as we can see on in the middle subfigure of Figure 2.

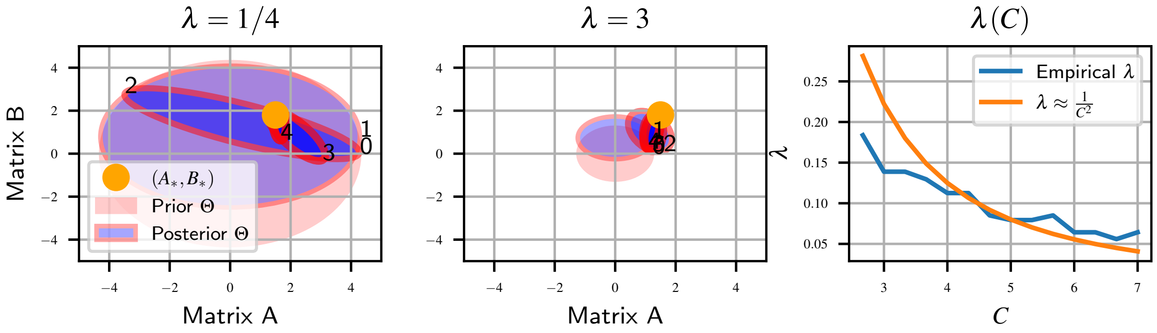

Figure 2: When credibility region Θ_i is small enough SDP (8) finds a controller which stabilizes every system in Θ_i. If we choose a Gaussian prior with too small covariance matrix the credibility regions Θ_i might not contain the true system and the resulting controller will not stabilize it. If we know a bound C on the Frobenious norm of the system, experiments show that a reasonable choice for λ is $\lambda \approx \frac{1}{C^2}$.

This means that selecting λ is a problem-dependent quantity (as any prior). However, if we assume that we have of a constant C such that $\|(A_* \; B_*)\|_F \leq C$, then in the Appendix I of the extended paper (Treven et al., 2020) we suggest that selecting $\lambda \approx 1/C^2$, results in the credibility regions which empirically contain the true system with probability at least $1-\delta$. On the right most subfigure of Figure 2 we plot the largest λ for which we empirically observe that the one dimensional systems with $\|(A_* \; B_*)\| \leq C$ are inside regions Θ_i with empirical probability at least $1 - \delta$.

5.2. EXPLORATION Performance

Next we will illustrate the cost suffered and time until we find a stabilizing controller on a system

$$A_* = \begin{pmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{pmatrix}, \quad B_* = I \tag{15}$$

introduced by Dean et al. (2019), here we use $\lambda = 1$. We show a sample run with different heuristics in the Appendix H of the extended paper (Treven et al., 2020), here we show in Table 1 more in-depth analysis of the EXPLORATION performances on system (15).

During exploration the agent suffers quadratic costs Equation (2) and at the eXploration termination leaves the system in state x_T. We define the total cost of exploration as

$$\text{Cost} = \sum_{i=1}^{T} \left(x_i^\top Q x_i + u_i^\top R u_i \right) + x_T^\top P x_T, \tag{16}$$

where $P = \text{DARE}(A_*, B_*, Q, R)$, and T is the termination time of eXploration. Since the cost can grow exponential with time we will report the logarithm of the Cost.

We compare robust synthesis with ellipsoidal bounds to two benchmarks. The first is robust synthesis with 2-ball estimation bounds of Dean et al. (2019). For the second benchmark we compare robust controller to the CEC which was used as a stabilizing controller in e.g. Faradonbeh et al. (2018b); Simchowitz and Foster (2020). In particular we analyze how large region around estimates they stabilize. For both controllers we use the tightest ellipsoidal bounds. To compute the stopping time when CEC stabilizes all systems inside the ellipsoidal bound we sample 1000 systems from the ellipsoid boundary and if CEC stabilizes all we stop eXploration[4].

Using ellipsoidal compared to 2-ball bounds significantly reduces the number of steps of eXploration. Consequently we also suffer much less exploration cost. CEC naturally stabilizes some region around the estimates, however as we can see on the Table 1 CEC stabilizes smaller region compared to the robust controller, which is paramount in the case when the cost grows exponentially.

Table 1: Using ellipsoidal confidence regions significantly shortens the exploration time and consequently the cost. Robust controller stabilizes larger region than CEC which is crucial when the cost grows exponentially. We report median ± standard deviation.

	Ellipsoidal region		2-ball region		CEC as stopping time	
	Steps	log(Cost)	Steps	log(Cost)	Steps	log(Cost)
Vanilla	$\mathbf{44 \pm 8.2}$	$\mathbf{8.7 \pm 0.58}$	84 ± 18	11 ± 1.1	50 ± 12	9.3 ± 0.73
CEC	$\mathbf{25 \pm 8.5}$	$\mathbf{6.2 \pm 1.1}$	110 ± 32	7.4 ± 0.62	53 ± 14	6.8 ± 0.56
MinMax	$\mathbf{24 \pm 9}$	$\mathbf{7.1 \pm 3.2}$	120 ± 39	7.9 ± 3.1	58 ± 22	7.8 ± 7.9
RelaxedSLS	$\mathbf{26 \pm 8.9}$	$\mathbf{7.2 \pm 3}$	170 ± 86	9.1 ± 2.9	78 ± 36	8.6 ± 3.2

6. Discussion and Conclusions

In Section 3 we presented two seemingly different relaxation techniques for solving Riccati equation under uncertainty. The two relaxations one in Z-transform space and the other convex relaxation in the SDP formulation lead to the same feasible regions. We instantiated our stopping rule with uncertainty regions over the system matrices constructed via Bayesian means, however the stopping rule is more general and can be used with confidence estimates constructed without prior assumptions once we can guarantee uncertainty sets otherwise. To the best of our knowledge, provable anytime adaptive consistent confidence estimates for unstable linear systems are not known, but should these be constructable our stopping rule can be used with them.

[4] Note that it can still happen that CEC does not stabilize all systems inside ellipsoid, however already with this approximation robust controller stabilizes larger ellipsoidal region.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No 815943, as well as from the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v19/abbasi-yadkori11a.html.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *Proceedings of Machine Learning and Systems 2020*, pages 7388–7396. 2020.

MOSEK ApS. *MOSEK Optimizer API for Python 9.2.4*, 2020. URL https://docs.mosek.com/9.2/pythonapi/index.html.

Harm Bart, Sanne ter Horst, André C.M. Ran, and Hugo J. Woerdeman, editors. *Operator Theory, Analysis and the State Space Approach*. Springer International Publishing, 2018. doi: 10.1007/978-3-030-04269-1. URL https://doi.org/10.1007%2F978-3-030-04269-1.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000. ISBN 1886529094.

Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. volume 80 of *Proceedings of Machine Learning Research*, pages 1029–1038, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/cohen18b.html.

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

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, Aug 2019. ISSN 1615-3383. doi: 10.1007/s10208-019-09426-y. URL https://doi.org/10.1007/s10208-019-09426-y.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018a.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64(8):3498–3505, 2018b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.

Wassim M. Haddad, VijaySekhar Chellaboina, and Sergey G. Nersesov. *Thermodynamics: A Dynamical Systems Approach*. Princeton University Press, 2005. ISBN 9780691123271. URL http://www.jstor.org/stable/j.ctt7s1k3.

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

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*, 2020.

Zhi-Quan Luo, Jos F. Sturm, and Shuzhong Zhang. Multivariate nonnegative quadratic mappings. *SIAM Journal on Optimization*, 14(4):1140–1162, 2004. doi: 10.1137/S1052623403421498. URL https://doi.org/10.1137/S1052623403421498.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10154–10164. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9205-certainty-equivalence-is-efficient-for-linear-quadratic-control.pdf.

Fernando Ribeiro, Gil Lopes, Tiago Maia, Hélder Ribeiro, Pedro Osório, Ricardo Roriz, and Nuno Ferreira. Motion control of mobile autonomous robots using non-linear dynamical systems approach. In Paulo Garrido, Filomena Soares, and António Paulo Moreira, editors, *CONTROLO 2016*, pages 409–421, Cham, 2017. Springer International Publishing. ISBN 978-3-319-43671-5.

Michael Safonov and Michael Athans. Gain and phase margin for multiloop lqg regulators. *IEEE Transactions on Automatic Control*, 22(2):173–179, 1977.

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

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

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

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

A. Tornambè, G. Conte, and A.M. Perdon. *Theory and Practice of Control and Systems: Proceedings of the 6th IEEE Mediterranean Conference, Alghero, Sardinia, Italy, 9-11 June 1998*. World Scientific, 1998. ISBN 9789810236687. URL https://books.google.ch/books?id=BkGYGwAACAAJ.

H. Trentelman, A.A. Stoorvogel, and M. Hautus. *Control Theory for Linear Systems*. Communications and Control Engineering. Springer London, 2001. ISBN 9781852333164. URL https://books.google.si/books?id=1KmPMQEACAAJ.

Lenart Treven, Sebastian Curi, Mojmir Mutny, and Andreas Krause. Learning controllers for unstable linear quadratic regulators from a single trajectory. *arXiv preprint arXiv:2006.11022*, 2020.

Jack Umenberger, Mina Ferizbegovic, Thomas B Schön, and Hå kan Hjalmarsson. Robust exploration in linear quadratic reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 15336–15346. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9668-robust-exploration-in-linear-quadratic-reinforcement-learning.pdf.

Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. A system-level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 64(10):4079–4093, 2019.

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