ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Online Learning of Linear-Quadratic Regulators

Master Thesis

Lenart Treven

Monday 28th September, 2020

Advisors: Sebastian Curi, Mojmir Mutny, Prof. Dr. A. Krause
Department of Computer Science, ETH Zürich

Abstract

We present different approaches for learning stabilizing controllers for Linear Quadratic Regulators (LQR) with unknown system matrices. Assuming Gaussian prior over system parameter we derive consistent data dependent estimation error upper bounds. Given estimates of system matrices and a ellipsoid confidence region around them we derive 2 convex semi-definite programs (SDPs) which feasible solutions stabilize all systems in the confidence region. To derive the first SDP we start from a sufficient condition for stabilization obtained from System Level Synthesis (SLS) by Dean et al. (2017) and then sequentially apply robust S-lemma and Kalman-Yakubovich-Popov (KYP) lemma to transform the sufficient condition to a convex SDP. The second SDP we obtain by applying S-lemma to a version of SDP which finds the optimal solution in the case we know system matrices. We further show that the obtained SDPs are equivalent in the sense that as soon as one SDP is feasible the other is feasible as well. Next we introduce an algorithm eXploration which provably finds a stabilizing controller for regular systems in finite time. We show how we can use eXploration as an initialization to algorithms which achieve $\mathcal{O}(\sqrt{T})$ regret but need a stabilizing controller as an input. We further propose different heuristics which try to stabilize the system even before we have a guarantee for that.

Acknowledgments

First and foremost, I would like to thank my advisers Sebastian Curi and Mojmir Mutny for their advice and support throughout my Master Thesis journey. Your ideas, discussions we had and feedback you gave me made the thesis possible. Thank you a lot! Further I would like to thank Prof. Dr. Andreas Krause for help and suggestions on the presentation of ideas.

Next I would like to thank Johannes Bäumler and Metod Jazbec for the various discussion we had about problems I encountered during the thesis work. Your views on the problems gave me plenty of ideas.

Finally, I would like to thank my family for supporting and encouraging me during my studies.

Contents

Notation

Sets and Topology	
$\mathrm{Cl}(X)$	Closure of set X
$\mathrm{Int}(X)$	Interior of set X
$\mathbb{D}$	Unit complex ball i.e. $\{z \in \mathbb{C}; \lvert z\rvert < 1\}$
$\mathbb{N}, \mathbb{R}, \mathbb{C}$	Set of natural, real and complex numbers
∂X	Boundary of set X
$B_p^d(r)$	p-ball in $\mathbb{R}^d$ with radius r defined as $\{v \in \mathbb{R}^d; \lVert v\rVert_p < r\}$
S^{d-1}	Unit sphere in $\mathbb{R}^d$ i.e. $\{v \in \mathbb{R}^d; \lVert v\rVert_2 = 1\}$
Vectors	
$\lVert v\rVert_M$	Norm defined as $\lVert v\rVert_M = \sqrt{v^\top M v}$
$\lVert v\rVert_p$	ℓ_p norm of vector v
$x, u, v, \ldots$	Vectors
Matrices	
$\langle A, B\rangle$	$\mathrm{Tr}(A^\top B)$
$\kappa(A)$	Conditional number of matrix A
$\lambda_i(M)$	i-th largest eigenvalue of symmetric matrix M
$\mathbb{S}_{++}^d(\mathbb{S}_+^d)$	Set of $d \times d$ positive definite (semi-definite) matrices
$\lVert A\rVert_*$	Nuclear norm
$\lVert A\rVert_2$	Spectral norm
$\lVert A\rVert_F$	Frobenious norm
$\lVert A\rVert_M$	Matrix norm defined as $\lVert A\rVert_M = \max_{v \in S^{d-1}} \lVert Av\rVert_M$
$\mathrm{rank}(A)$	Rank of matrix A
$\rho(A)$	Spectral radius of matrix A
$\sigma_i(A)$	i-th largest singular value of matrix A
$\mathrm{Tr}(A)$	Trace of matrix A
$A \prec B$	$B - A \in \mathbb{S}_{++}^d$

$A \preceq B$	$B - A \in \mathbb{S}_+^d$
$A, B, \ldots$	Matrices
$A^\dagger$	Moore–Penrose inverse of matrix A
$A^\top$	Transpose of matrix A
A_*, B_*	System matrices
I_n	Identity matrix of size $n \times n$
Probability	
$\mathbb{E}[X]$	Expected value of X
$\mathbb{P}(A)$	Probability of event A
O notation	
$\mathcal{O}(f)$	$\{g : \mathbb{R} \to \mathbb{R};\ \exists M, x_0,\ \text{s.t. } \forall x \geq x_0 : g(x) \leq Mf(x)\}$
$\Omega(f)$	$\{g : \mathbb{R} \to \mathbb{R};\ \exists M, x_0,\ \text{s.t. } \forall x \geq x_0 : g(x) \geq Mf(x)\}$
$\omega(f)$	$\{g : \mathbb{R} \to \mathbb{R};\ \lim_{x\to\infty} \left\lvert \frac{f(x)}{g(x)} \right\rvert = 0\}$
$\Theta(f)$	$\{g : \mathbb{R} \to \mathbb{R};\ g \in \mathcal{O}(f) \cap \Omega(f)\}$
$o(f)$	$\{g : \mathbb{R} \to \mathbb{R};\ \lim_{x\to\infty} \left\lvert \frac{g(x)}{f(x)} \right\rvert = 0\}$

Abbreviations

CE	Certainty equivalent
CEC	Certainty equivalent controller
LQ	Linear Quadratic
LQR	Linear Quadratic Regulator
OFU	Optimism in the face of uncertainty
OLS	Ordinary Least Squares
OSLO	Optimistic Semi-definite programming for Lq cOntrol
RLS	Regularized Least Squares
SDP	Semi-definite program
SLS	System Level Synthesis

Chapter 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. In this thesis we deal with the question how to find a stabilizing controller with as little prior information as possible. We show that assuming independent Gaussian prior over system parameters is enough to show that we can find a controller which stabilizes the underlying system in constant time.

1.1 Problem setting

Even though in many applications the agent is driven around the space in non linear manner, the setting of Linear Quadratic Regulators assumes linear system dynamics. We assume the system is driven by recursive equation:

$$x_{i+1} = A_* x_i + B_* u_i + w_{i+1}, \quad x_0 = 0 \tag{1.1}$$

where $x_i \in \mathbb{R}^{d_x}$ is state of the system, $u_i \in \mathbb{R}^{d_u}$ is action which we choose at time i and w_i is an independent Gaussian noise – $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$. We denote by $W = \sigma_w^2 I$ and $d = d_x + d_u$. The transition matrices A_* and B_* are of the appropriate dimensions. We further denote by $\mathcal{F}_i = \sigma((w_j)_{j\leq i}, (x_j)_{j\leq i})$ the filtration generated by the noise and states up to time i and assume that $u_i \in \mathcal{F}_i$. At every time step the system incurs a loss $c(x_i, u_i)$ which is defined as

$$c(x_i, u_i) = x_i^\top Q x_i + u_i^\top R u_i,$$

where $Q \in \mathbb{R}^{d_x \times d_x}, R \in \mathbb{R}^{d_u \times d_u}$ are symmetric positive semi-definite matrices. We call any mapping π, which maps at every time step i the history $\{(x_j)_{j\leq i}, (u_j)_{j<i}\}$ to an action u_i, a policy. Playing policy π for s steps the total cost we suffer is equal to:

$$C(s, \pi) = \sum_{i\leq s} c(x_i, u_i).$$

Due to the noise $(w_i)_{i\leq s}$ and potential randomness in action selection, $C(s, \pi)$ is a random variable. We are driven by the idea that we would like to suffer as small cost as possible. Hence we define different measures to asses goodness of a policy. The first measure which comes to our mind is expected s-steps cumulative cost $\mathbb{E}_{\boldsymbol{w},\boldsymbol{u}} C(s, \pi)$. Another measure for policy goodness is an average expected per step cost if we would run the system forever defined as $J(\pi) = \limsup_{s\to\infty} \frac{1}{s} \mathbb{E}_{\boldsymbol{w},\boldsymbol{u}} C(s, \pi)$.

In the setting of the presented work system matrices A_*, B_* are unknown at the beginning. However, we would like to learn policies which, despite limited knowledge of the system, choose such actions that the cost suffered until time s is as small as possible. To compare the goodness of the policy π we will define another measure $R(s, \pi)$ called *regret*:

$$R(s, \pi) = C(s, \pi) - sJ_*, \tag{1.2}$$

where J_* is the minimal infinite horizon expected per step cost attained by any policy, i.e. $J_* = \inf_\pi J(\pi)$. Note again that $R(s, \pi)$ is random variable. We will further limit ourselves to *stabilizable* systems A_*, B_*.

Definition 1.1 *The system A_*, B_* is called* stabilizable *if there exist a matrix $K \in \mathbb{R}^{d_u \times d_x}$ with $\rho(A_* + B_* K) < 1$.*

For stabilizable systems J_* is finite and is achieved by policy π_* which chooses actions as a fixed linear map of states: $u_i = K_* x_i$. We will show this result in section 2.1.2. We say that the policy has *sublinear regret* or *no regret* with probability $1 - \delta$ if the random variable $R(s, \pi) = o(s)$ with probability at least $1 - \delta$. To support the idea why we would like to have a policy with no regret we state next lemma.

Lemma 1.2 *If $R(s, \pi) = o(s)$, then $J(\pi) = J_*$.*

Proof Since $R(s, \pi) = o(s)$ we have

$$\lim_{s \to \infty} \frac{1}{s} R(s, \pi) = 0.$$

Rewriting the above equation using eq. (1.2) we obtain:

$$0 = \lim_{s \to \infty} \frac{1}{s} R(s, \pi) = \lim_{s \to \infty} \frac{1}{s} C(s, \pi) - J_*.$$

From there we conclude:

$$J_* = \lim_{s \to \infty} \frac{1}{s} C(s, \pi) = J(\pi) \qquad \square$$

Lemma 1.2 shows that if we are on the event where $R(s, \pi) = o(s)$, then the average per step cost converges towards the optimal one.

Recent works revealed that the optimal regret in the setting where we do not know matrices A_*, B_*, achieved by any policy π, scales as $R(T, \pi) = \Theta(\sqrt{d_u^2 d_x T})$, where T is the time horizon. However, the algorithm which achieves optimal regret needs a stabilizing controller as an input. How to find a stabilizing controller is addressed in the presented thesis. The setting addressed in the thesis is the following:

1. We run the experiment in a single trajectory for T steps.
2. System matrices A_*, B_* are unknown at the beginning and we do not know a stabilizing controller.
3. We have some knowledge about system matrices – we either assume a priori that our belief about the system is Gaussian, or we assume the upper bound on the system norm.
4. We propose algorithm *eXploration* which provably finds a stabilizing controller in constant time (constant in T).
5. Algorithm eXploration is used as an initialization for algorithms which require a stabilizing controller as an input.

1.2 Related Work

Regret evolution for stochastic LQR 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. Abbasi-Yadkori et al. (2011) show an algorithm which achieves $\mathcal{O}(d^d\sqrt{T})$ regret for *controllable*[1] systems in single trajectory, where T is the time horizon. Their algorithm, based on the optimism in the face of uncertainty (OFU) principle, was inefficient and the dependence on state dimension was exponential. Ibrahimi et al. (2012) improved the factor of state dimension dependence in regret to linear. Later, Cohen et al. (2019) introduced an OFU algorithm called OSLO which is based on semi-definite relaxation. Their OSLO algorithm is efficient and suffers regret $\mathcal{O}(poly(d)\sqrt{T})$ which depends polynomially on the system dimension. However, to run algorithm OSLO we need a stabilizing controller. How to obtain one was discussed by Dean et al. (2017). Dean et al. (2017) considered multiple trajectory setting and used SLS framework to derive robust control synthesis problem. In the follow up work Dean et al. (2018) show that using the robust control synthesis we obtain $\mathcal{O}(poly(d)T^{3/2})$ regret in the single trajectory setting. Their analysis again needed a stabilizing controller as an input to the algorithm. Then Mania et al. (2019) showed that we can obtain $\mathcal{O}(poly(d)\sqrt{T})$ regret also by playing ε-greedily. Their idea was to consider the estimates $\widehat{A}, \widehat{B}$ of the systems matrices A_*, B_* as true matrices - they computed the controller $\widehat{K}$ as the optimal infinite horizon controller if the true system was $\widehat{A}, \widehat{B}$. [2] The ε-greedy approach was further studied by Simchowitz and Foster (2020). They show that the regret in the single trajectory setting is lower bounded by $\Omega(\sqrt{d_u^2 d_x T})$. They further showed that by playing ε-greedily we also achieve the regret $\mathcal{O}(\sqrt{d_u^2 d_x T})$, hence showing that the optimal regret for LQR in single trajectory setting scales as $\Theta(\sqrt{d_u^2 d_x T})$. To run their algorithm we again need a stabilizing controller as an input. A concurrent work of Abeille and Lazaric (2020) achieved with new OFU algorithm slightly worse dependence on system dimension, namely $\sqrt{d_x(d_x+d_u)^2}$, however with better other system parameters. Most of the aforementioned algorithms require a stabilizing controller as an input. In the thesis we develop and analyze an algorithm which finds a stabilizing controller in the single trajectory setting in finite time and hence can be used as an initialization to the existing algorithms without changing the regret rate.

System identification One way to synthesize a stabilizing controller is to find tight estimates $\widehat{A}, \widehat{B}$ of the true system matrices A_*, B_*. Shirani Faradonbeh et al. (2018) showed that for *regular* systems with eigenvalues every-

[1] System A, B is called controllable if $(B \ AB \ \cdots \ A^{d_x-1}B)$ has full row rank.

[2] This greedy approach is called certainty equivalent (CE). We further denote certainty equivalent control by CEC.

where but unit circle the ordinary least squares (OLS) estimator is consistent. Further Sarkar and Rakhlin (2018) showed OLS consistency for all regular systems. They show that the estimation error scales as $\mathcal{O}(1/\sqrt{s})$, for general systems, where s is the number of steps. At the same time Umenberger et al. (2019) introduced an ellipsoid region around OLS estimates where the true system lies with high probability. The computation of their ellipsoid region is based on maximum likelihood estimation of system parameters. In the thesis we move to the Bayesian setting and assume Gaussian prior for system matrices. We derive ellipsoid confidence region around maximum a posterior (MAP) estimator, which turns out to be regularized least squares (RLS) estimator. Using the results of Sarkar and Rakhlin (2018) we further show that the ″radius″ of the derived confidence region decrease linearly with time.

Controller synthesis Using SLS synthesis Dean et al. (2017) derived a SDP which optimal solution results in a controller which stabilizes the true system. They used the proposed SDP with the multiple trajectories system identification. Faradonbeh et al. (2018) introduced a procedure which finds a stabilizing controller based on multiple random controllers. Their approach runs in a single trajectory however the time it takes to find a stabilizing controller depends on the properties of Jordan transition matrix of the true system, which are intractable if the system is unknown. Recently Lale et al. (2020a) showed that the modification of the algorithm proposed by Abbasi-Yadkori and Szepesvari (2011), where we explore more in the early stage of the algorithm, also achieves $\mathcal{O}(poly(d)\sqrt{T})$ and not only for the controllable systems but for more general stabilizable systems. While Dean et al. (2017) derived an SDP which feasible solution stabilizes every system within some ″square″ confident region, we derived an SDP (starting from the same SLS synthesis) which stabilizes every system within some ellipsoid confidence region, which goes hand in hand with the confidence region obtained from system identification procedure. We derived another SDP which is based on the robust version of the SDP that, using the true system matrices, can be used to compute the optimal policy (c.f. (Cohen et al., 2018)). We show the feasibility equivalence of the derived SDPs.

Partial observation setting Another line of work deals with the systems where we do not observe the state x_i but only $y_i = C_* x_i + v_i$, where C_* is unknown matrix and v_i independent Gaussian noise. System identification in this case is not uniquely defined, since for any unitary matrix U we would observe the same response if the system was $A_*' = UA_*U^{-1}, B_*' = UB_*, C_*' = C_*U^{-1}$. Works of Oymak and Ozay (2019), Sarkar et al. (2019), Simchowitz et al. (2019) describes how we can obtain one such representative in the case of stable or marginally stable A_*. How to obtain the stabilizing controller when matrix A_* is potentially unstable in single trajectory is not known. However, Lale et al. (2020b) showed an algorithm that in the setting with $\rho(A_*) < 1$ achieves $\mathcal{O}(poly(\log T))$ regret.

Adversarial noise Another line of work which is closely related to our setting is when w_i and v_i are not necessary stochastic but could be chosen by an adversary. Simchowitz et al. (2020), Simchowitz (2020) considers this setting and show that the regret scales also as $\mathcal{O}(\sqrt{T})$ in both – partially and fully observed – cases, given a stabilizing controller. How to find a stabilizing controller in this setting is described in the concurrent work of Chen and Hazan (2020).

1.3 Structure of the thesis

We start the thesis by LQR problem description and a survey of the related work in Chapter 1.

In Chapter 2 we present the results on top of which we then build in the following chapters. We show the derivation of optimal policy, derive discrete algebraic Riccati equations, present results from robust control synthesis, introduce handy results from linear algebra and probability theory, show finite time identification of linear dynamical systems results of Sarkar and Rakhlin (2018) and show how we can compute the optimal infinite horizon policy via semi-definite program (SDP).

In Chapter 3 we first show how we can find data dependent estimation errors for regularized least squares estimates. Next we propose different robust SDPs which solutions yield a controller which stabilizes the true system with high probability. We also introduce an algorithm eXploration and prove that it finds a stabilizing controller in finite time.

How to use eXploration as an initialization to algorithms which need a stabilizing controller (OSLO, CEC) as an input is discussed in Chapter 4.

In Chapter 5 we propose different heuristics for action selection which empirically results in faster controller synthesis and smaller blow up of the system's state. We further show that with suitable actions selection it can happen that the estimation procedure can be inconsistent.

In Chapter 6 we show various numerical experiments where we compare different approaches derived in the Chapter 3. We compare the time it takes them to find a stabilizing controller and the cost we suffer until that happens. We further compare robust and CE controller and their stabilizing regions in the case when system matrices are one dimensional.

We discuss obtained results in Chapter 7. We further propose different open question which we believe are interesting and relevant and could be addressed in the future work.

Chapter 2

Preliminaries

In this chapter we introduce the background theory which we extensively use in chapters 3 to 5.

2.1 Benchmark class

The goal of this section is to derive the optimal policies for finite and infinite horizon in the case when we know matrices A_*, B_*. We will see that in both cases the best policy is to choose actions which are linear maps of state.

2.1.1 Value function

First we consider finite horizon setting, where we would like to find actions $(u_i)_{i<T}$ such that the cost:

$$C(T,\pi) = \mathbb{E}\left[\sum_{i=0}^{T-1} x_i^\top Q x_i + u_i^\top R u_i + x_T^\top Q_F x_T\right] \tag{2.1}$$

is minimal. Here $Q_F \in \mathbb{R}^{d_x \times d_x}$ is a positive semi-definite matrix. In the setting introduced above Q_F was equal to Q, here we consider a bit more general setting. In order to find the policy which minimize the $C(T,\pi)$ given by eq. (2.1) we define, following Boyd (2009), for $t = 0, \dots, T$ value functions $V_t : \mathbb{R}^{d_x} \to \mathbb{R}$ as:

$$\begin{aligned} V_t(z) = \min_{(u_j)_{j=t}^{T-1}} \mathbb{E}\left[\sum_{i=t}^{T-1} x_i^\top Q x_i + u_i^\top R u_i + x_T^\top Q_F x_T\right] \\ \text{s.t. } x_t = z, \quad x_{i+1} = A_* x_i + B_* u_i + w_{i+1}, \quad i = t, \dots, T-1. \end{aligned} \tag{2.2}$$

The goal is to obtain $V_0(0)$ since this is the minimal expected T-step cost defined in eq. (2.1). In order to find V_0 we will first find V_T and then recursively find V_t running backwards in time. First observe that $V_T(z) = z^\top Q_F z$, since we do not have anything to optimize over. Next observe the recursive nature of the problem:

$$V_t(z) = z^\top Qz + \min_u \left(u^\top Ru + \mathbb{E}V_{t+1}(A_*z + B_*u + w_{t+1}) \right). \tag{2.3}$$

We make an induction hypothesis, where we claim that V_t is of the from $V_t(z) = z^\top P_t z + q_t$ for $t = 0, \ldots, T$. The explicit values of P_t and q_t will be given during calculations. The base of the induction is fulfilled with $P_T = Q_F$ and $q_T = 0$. Assume now $V_{t+1}(z) = z^\top P_{t+1} z + q_{t+1}$. We will find the minimal value given in eq. (2.3). We first use induction hypothesis to obtain:

$$\begin{aligned} &u^\top Ru + \mathbb{E}V_{t+1}(A_*z + B_*u + w_{t+1}) \\ =&u^\top Ru + (A_*z + B_*u)^\top P_{t+1}(A_*z + B_*u) + \langle P_{t+1}, W \rangle + q_{t+1} \end{aligned}$$

To find the minimal value we set the derivative of the last expression over u to zero and obtain:

$$\begin{aligned} 2Ru + 2B_*^\top P_{t+1}B_*u + 2B_*^\top P_{t+1}A_*z = 0 \\ \implies u = -(R + B_*^\top P_{t+1}B_*)^{-1}B_*^\top P_{t+1}A_*z \end{aligned}$$

Plugging the optimal u to the expression given by eq. (2.3) we obtain after some elementary calculations given by lemma A.1:

$$V_t(z) = z^\top P_t z + q_t,$$

where:

$$\begin{aligned} &P_t = Q + A_*^\top P_{t+1}A_* - A_*^\top P_{t+1}B_*(R + B_*^\top P_{t+1}B_*)^{-1}B_*^\top P_{t+1}A_*, \\ &q_t = \langle P_{t+1}, W \rangle + q_{t+1}. \end{aligned} \tag{2.4}$$

With that we finished the induction step. We showed that in order to minimize the expected T-step cost $C(T, \pi)$ it is optimal to choose actions

$$u_t = K_t x_t, \tag{2.5}$$

where

$$\begin{aligned} K_t &= -(R + B_*^\top P_{t+1}B_*)^{-1}B_*^\top P_{t+1}A_*z, \\ P_t &= Q + A_*^\top P_{t+1}A_* - A_*^\top P_{t+1}B_*(R + B_*^\top P_{t+1}B_*)^{-1}B_*^\top P_{t+1}A_*, \quad P_T = Q_F \end{aligned}$$

At the same time it follows from eq. (2.4) that the minimal expected T-step cost $C(T, \pi)$ is equal to:

$$C_\pi(T) = V_0(0) = q_0 = \sum_{i=1}^{T} \langle P_i, W \rangle$$

Remark 2.1 *As a byproduct of the presented analysis we obtain that for every $z \in \mathbb{R}^{d_x}$:*

$$\min_K z^\top \left(Q + K^\top R K + (A_* + B_* K)^\top P_{t+1} (A_* + B_* K) \right) z = z^\top P_t z,$$

and minimum is attained at $K = K_t$.

2.1.2 Infinite horizon

In section 2.1.1 we ran the computation backwards – we first computed P_T and then we went backwards in time and compute P_t for $t = T-1, T-2, \ldots, 1, 0$. The idea of this section is to send T towards the infinity and study the behavior of P_0. The problem is equivalent to the problem where we set $P_0 = Q_F$, define

$$P_{t+1} = Q + A_*^\top P_t A_* - A_*^\top P_t B_* (R + B_*^\top P_t B_*)^{-1} B_*^\top P_t A_*, \tag{2.6}$$

run the recursion forward in time and analyze the behavior of $\lim_{t \to \infty} P_t$. In this section we denote $K_t = -(R + B_*^\top P_t B_*)^{-1} B_*^\top P_t A_*$.The results in this section mainly follows the discussion in chapter 4.3 of Anderson and Moore (1979). We will show the following theorem.

Theorem 2.2 *Let (A_*, B_*) be stabilizable and $Q \succ 0$. Further let P_0 be arbitrary positive semidefinite matrix and let P_t evolve via eq. (2.6). Then we have:*

1. *There exist positive definite matrix P_* with: $\lim_{t \to \infty} P_t = P_*$.*
2. *P_* is the unique positive definite solution of the discrete algebraic Riccati equation:*

$$P_* = Q + A_*^\top P_* A_* - A_*^\top P_* B_* (R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*.$$

3. *Denote by $K_* = -(R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*$, then $\rho(A_* + B_* K_*) < 1$ and*

$$\|P_t - P_*\|_2 \leq \mathcal{O}\left(\rho(A_* + B_* K_*)^t\right).$$

As we can see the speed of convergence is exponential. The latter observation will be crucial in the comparison between best infinite horizon policy with the finite one.

Proof The outline of the proof is the following. We first show boundedness of P_t for any fixed initial matrix P_0 (the bound indeed depends on the P_0). Then we show that if we choose $P_0 = 0$, the sequence P_t is monotonically increasing (in Loewner order) and hence it has a limit. Then we show that in the limit we have $\rho(A_* + B_* K_*) < 1$ and that the convergence is exponential. We finish the proof by showing the convergence for arbitrary initial positive semidefinite matrix P_0.

To show boundedness of P_t take controller K with $\rho(A_* + B_*K) < 1$ and consider the sequence of matrices $(P'_t)_{t\geq 0}$ defined as:

$$\begin{aligned} P'_0 &= P_0, \\ P'_{t+1} &= Q + K^\top RK + (A_* + B_*K)^\top P'_t(A_* + B_*K), \quad t = 0, \dots. \end{aligned} \tag{2.7}$$

We will show that $P_t \preceq P'_t$ by induction. Since $P_0 = P'_0$ the base of the induction holds. By remark 2.1 we have for every $z \in \mathbb{R}^{d_x}$:

$$\begin{aligned} z^\top P'_{t+1} z &= z^\top \left(Q + K^\top RK + (A_* + B_*K)^\top P'_t(A_* + B_*K)\right) z \\ &\overset{I.H.}{\geq} z^\top \left(Q + K^\top RK + (A_* + B_*K)^\top P_t(A_* + B_*K)\right) z \\ &\geq \min_K z^\top \left(Q + K^\top RK + (A_* + B_*K)^\top P_t(A_* + B_*K)\right) z \\ &= z^\top P_{t+1} z. \end{aligned}$$

Since z was arbitrary we showed that for every t we have $P_t \preceq P'_t$. Since $\rho(A_* + B_*K) < 1$ and we have the recursive relation for P'_t given by eq. (2.7) it follows that there exist M such that for every t we have $\|P_t\|_2 \leq \|P'_t\|_2 \leq M$. Now we prove by induction that if we choose $P_0 = 0$, then $P_t \preceq P_{t+1}$ for every t. Since $P_0 = 0$ and $P_1 = Q$, we have $P_0 \preceq P_1$, hence the induction base holds. Since for every $z \in \mathbb{R}^{d_x}$:

$$\begin{aligned} z^\top P_{t+1} z &= \min_K z^\top \left(Q + K^\top RK + (A_* + B_*K)^\top P_t(A_* + B_*K)\right) z \\ &\overset{I.H.}{\succeq} \min_K z^\top \left(Q + K^\top RK + (A_* + B_*K)^\top P_{t-1}(A_* + B_*K)\right) z \\ &= z^\top P_t z, \end{aligned}$$

we conclude $P_{t+1} \succeq P_t$. Since the sequence of matrices $(P_t)_{t\geq 0}$ is increasing (in Loewner order) and is bounded it has a limit which we denote by P_*. Taking the limit in eq. (2.6) we obtain:

$$P_* = Q + A_*^\top P_* A_* - A_*^\top P_* B_*(R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*. \tag{2.8}$$

Now we will show that $\rho(A_* + B_*K_*) < 1$. For that consider the equation equivalent to eq. (2.8):

$$P_* = Q + K_*^\top RK_* + (A_* + B_*K_*)^\top P_*(A_* + B_*K_*)$$

Take the eigenpair (λ, v) of matrix $A_* + B_*K_*$ with the largest absolute eigenvalue. We have:

$$\begin{aligned} v^H P_* v &= v^H Q v + v^H K_*^\top RK_* v + |\lambda|^2 v^H P_* v \\ \iff (1 - |\lambda|^2) v^H P_* v &= v^H Q v + v^H K_*^\top RK_* v > 0. \end{aligned}$$

Hence $|\lambda| < 1$. Next we will show that for any $P_0 = \rho I$ we have $\lim_{t\to\infty} P_t = P_*$. For that observe:

$$\begin{aligned}
P_{t+1} &= Q + K_t^\top R K_t + (A_* + B_* K_t)^\top P_t (A_* + B_* K_t) \\
&\succeq (A_* + B_* K_t)^\top P_t (A_* + B_* K_t) \\
&\succeq (A_* + B_* K_t)^\top (A_* + B_* K_{t-1})^\top P_{t-1} (A_* + B_* K_{t-1})(A_* + B_* K_t) \\
&\vdots \\
&\succeq \Psi_t^\top P_0 \Psi_t = \rho \Psi_t^\top \Psi_t,
\end{aligned}$$

where $\Psi_t = (A_* + B_* K_0)(A_* + B_* K_1) \cdots (A_* + B_* K_t)$. Since $\|P_{t+1}\|_2 \leq M$ for every t, the elements of the sequence $(\Psi_t)_{t\geq 0}$ are bounded. Now we will show that the sequence $(P_t)_{t\geq 0}$ starting at $P_0 = \rho I$ converges towards P_*. For that observe the following relation which is proved in lemma A.2:

$$\begin{aligned}
P_{t+1} - P_* &= (A_* + B_* K_*)^\top (P_t - P_*)(A_* + B_* K_t) \\
&= \left((A_* + B_* K_*)^t\right)^\top (P_0 - P_*)\Psi_t.
\end{aligned}$$

Since $(P_0 - P_*)\Psi_t$ is bounded and $\rho(A_* + B_* K_*) < 1$ we have

$$\lim_{t\to\infty} \|P_{t+1} - P_*\|_2 = \lim_{t\to\infty} \mathcal{O}\left(\rho(A_* + B_* K_*)^t\right) = 0.$$

We will finish the proof by showing the convergence of sequence $(P_t)_{t\geq 0}$ for arbitrary positive semidefintie matrix P_0 towards P_*. For that consider sequences: $(P_t^l)_{t\geq 0}, (P_t)_{t\geq 0}, (P_t^u)_{t\geq 0}$, where $P_0^l = 0, P_0 = P_0, P_0^u = \rho I$, where $\rho = \|P_0\|_2$. All sequences follow the recursive equation given by eq. (2.6). We will show by induction that for every t we have: $P_t^l \preceq P_t \preceq P_t^u$. The base case $P_0^l \preceq P_0 \preceq P_0^u$ follows from definition. Using the relation from remark 2.1 we obtain for every $z \in \mathbb{R}^{d_x}$:

$$\begin{aligned}
z^\top P_{t+1}^l z &= \min_K z^\top \left(Q + K^\top R K + (A_* + B_* K)^\top P_t^l (A_* + B_* K)\right) z \\
&\leq \min_K z^\top \left(Q + K^\top R K + (A_* + B_* K)^\top P_t (A_* + B_* K)\right) z = z^\top P_{t+1} z \\
&\leq \min_K z^\top \left(Q + K^\top R K + (A_* + B_* K)^\top P_t^u (A_* + B_* K)\right) z = z^\top P_{t+1}^u z.
\end{aligned}$$

Therefore we have $P_t^l \preceq P_t \preceq P_t^u$ for every t. Since P_t^l, P_t^u both converge to P_* also P_t converges to P_*. □

Since by letting T towards infinity the expected T-step cost $C(T, \pi)$ diverges we use the notion of expected average per step cost $J(\pi)$ to measure the performance in this regime. From theorem 2.2 follows that the policy, where we play $u_i = K_* u_i$, minimizes the infinite horizon expected per step cost $J(\pi)$.

2.1.3 Cost comparison

In the setting where the matrices A_*, B_* are unknown we measure the performance of the policy via regret. In the eq. (1.2) we defined the regret $R(s, \pi)$ as the difference between the cost suffered by the policy π and sJ_*. However as we have seen in section 2.1.1, when the system evolves for s steps it is optimal to play $u_i = K_i x_i$ and the minimal expected incurred cost is equal to $\sum_{i=1}^{s} \langle P_i, W \rangle$. Hence one could ask why don't we compare in the definition of the regret the cost incurred by the policy with the minimal expected s-step cost. The answer lies in the following computation:

$$\begin{aligned} s \langle P_*, W \rangle - \sum_{i=1}^{s} \langle P_i, W \rangle &= \sum_{i=1}^{s} \langle P_* - P_i, W \rangle \\ &\leq \sum_{i=1}^{s} \|P_* - P_i\|_2 \|W\|_* \\ &\leq \|W\|_* \sum_{i=0}^{\infty} \mathcal{O}\left(\rho(A_* + B_* K_*)^i\right) \\ &= \mathcal{O}\left(\frac{\|W\|_*}{1 - \rho(A_* + B_* K_*)}\right). \end{aligned}$$

We see that the difference between the minimal expected s-steps cost and sJ_* is of order $\frac{\|W\|_*}{1-\rho(A_*+B_*K_*)}$ and hence constant. Since we are usually interested in the order of the regret i.e. $\mathcal{O}(\sqrt{s})$ or $\mathcal{O}(\log s)$ and not particularly in the constants and the analysis of the regret is smoother in the case when we compare incurred cost to sJ_* we defined regret in this way.

2.2 System Level Synthesis

This section mainly follows and adapts the results which were discussed by Dean et al. (2017) and Wang et al. (2019). They use the so called *z-transform* to lift the analysis in the infinite dimensional Hilbert space, where the analysis becomes linear, obtain the results and take the results back down to the finite dimensional space. Let us first acquaint ourselves with the notation necessary for this kind of analysis.

Definition 2.3 *For a sequence of vectors $(x_k)_{k=0}^{\infty}$ and for $z \in \mathbb{C}$ define $\mathbf{x}(z) = \sum_{k=0}^{\infty} z^{-k} x_k$ whenever the sum exist. Similarly for a sequence of matrices $(M_k)_{k=0}^{\infty}$ and for $z \in \mathbb{C}$ define $\mathbf{M}(z) = \sum_{k=0}^{\infty} z^{-k} M_k$ whenever the sum exist.*

We will call any such $\mathbf{x}(z), \mathbf{M}(z)$ a transfer function. For a matrix-valued transfer function we define the following norm:

$$\|\mathbf{M}(z)\|_{\mathcal{H}_\infty} = \sup_{z \in \partial \mathbb{D}} \|\mathbf{M}(z)\|_2 .$$

We will be mostly interested in matrix transfer functions from the subset of so called set of *real-rational proper transfer matrices* denoted by $\mathcal{RH}_\infty$. Let us now formally[1] define this space.

Definition 2.4 *Let $\mathbf{M}(z)$ be a matrix whose entries are functions in variable z. We say that $\mathbf{M}(z)$ is* real-rational proper *transfer matrix if its entries are rational functions in variable z with real coefficients and $\lim_{z\to\infty}\mathbf{M}(z)$ exist. Denote the class of all real-rational proper transfer matrices with $RP_{\mathbb{R}}$. For a transfer matrix $\mathbf{M}(z)$ define the set of poles (matrix has a pole at z_0 if at least one of its entries has a pole at z_0) as $P(\mathbf{M})$. Let us further define class $\mathcal{RH}_\infty$ as:*

$$\mathcal{RH}_\infty = \{\mathbf{M}(z)|\mathbf{M}(z) \in PR_{\mathbb{R}},\ P(\mathbf{M}) \subset \mathbb{D}\}$$

We say that $\mathbf{M}(z) \in \frac{1}{z}\mathcal{RH}_\infty$ if $z\mathbf{M}(z) \in \mathcal{RH}_\infty$. For a matrix M define the resolvent of M as $\mathfrak{R}_M(z) = (zI - M)^{-1}$. Further we call the resolvent of the system (A_*, B_*) with the controller K the resolvent of a matrix $A_* + B_*K$.

2.2.1 LQR in the language of System Level Synthesis

Let us now motivate why should we consider transfer functions. Assume that we choose to use a fixed controller K. We can express the state x_i and action u_i as

$$x_i = \sum_{k=1}^{i}(A_* + B_*K)^{i-k}w_k,$$
$$u_i = \sum_{k=1}^{i}K(A_* + B_*K)^{i-k}w_k.$$

We observe that the connection between x_i, u_i and matrices A_*, B_* and K is not linear. Denoting $\Phi_x(k) = (A_* + B_*K)^{k-1}$ and $\Phi_u(k) = K(A_* + B_*K)^{k-1}$ we can rewrite that as

$$\begin{pmatrix} x_i \\ u_i \end{pmatrix} = \sum_{k=1}^{i}\begin{pmatrix} \Phi_x(i-k+1) \\ \Phi_u(i-k+1) \end{pmatrix} w_k.$$

Hence x_i, u_i have a linear dependence on $(\Phi_x(k))_{k\geq 1}, (\Phi_u(k))_{k\geq 1}$ and since dealing with linear systems is much easier we consider such notation/transformation. From the definition it follows that we have:

$$\Phi_x(k) = A_*\Phi_x(k-1) + B_*\Phi_u(k-1), \forall k \geq 1,\ \Phi_x(1) = I. \tag{2.9}$$

Denoting $\mathbf{\Phi}_x(z) = \sum_{i=1}^{\infty} z^{-i}\Phi_x(i)$ and $\mathbf{\Phi}_u(z) = \sum_{i=1}^{\infty} z^{-i}\Phi_u(i)$ the constraints given by eq. (2.9) can be equivalently written as:

$$\begin{pmatrix} zI - A_* & -B_* \end{pmatrix}\begin{pmatrix} \mathbf{\Phi}_x(z) \\ \mathbf{\Phi}_u(z) \end{pmatrix} = I \tag{2.10}$$

[1] The definition was taken from Stoorvogel (1992).

With such a notation we also obtain $\mathbf{u}(z) = K\mathbf{x}(z)$ and $K = \mathbf{\Phi}_u(z)\mathbf{\Phi}_x(z)^{-1}$. Bearing this in mind we will turn now to non-static controllers, represent the theory from Dean et al. (2017) and then use the derived theory to obtain a robust static controller. For $\mathbf{\Phi}_u, \mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty$ define a controller as $\mathbf{K} = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}$ and its corresponding response $\mathbf{u} = \mathbf{Kx}$. Let us now derive the so called *system response* for such a controller. Multiply equation

$$x_{i+1} = A_* x_i + B_* u_i + w_{i+1}$$

by z^{-i-1} and sum from 0 to ∞ in i to obtain:

$$\mathbf{x}(z) = \frac{1}{z}\left(A_*\mathbf{x}(z) + B_*\mathbf{u}(z) + \mathbf{w}(t)\right),$$

where we denoted $\mathbf{w}(z) = \sum_{k\geq 0} z^{-k} w_{k+1}$. Inserting $\mathbf{u} = \mathbf{Kx}$ and rearranging the terms yields:

$$\left(zI - A_* - B_*\mathbf{K}\right)\mathbf{x} = \mathbf{w}.$$

Assuming that the inverse exist and using the relation $\mathbf{u} = \mathbf{Kx}$ we arrive at the *system response* equation:

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\left(zI - A_* - B_*\mathbf{K}\right)^{-1}\\ \mathbf{K}\left(zI - A_* - B_*\mathbf{K}\right)^{-1}\end{pmatrix}\mathbf{w} \tag{2.11}$$

In the case of a static controller the notion of stability is clear – a controller K stabilizes the system A_*, B_* if $\rho(A_* + B_*K) < 1$. Let us now define stability also for the dynamic controllers.

Definition 2.5 (Stability of general controller) *A controller* $\mathbf{K}$ *is stable if starting from arbitrary* x_0*, playing* $\mathbf{K}$ *and further assuming* $w_i = 0$ *for every* i *results in:*

$$\lim_{k\to\infty} x_k = 0.$$

Next theorem will connect the system response equation with the affine conditions which we introduced in (2.10).

Theorem 2.6 (Theorem 3.1 in Dean et al. (2017)) *We have:*

1. *The affine space defined by*

$$\begin{pmatrix} zI - A_* & -B_* \end{pmatrix}\begin{pmatrix}\mathbf{\Phi_x}\\ \mathbf{\Phi_u}\end{pmatrix} = I, \quad \mathbf{\Phi_x}, \mathbf{\Phi_u} \in \frac{1}{z}\mathcal{RH}_\infty \tag{2.12}$$

parametrizes all system responses (2.11) *achievable by a stabilizing state-feedback controller* $\mathbf{K}$.

2. *For any transfer matrices $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$ satisfying* (2.12) *the controller* $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ *is stabilizing and achieves system response* (2.11).

Using the relation $\mathbf{\Phi}_u = \mathbf{K}\mathbf{\Phi}_x$ and equation (2.12) we have for any $\mathbf{\Phi}_\mathbf{x}, \mathbf{\Phi}_\mathbf{u} \in \frac{1}{z}\mathcal{RH}_\infty$: $\mathbf{\Phi}_\mathbf{x} = (zI - A_* - B_*\mathbf{K})^{-1}$ and $\mathbf{\Phi}_\mathbf{u} = \mathbf{K}\,(zI - A_* - B_*\mathbf{K})^{-1}$. Hence $\mathbf{x} = \mathbf{\Phi}_\mathbf{x}\mathbf{w}$ and $\mathbf{u} = \mathbf{\Phi}_\mathbf{u}\mathbf{w}$.

The usual setting in the online learning of the stabilizing controller is that we have estimates $\widehat{A}, \widehat{B}$ and we know a neighborhood around the $\widehat{A}, \widehat{B}$ where the true underlying system A_*, B_* lies with high probability. We wish to find a controller which stabilizes everything inside this neighborhood. The next lemma shows a condition for controller $\mathbf{K}$, which stabilizes $\widehat{A}, \widehat{B}$, to stabilize also the system A, B, where $A = \widehat{A} + \Delta_A, B = \widehat{B} + \Delta_B$.

Lemma 2.7 (Sufficient condition, Lemma 3.4 in Dean et al. (2017)) *Assume controller* $\mathbf{K}$ *stabilize* $(\widehat{A}, \widehat{B})$ *and let* $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ *be its corresponding system response. Denoting* $\widehat{\mathbf{\Delta}} = (\Delta_A + \Delta_B\mathbf{K})\mathfrak{R}_{\widehat{A}+\widehat{B}\mathbf{K}}$, *a sufficient condition for* $\mathbf{K}$ *to stabilize* (A, B) *is* $\left\|\widehat{\mathbf{\Delta}}\right\|_{\mathcal{H}_\infty} < 1$.

Remark 2.8 *Expanding the resolvent one can see that* $(\Delta_A + \Delta_B\mathbf{K})\mathfrak{R}_{\widehat{A}+\widehat{B}\mathbf{K}} = \begin{pmatrix}\Delta_A & \Delta_B\end{pmatrix}\begin{pmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{pmatrix}$.

Next we will rewrite the sufficient condition $\left\|\widehat{\Delta}\right\|_{\mathcal{H}_\infty} < 1$ in an equivalent form and show a sufficient condition when there indeed exist a controller which stabilizes a every system with $\|\Delta_A\|_2 \le \varepsilon_A$, $\|\Delta_B\|_2 \le \varepsilon_B$.

We will limit ourselves to static controllers, which means that $\mathbf{K} = K$. For such condition the sufficient condition $\left\|\widehat{\Delta}\right\|_{\mathcal{H}_\infty} < 1$ is equivalent to:

$$\left\|(\Delta_A\ \Delta_B)\begin{pmatrix}I \\ K\end{pmatrix}\left(zI - \widehat{A} - \widehat{B}K\right)^{-1}\right\|_{\mathcal{H}_\infty} < 1. \tag{2.13}$$

Next lemma shows that if $\varepsilon_A, \varepsilon_B$ are small enough, then there exist a static controller for which the constraint given by eq. (2.13) is satisfied. We state the lemma with the notation presented in the described setting.

Lemma 2.9 (Fulfilled Sufficient condition, Lemma 4.2 in Dean et al. (2017)) *Let K be a controller which stabilizes* (A_*, B_*). *Assume that* $\varepsilon_A, \varepsilon_B$ *are small enough that for* ζ *defined as* $\zeta = (\varepsilon_A + \varepsilon_B\,\|K\|_2)\,\|\mathfrak{R}_{A_*+B_*K}\|_{\mathcal{H}_\infty}$ *we have* $\zeta \le (1+\sqrt{2})^{-1}$. *Then K satisfies the constraint given by eq.* (2.13).

The constraint given by eq. (2.13) is infinite dimensional one and hard to compute is this from. However there exist a technique which translates the constraint given by eq. (2.13) to a SDP program. The technique is called

Kalman-Yakubovich-Popov (KYP) Lemma. To obtain constraint in the form of semidefinite inequality we will use a version of discrete time KYP Lemma. Let us first introduce a useful notation: for matrices X, Y, Z and W of appropriate dimensions define

$$\left(\begin{array}{c|c} X & Y \\ \hline Z & W \end{array}\right)(z) = Z(zI - X)^{-1}Y + W.$$

With the introduced notation we can state KYP lemma:

Lemma 2.10 (KYP Lemma, taken from Theorem 1.1 in Bart et al. (2018)) *Let* $M(z) = \left(\begin{array}{c|c} X & Y \\ \hline Z & W \end{array}\right)(z) \in \mathcal{RH}_\infty$. *Then the following are equivalent:*

1. $\|M(z)\|_{\mathcal{H}_\infty} < 1$
2. $\exists P \succ 0$ *such that*

$$\begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} X & Y \\ Z & W \end{pmatrix}\begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} X & Y \\ Z & W \end{pmatrix}^\top \succ 0$$

2.2.2 $\mathcal{H}_\infty$ constraint to semi-definite constraint

In this section we will present the result of Dean et al. (2017), where they transform the sufficient condition eq. (2.13) to a semi-definite constraint. Assuming that the $\|\Delta_A\|_2 \leq \varepsilon_A, \|\Delta_B\|_2 \leq \varepsilon_B$ they first showed that the constraint given by eq. (2.13) is satisfied if the following holds:

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K \end{pmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1}) \right\|_{\mathcal{H}_\infty} \leq \gamma, \quad \gamma \in [0, 1) \tag{2.14}$$

Using KYP lemma and some matrix manipulations Dean et al. (2017) then reformulates eq. (2.14) to the semi-definite constraint:

$$\begin{gathered} \exists s \in [0,1), P \in \mathbb{R}^{d_x \times d_x} \text{ with } P \succ 0 \text{ and } S \in \mathbb{R}^{d_u \times d_x} \text{ s.t.:} \\ \begin{pmatrix} P - I & \widehat{A}P + \widehat{B}S & 0 \\ (\widehat{A}P + \widehat{B}S)^\top & P & \begin{pmatrix} \varepsilon_A P \\ \varepsilon_B S \end{pmatrix}^\top \\ 0 & \begin{pmatrix} \varepsilon_A P \\ \varepsilon_B S \end{pmatrix} & \frac{1}{2} s I \end{pmatrix} \succeq 0. \end{gathered} \tag{2.15}$$

From their derivation it follows that from any triple (s, P, S), which satisfies constraints given by eq. (2.15), we can synthesize the controller K as $K = SP^{-1}$ and for controller K then holds:

1. $\forall A, B$ with $\left\|A - \widehat{A}\right\|_2 \leq \varepsilon_A, \left\|B - \widehat{B}\right\|_2 \leq \varepsilon_B$ we have $\rho(A + BK) < 1$,

2.

$$\left\| \begin{pmatrix} \sqrt{2}\varepsilon_A I \\ \sqrt{2}\varepsilon_B K \end{pmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1}) \right\|_{\mathcal{H}_\infty} \leq \sqrt{s} \tag{2.16}$$

Since the derivation of constraint given by eq. (2.15) is very similar to the derivation presented in section 3.2.1 we omit it here.

2.3 Tools from Probability and Linear Algebra

2.3.1 Probabilistic bounds

Definition 2.11 (Sub-Gaussian random variables) *A random variable $X \in \mathbb{R}$ is said to to be sub-Gaussian with variance proxy σ^2 if we have:*

$$\mathbb{E}X = 0 \textit{ and } \forall s \in \mathbb{R} : \mathbb{E}e^{sX} \leq e^{\frac{s^2\sigma^2}{2}}.$$

We write $X \sim \mathrm{subG}(\sigma^2)$. For a random vector $X \in \mathbb{R}^n$ we say that is sub-Gaussian with variance proxy σ^2 and write $X \sim \mathrm{subG}_n(\sigma^2)$ if $\mathbb{E}X = 0$ and for every $u \in S^{n-1}$ we have $u^\top X \sim \mathrm{subG}(\sigma^2)$. For a random matrix $X \in \mathbb{R}^{n\times m}$ we say that is sub-Gaussian with variance proxy σ^2 and write $X \sim \mathrm{subG}_{n\times m}(\sigma^2)$ if $\mathbb{E}X = 0$ and for every $u \in S^{n-1}$ and $v \in S^{m-1}$ we have $u^\top X v \sim \mathrm{subG}(\sigma^2)$.

Remark 2.12 *Since for $X \sim \mathcal{N}(0, \sigma^2)$ holds $\mathbb{E}e^{sX} = e^{\frac{s^2\sigma^2}{2}}$ we have $X \sim \mathrm{subG}(\sigma^2)$.*

Definition 2.13 (Degenerate random variable) *We say that a random variable $X \in \mathbb{R}^n$ is* degenerate *if its support lies in the space with strictly lower dimension. If the support of X has dimension n then X is* non-degenerate. *If X is Gaussian, X is degenerate if and only if its covariance matrix is not invertible.*

Theorem 2.14 (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 $z > 0$ we have:*

$$\mathbb{P}\left(\|Ax\|_2^2 > \mathrm{Tr}(\Sigma) + 2\sqrt{\mathrm{Tr}(\Sigma^2)z} + 2\|\Sigma\|_2 z \right) \leq e^{-z}$$

Corollary 2.15 *Let $x \sim \mathcal{N}(0, \Sigma)$. Then for any $\delta \in (0, \frac{1}{e})$ we have with probability at least $1 - \delta$:*

$$\|x\|_2^2 \leq 5\,\mathrm{Tr}(\Sigma) \log \frac{1}{\delta}$$

Proof Since $\delta \in (0, \frac{1}{e})$ we have $\log \frac{1}{\delta} > 1$. Hence if we set $z = \log \frac{1}{\delta}$ we have $\sqrt{z} \leq z$. Inserting $z = \log \frac{1}{\delta}$ to theorem 2.14 we have w.p at least $1 - \delta$:

$$\|x\|_2^2 \leq \mathrm{Tr}(\Sigma) + (2\|\Sigma\|_F + 2\|\Sigma\|_2) \log \frac{1}{\delta}.$$

Hence it is enough to show that $\|\Sigma\|_2, \|\Sigma\|_F \leq \mathrm{Tr}(\Sigma)$. 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 $\|\Sigma\|_2 \leq \|\Sigma\|_F$, hence we showed that $\|\Sigma\|_2, \|\Sigma\|_F \leq \mathrm{Tr}(\Sigma)$. □

It often happens that we would like to show that some property holds with high probability for every point on some manifold in $\mathbb{R}^n$. However we have only a bound for a single point. We can usually extend this property to finite number of points by use of union bound. But how to extend it to uncountable sets? We can make use of the so called *ε-nets*[2] as we will see in the proof of proposition 2.18.

Definition 2.16 (ε-net, covering number) *Let (X, d) be a metric space and let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon$ is called an ε-net if $\forall x \in X \; \exists y \in \mathcal{N}_\varepsilon$ such that $d(x, y) \leq \varepsilon$. The minimal cardinality (finite) of an ε-net is denoted by $\mathcal{N}(X, \varepsilon)$ and is called covering number.*

Lemma 2.17 (Covering number of Unit ball) *For $\mathbb{R}^n$ equipped with euclidean metric we have: $\mathcal{N}(S^{n-1}, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^n$.*

Proof Let $\mathcal{N}_\varepsilon$ be maximal ε-set for which we have $\forall x, y \in \mathcal{N}_\varepsilon : \|x - y\|_2 \geq \varepsilon$. Since $\mathcal{N}_\varepsilon$ is maximal for any other point $z \in S^{n-1}$ we have that there exist $x \in \mathcal{N}_\varepsilon$ such that $\|x - z\|_2 \leq \varepsilon$. The balls with radius $\frac{\varepsilon}{2}$ around the points in $\mathcal{N}_\varepsilon$ are disjoint and inside the ball centered at zero with radius $1 + \frac{\varepsilon}{2}$. Hence we have: $vol\left(\frac{\varepsilon}{2}B\right)|\mathcal{N}_\varepsilon| \leq vol\left((1 + \frac{\varepsilon}{2})B\right)$, where B is the unit ball. Since $vol(aB) = a^n vol(B)$ we obtain: $\left(\frac{\varepsilon}{2}\right)^n |\mathcal{N}_\varepsilon| \leq \left(1 + \frac{\varepsilon}{2}\right)^n$. Rearranging the terms yields the bound: $|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^n$. □

Using ε-net argument the next proposition was proven. Proposition 2.18 we will use later in order to prove an estimation error upper bound.

Proposition 2.18 (Proposition 8.1 in Sarkar and Rakhlin (2018)) *Let $M \in \mathbb{R}^{n \times d}$ be a random matrix. Then for any $\varepsilon \in (0, 1)$ there exist $w \in S^{d-1}$ such that we have:*

$$\mathbb{P}\left(\|M\|_2 > z\right) \leq \left(1 + \frac{2}{\varepsilon}\right)^d \mathbb{P}\left(\|Mw\|_2 > (1 - \varepsilon)z\right).$$

[2] For a more in depth discussion about ε-nets see section 4.2 in Vershynin (2018).

Proof Let $\mathcal{N}_\varepsilon$ be minimal (by cardinality) ε-net of S^{d-1}. We will first prove that it holds $\|M\|_2 \leq \frac{1}{1-\varepsilon} \max_{w \in \mathcal{N}_\varepsilon} \|Mw\|_2$. Let $x \in S^{d-1}$ such that $\|M\|_2 = \|Mx\|_2$ and $y \in \mathcal{N}_\varepsilon$ such that $\|x-y\|_2 \leq \varepsilon$. Since $\|M\|_2 = \|Mx\|_2 \leq \|M(x-y)\|_2 + \|My\|_2 \leq \|M\|_2 \varepsilon + \|My\|_2$, we have: $\|M\|_2 \leq \frac{1}{1-\varepsilon} \|My\|_2$. Therefore we have: $\|M\|_2 \leq \frac{1}{1-\varepsilon} \max_{w \in \mathcal{N}_\varepsilon} \|Mw\|_2$. With this in hand we can bound:

$$\mathbb{P}\left(\|M\|_2 > z\right) \leq \mathbb{P}\left(\max_{w \in \mathcal{N}_\varepsilon} \|Mw\|_2 > (1-\varepsilon)z\right).$$

Using union bound we obtain:

$$\begin{aligned}\mathbb{P}\left(\|M\|_2 > z\right) &\leq \sum_{w \in \mathcal{N}_\varepsilon} \mathbb{P}\left(\|Mw\|_2 > (1-\varepsilon)z\right) \\ &\leq |\mathcal{N}_\varepsilon| \, \mathbb{P}\left(\|Mw_m\|_2 > (1-\varepsilon)z\right),\end{aligned}$$

where w_m is such that the probability is maximal among $w \in \mathcal{N}_\varepsilon$. Since by lemma 2.17 $|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^d$ we proved the proposition. □

2.3.2 Results from Linear Algebra

When we will derive the semi-definite constraint in section 3.2.1 we will extensively use Schur's complement lemma. Here we will state the lemma and the results which are closely related to it and will also help us with derivation.

Lemma 2.19 (Schur complement lemma) *Let X be symmetric matrix given by $X = \begin{pmatrix} M & N \\ N^\top & O \end{pmatrix}$. The following holds:*

1. *If $M \succ 0$, then $X \succ 0 \Leftrightarrow O - N^\top M^{-1} N \succ 0$.*
2. *If $O \succ 0$, then $X \succ 0 \Leftrightarrow M - NO^{-1}N^\top \succ 0$.*

Definition 2.20 (Matrix inertia) *Associate with every symmetric matrix $M \in \mathbb{R}^{d \times d}$ a triple (μ, z, p), where μ is the number of positive eigenvalues, z the number of zero eigenvalues and p the number of negative eigenvalues ($\mu + z + p = d$). We call the triple (μ, z, p) the* inertia *of matrix M.*

Definition 2.21 (Conjugation) *For a nonsingular matrix X we say that matrices $X^\top MX$ and M are* congruent. *By saying that we conjugate a matrix M with a matrix X we mean that we perform the mapping: $M \to X^\top MX$.*

Theorem 2.22 (Sylvester) *Congruent symmetric matrices have the same inertia.*

When we will derive the region around RLS estimates where A_*, B_* lies with high probability we will extensively use kronecker product $\otimes$ and vectorization $\mathrm{vec}(\cdot)$. We state the following lemma for smoother reading of that analysis.

Lemma 2.23 *Let matrices M, N, O, P be of appropriate dimensions and invertible when we would like to take the inverse. The followig holds:*

1. $(M \otimes N)(O \otimes P) = MO \otimes NP$,
2. $\operatorname{vec}(MN) = (N^\top \otimes I) \operatorname{vec}(M)$,
3. $(M \otimes N)^\top = M^\top \otimes N^\top$,
4. $(M \otimes N)^{-1} = M^{-1} \otimes N^{-1}$,
5. $\operatorname{Tr}(M^\top NM) = \operatorname{vec}(M^\top)^\top (N \otimes I) \operatorname{vec}(M^\top)$,
6. $M \otimes (N + O) = M \otimes N + M \otimes O$ *and* $(M + N) \otimes O = MO \otimes NO$.

Another tool which will come handy is the so called *S-lemma*. We present the result which was obtained by Luo et al. (2004).

Theorem 2.24 (Proposition 3.4 of Luo et al. (2004)) *Let M, N, O, P be matrices of appropriate dimensions where M, O are symmetric and P is positive semi-definite. The following are equivalent:*

1. *$\forall X$ with $X^\top PX \preceq I$ we have:*

$$O + X^\top N + N^\top X + X^\top MX \succeq 0$$

2. *$\exists t \geq 0$ s.t.:*

$$\begin{pmatrix} O - tI & N^\top \\ N & M + tP \end{pmatrix} \succeq 0$$

Theorem 2.25 (Proposition 3.6 of Luo et al. (2004)) *Let M, N, O, P, S, T, U be matrices of appropriate dimensions where M, O, U are symmetric and P is positive semi-definite. The following are equivalent:*

1. *$\forall X$ with $X^\top PX \preceq I$ we have:*

$$\begin{pmatrix} U & S + TX \\ (S + TX)^\top & O + X^\top N + N^\top X + X^\top MX \end{pmatrix} \succeq 0$$

2. *$\exists t \geq 0$ s.t.:*

$$\begin{pmatrix} U & S & T \\ S^\top & O - tI & N^\top \\ T^\top & N & M + tP \end{pmatrix} \succeq 0.$$

It is a well known fact that for a matrix A with $\rho(A) < 1$ there exist constants M, γ with $\gamma < 1$ such that for every $k \geq 0$ we have $\left\|A^k\right\|_2 \leq M\gamma^k$. To quantify M and γ we will use the following results.

Theorem 2.26 (Theorem 2.16 from Dowler (2013)) *Let $M \in \mathbb{R}^{d \times d}$ be a square matrix and let Γ be a positively oriented Jordan curve in the complex plane which contains the ball $B(\rho(M))$ in its interior. Then we have:*

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

Theorem 2.27 (Theorem 2.3 in Gil′ (2014)) *Let $M \in \mathbb{R}^{d \times d}$ and $z \notin \sigma(M)$. Denote $\rho(M, z) = \min_{k=1}^{d} |\lambda_k(M) - z|$, then we have:*

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

Corollary 2.28 *Let $M \in \mathbb{R}^{d \times d}$ and $\rho(M) < 1$, then we have:*

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

In the following we define the notion of *regular* matrix.

Definition 2.29 (Regular matrix) *Matrix $M \in \mathbb{R}^{n \times n}$ is regular if for every eigenvalue μ of M with $|\mu| > 1$ we have* $\mathrm{rank}(M - \mu I) = n - 1$.

Example 2.30 *Let M_1, M_2 be matrices defined as:*

$$M_1 = \begin{pmatrix} 1.1 & 1 & 0 \\ 0 & 1.1 & 0 \\ 0 & 0 & 0.8 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1.1 & 0 & 0 \\ 0 & 1.1 & 2.1 \\ 0 & 0 & 0.8 \end{pmatrix}.$$

Then matrix M_1 is regular, whereas matrix M_2 is not.

2.4 Linear system identification

In this section we will present results of Sarkar and Rakhlin (2018) which deals with the linear system identification. To estimate matrices A_*, B_* after running the system for s steps we will use regularized least squares estimator defined as

$$A_s, B_s = \underset{A,B}{\mathrm{argmin}} \sum_{i=0}^{s-1} \left\| x_{i+1} - (A\ B) \begin{pmatrix} x_i \\ u_i \end{pmatrix} \right\|_2^2 + \lambda \left\| (A\ B) \right\|_F^2. \tag{2.17}$$

In the following we will show, based on the results of Sarkar and Rakhlin (2018), that when the matrix A_* is regular and if we choose $(u_i)_{i \leq s} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2 I)$, the RLS estimates are consistent and the convergence towards the true parameters scales as $\mathcal{O}\left(\frac{1}{\sqrt{s}}\right)$.

Corollary 2.31 (based on Theorem 2 of Sarkar and Rakhlin (2018)) *Suppose the system evolves via* (1.1) *and the chosen actions* $(u_i)_{i\geq 0} \overset{i.i.d.}{\sim} \mathrm{subG}_k(\sigma_u^2)$ *come from non-degenerate distribution and are 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 defined by eq.* (2.17) *we have:*

$$\max\left(\|A_s - A_*\|_2, \|B_s - B_*\|_2\right) \leq \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}},$$

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

Proof Denoting $z_i = (x_i^\top \; u_i^\top)^\top$ we obtain that from the definition of RLS estimators it follows that A_s, B_s minimize the expression

$$\left\|\begin{pmatrix} x_1 & \dots & x_s \end{pmatrix} - (A \; B)\begin{pmatrix} z_0 & \dots & z_{s-1} \end{pmatrix}\right\|_F^2 + \lambda \left\|(A \; B)\right\|_F^2 \tag{2.18}$$

in variables A, B. Deriving eq. (2.18) 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:

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

Next rewrite eq. (1.1) as:

$$\begin{pmatrix} x_{i+1} \\ u_{i+1} \end{pmatrix} = \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_i \\ u_i \end{pmatrix} + \begin{pmatrix} w_{i+1} \\ u_{i+1} \end{pmatrix}. \tag{2.19}$$

Since $v_{i+1} := \begin{pmatrix} w_{i+1} \\ u_{i+1} \end{pmatrix}$ are $\mathrm{subG}_{d+k}(\sigma_*^2)$, where $\sigma_* = \max(\sigma_w, \sigma_u)$, and since by further denoting $A = \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix}$ the eq. (2.19) can be rewritten as $z_{i+1} = Az_i + v_{i+1}$, we are in the setting analyzed by Sarkar and Rakhlin (2018). 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 presented by Sarkar and Rakhlin (2018) follows that $\left\|V_s^{-1/2} S'_s\right\|_2 \leq poly(\log s, \log \frac{1}{\delta})$ and $V_s \succeq \Omega(s) I$ (the latter can be observed from eq. (116) in (Sarkar and Rakhlin, 2018)). Since $\left\|(V_s + \lambda I)^{-1/2} S'_s\right\|_2 \leq \left\|V_s^{-1/2} S'_s\right\|_2$ and $V_s + \lambda I \succeq V_s \succeq \Omega(s) I$ we obtain:

$$\begin{aligned}
\|(A_s \; B_s) - (A_* \; B_*)\|_2 &= \left\|\begin{pmatrix} I_d & 0 \end{pmatrix} \left(\begin{pmatrix} A_s & B_s \\ * & * \end{pmatrix} - \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix} \right)\right\|_2 \\
&\leq \left\| \begin{pmatrix} A_s & B_s \\ * & * \end{pmatrix} - \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix} \right\|_2 \\
&\leq \left\|(V_s + \lambda I)^{-1/2}\right\|_2 \left\|(V_s + \lambda I)^{-1/2} S'_s\right\|_2 \\
&+ \lambda \left\|(V_s + \lambda I)^{-1}\right\|_2 \left\| \begin{pmatrix} A_* & B_* \\ 0 & 0 \end{pmatrix} \right\|_2 \\
&\leq \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}} + \mathcal{O}\left(\frac{1}{\sqrt{s}}\right) = \frac{poly(\log s, \log \frac{1}{\delta})}{\sqrt{s}}.
\end{aligned}$$

We finish the proof with the observation $\max\left(\|A_s - A_*\|_2, \|B_s - B_*\|_2\right) \leq \|(A_s \; B_s) - (A_* \; B_*)\|_2$. □

2.5 Optimal infinite horizon policy via SDP

In section 2.1.2 we showed that the optimal infinite horizon policy is to choose actions $u_i = K_* x_i$, where

$$\begin{aligned}
K_* &= -(R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*, \\
P_* &= A_*^\top P_* A_* - A_*^\top P_* B_* (R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_*.
\end{aligned}$$

In this section we will motivate and show a different approach how we can find K_* based on a SDP presented by Cohen et al. (2018). Assume that we play $u_i = K x_i$ where K stabilizes the underlying system and that the states converge in distribution towards x. Denote by Σ_{xx} its limiting covariance matrix. The $J(\pi)$ of this policy will be equal to:

$$\begin{aligned}
J(\pi) &= \mathbb{E}\left[x^\top Q x + x^\top K^\top R K x\right] \\
&= \operatorname{Tr}\left((Q + K^\top R K)\Sigma_{xx}\right) \\
&= \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx} K^\top \\ K\Sigma_{xx} & K\Sigma_{xx} K^\top \end{pmatrix} \right\rangle.
\end{aligned}$$

At the same time as system follows eq. (1.1) we have:

$$\begin{aligned}
\Sigma_{xx} &= (A_* + B_* K)\Sigma_{xx}(A_* + B_* K)^\top + \sigma_w^2 I \\
&= (A_* \; B_*) \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx} K^\top \\ K\Sigma_{xx} & K\Sigma_{xx} K^\top \end{pmatrix} (A_* \; B_*)^\top + \sigma_w^2 I.
\end{aligned}$$

With this in mind Cohen et al. (2018) showed that the optimal infinite horizon controller K_* can be obtained by first solving the SDP:

$$\begin{aligned} \min_{\Sigma \succeq 0} \quad & \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle \\ \text{s.t.} \quad & \Sigma_{xx} = (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I, \end{aligned} \tag{2.20}$$

where $\Sigma \in \mathbb{R}^{(d_x+d_u)\times(d_x+d_u)}$ has a block structure $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix}$. The optimal controller is then obtained as:

$$K_* = \Sigma_{ux}\Sigma_{xx}^{-1}.$$

Chapter 3

Find stabilizing controller

In this section we will show an algorithm which we can use to find a controller which stabilizes the system A_*, B_*. Let us first give a pseudo-code for the algorithm, which parts we will then disentangle and describe in depth in this chapter.

Algorithm 1 eXploration

```
1: Input: $x_0 = 0$, other parameters depending on initialization
2: for $s = 1, \ldots$ do
3:     Play action $u_{s-1}$ and observe state $x_s$
4:     Let $(A_s, B_s)$ be a RLS estimators
5:     Compute high probability region $\Theta$ around $(A_s, B_s)$
6:     Try to synthesize controller $K_0$ which stabilizes every system in $\Theta$
7:     if we find $K_0$ return $K_0$
8: end for
```

When the algorithm terminates it will return a controller which, with high probability, stabilizes the underlying system A_*, B_*. We can use this algorithm to initialize algorithms, such as OSLO (Cohen et al., 2019) or CEC (Simchowitz and Foster, 2020), which need a stabilizing controller as an input. We will show that using specific initialization the algorithm 1 terminates with high probability in constant time. Since we finish in constant time the regret we suffer is also constant (could be exponential in system parameters) in time horizon T. Hence if we initialize an algorithm, which needs a stabilizing controller as an input, with algorithm 1 the new algorithm has the same order of regret should we be given a stabilizing controller. In the next sections we will describe possible ways how to execute lines 3, 5 and 6 of algorithm 1.

3.1 Data driven estimation error

In this section we will show how to obtain regions around RLS estimates A_s, B_s where A_*, B_* lies with high probability. The regions will be either of the form

$$\{(A,B)|\ \|A - A_s\|_2 \leq \varepsilon_A, \|B - B_s\|_2 \leq \varepsilon_B\}, \tag{3.1}$$

where $\varepsilon_A, \varepsilon_B$ are data dependent error upper bounds, or of the form

$$\left\{(A,B)|X^\top DX \preceq I, X^\top = (A\ B) - (A_s\ B_s)\right\}, \tag{3.2}$$

where D is a positive definite matrix which depends on the observed past states and played actions. The region given by eq. (3.1) consists of two balls, one around A_s with radius ε_A and the other around B_s with radius ε_B, whereas the region given by eq. (3.2) is an ellipsoid around $(A_s\ B_s)$. We will present how we can obtain the high probability regions in two different, in particular Bayesian and non-Bayesian, settings.

3.1.1 Bayesian view

In this section we will describe how we can obtain $\varepsilon_A, \varepsilon_B$ or D, introduced in the discussion given in section 3.1, from the Bayesian perspective. We assume that a priori our belief about $\vartheta_* = \mathrm{vec}((A_*\ B_*))$ is given as $\mathrm{vec}((A\ B)) \sim \mathcal{N}(0, \frac{\sigma_w^2}{\lambda}I)$. Running the system for s steps we observe $(x_i)_{1\leq i\leq s}$ and we play actions $(u_i)_{0\leq i\leq s-1}$. We denote by $\mathcal{D} = \{(x_i)_{1\leq i\leq s}, (u_i)_{0\leq i\leq s-1}\}$ the data which is produced by running the system. Here we compute the posterior belief of $\mathrm{vec}((A_*\ B_*))$ namely $p(\vartheta|\mathcal{D})$ and derive consistent data dependent error upper bounds for matrices A_*, B_* and their RLS estimates. First we show lemma which converts the eq. (1.1) to a form which we will use in the derivation of a posterior belief.

Lemma 3.1 *Denoting $\Phi_i = (x_i^\top\ u_i^\top) \otimes I_{d_x}$ we can rewrite eq.* (1.1) *as:*

$$x_{i+1} = \Phi_i \vartheta_* + w_{i+1} \tag{3.3}$$

Proof Compute:

$$\begin{aligned} x_{i+1} &= (A_*\ B_*) \begin{pmatrix} x_i \\ u_i \end{pmatrix} + w_{i+1} \\ &= \mathrm{vec}\left((A_*\ B_*) \begin{pmatrix} x_i \\ u_i \end{pmatrix}\right) + w_{i+1} \\ &= \left((x_i^\top\ u_i^\top) \otimes I_{d_x}\right) \mathrm{vec}((A_*\ B_*)) + w_{i+1} \\ &= \Phi_i \vartheta_* + w_{i+1} \end{aligned}$$

□

Computation of the exact posterior To compute the posterior distribution we first observe:

$$p(\vartheta|\mathcal{D}) \propto p(\mathcal{D}|\vartheta)p(\vartheta).$$

We first compute $p(\mathcal{D}|\vartheta)$. By product rule we have:

$$p(\mathcal{D}|\vartheta) \propto \prod_{i=1}^{s} p(x_i|x_{i-1}, u_{i-1}, \vartheta).$$

Since $p(x_i|x_{i-1}, u_{i-1}, \vartheta)$ is the density of $\mathcal{N}(\Phi_{i-1}\vartheta, \sigma_w^2 I)$ we further have:

$$\begin{aligned} p(\mathcal{D}|\vartheta) &\propto \prod_{i=1}^{s} e^{-\frac{1}{2\sigma_w^2}(x_i-\Phi_{i-1}\vartheta)^\top(x_i-\Phi_{i-1}\vartheta)} \\ &= \exp\left(-\frac{1}{2\sigma_w^2}\sum_{i=1}^{s}\|x_i - \Phi_{i-1}\vartheta\|^2\right) \\ &\propto \exp\left(-\vartheta^\top\left(\frac{1}{2\sigma_w^2}\sum_{i=1}^{s}\Phi_{i-1}^\top\Phi_{i-1}\right)\vartheta + \left(\frac{1}{\sigma_w^2}\sum_{i=1}^{s}x_i^\top\Phi_{i-1}\right)\vartheta\right). \end{aligned}$$

Together with prior

$$p(\vartheta) \propto \exp\left(-\vartheta^\top\left(\frac{\lambda}{2\sigma_w^2}I\right)\vartheta\right)$$

we obtain:

$$p(\vartheta|\mathcal{D}) \propto \exp\left(-\vartheta^\top\left(\frac{1}{2\sigma_w^2}\sum_{i=1}^{s}\Phi_{i-1}^\top\Phi_{i-1} + \frac{\lambda}{2\sigma_w^2}I\right)\vartheta + \left(\frac{1}{\sigma_w^2}\sum_{i=1}^{s}x_i^\top\Phi_{i-1}\right)\vartheta\right).$$

Matching the coefficients we obtain that $\vartheta|\mathcal{D} \sim \mathcal{N}(\mu, \Sigma)$, where:

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\sigma_w^2}\sum_{i=1}^{s}\Phi_{i-1}^\top\Phi_{i-1} + \frac{\lambda}{\sigma_w^2}I \\ \mu &= \left(\frac{1}{\sigma_w^2}\sum_{i=1}^{s}\Phi_{i-1}^\top\Phi_{i-1} + \frac{\lambda}{\sigma_w^2}I\right)^{-1}\left(\frac{1}{\sigma_w^2}\sum_{i=1}^{s}\Phi_{i-1}^\top x_i\right) \\ &= \left(\sum_{i=1}^{s}\Phi_{i-1}^\top\Phi_{i-1} + \lambda I\right)^{-1}\left(\sum_{i=1}^{s}\Phi_{i-1}^\top x_i\right) \end{aligned}$$

For the estimators $\widehat{A}, \widehat{B}$ we than take MAP, for which we have $\operatorname{vec}((\widehat{A}\ \widehat{B})) = \mu$. We will now derive explicit value of $(\widehat{A}\ \widehat{B})$. We denote by $z_i = (x_i^\top\ u_i^\top)^\top$

and compute:

$$\begin{aligned}
\operatorname{vec}((\widehat{A}\ \widehat{B})) &= \left(\sum_{i=1}^{s} \Phi_{i-1}^{\top}\Phi_{i-1} + \lambda I\right)^{-1} \left(\sum_{i=1}^{s} \Phi_{i-1}^{\top} x_i\right) \\
&= \left(\sum_{i=1}^{s} (z_{i-1} \otimes I_{d_x})(z_{i-1}^{\top} \otimes I_{d_x}) + \lambda I\right)^{-1} \left(\sum_{i=1}^{s} (z_{i-1} \otimes I_{d_x}) x_i\right) \\
&= \left(\left(\sum_{i=1}^{s} z_{i-1} z_{i-1}^{\top} + \lambda I_{d_x+d_u}\right) \otimes I_{d_x}\right)^{-1} \operatorname{vec}\left(\sum_{i=1}^{s} x_i z_{i-1}^{\top}\right) \\
&= \left(\left(\sum_{i=1}^{s} z_{i-1} z_{i-1}^{\top} + \lambda I_{d_x+d_u}\right)^{-1} \otimes I_{d_x}\right) \operatorname{vec}\left(\sum_{i=1}^{s} x_i z_{i-1}^{\top}\right) \\
&= \operatorname{vec}\left(\left(\sum_{i=1}^{s} x_i z_{i-1}^{\top}\right)\left(\sum_{i=1}^{s} z_{i-1} z_{i-1}^{\top} + \lambda I_{d_x+d_u}\right)^{-1}\right).
\end{aligned}$$

Hence we obtained that MAP estimator satisfy:

$$(\widehat{A}\ \widehat{B}) = \left(\sum_{i=1}^{s} x_i z_{i-1}^{\top}\right)\left(\sum_{i=1}^{s} z_{i-1} z_{i-1}^{\top} + \lambda I_{d_x+d_u}\right)^{-1}$$

which is the same if we would compute the RLS estimator with regularizing parameter λ.

High probability regions In this section we will first derive D and its corresponding ellipsoid region where A_*, B_* lies with high probability. Then we will derive error bounds $\varepsilon_A, \varepsilon_B$ such that $\left\|A_* - \widehat{A}\right\|_2 \le \varepsilon_A$ and $\left\|B_* - \widehat{B}\right\|_2 \le \varepsilon_B$ w.p. at least $1-\delta$. Using the exact posterior distribution computed in section 3.1.1 we obtain that $(A_*, B_*) \in \Theta$ w.p. at least $1-\delta$, where

$$\Theta = \{(A,B) | \theta = \operatorname{vec}((A\ B)), (\theta-\mu)^{\top}\Sigma^{-1}(\theta-\mu) \le c_\delta\}.$$

Here c_δ is chosen in such a way that for $Z \sim \chi^2_{d_x^2+d_x d_u}$ we have: $\mathbb{P}(Z \ge c_\delta) = \delta$. For matrices A, B denote by $X^{\top} = (A\ B) - (\widehat{A}\ \widehat{B})$. For $(A,B) \in \Theta$ we have:

$$\begin{aligned}
1 &\ge \operatorname{vec}(X^{\top})^{\top}\left(\frac{1}{\sigma_w^2 c_\delta}\sum_{i=1}^{s} \Phi_{i-1}^{\top}\Phi_{i-1} + \frac{\lambda}{c_\delta \sigma_w^2} I\right)\operatorname{vec}(X^{\top}) \\
&= \operatorname{vec}(X^{\top})^{\top}\left(\left(\frac{1}{\sigma_w^2 c_\delta}\sum_{i=1}^{s} z_{i-1} z_{i-1}^{\top} + \frac{\lambda}{c_\delta \sigma_w^2} I_{d_x+d_u}\right) \otimes I_{d_x}\right)\operatorname{vec}(X^{\top}) \\
&= \operatorname{Tr}(X^{\top} D_s X) \ge \lambda_{\max}(X^{\top} D_s X).
\end{aligned}$$

where $D_s = \frac{1}{c_\delta \sigma_w^2}\left(\sum_{i=1}^{s} z_{i-1} z_{i-1}^\top + \lambda I_{d_x+d_u}\right)$. Hence with probability at least $1-\delta$ matrix (A_*, B_*) lies in the set $\{(A,B) | X^\top D_s X \preceq I, X^\top = (A\ B) - (\widehat{A}\ \widehat{B})\}$. To compute ε_A and ε_B we solve two maximization problems

$$\begin{aligned} &\max t \\ &\text{s.t.} \begin{pmatrix} tI_{d_x} & 0 \\ 0 & 0 \end{pmatrix} \preceq D_s \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} &\max t \\ &\text{s.t.} \begin{pmatrix} 0 & 0 \\ 0 & tI_{d_u} \end{pmatrix} \preceq D_s \end{aligned} \tag{3.5}$$

and set $\varepsilon_A = \frac{1}{\sqrt{t_A}}, \varepsilon_B = \frac{1}{\sqrt{t_B}}$, where t_A, t_B are optimal values of the problems given by eq. (3.4), eq. (3.5) respectively. If for $X^\top = (A_*\ B_*) - (\widehat{A}\ \widehat{B})$ we have $X^\top D_s X \preceq I$, than also:

$$X^\top \begin{pmatrix} \frac{1}{\varepsilon_A^2} I_{d_x} & 0 \\ 0 & 0 \end{pmatrix} X \preceq I$$

Since

$$X^\top \begin{pmatrix} \frac{1}{\varepsilon_A^2} I_{d_x} & 0 \\ 0 & 0 \end{pmatrix} X = \frac{1}{\varepsilon_A^2}(A_* - \widehat{A})(A_* - \widehat{A})^\top \preceq I,$$

we have that $\left\| A_* - \widehat{A} \right\|_2 \leq \varepsilon_A$. In a similar way we obtain $\left\| B_* - \widehat{B} \right\|_2 \leq \varepsilon_B$. Hence w.p. at least $1-\delta$ we have $\left\| A_* - \widehat{A} \right\|_2 \leq \varepsilon_A$ and $\left\| B_* - \widehat{B} \right\|_2 \leq \varepsilon_B$.

Explicit formulas for ε_A and ε_B In section 3.1.1 we showed how we can find $\varepsilon_A, \varepsilon_B$. In this section we will derive the explicit formulas for $\varepsilon_A, \varepsilon_B$, which we than use in the computation. From the definition it follows that ε_A is the smallest scalar which satisfies:

$$\begin{pmatrix} \frac{1}{\varepsilon_A^2} I_{d_x} & 0 \\ 0 & 0 \end{pmatrix} \preceq D_s. \tag{3.6}$$

Since D_s is symmetric positive definite matrix it has a spectral value decomposition $D_s = U^\top S U$, where U is orthogonal matrix, and $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_{d_x+d_u})$, with $\sigma_1 \geq \cdots \geq \sigma_{d_x+d_u} > 0$. Denoting $U = (U_1\ U_2)$, where $U_1 \in \mathbb{R}^{(d_x+d_u)\times d_x}$ and $U_2 \in \mathbb{R}^{(d_x+d_u)\times d_u}$, and conjugating eq. (3.6) with U we obtain that eq. (3.6) is equivalent to:

$$\frac{1}{\varepsilon_A^2} U_1 U_1^\top \preceq S. \tag{3.7}$$

Further denoting $P = S^{-1/2}$, conjugating eq. (3.7) with P and multiplying eq. (3.7) with ε_A^2 on both sides, eq. (3.7) is equivalent to:

$$PU_1U_1^\top P^\top \preceq \varepsilon_A^2 I$$

Since ε_A is the smallest such scalar we have: $\varepsilon_A^2 = \left\|PU_1U_1^\top P^\top\right\|_2$, from where we finally obtain the explicit formula for ε_A:

$$\varepsilon_A = \left\|PU_1\right\|_2.$$

With similar derivation we obtain also the explicit formula for ε_B:

$$\varepsilon_B = \left\|PU_2\right\|_2.$$

To conclude section 3.1.1 we will show that if we choose actions $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2 I)$ the estimation regions are consistent i.e. we will show that in this case the error upper bounds $\varepsilon_A, \varepsilon_B$ which we derived converge towards 0. For that we first look at the tail of χ^2 distribution.

χ^2 tail Let $Z \sim \chi_n^2$, where n are the degrees of freedom. The result of Laurent and Massart (2000) states that we have:

$$\mathbb{P}(Z \geq n + 2\sqrt{nz} + 2z) \leq e^{-z} \tag{3.8}$$

Since $5z \geq n + 2\sqrt{nz} + 2z$ for $z \geq n$, we have for $z \geq n$:

$$\Pr(Z \geq z) \leq e^{-\frac{z}{5}}. \tag{3.9}$$

In the next section we will use c_{δ_i} with property:

$$\Pr(Z \geq c_{\delta_i}) = \frac{\delta}{2i^2},$$

where we will send i towards infinity. Here we search for an upper bound for c_{δ_i}. Using eq. (3.9) we obtain that for $c_{\delta_i} \geq n$ we have:

$$\frac{\delta}{2i^2} = \Pr(Z \geq c_{\delta_i}) \leq e^{-\frac{c_{\delta_i}}{5}},$$

from where we obtain:

$$c_{\delta_i} \leq 5 \log \frac{2i^2}{\delta}.$$

Hence for every i we have:

$$c_{\delta_i} \leq n \vee 5 \log \frac{2i^2}{\delta} = \mathcal{O}(\log i)$$

Error bounds consistency First since we would like to have an algorithm which probability of failure is bounded above by δ, we define δ_i as

$$\delta_i = \frac{\delta}{2i^2}.$$

We understand δ_i as a probability of failure at step i. By union bound the total probability of failure is then bounded by:

$$\sum_{i \geq 1} \frac{\delta}{2i^2} = \delta \frac{\pi^2}{12} \leq \delta$$

Hence at step i we will use the constant c_{δ_i} to create matrix D_i. By the properties of the tail of χ^2 distribution derived in section 3.1.1 we have that $\frac{1}{c_{\delta_i}} \geq \Omega(\frac{1}{\log(i)})$.

Sarkar and Rakhlin (2018) showed that for every regular system A_*, B_* the empirical covariance matrix $V_s = \sum_{i=1}^{s} z_{i-1} z_{i-1}^\top \succeq \Omega(s) I$ if we play $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2 I)$.

Combining those two results together we obtain that D_s scales as $D_s \succeq \Omega(\frac{s}{\log(s)}) I$. Since we obtained $\varepsilon_A, \varepsilon_B$ by maximization, we have:

$$\frac{1}{\varepsilon_A^2}, \frac{1}{\varepsilon_B^2} \geq \Omega\left(\frac{s}{\log s}\right),$$

which is equivalent to:

$$\varepsilon_A, \varepsilon_B \leq \mathcal{O}\left(\sqrt{\frac{\log s}{s}}\right).$$

As $\mathcal{O}\left(\sqrt{\frac{\log s}{s}}\right) \overset{s \to \infty}{\to} 0$, we derived the consistency of the estimation error upper bounds.

3.1.2 Self Normalizing Processes

In section 3.1.1 we assumed a priori a belief about the true system A_*, B_* and build confidence regions based on this. In this section we will instead assume that we know an upper bound ϑ with $\|(A_* \; B_*)\|_2 \leq \vartheta$ and use the theory of multivariate Self Normalizing Processes studied by Peña et al. (2008).

For the RLS estimator we derived that:

$$((A_s \; B_s) - (A_* \; B_*))^\top = (V_s + \lambda I)^{-1} S_s - \lambda (V_s + \lambda I)^{-1} (A_* \; B_*)^\top, \quad (3.10)$$

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 $\|(A_* \; B_*)\|_2 \leq \vartheta$ 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 $\left\|(V_s + \lambda I)^{-\frac{1}{2}} S_s\right\|_2$. The result builds on ideas from Abbasi-Yadkori et al. (2011) and Sarkar and Rakhlin (2018). Since we will use the result of Self Normalzing process also in section 5.2, we take here a bit more general approach.

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 subG$_l(\sigma^2)$, meaning that we have for every $\|u\|_2 = 1$, $\gamma \geq 0$ 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 3.2 *Let us be in the aforementioned setting, then we have w.p. at least $1-\delta$:*

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

Proof For $\varepsilon \in (0,1)$ we obtain from Proposition 2.18:

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

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 subG$_l(\sigma^2)$ random variables, $w_s^\top u$ are independent subG(σ^2) random variables. Hence we can apply Theorem 3 of Abbasi-Yadkori et al. (2011). Setting

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

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

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

Using the lemma 3.2 the bound on $\left\|(V_s + \lambda I)^{-\frac{1}{2}} S_s\right\|_2$ follows.

Proposition 3.3 *In the aforementioned setting let $\varepsilon \in (0, 1)$ arbitrary. Then we have w.p. at least $1 - \delta$:*

$$\forall s \geq 0 : \left\| (V_s + \lambda I)^{-\frac{1}{2}} S_s \right\|_2^2 \leq \frac{2\sigma^2}{(1-\varepsilon)^2} \log \left(\frac{\det(V_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\varepsilon}\right)^d}{\delta} \right)$$

Proof Denote by $\mathcal{F}_i = \sigma((u_j)_{j \leq i}, (w_j)_{j \leq i})$. Apply Lemma 3.2 with $\mathcal{S}_s = S_s$ and $\mathcal{V}_s = V_s$ and the result follows. □

Collecting the results together, we arrive at the data dependent $\varepsilon_A, \varepsilon_B$ which we can use in algorithm 1.

Corollary 3.4 *For the RLS estimates we have w.p. at least $1 - \delta$ for every $s \geq 0$:*

$$\begin{aligned} \|A_s - A_*\|_2 \leq{}& \frac{\sigma_w}{1-\varepsilon} \sqrt{2 \log \left(\frac{\det(V_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\varepsilon}\right)^d}{\delta} \right)} \left\| (I_d \; 0)(V_s + \lambda I)^{-1/2} \right\|_2 \\ &+ \lambda \left\| (I_d \; 0)(V_s + \lambda I)^{-1} \right\|_2 \vartheta \\ \|B_s - B_*\|_2 \leq{}& \frac{\sigma_w}{1-\varepsilon} \sqrt{2 \log \left(\frac{\det(V_s + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\varepsilon}\right)^d}{\delta} \right)} \left\| (0 \; I_k)(V_s + \lambda I)^{-1/2} \right\|_2 \\ &+ \lambda \left\| (0 \; I_k)(V_s + \lambda I)^{-1} \right\|_2 \vartheta \end{aligned}$$

Proof Since the analysis for the matrix A_s is very much the same as for the matrix B_s we will do it just for A_s. First observe that we have:

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

Next using triangle inequality we obtain:

$$\|A_s - A\|_2 \leq I_1 + I_2,$$

where $I_1 = \left\| (I_d \; 0) \, (V_k + \lambda I)^{-1} \, S_k \right\|_2$ and $I_2 = \left\| \lambda (I_d \; 0) \, (V_k + \lambda I)^{-1} \, (A \; B) \right\|_2$. The first term is by Lemma 3.2 bounded w.p. at least $1 - \delta$:

$$\begin{aligned} I_1 &\leq \left\| (I_d \; 0) \, (V_k + \lambda I)^{-\frac{1}{2}} \right\|_2 \left\| (V_k + \lambda I)^{-\frac{1}{2}} \, S_k \right\|_2 \\ &\leq \left\| (I_d \; 0) \, (V_k + \lambda I)^{-\frac{1}{2}} \right\|_2 \frac{\sigma_w}{1-\varepsilon} \sqrt{2 \log \left(\frac{\det(V_t + \lambda I)^{\frac{1}{2}}}{\det(\lambda I)^{\frac{1}{2}}} \frac{\left(1 + \frac{2}{\varepsilon}\right)^d}{\delta} \right)}. \end{aligned}$$

With the bound on I_2 term:

$$I_2 = \left\| \lambda (I_d \; 0) \, (V_k + \lambda I)^{-1} \, (A \; B) \right\|_2 \leq \lambda \left\| (I_d \; 0) \, (V_k + \lambda I)^{-1} \right\|_2 \vartheta,$$

we conclude the proof. □

Since the upper bound in Corollary 3.4 holds for every $\varepsilon \in (0,1)$, we optimize over ε to obtain the best possible bound while running algorithm 1. We showed in section 3.1.1 that data dependent upper bounds which we derived in Bayesian setting are consistent. Experiments in section 6.1 show that errors which we obtain from the theory of multivariate Self Normalizing Processes perform comparable to the errors from Bayesian setting, however we do not have a guarantee that they are consistent.

3.2 Robust control synthesis

In section 3.1 we demonstrated how we can obtain region Θ around RLS estimates A_s, B_s such that $(A_*, B_*) \in \Theta$ with probability at least $1-\delta$. In this section we will show different approaches how we can search for controller K which stabilizes every system in region Θ. If the region Θ is large the controller with such properties usually does not exist. However as the estimation becomes more accurate region Θ shrinks. We will further give a sufficient condition on the size of Θ when we have a guarantee that we can synthesize a controller which stabilizes every system inside Θ.

3.2.1 Robust controller from SLS

We will start with the results which we presented in section 2.2.2. Dean et al. (2017) showed that from any feasible solution of eq. (2.15) we obtain a controller which stabilizes all systems A, B with $\left\|A-\widehat{A}\right\|_2 \le \varepsilon_A$, $\left\|B-\widehat{B}\right\|_2 \le \varepsilon_B$. Since by running algorithm 1 we update RLS estimates and their estimation errors after every step we can try to solve at every time step a SDP

$$\begin{aligned} \min_{s\in[0,1), P\succ 0, S} \quad & s \\ \text{s.t.:} \quad & \begin{pmatrix} P-I & \widehat{A}P+\widehat{B}S & 0 \\ (\widehat{A}P+\widehat{B}S)^\top & P & \begin{pmatrix}\varepsilon_A P\\ \varepsilon_B S\end{pmatrix}^\top \\ 0 & \begin{pmatrix}\varepsilon_A P\\ \varepsilon_B S\end{pmatrix} & \frac{1}{2}sI \end{pmatrix} \succeq 0. \end{aligned} \tag{3.11}$$

The first time we find a feasible solution to SDP (3.11) we obtain the controller K which stabilizes system A_*, B_*, with probability at least $1-\delta$, as $K = SP^{-1}$. As we have seen in section 3.1.1 we derived $\varepsilon_A, \varepsilon_B$ based on ellipsoid associated with positive definite matrix D_s. In the following we will derive a semi-definite constraint which deals with ellipsoidal region Θ. To formalize, we denote $X^\top = (A\ B) - (\widehat{A}\ \widehat{B})$ and would like to solve the following problem:

$$\begin{aligned}&\text{find } K\\&\text{s.t.}\forall(A,B) \text{ with } X^\top DX \preceq I:\\&\qquad \|\Delta\|_{\mathcal{H}_\infty} < 1.\end{aligned}\tag{3.12}$$

Using the explicit value of Δ, presented in eq. (2.13), the eq. (3.12) is equivalent to:

$$\begin{aligned}&\text{find } K\\&\forall X \text{ with } X^\top DX \preceq I:\\&\qquad \left\|X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1}\right\|_{\mathcal{H}_\infty} < 1.\end{aligned}\tag{3.13}$$

In the problem posed by eq. (3.13) there are two issues which we need to solve. First there is the condition that we would like to find controller K for which a constraint holds for every X with $X^\top DX \preceq I$. We will solve this by application of S-lemma (c.f. theorem 2.25). The other issue is the $\mathcal{H}_\infty$ norm which we will transform to semi-definite constraint by application of KYP lemma (c.f. lemma 2.10).

First S then KYP lemma The constraint

$$\left\|X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1}\right\|_{\mathcal{H}_\infty} < 1$$

is equivalent to the constraint that for every $z \in \partial\mathbb{D}$:

$$\left\|X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1}\right\|_2 < 1.$$

The latter constraint is equivalent to:

$$X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1}(zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I\\ K\end{pmatrix}^\top X \prec I$$

which is further equivalent to:

$$\begin{pmatrix} I & (zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I\\ K\end{pmatrix}^\top X\\ X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & I\end{pmatrix} \succ 0.$$

The problem given by eq. (3.13) is therefore equivalent to:

$$\begin{aligned}&\forall X \text{ with } X^\top DX \preceq I, \forall z \in \partial\mathbb{D}:\\&\begin{pmatrix} I & (zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I\\ K\end{pmatrix}^\top X\\ X^\top \begin{pmatrix} I\\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & I\end{pmatrix} \succ 0.\end{aligned}\tag{3.14}$$

By theorem 2.25, eq. (3.14) is further equivalent to:

$$\forall z \in \partial\mathbb{D}, \exists t \in (0,\infty) \text{ s.t. :}$$
$$\begin{pmatrix} I & 0 & (zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I \\ K\end{pmatrix}^{\top} \\ 0 & (1-t)I & 0 \\ \begin{pmatrix} I \\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & 0 & tD \end{pmatrix} \succ 0. \tag{3.15}$$

Observe that eq. (3.15) is then equivalent to:

$$\forall z \in \partial\mathbb{D}, \exists t \in (0,1) \text{ s.t. :}$$
$$\begin{pmatrix} I & (zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I \\ K\end{pmatrix}^{\top} \\ \begin{pmatrix} I \\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & tD \end{pmatrix} \succ 0. \tag{3.16}$$

Next observe that in eq. (3.16) if the positive definite constraint holds for one t it will also hold for all $t' \in [t,1)$. Therefore instead of searching at every $z \in \partial\mathbb{D}$ for suitable t we can equivalently search uniformly in t – we can take the supremum. Hence eq. (3.16) is equivalent to:

$$\exists t \in (0,1) \text{ s.t. } \forall z \in \partial\mathbb{D} \text{ :}$$
$$\begin{pmatrix} I & (zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I \\ K\end{pmatrix}^{\top} \\ \begin{pmatrix} I \\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & tD \end{pmatrix} \succ 0. \tag{3.17}$$

Matrix D is positive definite hence $D^{-\frac{1}{2}}$ exists. Observe that conjugating the positive definite constraint in eq. (3.17) with matrix $\operatorname{diag}(I, D^{-\frac{1}{2}})$ the eq. (3.17) is equivalent to:

$$\exists t \in (0,1) \text{ s.t. } \forall z \in \partial\mathbb{D} \text{ :}$$
$$\begin{pmatrix} I & \frac{1}{\sqrt{t}}(zI-\widehat{A}-\widehat{B}K)^{-\top}\begin{pmatrix} I \\ K\end{pmatrix}^{\top} D^{-\frac{1}{2}} \\ \frac{1}{\sqrt{t}}D^{-\frac{1}{2}}\begin{pmatrix} I \\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1} & I \end{pmatrix} \succ 0, \tag{3.18}$$

which is by using lemma 2.19 further equivalent to:

$$\exists t \in (0,1) \text{ s.t. } \forall z \in \partial\mathbb{D} \text{ :}$$
$$\left\| \frac{1}{\sqrt{t}} D^{-\frac{1}{2}} \begin{pmatrix} I \\ K\end{pmatrix}(zI-\widehat{A}-\widehat{B}K)^{-1}\right\|_2 < 1. \tag{3.19}$$

By the definition of $\mathcal{H}_\infty$ norm this is further equivalent to:

$$\begin{aligned}&\exists t \in (0,1) \text{ s.t. :}\\ &\quad \left\| \frac{1}{\sqrt{t}} D^{-\frac{1}{2}} \begin{pmatrix} I \\ K \end{pmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1} \right\|_{\mathcal{H}_\infty} < 1.\end{aligned} \tag{3.20}$$

Now we are in the position to apply lemma 2.10. This yields that the eq. (3.20) is equivalent to:

$$\begin{aligned}&\exists t \in (0,1), \exists P \succ 0 \text{ s.t. :}\\ &\begin{pmatrix} \widehat{A} + \widehat{B}K & I \\ \frac{1}{\sqrt{t}} D^{-\frac{1}{2}} \begin{pmatrix} I \\ K \end{pmatrix} & 0 \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \widehat{A} + \widehat{B}K & I \\ \frac{1}{\sqrt{t}} D^{-\frac{1}{2}} \begin{pmatrix} I \\ K \end{pmatrix} & 0 \end{pmatrix}^\top \preceq \begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix}.\end{aligned} \tag{3.21}$$

Applying lemma 2.19 we observe that the positive definite constraint given by eq. (3.21) is equivalent to:

$$\begin{pmatrix} P & 0 & \widehat{A} + \widehat{B}K & I \\ 0 & I & \frac{1}{\sqrt{t}} D^{-\frac{1}{2}} \begin{pmatrix} I \\ K \end{pmatrix} & 0 \\ (\widehat{A} + \widehat{B}K)^\top & \frac{1}{\sqrt{t}} \begin{pmatrix} I \\ K \end{pmatrix}^\top D^{-\frac{1}{2}} & P^{-1} & 0 \\ I & 0 & 0 & I \end{pmatrix} \succeq 0 \tag{3.22}$$

Conjugating with matrix $\operatorname{diag}(I, \sqrt{t}I, P, I)$ and denoting $S = KP$ we obtain that eq. (3.22) is equivalent to:

$$\begin{pmatrix} P & 0 & \widehat{A}P + \widehat{B}S & I \\ 0 & tI & D^{-\frac{1}{2}} \begin{pmatrix} P \\ S \end{pmatrix} & 0 \\ (\widehat{A}P + \widehat{B}S)^\top & \begin{pmatrix} P \\ S \end{pmatrix}^\top D^{-\frac{1}{2}} & P & 0 \\ I & 0 & 0 & I \end{pmatrix} \succeq 0 \tag{3.23}$$

Taking lemma 2.19 again we obtain that the eq. (3.23) is equivalent to:

$$\begin{pmatrix} P - I & 0 & \widehat{A}P + \widehat{B}S \\ 0 & tI & D^{-\frac{1}{2}} \begin{pmatrix} P \\ S \end{pmatrix} \\ (\widehat{A}P + \widehat{B}S)^\top & \begin{pmatrix} P \\ S \end{pmatrix}^\top D^{-\frac{1}{2}} & P \end{pmatrix} \succeq 0 \tag{3.24}$$

Conjugating by matrix

$$\begin{pmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{pmatrix}$$

we obtain that eq. (3.24) is further equivalent to:

$$\begin{pmatrix} P-I & \widehat{A}P+\widehat{B}S & 0 \\ (\widehat{A}P+\widehat{B}S)^\top & P & \begin{pmatrix} P \\ S \end{pmatrix}^\top D^{-\frac{1}{2}} \\ 0 & D^{-\frac{1}{2}}\begin{pmatrix} P \\ S \end{pmatrix} & tI \end{pmatrix} \succeq 0. \tag{3.25}$$

We derived that the problem given by eq. (3.12) is equivalent to:

$$\begin{aligned} &\text{find } t \in (0,1), P \succ 0, S \\ &\text{s.t.} \begin{pmatrix} P-I & \widehat{A}P+\widehat{B}S & 0 \\ (\widehat{A}P+\widehat{B}S)^\top & P & \begin{pmatrix} P \\ S \end{pmatrix}^\top D^{-\frac{1}{2}} \\ 0 & D^{-\frac{1}{2}}\begin{pmatrix} P \\ S \end{pmatrix} & tI \end{pmatrix} \succeq 0. \end{aligned} \tag{3.26}$$

Hence we can solve for example SDP:

$$\begin{aligned} &\min_{t\in(0,1), P\succ 0, S} \quad t \\ &\text{s.t.} \begin{pmatrix} P-I & \widehat{A}P+\widehat{B}S & 0 \\ (\widehat{A}P+\widehat{B}S)^\top & P & \begin{pmatrix} P \\ S \end{pmatrix}^\top D^{-\frac{1}{2}} \\ 0 & D^{-\frac{1}{2}}\begin{pmatrix} P \\ S \end{pmatrix} & tI \end{pmatrix} \succeq 0. \end{aligned} \tag{3.27}$$

From the optimal solution of SDP given by eq. (3.27) we obtain the stabilizing controller via $K = SP^{-1}$.

3.2.2 Robust controller from SDP

In this section we will derive the robust controller synthesis based on the SDP program described in section 2.5. First observe that the SDP given by eq. (2.20) can be equivalently written in the form:

$$\begin{aligned} &\min_{\Sigma\succeq 0} \quad \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle \\ &\quad\text{s.t.} \quad \Sigma_{xx} \succeq (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I, \end{aligned} \tag{3.28}$$

Lemma 3.5 *For the optimal* Σ *of SDP given by eq.* (3.28) *we have:*

$$\Sigma_{xx} = (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I.$$

Proof Assume that

$$\Sigma_{xx} = (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I + E,$$

where $E \succeq 0$ and $E \neq 0$. Since

$$\begin{aligned}\Sigma_{xx} - E &= (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I \\ &\succeq (A_* \; B_*)\left(\Sigma - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}\right)(A_* \; B_*)^\top + \sigma_w^2 I,\end{aligned}$$

also

$$\Sigma - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}$$

is feasible solution. And since Q is positive semi-definite its cost is smaller than the one of Σ. □

We will now write the SDP given by eq. (3.28) in a robust variant. Here we denote $\Theta = \{(A, B) | X^\top D X \preceq I, X^\top = (A \; B) - (\widehat{A} \; \widehat{B})\}$:

$$\begin{aligned}&\min_{\Sigma \succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle \\ &\text{s.t. } \forall (A, B) \in \Theta : \; \Sigma_{xx} \succeq (A \; B)\Sigma(A \; B)^\top + \sigma_w^2 I\end{aligned} \tag{3.29}$$

Next we show that from any feasible solution Σ of the SDP given by eq. (3.29) we can synthesize a controller K which stabilizes every system in Θ.

Lemma 3.6 *Let Σ be a feasible solution of SDP given by eq.* (3.29). *Then we have:*

1. *Σ' of the form*

   $$\Sigma' = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx} K^\top \\ K\Sigma_{xx} & K\Sigma_{xx}K^\top \end{pmatrix},$$

   where $K = \Sigma_{ux}\Sigma_{xx}^{-1}$, is also feasible solution of the SDP given by (3.29) *with cost at most that of* Σ.

2. *For $K = \Sigma_{ux}\Sigma_{xx}^{-1}$ we have: $\forall (A, B) \in \Theta : \; \rho(A + BK) < 1$.*

Proof Since

$$\Sigma - \Sigma' = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{uu} - \Sigma_{ux}\Sigma_{xx}^{-1}\Sigma_{xu} \end{pmatrix}$$

and $\Sigma_{uu} - \Sigma_{ux}\Sigma_{xx}^{-1}\Sigma_{xu}$ is Schur complement of Σ we have $\Sigma_{uu} - \Sigma_{ux}\Sigma_{xx}^{-1}\Sigma_{xu} \succeq 0$ and consequently $\Sigma \succeq \Sigma'$. Now fix aribtrary $(A, B) \in \Theta$. We have $\Sigma_{xx} \succeq$

$(A\ B)\Sigma(A\ B)^\top + \sigma_w^2 I \succeq (A\ B)\Sigma'(A\ B)^\top + \sigma_w^2 I$, therefore Σ' is feasible. Next we will show $\rho(A+BK) < 1$. The semi-definite inequality

$$\Sigma_{xx} \succeq (A\ B)\Sigma'(A\ B)^\top + \sigma_w^2 I$$

is equivalent to:

$$\Sigma_{xx} \succeq (A+BK)\Sigma_{xx}(A+BK)^\top + \sigma_w^2 I.$$

Let μ, v be eigenpair of $(A+BK)^\top$. We have:

$$v^H\Sigma_{xx}v \geq |\mu|^2 v^H\Sigma_{xx}v + \sigma_w^2 \|v\|^2 > |\mu|^2 v^H\Sigma_{xx}v.$$

Hence $|\mu| < 1$. □

In the following we will rewrite SDP given by eq. (3.29) to a convex SDP using theorem 2.24. Inserting $(A\ B) = X^\top + (\widehat{A}\ \widehat{B})$ to eq. (3.29) we obtain that SDP given by eq. (3.29) is equivalent to:

$$\begin{aligned}\min_{\Sigma\succeq 0} & \left\langle \begin{pmatrix} Q & 0 \\ 0 & R\end{pmatrix}, \Sigma \right\rangle \\ & \text{s.t. } \forall (A,B) \in \Theta : \\ & \Sigma_{xx} - \sigma_w^2 I - X^\top\Sigma X - X^\top\Sigma(\widehat{A}\ \widehat{B})^\top - (\widehat{A}\ \widehat{B})\Sigma X - (\widehat{A}\ \widehat{B})\Sigma(\widehat{A}\ \widehat{B})^\top \succeq 0\end{aligned} \tag{3.30}$$

The latter is by theorem 2.24 equivalent to:

$$\begin{aligned}\min_{\Sigma\succeq 0, t\geq 0} & \left\langle \begin{pmatrix} Q & 0 \\ 0 & R\end{pmatrix}, \Sigma \right\rangle \\ \text{s.t. } & \begin{pmatrix} \Sigma_{xx} - (\widehat{A}\ \widehat{B})\Sigma(\widehat{A}\ \widehat{B})^\top - (t+\sigma_w^2)I & (\widehat{A}\ \widehat{B})\Sigma \\ \Sigma(\widehat{A}\ \widehat{B})^\top & tD - \Sigma \end{pmatrix} \succeq 0.\end{aligned} \tag{3.31}$$

This is a convex formulation of SDP and we can solve it using e.g. MOSEK (ApS, 2020).

3.3 But they are the same

As we have seen in section 3.2 any controller which we synthesize from SDP (3.27) or SDP (3.31) stabilizes every system inside ellipsoidal region around estimates $\widehat{A}, \widehat{B}$ given by $\{(A,B) | X^\top D X \preceq I, X^\top = (A\ B) - (\widehat{A}\ \widehat{B})\}$. Even though the way we obtained SDP (3.27) and SDP (3.31) are different we show in this section that in fact they are the same, meaning that as soon as one SDP is feasible the other is feasible as well. To see this first note that semi-definite constraint in eq. (3.27) ca be rewritten as:

$$\begin{pmatrix} P - I & (\widehat{A}\ \widehat{B}) \begin{pmatrix} I \\ K \end{pmatrix} P & 0 \\ P \begin{pmatrix} I \\ K \end{pmatrix}^\top (\widehat{A}\ \widehat{B})^\top & P & P \begin{pmatrix} I \\ K \end{pmatrix}^\top \\ 0 & \begin{pmatrix} I \\ K \end{pmatrix} P & tD \end{pmatrix} \succeq 0. \tag{3.32}$$

Conjugating by matrix

$$\begin{pmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{pmatrix}$$

we obtain that eq. (3.32) is equivalent to:

$$\begin{pmatrix} P - I & 0 & (\widehat{A}\ \widehat{B}) \begin{pmatrix} I \\ K \end{pmatrix} P \\ 0 & tD & \begin{pmatrix} I \\ K \end{pmatrix} P \\ P \begin{pmatrix} I \\ K \end{pmatrix}^\top (\widehat{A}\ \widehat{B})^\top & P \begin{pmatrix} I \\ K \end{pmatrix}^\top & P \end{pmatrix} \succeq 0. \tag{3.33}$$

We can rewrite eq. (3.33) using lemma 2.19 to:

$$\begin{pmatrix} P - I & 0 \\ 0 & tD \end{pmatrix} - \begin{pmatrix} (\widehat{A}\ \widehat{B}) \begin{pmatrix} I \\ K \end{pmatrix} P \\ \begin{pmatrix} I \\ K \end{pmatrix} P \end{pmatrix} P^{-1} \left(P \begin{pmatrix} I \\ K \end{pmatrix}^\top (\widehat{A}\ \widehat{B})^\top \quad P \begin{pmatrix} I \\ K \end{pmatrix}^\top \right) \succeq 0, \tag{3.34}$$

which is, by multiplying the matrices, further equivalent to:

$$\begin{pmatrix} P - (\widehat{A}\ \widehat{B}) \begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top (\widehat{A}\ \widehat{B})^\top - I & -(\widehat{A}\ \widehat{B}) \begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top \\ -\begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top (\widehat{A}\ \widehat{B})^\top & tD - \begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top \end{pmatrix} \succeq 0 \tag{3.35}$$

We know by lemma 3.6 that the optimal solution of SDP (3.31) is parametrized as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx} K^\top \\ K\Sigma_{xx} & K\Sigma_{xx}K^\top \end{pmatrix} = \begin{pmatrix} I \\ K \end{pmatrix} \Sigma_{xx} \begin{pmatrix} I \\ K \end{pmatrix}^\top$$

Hence by denoting $U = \begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top$ we obtain that eq. (3.35) can be rewritten as:

$$\begin{pmatrix} U_{xx} - (\widehat{A}\ \widehat{B})U(\widehat{A}\ \widehat{B})^\top - I & -(\widehat{A}\ \widehat{B})U \\ -U(\widehat{A}\ \widehat{B})^\top & tD - U \end{pmatrix} \succeq 0 \tag{3.36}$$

By lemma A.3 we further obtain that eq. (3.36) is equivalent to:

$$\begin{pmatrix} U_{xx} - (\widehat{A}\ \widehat{B})U(\widehat{A}\ \widehat{B})^\top - I & (\widehat{A}\ \widehat{B})U \\ U(\widehat{A}\ \widehat{B})^\top & tD - U \end{pmatrix} \succeq 0 \tag{3.37}$$

To show that SDP (3.27) is feasible if and only if SDP (3.31) is feasible is then equivalent to show that

$$\begin{aligned} &\exists U \succeq 0, s \in (0,1) \text{ s.t.:} \\ &\qquad \begin{pmatrix} U_{xx} - (\widehat{A}\ \widehat{B})U(\widehat{A}\ \widehat{B})^\top - I & (\widehat{A}\ \widehat{B})U \\ U(\widehat{A}\ \widehat{B})^\top & sD - U \end{pmatrix} \succeq 0 \end{aligned} \tag{3.38}$$

is equivalent to:

$$\begin{aligned} &\exists \Sigma \succeq 0, t \geq 0 \text{ s.t.:} \\ &\qquad \begin{pmatrix} \Sigma_{xx} - (\widehat{A}\ \widehat{B})\Sigma(\widehat{A}\ \widehat{B})^\top - (t + \sigma_w^2)I & (\widehat{A}\ \widehat{B})\Sigma \\ \Sigma(\widehat{A}\ \widehat{B})^\top & tD - \Sigma \end{pmatrix} \succeq 0 \end{aligned} \tag{3.39}$$

Assume that we have eq. (3.38). Multiply semi-definite constraint in eq. (3.38) with $\frac{\sigma_w^2}{1-s}$ and denote $t = \frac{s\sigma_w^2}{1-s}, \Sigma = \frac{\sigma_w^2}{1-s}U$. With such a notation we have:

$$\begin{pmatrix} \Sigma_{xx} - (\widehat{A}\ \widehat{B})\Sigma(\widehat{A}\ \widehat{B})^\top - (t + \sigma_w^2)I & (\widehat{A}\ \widehat{B})\Sigma \\ \Sigma(\widehat{A}\ \widehat{B})^\top & tD - \Sigma \end{pmatrix} \succeq 0. \tag{3.40}$$

Since $\Sigma = \frac{\sigma_w^2}{1-s}U \succeq 0$ and $t = \frac{s\sigma_w^2}{1-s} \geq 0$ we see that condition given by eq. (3.39) is satisfied. To show the equivalence in other direction assume that we have eq. (3.39). Multiplying semi-definite constraint in eq. (3.39) with $\frac{1}{t+\sigma_w^2}$ and denoting $s = \frac{t}{t+\sigma_w^2}, U = \frac{1}{t+\sigma_w^2}\Sigma$ we obtain:

$$\begin{pmatrix} U_{xx} - (\widehat{A}\ \widehat{B})U(\widehat{A}\ \widehat{B})^\top - I & (\widehat{A}\ \widehat{B})U \\ U(\widehat{A}\ \widehat{B})^\top & sD - U \end{pmatrix} \succeq 0. \tag{3.41}$$

Since $U = \frac{1}{t+\sigma_w^2}\Sigma \succeq 0$ and $s = \frac{t}{t+\sigma_w^2} < 1$ we obtain that eq. (3.38) is satisfied. Hence we obtained that as soon as one of the SDP eq. (3.27) or SDP eq. (3.31) is feasible, the other is feasible as well.

3.4 Minimize spectral norm of closed loop system

In this section we show how to synthesize a controller K which minimizes the maximal closed loop system norm for systems in

$$\Theta = \{(A\ B) | X^\top D X \preceq I, X^\top = (\widehat{A}\ \widehat{B}) - (A\ B)\}. \tag{3.42}$$

We formulate the problem as:

$$\begin{aligned} &\min_{t \geq 0, K} t \\ &\quad \text{s.t. } \forall (A\ B) \in \Theta: \quad \|A + BK\|_2 \leq t. \end{aligned} \tag{3.43}$$

If the optimal solution of the problem eq. (3.43) is less than 1, then controller K stabilizes every system inside region Θ. Note that with this approach it can happen that even if the Θ goes to zero, such a controller does not exist, since $\rho(A) \leq \|A\|_2$ and the difference can be arbitrarily large. To be specfic, consider the system:

$$A_* = \begin{pmatrix} 0.8 & 0 & 0 \\ 0 & 0.5 & 5 \\ 0 & 0 & 0.5 \end{pmatrix}, \quad B_* = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

The system is stabilizable since for controller $K = 0$ we have $\rho(A_* + B_* K) = 0.8 < 1$, however for any controller $K \in \mathbb{R}^{1 \times 3}$ we have $5 \leq \|A_* + B_* K\|_2$.

Next we will transform problem eq. (3.43) to convex SDP using theorem 2.25. To reformulate the problem in such a way observe first that the constraint $\|A + BK\|_2 \leq t$ can be rewritten using lemma 2.19 as:

$$\begin{aligned} &\|A + BK\|_2 \leq t \\ \iff &(A + BK)^\top (A + BK) \preceq t^2 I \\ \overset{\text{lemma 2.19}}{\iff} &\begin{pmatrix} tI & (A + BK)^\top \\ A + BK & tI \end{pmatrix} \succeq 0 \end{aligned}$$

Using the notation from the definition of high probability region given in eq. (3.42) we reformulate the minimization problem given by eq. (3.43) to:

$$\begin{aligned} &\min_{t \geq 0, K} t \\ &\quad \text{s.t. } \forall X \text{ with } X^\top D X \preceq I: \\ &\quad \begin{pmatrix} tI & (\widehat{A} + \widehat{B}K)^\top - \begin{pmatrix} I \\ K \end{pmatrix}^\top X \\ \left((\widehat{A} + \widehat{B}K)^\top - \begin{pmatrix} I \\ K \end{pmatrix}^\top X \right)^\top & tI \end{pmatrix} \succeq 0 \end{aligned} \tag{3.44}$$

Applying theorem 2.25 we obtain that the eq. (3.44) is equivalent to:

$$\begin{aligned}&\min_{t\geq 0,\lambda\geq 0,K} \ t\\ &\text{s.t.} \begin{pmatrix} tI & (\widehat{A}+\widehat{B}K)^\top & -\begin{pmatrix} I\\ K\end{pmatrix}^\top \\ \widehat{A}+\widehat{B}K & (t-\lambda)I & 0 \\ -\begin{pmatrix} I\\ K\end{pmatrix} & 0 & \lambda D\end{pmatrix} \succeq 0\end{aligned} \tag{3.45}$$

which is a convex SDP. At the same time we can also use SDP (3.45) to bound for a given controller K' the norm of associated closed loop matrix:

$$\begin{aligned}&\min_{t\geq 0,\lambda\geq 0} \ t\\ &\text{s.t.} \begin{pmatrix} tI & (\widehat{A}+\widehat{B}K')^\top & -\begin{pmatrix} I\\ K'\end{pmatrix}^\top \\ \widehat{A}+\widehat{B}K' & (t-\lambda)I & 0 \\ -\begin{pmatrix} I\\ K'\end{pmatrix} & 0 & \lambda D\end{pmatrix} \succeq 0\end{aligned} \tag{3.46}$$

For the optimal t which we obtain from the solution of SDP (3.46) we have that with probability at least $1-\delta$:

$$\left\|A_* + B_*K'\right\|_2 \leq t.$$

3.5 Feasibility conditions

In section 3.2 we described how we can obtain a controller which stabilizes every system inside some region Θ. The controller synthesis introduced requires solving a SDP. As we start evolving the system usually the proposed SDPs are not feasible and become feasible only after the region Θ shrinks enough. In this section we will provide a sufficient condition for feasibility of the proposed SDPs.

From lemma 2.9 follows that if $(\varepsilon_A + \varepsilon_B \left\|K\right\|_2) \left\|\mathfrak{R}_{A_*+B_*K}\right\|_{\mathcal{H}_\infty} \leq (1+\sqrt{2})^{-1}$ then the SDP given by eq. (3.11) is feasible. Denote by $\varepsilon = \varepsilon_A \vee \varepsilon_B$. Then a sufficient condition for feasibility of SDP given by eq. (3.27) is:

$$\frac{1}{\sqrt{\lambda_{\min}(D)}} \leq \frac{1}{(1+\left\|K\right\|_2) \left\|\mathfrak{R}_{A_*+B_*K}\right\|_{\mathcal{H}_\infty}(1+\sqrt{2})},$$

which is equivalent to:

$$\mathcal{O}\left(\left((1+\left\|K\right\|_2) \left\|\mathfrak{R}_{A_*+B_*K}\right\|_{\mathcal{H}_\infty}\right)^2\right) \leq \lambda_{\min}(D). \tag{3.47}$$

To rewrite the condition eq. (3.47) to a form without $\mathcal{H}_\infty$ norm we first show the following lemma:

Lemma 3.7 *For any square matrix A with $\rho(A) < 1$ and for a positive definite matrix P, which is the solution of the equation:*

$$P = A^\top P A + I,$$

we have:

$$\left\|(zI - A)^{-1}\right\|_{\mathcal{H}_\infty} \leq 2\sqrt{\kappa(P)}\, \|P\|_2 .$$

Proof Since

$$(zI - A)^{-1} = \sum_{i=0}^{\infty} \frac{1}{z^{i+1}} A^i,$$

we have:

$$\begin{aligned}
\left\|(zI - A)^{-1}\right\|_{\mathcal{H}_\infty} &\leq \sum_{i\geq 0} \left\|A^i\right\|_2 = \sum_{i\geq 0} \sqrt{\|(A^i)^\top A^i\|_2} \\
&\leq \sum_{i\geq 0} \sqrt{\frac{\|(A^i)^\top P A^i\|_2}{\sigma_{\min}(P)}} \\
&\leq \sqrt{\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}} \sum_{i\geq 0} \sqrt{\left(1 - \frac{1}{\|P\|_2}\right)^i} \\
&= \sqrt{\kappa(P)} \frac{1}{1 - \sqrt{1 - \frac{1}{\|P\|_2}}} \leq 2\sqrt{\kappa(P)}\, \|P\|_2 ,
\end{aligned}$$

where we used the fact that $P \succeq I$ on multiple steps. □

Using SDP (2.20) we obtain:

$$\Sigma_{xx}^* = (A_* + B_* K_*) \Sigma_{xx}^* (A_* + B_* K_*)^\top + \sigma_w^2 I.$$

The by lemma 3.7 follows that a sufficient condition for SDP (3.27) to be feasible is:

$$\mathcal{O}\left(\frac{(1 + \|K_*\|_2)^2 \left\|\Sigma_{xx}^*\right\|_2^2 \kappa(\Sigma_{xx}^*)}{\sigma_w^4}\right) \leq \lambda_{\min}(D).$$

3.6 eXploration termination

To end this chapter will will show that if we use data driven estimation errors which we obtain from Bayesian view on the problem, then the algorithm 1 terminates with high probability. The termination time depends solely on the system parameters as we will see in the following theorem.

Theorem 3.8 *Let (A_*, B_*) be regular stabilizable system and let us use the notation from above. Further assume that we play actions $u_i \sim \mathcal{N}(0, \sigma_u^2 I))$, use estimation region $\Theta = \{(A,B) | X^\top D_s X, X^\top = (A\ B) - (A_s\ B_s)\}$ obtained from the Bayesian setting and synthesize the controller via SDP* (3.27) *or* (3.31)*. Then Algorithm 1 will terminate with probability $1-\delta$ in time*

$$\mathcal{O}\left(\frac{(1 + \|K_*\|_2)^4 \, \|\Sigma_{xx}^*\|_2^4 \, \kappa(\Sigma_{xx})^2}{\sigma_w^8}\right).$$

Proof From the results in section 3.5 we have that a sufficient condition to successfully synthesize a controller via SDP (3.27) or SDP (3.31) is

$$\mathcal{O}\left(\frac{(1 + \|K_*\|_2)^2 \, \|\Sigma_{xx}^*\|_2^2 \, \kappa(\Sigma_{xx})}{\sigma_w^4}\right) \leq \lambda_{\min}(D_s).$$

At the same time using the results from section 3.1.1 we see that running the system for s steps we have $D_s \succeq \Omega(\frac{s}{\log s})I$. Hence we need to run the system for s steps so that we have:

$$\mathcal{O}\left(\frac{(1 + \|K_*\|_2)^2 \, \|\Sigma_{xx}^*\|_2^2 \, \kappa(\Sigma_{xx})}{\sigma_w^4}\right) \leq \frac{s}{\log s}$$

Since we have $\sqrt{s} \leq \frac{s}{\log s}$ we obtain that if we run the system for

$$\mathcal{O}\left(\frac{(1 + \|K_*\|_2)^4 \, \|\Sigma_{xx}^*\|_2^4 \, \kappa(\Sigma_{xx})^2}{\sigma_w^8}\right)$$

steps the the Algorithm 1 will terminate with probability at least $1-\delta$. □

As we have seen in theorem 3.8 algorithm 1 terminates in time which depends solely on system parameters, in particular it depends on the norm of optimal controller K_*, norm and conditional number of the steady state covariance matrix Σ_{xx}^* and on the noise scale σ_W. Using the obtained results we immediately see that the cost and with that also regret of the algorithm 1 is upper bounded by

$$\|A_*\|^{\mathcal{O}\left(\frac{(1+\|K_*\|_2)^4 \|\Sigma_{xx}^*\|_2^4 \kappa(\Sigma_{xx})^2}{\sigma_w^8}\right)}.$$

While this result tells us that if we use eXploration as an initialization for the algorithms which need a stabilizing controller as an input we additionally suffer only constant regret, this constant can be exponentially large. Also note that the obtained bound is not the tightest bound possible. With just a small modification where we would for any $\varepsilon > 0$ use the fact that $s^{1-\varepsilon} \leq \frac{s}{\log s}$ for s large enough, we could already obtain a bound with better parameters. The question if we can find a stabilizing controller in the single trajectory setting without suffering cost, which is exponential in system parameters is further addressed in chapters 5 and 7. Instead of playing zero mean Gaussian actions we propose to use controllers based on the observed data which heuristically reduce the norm of the state and consequently the suffered cost.

Chapter 4

eXploration as initialization

In chapter 3 we have shown how we can find a controller which with high probability stabilizes the system A_*, B_* in time which depends only on the system parameters. In this chapter we will show how we can initialize the existing algorithms, such as OSLO (Cohen et al., 2019) or CEC (Simchowitz and Foster, 2020), which require a stabilizing controller as an input, with eXploration. Both algorithms, OSLO and CEC, consist of two parts. In the first part, which we call *warm up phase*, they utilize the stabilizing controller to obtain tight estimates of system matrices A_*, B_*, which knowledge they utilize in the second part, where they choose actions optimistically (OSLO) or greedily (CEC). Together with eXploration as initialization we obtain two 3-phased algorithms which we call *X-OSLO* and *X-CEC*.

The second phase of X-OSLO and X-CEC is given in algorithm 2. Parameter σ^2_{init} is different for both algorithms, also the number of steps we run algorithm 2 differs between OSLO and CEC.

Algorithm 2 Utilize the stabilizing controller

```
Input: Controller $K$ with $\rho(A_* + B_*K) < 1$
for $i = 1, \dots$ do
    observe state $x_i$
    play $u_i \sim \mathcal{N}(Kx_i, \sigma^2_{init} I)$
end for
```

4.1 Initialization of OSLO

In the second phase of X-OSLO we set $\sigma^2_{init} = 2\sigma^2_w \kappa^2_0$, where κ_0 is the first of the so called *strongly stable* parameters of the controller K.

Definition 4.1 *A controller K is* (κ, γ)*-strongly stable for* $0 < \gamma \leq 1$ *if:*

1. $\|K\|_2 \le \kappa$
2. $A_* + B_*K = HLH^{-1}$, *with* $\|L\|_2 \le 1 - \gamma$ *and* $\|H\|_2 \left\|H^{-1}\right\|_2 \le \kappa$.

Here we call κ and γ the first and the second strongly stable parameter respectively.

In the rest of this section we will show we can obtain strongly stable parameter from the controller which we obtain from SDP eq. (3.27) or SDP (3.31) and how we can modify the analysis of Cohen et al. (2019) to utilize the knowledge of η with $\rho(A_* + B_*K) \le \eta < 1$, where K is the controller which we obtained either from SDP (3.11) or (3.27).

4.1.1 Strong stability parameters from robust SDP

In the following we denote $K = \Sigma_{ux}\Sigma_{xx}^{-1}$, where Σ is the optimal solution of SDP (3.31).

Lemma 4.2 *Let us be in the aforementioned setting and denote by $\nu = \mathrm{Tr}(\Sigma)$ and $\kappa^2 = \frac{\nu}{\sigma_w^2}$. Then controller K is $(\kappa, \frac{1}{2\kappa^2})$-strongly stable.*

Proof Since for every $(A, B) \in \Theta$ (also for (A_*, B_*)) holds $\Sigma_{xx} \succeq (A + BK)\Sigma_{xx}(A + BK)^\top + \sigma_w^2 I$ we have $\sigma_w^2 I \preceq \Sigma_{xx}$. Since we know Σ, we can compute its trace $\nu = \mathrm{Tr}(\Sigma_{xx}) + \mathrm{Tr}(\Sigma_{uu})$. With such a notation we have: $\sigma_w^2 I \preceq \Sigma_{xx} \preceq \nu I$. Denote by $L = \Sigma_{xx}^{-1/2}(A_* + B_*K)\Sigma_{xx}^{1/2}$. Multiplying equation

$$\Sigma_{xx} \succeq (A_* + B_*K)\Sigma_{xx}(A_* + B_*K)^\top + \sigma_w^2 I$$

from left and right with $\Sigma_{xx}^{-1/2}$ we obtain:

$$I \succeq LL^\top + \sigma_w^2\Sigma_{xx}^{-1} \succeq LL^\top + \frac{\sigma_w^2}{\nu} I.$$

From there it follows:

$$LL^\top \preceq \left(1 - \frac{\sigma_w^2}{\nu}\right) I,$$

which yields: $\|L\|_2 \le \sqrt{1 - 1/\kappa^2} \le 1 - \frac{1}{2\kappa^2}$. In the notation of Definition 4.1 we have $H = \Sigma_{xx}^{1/2}$. Since $\sigma_w^2 I \preceq \Sigma_{xx} \preceq \nu I$ we have: $\|\Sigma_{xx}^{1/2}\|_2 \|\Sigma_{xx}^{-1/2}\|_2 \le \sqrt{\nu}\frac{1}{\sigma_w} = \kappa$. To finish the proof observe:

$$\sigma_w^2 \|K\|_F^2 \le \mathrm{Tr}(K\Sigma_{xx}K^\top) = \mathrm{Tr}(\Sigma_{uu}) \le \nu,$$

from where we conclude: $\|K\|_2 \le \|K\|_F \le \kappa$. □

From the discussion in section 3.3 we see that we can obtain strong stability parameters also from the solution of SDP (3.27). If we define

$$\Sigma' = \frac{\sigma_w^2}{1-t} \begin{pmatrix} I \\ K \end{pmatrix} P \begin{pmatrix} I \\ K \end{pmatrix}^\top, \quad t' = \frac{t\sigma_w^2}{1-t}$$

then from the reformulation of SDP (3.27) given in section 3.3 follows that Σ', t' are feasible solution of SDP (3.31) and hence the following lemma holds:

Lemma 4.3 *Let P, K, t be the paramters of the optimal soluton of SDP* (3.27). *Then for $\kappa^2 = \frac{1}{1-t} \operatorname{Tr}(P(I + K^\top K))$ controller K is $(\kappa, \frac{1}{2\kappa^2})$ strongly stable.*

4.1.2 Adjusted Warm Up

As we have seen in the section 4.1.1 we can calculate the strong stability parameters of the synthesized controller K and hence start the OSLO algorithm. In the case when we synthesize the controller with SDPs (eq. (3.11), eq. (3.27)) which we derived using SLS framework we, instead of the strong stability parameters can calculate parameter η with $\rho(A_* + B_*K) \leq \eta < 1$. In this section we will first show how we can calculate parameter η and then modify the warm-up phase of X-OSLO algorithm to the extent that we start it with knowledge of η instead of the knowledge of strong stability parameter κ.

Compute η with $\rho(A_* + B_*K) \leq \eta < 1$ In this section we will show in depth how we can obtain parameter η for SDP (3.11) introduced by Dean et al. (2017). For the SDP given by eq. (3.27) the calculations are very similar and we show only the direction how to obtain η in this case.

We first show an upper bound on the norm of resolvent of perturbed matrix.

Lemma 4.4 *Let $D \in \mathbb{R}^{d\times d}$ with $\rho(D) < 1$. Then for $\varepsilon \leq \frac{1-\rho(D)}{2\rho(D)}$ we have:*

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

Proof By Corollary 2.28 we have:

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

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

In what comes next we will denote

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

Proposition 4.5 *Let s^* be the minimal value, $\widehat{A}, \widehat{B}$ the estimates, $\varepsilon_A, \varepsilon_B$ upper bounds and K synthesized from SDP* (3.11). *Denote $D = \widehat{A} + \widehat{B}K$. Then we have:*

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

where $\varepsilon = \min\left(\frac{1-\rho(D)}{2\rho(D)}, \frac{\sqrt{1+(1/s^*-1)\|D\|_2 f(D)}-1}{2\|D\|_2 f(D)}\right)$.

Proof Observe that the condition $\rho(A_* + B_*K) < \frac{1}{1+\varepsilon}$ is equivalent to $\rho(A'_* + B'_*K) < 1$, 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 eq. (2.14) we obtain that the sufficient condition for $\rho(A'_* + B'_*K) < 1$ is:

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

which is equivalent to:

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

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

$$\begin{aligned}\left\|C\,(zI-(1+\varepsilon)D)^{-1}\right\|_{\mathcal{H}_\infty} &= \left\|C\left((zI-D)^{-1} + \varepsilon\,(zI-D)^{-1}D\,(zI-(1+\varepsilon)D)^{-1}\right)\right\|_{\mathcal{H}_\infty}\\ &\le \left\|C\,(zI-D)^{-1}\right\|_{\mathcal{H}_\infty}\left(1+\varepsilon\,\|D\|_2\left\|(zI-(1+\varepsilon)D)^{-1}\right\|_{\mathcal{H}_\infty}\right),\end{aligned}$$

where we used the equality $(X+Y)^{-1} = X^{-1} + X^{-1}Y(X+Y)^{-1}$. Then by eq. (2.16) and lemma 4.4 follows:

$$\left\|C\,(zI-(1+\varepsilon)D)^{-1}\right\|_{\mathcal{H}_\infty} \le \sqrt{s_*}\,(1+\varepsilon\,\|D\|_2\,f(D)).$$

The right hand side is smaller than $1/(1+\varepsilon)$ by setting

$$\varepsilon = \min\left(\frac{1-\rho(D)}{2\rho(D)}, \frac{\sqrt{1+(1/s^*-1)\,\|D\|_2\,f(D)}-1}{2\,\|D\|_2\,f(D)}\right).$$

For such a choice of ε then follows:

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

Since matrix $\widehat{A}+\widehat{B}K$ and scalar s^* are known after we synthesize the controller, we can compute ε given in Proposition 4.5 and hence we found η, defined as $\eta = \frac{1}{1+\varepsilon}$, which we can compute, and for which we have $\rho(A_* + B_*K) \le \eta < 1$.

In the case of SDP (3.27) we can obtain that for optimal solution t_* we have:

$$\left\| X^\top \begin{pmatrix} I \\ K \end{pmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1} \right\|_{\mathcal{H}_\infty} < \sqrt{t_*}.$$

We than find η with property $\rho(A_* + B_*K) \le \eta < 1$ with a similar derivation as presented for the case of SDP (3.11).

Refined analysis of OSLO's warm-up phase To analyze the case when we would like to utilize knowledge of η with $\rho(A_* + B_*K) \le \eta < 1$ we further need to know how to bound the norm of power of closed loop matrix. The following lemma will come handy.

Lemma 4.6 (Matrix power norm bound) *Let $A \in \mathbb{R}^{d\times d}$ with $\rho(A) < 1$. Then we have:*

$$\left\|A^k\right\|_2 \le \left(\frac{1+\rho(A)}{2}\right)^{k+1} \frac{2}{1-\rho(A)} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}}$$

Proof Since $\rho(A) < 1$ the curve which parametrizes the circle $\partial B_2^d\left(\frac{1+\rho(A)}{2}\right)$ in the positive way contains in its interior all the eigenvalues of A. Hence we can use Theorem 2.26 and compute

$$\begin{aligned}
\left\|A^k\right\|_2 &\le \frac{1}{2\pi}\int_{\partial B_2^d\left(\frac{1+\rho(A)}{2}\right)} |z|^k \left\|\mathfrak{R}_A(z)\right\|_2 dz \\
&\le \frac{1}{\pi(1-\rho(A))}\left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}} \int_{\partial B_2^d\left(\frac{1+\rho(A)}{2}\right)} |z|^k dz \\
&= \left(\frac{1+\rho(A)}{2}\right)^{k+1} \frac{2}{1-\rho(A)} \left(1 + \frac{1}{d-1}\left(1 + \frac{4(\|A\|_F^2 - \left|\mathrm{Tr}(A^2)\right|)}{(1-\rho(A))^2}\right)\right)^{\frac{d-1}{2}},
\end{aligned}$$

where we used Theorem 2.27 in the second inequality. □

Now we present the refined analysis of Cohen et al. (2019) where we leverage the knowledge of η with $\rho(A_* + B_*K) \le \eta < 1$. In the rest of this section we denote by

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

Using data dependent upper bounds we can compute a high probability upper bound ϑ for $\|(A\ B)\|_2$. Further with the use of ϑ we can compute an upper bound for C_0.

Lemma 4.7 *Let $x_0, x_1, \ldots$ be a sequence of states starting from state x_0 and generated by dynamics* (1.1) *following a policy K, synthesized from SDP* (3.11)*. Then we have:*

$$\|x_i\|_2 \leq C_0 \left(\frac{1+\eta}{2}\right)^{i+1} \|x_0\|_2 + \frac{2C_0}{1-\eta} \max_{j=0}^{i-1} \|B\zeta_j + w_{j+1}\|_2$$

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

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

Using first triangle inequality and then Lemma 4.6 we obtain:

$$\begin{aligned}\|x_i\|_2 &\leq \left\|(A+BK)^i\right\|_2 \|x_0\|_2 + \sum_{j=0}^{i-1} \left\|(A+BK)^{i-j-1}\right\|_2 \|B\zeta_j + w_{j+1}\|_2 \\ &\leq C_0 \left(\frac{1+\eta}{2}\right)^{i+1} \|x_0\|_2 + C_0 \max_{j=0}^{i-1} \|B\zeta_j + w_{j+1}\|_2 \sum_{i=0}^{\infty} \left(\frac{1+\eta}{2}\right)^i \\ &= C_0 \left(\frac{1+\eta}{2}\right)^{i+1} \|x_0\|_2 + \frac{2C_0}{1-\eta} \max_{j=0}^{i-1} \|B\zeta_j + w_{j+1}\|_2 \qquad \square\end{aligned}$$

In the next lemma we will apply a corollary 2.15 of Hanson-Wright inequality and bound the maximal norm of the noise.

Lemma 4.8 *Let $\delta \in (0, \frac{1}{e})$. With probability at least $1-\delta$ for all $i = 1, \ldots T_0$ we have:*

$$\|x_i\|_2 \leq C_0 \left(\frac{1+\eta}{2}\right)^{i+1} \|x_0\|_2 + \frac{2\sqrt{5}C_0\sigma_w}{1-\eta} \sqrt{(d_x + 2d_u\kappa_0^2\vartheta^2)\log\frac{T_0}{\delta}}$$

Proof In order to use Lemma 4.7 we need to bound $\max_{j=0}^{T_0-1} \|B\zeta_j + w_{j+1}\|_2$. Since $B\zeta_j + w_{j+1} \sim \mathcal{N}(0, 2\sigma_w^2\kappa_0^2 BB^\top + \sigma_w^2 I)$ we can use Corollary 2.15. For every $0 \leq j \leq T_0 - 1$ we have w.p. at least $1 - \frac{\delta}{T_0}$:

$$\|B\zeta_j + w_{j+1}\|_2^2 \leq 5\sigma_w^2(d_x + 2\kappa_0^2 \|B\|_F^2)\log\frac{T_0}{\delta}.$$

Using union bound and Lemma 4.7 we obtain that we have w.p. at least $1-\delta$:

$$\|x_i\|_2 \leq C_0 \left(\frac{1+\eta}{2}\right)^{i+1} \|x_0\|_2 + \frac{2\sqrt{5}C_0\eta}{1-\eta} \sqrt{(d_x + 2d_u\kappa_0^2\vartheta^2)\log\frac{T_0}{\delta}},$$

which finishes the proof. □

The rest of the analysis which shows that running phase 2 for $\widetilde{\mathcal{O}}(\sqrt{T})$ rounds yields a controller with tight enough estimates to start phase 3 is very similar to the analysis presented in the proof of Theorem 20 in (Cohen et al., 2019) and hence we omit it here.

Optimal infinite horizon cost upper bound We can start with optimistic phase of OSLO when we have estimates $\widehat{A}, \widehat{B}$ with $\|(\widehat{A}\ \widehat{B}) - (A_*\ B_*)\|_F^2 \leq c\frac{\alpha_0^5\sigma_w^{10}}{\nu^5\vartheta\sqrt{T}}$. Here $\alpha_0 = \min(\lambda_{min}(Q), \lambda_{min}(R))$, c universal constant, σ_w, ϑ, T as defined above and ν an upper bound for the optimal expected infinite horizon cost J_*. We will now show how we can compute ν from the optimal solution of SDP with which we finish phase 1 of X-OSLO.

If we choose action $u_i = Kx_i$, where $K \in \mathbb{R}^{k\times d}$ is a 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)):

$$\begin{aligned}&\min_{X\succeq 0} \left\langle Q + K^\top RK, P\right\rangle,\\ &\quad\text{s.t. } P = (A_* + B_*K)P(A_* + B_*K)^\top + \sigma_w^2 I.\end{aligned} \tag{4.1}$$

Note that SDP (4.1) is just a non convex formulataion of SDP (2.20). We denote the expected infinite horizon cost for such a policy with $J(A_*, B_*, K)$. With the next lemma we reformulate lemma 3.5.

Lemma 4.9 *Let ν^* be minimal value of* (4.1) *and ν' minimal value of*

$$\begin{aligned}&\min_{P\succeq 0} \left\langle Q + K^\top RK, P\right\rangle,\\ &\quad s.t.\ P \succeq (A_* + B_*K)P(A_* + B_*K)^\top + \sigma_w^2 I.\end{aligned} \tag{4.2}$$

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

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

Lemma 4.10 *The minimal value of the optimization problem* (4.1) *is equal to the optimal value of:*

$$\begin{aligned}&\min_{P\succeq 0} \sigma_w^2 \left\langle Q + K^\top RK, P\right\rangle,\\ &\quad s.t.\ P = (A_* + B_*K)P(A_* + B_*K)^\top + I.\end{aligned} \tag{4.3}$$

Proof First notice that the optimal value of (4.1) is equal to:

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

Since $u_i = Kx_i$ we obtain:

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

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

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}[x_ix_i^\top] &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}\left[\sum_{j=1}^{i}(A_* + B_*K)^{i-j}w_jw_l^\top((A_* + B_*K)^\top)^{i-j}\right]\\ &= \sigma_w^2\lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}\left[\sum_{j=1}^{i}(A_* + B_*K)^{i-j}\frac{w_j}{\sigma_w}\frac{w_l}{\sigma_w}^\top((A_* + B_*K)^\top)^{i-j}\right].\end{aligned}$$

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

To finish we will first show a result presented by Dean et al. (2017) and then use it to to arrive at an upper bound for J_*.

Lemma 4.11 *Let $P, K = SP^{-1}, s$ be a feasible solution of SDP* (3.11)*. Then we have:*

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

Lemma 4.12 *Let s^*, P, K be parameters of the optimal solution of SDP* (3.11)*. Then we have:*

$$J^* \leq \frac{\sigma_w^2}{1-\sqrt{s^*}}\left\langle Q + K^\top RK, P\right\rangle$$

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^*, P, K are optimal solution to SDP (3.11) they are also feasible solution and hence we have:

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

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

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

Reordering the terms we obtain:

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

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

$$J(\widehat{A}, \widehat{B}, K) \leq \sigma_w^2 \left\langle Q + K^\top RK, P \right\rangle.$$

To finish the proof we use lemma 4.11:

$$J_* \leq J(A_*, B_*, K) \leq \frac{1}{1 - \sqrt{s^*}} J(\widehat{A}, \widehat{B}, K) \leq \frac{\sigma_w^2}{1 - \sqrt{s^*}} \left\langle Q + K^\top RK, P \right\rangle. \quad \square$$

In the latter derivation we show how we obtain the bound $J_* \leq \nu$ from SDP (3.11) or SDP (3.27). For SDP (3.31) we have that the optimal solution is already the upper bound for J_*. Hence we showed that by first identifying the stable controller with any SDP among (3.11), (3.27), (3.31) we can then start OSLO algorithm. Using the data dependent upper bounds which we obtained from the Bayesian setting the phase of identifying the stable controller finishes in constant time and hence adds constant albeit exponentially large cost in system parameters to the total regret. Using Corollary 5 of Cohen et al. (2019) we obtain the following theorem.

Theorem 4.13 *Suppose the system matrices A_*, B_* are stabilizable and regular, cost matrices Q, R are positive definite and time horizon is T. Then by first running eXploration, where we synthesize the controller with any SDP* (3.11), (3.27), (3.31), *using data dependent upper bounds from the Bayesian setting, and then OSLO algorithm, the total regret we suffer is upper bounded with probability at least $1 - \delta$ as:*

$$R(T, X\text{-}OSLO) = \mathcal{O}\left(\sqrt{T}\log^2 T\right).$$

4.2 Initialization of CEC

Initialization of CEC requires only the stabilizing controller K. Hence we can directly state the theorem.

Theorem 4.14 *Suppose the system matrices A_*, B_* are stabilizable and regular, cost matrices Q, R are positive definite, time horizon is T and probability of failure is $\delta \in (0, \frac{1}{T})$. Then by first running eXploration, where we synthesize the controller with any SDP* (3.11), (3.27), (3.31), *using data dependent upper bounds* *from the Bayesian setting, and then CEC algorithm, the total regret we suffer is upper bounded with probability at least* $1-\delta$ *as:*

$$R(T, X\text{-}CEC) = \mathcal{O}\left(\sqrt{T \log T}\right).$$

The proof follows directly from the Theorem 2 of Simchowitz and Foster (2020) and Theorem 3.8.

Chapter 5

Improved eXploration strategies

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., fig. 6.4), 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. Applying such a controller, we generally lose the theoretical guarantee that the Phase I will end. However, the upper bounds on estimation error from the Bayesian setting (and thus the validity of the stopping condition) still hold and we can run Algorithm 1. Here, we discuss four 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 $P = Q + A_i^\top P A_i - A_i^\top P B_i (R + B_i^\top P B_i)^{-1} B_i^\top P A_i$, i.e., we act using *certainty equivalent control*.

For the second K_i we consider SDP (3.45). At every time step we synthesize the controller for which we have that

$$\max_{(A,B) \in \Theta} \|A + BK\|_2$$

is minimal among all the controllers. We call this controller *MinMax* controller.

As third alternative we use relaxed version of SDP (3.27). We relax the constraint $t \in (0,1)$ to $t \geq 0$. With such a relaxed constraint the obtained SDP is always feasible. There are two possible interpretations for this relaxations. The first is that by allowing values $t \geq 1$ we try to stabilize smaller ellipsoidal region. Instead of trying to stabilize ellipsoidal region associated with positive definite matrix D, we are trying to stabilize ellipsoidal region associated with positive definite matrix tD. Another way of interpreting this result is that instead we are trying to find the smallest t such that we have a guarantee that with synthesized controller the maximal eigenvalue of the closed loop system associated with any system A, B in ellipsoidal region is at most $\sqrt{t}$. We call this control *RelaxedSDP* control.

Lastly, the fourth controller we consider is the so called *DeadBeat* controller. With the RLS estimates A_i, B_i we synthesize the controller K_i as:

$$K_i = \operatorname*{argmin}_K \|(A_i + B_i K)x_i\|_2 .$$

To compute K_i we use semi-definite program:

$$\min_{K,t} t \tag{5.1}$$

$$\text{s.t.} \begin{pmatrix} t & x_i^\top (A_i + B_i K)^\top \\ (A_i + B_i K)x_i & tI \end{pmatrix} \succeq 0. \tag{5.2}$$

The controller which minimizes SDP (5.1) is K_i.

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) we have for ordinary least squares (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_x with $d_x \geq 2$. We show now that when $d_x \geq 2$, 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 $d_x \geq 2$ this does not hold, and the estimation procedure might not be consistent anymore.

5.1 Inconsistency of OLS in case $d_x > 1$

The construction will be based on the inconsistency of OLS estimator. Nielsen (2008) and Phillips and Magdalinos (2013) show 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 actions 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:

$$(A_s^o - A_*)^\top = \left(\sum_{i=0}^{s-1} x_i x_i^\top\right)^{-1} \sum_{i=0}^{s-1} x_i w_{i+1}^\top \tag{5.3}$$

Since $2I_d$ is irregular matrix, the right hand side of the eq. (5.3) 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.

5.2 Convergence in the constrained case

As we have seen in section 5.1 OLS estimator does not converge for all measurable actions u_i. However, we can still prove convergence under some additional assumptions, as stated in the next theorem. From the conducted experiments we observed that usually we find a controller which stabilizes the underlying system before we have a guarantee for that. Theorem 5.1 shows that in such setting, even if the controller varies the estimation will be consistent.

Theorem 5.1 *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 we have: $\|x_i\|_2 \leq M_1$ and $\|A_* + K_i\|_2 \leq M_2$ for some constants M_1, M_2, which do not depend on time. Then at every time s we have for the RLS estimator A_s of the matrix A_* with probability at least $1 - \delta$:*

$$\|A_s - A_*\|_2 \leq \frac{\mathcal{O}(1)(d_x \log s + \log \frac{1}{\delta})}{\sqrt{s}}$$

The strategy to prove theorem 5.1 will be the following. We first use the explicit RLS formula and bound:

$$\|A_s - A_*\|_2 \leq \left\|(V_s + \lambda I)^{-\frac{1}{2}}\right\|_2 \left\|(V_s + \lambda I)^{-\frac{1}{2}} S_s\right\|_2 + \lambda \|A_*\|_2 \left\|(V_s + \lambda I)^{-1}\right\|_2,$$

and then show that $\left\|(V_s+\lambda I)^{-\frac{1}{2}}\right\|_2 = \mathcal{O}(1/\sqrt{s})$ and $\left\|(V_s+\lambda I)^{-\frac{1}{2}}S_s\right\|_2 = \mathcal{O}(1)\left(d_x \log s + \log \frac{1}{\delta}\right)\Big)$.

The toughest part is to show that $\left\|(V_s+\lambda I)^{-\frac{1}{2}}\right\|_2 = \mathcal{O}(1/\sqrt{s})$, which is equivalent to show that $V_s \succeq \Omega(s)I$. Let us begin with a simple lemma which was proven by Sarkar and Rakhlin (2018).

Lemma 5.2 *Let $P, Q \in \mathbb{R}^{d\times d}$ such that $P \succ 0$. Assume $\|Q\|_{P^{-1}} \le \gamma$. Then for every vector v for which it holds $v^\top P v = \alpha$ we have: $\left\|v^\top Q\right\|_2 \le \sqrt{\alpha}\gamma$*

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 5.3 *Let $y_i = (A_* + K_i)x_i$. Then for every $s \ge 1$ we have:*

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

Proof By inserting $V_s = \sum_{i=0}^{s-1} x_i x_i^\top$ and using 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 the middle term, normalized with the regularized first term of the V_s decomposition.

Lemma 5.4 *Let us be in setting of this section. Then we have w.p. at least $1-\delta$:*

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

Proof Denote by $\mathcal{F}_s = \sigma\left((w_i)_{i\le s}\right)$ and $\mathcal{F} = (\mathcal{F}_s)_{s\ge 0}$. With this notation $(y_i)_{i\ge 0}$ is stochastic process in $\mathbb{R}^{d_x}$ adopted to filtration $\mathcal{F}$. Further denote by $V = I$. Now we apply Lemma 3.2 with $\varepsilon = \frac{1}{2}$ and obtain:

$$\forall s \geq 1 : \quad \left\| \sum_{i=0}^{s-2} y_i w_{i+1}^\top \right\|^2_{\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)^{-1}} \leq 8\sigma_w^2 \log \left(\frac{\det\left(\sum_{i=0}^{s-2} y_i y_i^\top + I\right)}{\det(I)} \frac{5^{d_x}}{\delta} \right). \tag{5.4}$$

Since

$$\sum_{i=0}^{s-2} y_i y_i^\top \preceq \sum_{i=0}^{s-2} \|y_i\|_2^2 I \preceq \sum_{i=0}^{s-2} \|A_* + K_i\|_2^2 \|x_i\|_2^2 I \preceq M_2^2 M_1^2 (s-1) I,$$

we have

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

Therefore the upper bound from eq. (5.4) is upper bounded by

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

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 5.5 *For any $u \in S^{d_x-1}$ we have w.p at least $1-\delta$ for every $s \geq 1$:*

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

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 5.4 we have:

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

Denote by $u^\top (P+I)u = \alpha^2$. Then we have by Lemma 5.2:

$$\left\| u^\top Q \right\|_2 \leq \alpha \sqrt{8\sigma_w^2 d_x \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d_x}} \right)}$$

Hence:

$$\begin{aligned} u^\top P u + 2u^\top Q u &= u^\top (P+I)u + 2u^\top Q u - 1 \\ &\geq \alpha^2 - 2 \left\| u^\top Q \right\|_2 \|u\|_2 - 1 \\ &= \alpha^2 - 2 \left\| u^\top Q \right\|_2 - 1 \\ &\geq \alpha^2 - 2\alpha \sqrt{8\sigma_w^2 d_x \left(\log s + \log \frac{5M_1^2 M_2^2}{\delta^{1/d_x}} \right)} - 1 \end{aligned}$$

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

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

Plugging this expression for α we arrive at:

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

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 5.7. We will then join this result with the result from Lemma 5.5 to obtain $V_s \succeq \Omega(s)I$ in Proposition 5.8.

Theorem 5.6 (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 $1 - e^{-\frac{t^2}{2}}$ we have:*

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

Corollary 5.7 *Let $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$. Then for every $s \geq 1$ we have w.p. at least $1-\delta$:*

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

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

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

From Theorem 5.6 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 5.8 *Let us be in the setting of this section. Then we have for all $s \geq 1$ w.p. at least $1 - \delta$:*

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

Proof By Lemma 5.3 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 $u \in S^{d_x - 1}$ 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 5.5, Part 1 is lower bounded by $-8\sigma_w^2 d_x \left(\log s + \log \frac{5 M_1^2 M_2^2 2^{1/d_x}}{\delta^{1/d_x}} \right) - 1$ w.p. at least $1 - \frac{\delta}{2}$. By Corollary 5.7, Part 2 term is lower bounded w.p. at least $1 - \frac{\delta}{2}$ by $\sigma_w^2 \left(\sqrt{s-1} - \sqrt{d_x} - \sqrt{2 \log \frac{2}{\delta}} \right)^2$. Using union bound we obtain that w.p. at least $1 - \delta$ we have:

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

Since $u \in S^{d_x - 1}$ was arbitrary we have w.p. at least $1 - \delta$:

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

which concludes the proof. □

Since by Proposition 5.8 we have $\sigma_{d_x}(V_s) \geq \mathcal{O}(s)$ we also have: $\sigma_{d_x}(V_s + \lambda I) \geq \sigma_{d_x}(V_s) \geq \mathcal{O}(s)$. Now the proof of Theorem 5.1 easily follows. By application of Lemma 3.2 with $\varepsilon = \frac{1}{2}$ we further obtain that we have w.p. at least $1 - \delta$:

$$\begin{aligned}
\left\| (V_s + \lambda I)^{-\frac{1}{2}} S_s \right\|_2^2 &\leq 8\sigma_w^2 \log \left(\frac{\det\left(\sum_{i=0}^{s-1} x_i x_i^\top + \lambda I\right)}{\det(\lambda I)} \frac{5^{d_x}}{\delta} \right) \\
&\leq 8\sigma_w^2 \log \left(\frac{((s-1)M_1 + \lambda)^{d_x}}{\lambda^{d_x}} \frac{5^{d_x}}{\delta} \right) \\
&= 8\sigma_w^2 d_x \left(\log \frac{(s-1)M_1 + \lambda}{\lambda} + \log \frac{5}{\delta^{1/d_x}} \right)
\end{aligned}$$

Using union bound we have w.p. at least $1 - 2\delta$:

$$\begin{aligned}
\|A_s - A_*\|_2 \leq &\frac{8d_x \left(\log \frac{(s-1)M_1+\lambda}{\lambda} + \log \frac{5}{\delta^{1/d_x}} \right)}{\sqrt{\left(\left(\sqrt{s-1} - \sqrt{d_x} - \sqrt{2\log\frac{2}{\delta}} \right)^2 - 8d_x \left(\log s + \log \frac{5M_1^2 M_2^2 2^{1/d_x}}{\delta^{1/d_x}} \right) - 1 \right)}} \\
&+ \frac{\lambda \|A_*\|_2}{\sigma_w^2 \left(\left(\sqrt{s-1} - \sqrt{d_x} - \sqrt{2\log\frac{2}{\delta}} \right)^2 - 8d_x \left(\log s + \log \frac{5M_1^2 M_2^2 2^{1/d_x}}{\delta^{1/d_x}} \right) - 1 \right)}
\end{aligned}$$

Hence we have established that $\|A_s - A_*\|_2 \leq \frac{\mathcal{O}(1)(d \log s + \log \frac{1}{\delta})}{\sqrt{s}}$. The same analysis as in the proof of Theorem 3.8 then shows that in this setting eXploration finishes in constant (in T) time.

Chapter 6

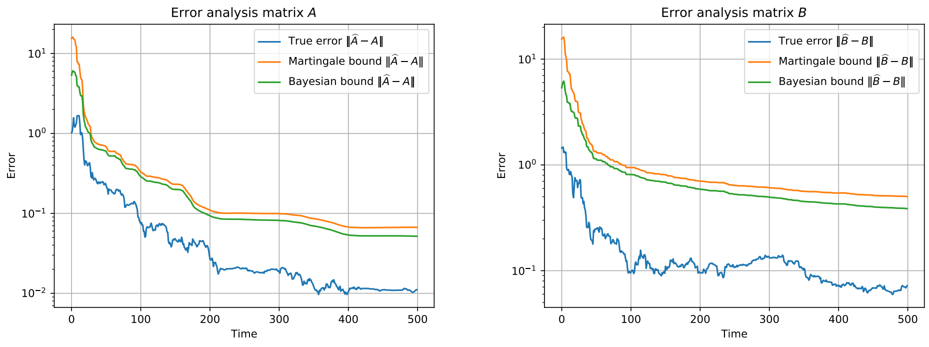
Numerical experiments

6.1 Comparison of error bounds

In this section we will compare the data dependent estimation error bounds $\varepsilon_A, \varepsilon_B$ obtained in section 3.1.1 and section 3.1.2. We will compare the errors on the system introduced by Dean et al. (2017):

$$A_* = \begin{pmatrix} 1.01 & 0.01 & 0.00 \\ 0.01 & 1.01 & 0.01 \\ 0.00 & 0.01 & 1.01 \end{pmatrix}, \quad B_* = I_3, \quad Q = R = I_3 \tag{6.1}$$

we further set $(w_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ and choose actions $(u_i)_{i\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$. The probability of failure is $\delta = 0.1$, regularizing parameter is set to $\lambda = 1$. We set the initial upper bound for the system norm to $\vartheta = 10$.



(a) Error analysis for matrix A_*

(b) Error analysis for matrix B_*

Figure 6.1: Data dependent bounds obtained from Bayesian setting are slightly tighter than the one obtained from the theory of Self Normalizing Processes.

As we have seen in fig. 6.1, data dependent upper bounds obtained from Bayesian setting perform better. This trend is observed on all the conducted experiments. Therefore in the upcoming experiments we will only focus on confidence regions obtained from the Bayesian setting. Since we obtained errors $\varepsilon_A, \varepsilon_B$ in the Bayesian setting by finding the smallest "box" which contains ellipsoid region, using $\varepsilon_A, \varepsilon_B$ we lose a lot of structure in the estimation. On fig. 6.2 we see an example of how much structure we can lose. Since we are trying to find a controller which stabilizes everything inside given confidence region we conclude that we will always find faster a controller which stabilizes everything inside given ellipsoid than if we would like to find a controller which stabilizes everything inside the associated "box" confidence region.

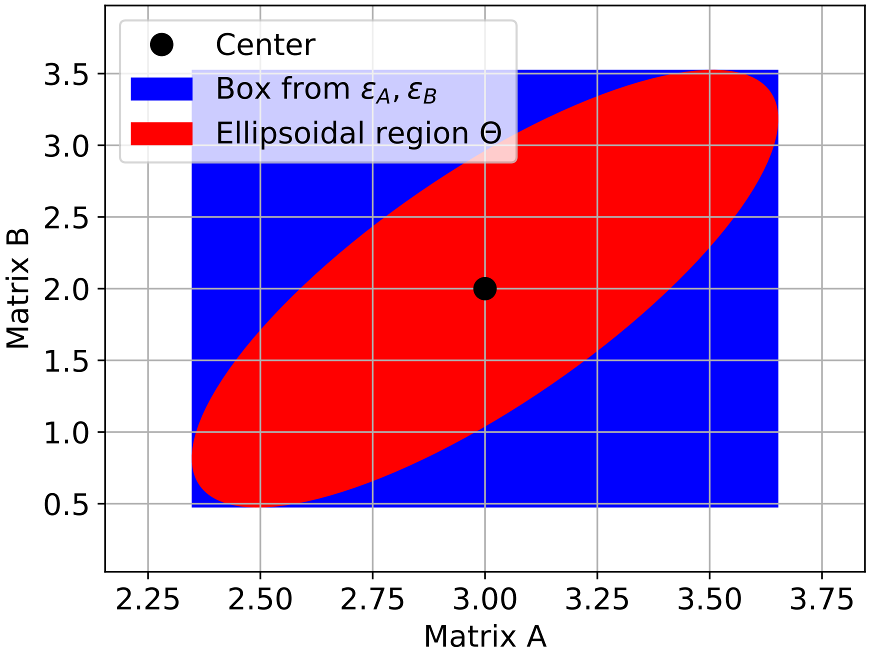
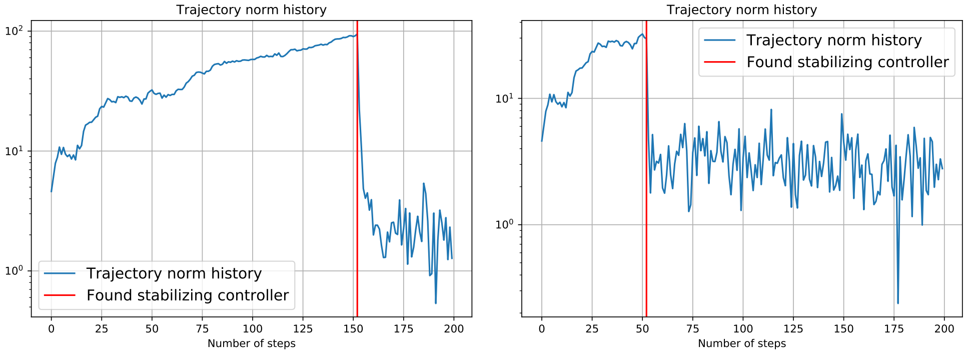


Figure 6.2: By using box confidence region we lose a lot of structure compared to ellipsoid confidence region.

On fig. 6.3 we present how much faster we find a stabilizing controller if we use ellipsoid confidence region compared to the box confidence region. We run the experiment on system given by eq. (6.1). To synthesize the controller with ellipsoid confidence region we used SDP given by eq. (3.27) and to synthesize the controller with box confidence region we used SDP given by eq. (3.11). Both SDPs were obtained on the same way – both have an equivalent formulation in SLS language, the only difference is which region we try to stabilize, hence the comparison make sense. We see that by using ellipsoid region we find a stabilizing controller around 3 times faster in this case.



(a) Number of steps we need to find a stabilizing controller using box confidence interval

(b) Number of steps we need to find a stabilizing controller using ellipsoid confidence interval

Figure 6.3: Using ellipsoid confidence region around the estimates the algorithm finds stabilizing controller a couple of times faster than by using box confidence regions.

We further run the experiment on the same system but with different random seeds. We tried 100 times, among them in 93 cases the stabilizing controller was found before 500 steps for both settings – if we use ellipsoid or box confidence region – and for them the statistics of number of steps which we need to take to find a stabilizing controller is presented in table 6.1.

	avg(#steps)	std(#steps)
Box region Θ	191.42	28.31
Ellipsoid region Θ	87.40	20.99

Table 6.1: On average with ellipsoid confidence region we find stabilizing controller 2 times faster for the system given by eq. (6.1).

6.2 Control before stabilization

As we have seen on fig. 6.3 if we choose actions $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2 I)$ the norm of the state can grow exponentially if $\rho(A_*) > 1$. This can be especially problematic if matrix A_* has an eigenvalue which is significantly larger than

1. To illustrate the blowup consider the system:

$$A_* = \begin{pmatrix} 1.5 & 1.0 & 0.4 & 2.3 \\ 0.0 & 1.3 & 1.3 & 1.1 \\ 0.0 & 0.0 & 1.0 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.8 \end{pmatrix}, \quad B_* = \begin{pmatrix} 0.6 & 0.7 & 0.3 \\ 0.8 & 1.1 & 1.1 \\ 1.2 & 0.2 & 2.3 \\ 2.1 & 0.4 & 0.4 \end{pmatrix}, \quad Q = I_4, R = I_3, \tag{6.2}$$

where noise follows $(w_i)_{i \geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$. We will consider the case when we try to synthesize a stabilizing controller with probability of failure $\delta = 0.1$ and regularizing parameter $\lambda = 1$. We first choose actions $(u_i)_{i \geq 0} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_3)$. As we can see on fig. 6.4 since we choose zero mean Gaussian actions the norm of the state grows exponentially until we find a stabilizing controller in step 31.

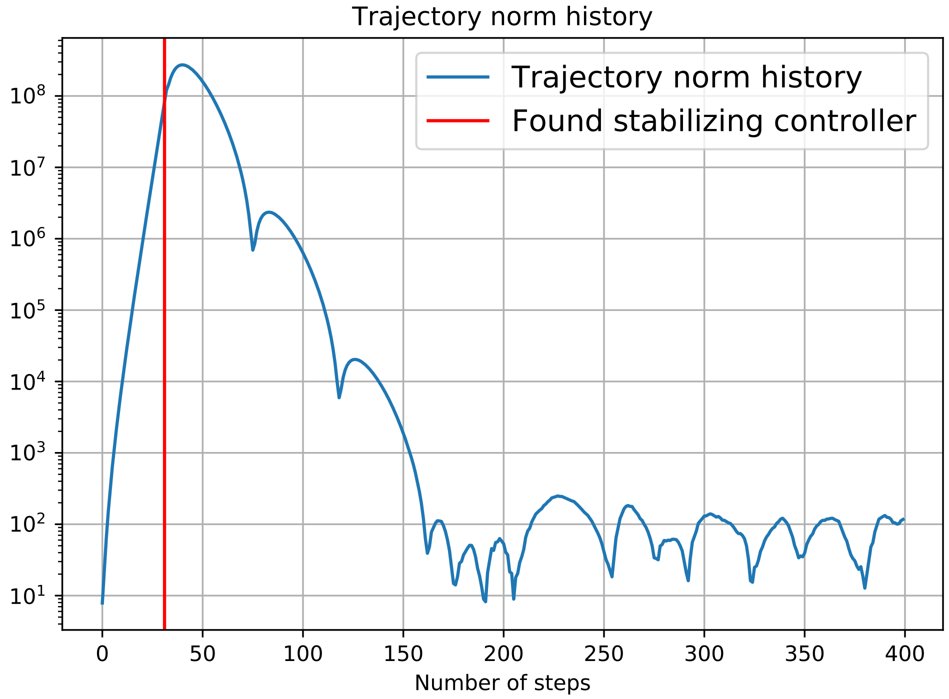
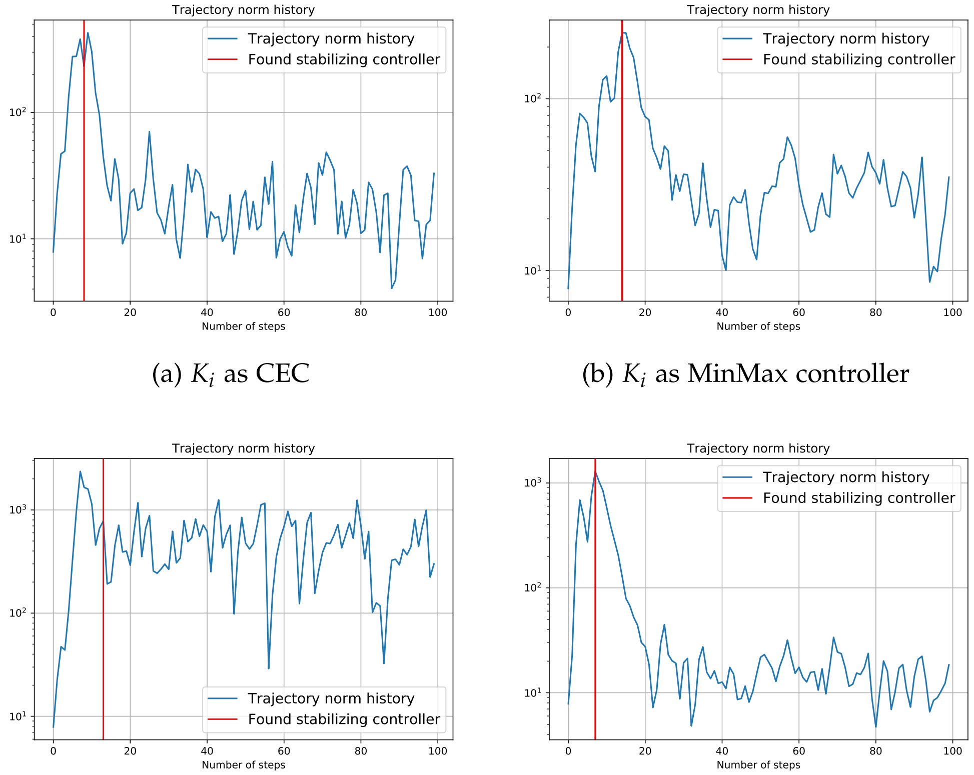


Figure 6.4: By choosing actions taken from zero mean Gaussian the state norm grow exponentially until we find a stabilizing controller.

Since for some systems large states could be prohibitive, we can instead of choosing zero mean Gaussian actions choose actions $u_i \sim \mathcal{N}(K_i x_i, \sigma_u^2 I)$ for different controllers K_i.



(a) K_i as CEC

(b) K_i as MinMax controller

(c) K_i as RelaxedSDP controller

(d) K_i as DeadBeat controller

Figure 6.5: Using different controllers before we find a stabilizing controller reduces the initial blowup. Since we do not choose zero mean actions we do not have a theoretical guarantee that eXploration will eventually terminate, however experiments show that using controller before stabilization does not harm the time it take to finish eXploration phase.

	Steps	Max state norm	Total cost
CEC	8.4 ± 1.9	$3.9 \cdot 10^2 \pm 4.4 \cdot 10^2$	$6.5 \cdot 10^5 \pm 1.6 \cdot 10^6$
MinMax	16 ± 6.9	$1.4 \cdot 10^4 \pm 4.2 \cdot 10^4$	$9.2 \cdot 10^8 \pm 3.8 \cdot 10^9$
RelaxedSdp	16 ± 14	$6 \cdot 10^2 \pm 9.3 \cdot 10^2$	$1.5 \cdot 10^6 \pm 3.1 \cdot 10^6$
DeadBeat	7.3 ± 0.66	$2.8 \cdot 10^4 \pm 9.7 \cdot 10^4$	$6.4 \cdot 10^{10} \pm 5.2 \cdot 10^{11}$

Table 6.2: Performance of different controllers if we apply them on system eq. (6.2) before we find a stabilizing controller.

Tables 6.2 and 6.3 show statistics of performances of different controllers which we apply before guaranteed stabilization. We can see that the number of steps it takes to find a stabilizing controller for system (6.2) is smaller compared to the system (6.1). The intuition behind this lies in the observation of Sarkar and Rakhlin (2018) who showed that the identification of explosive modes of the system happens exponentially fast where the base of the exponent is the mode of explosiveness. Since explosive modes 1.5 and 1.3 of system (6.2) are significantly larger than the explosive modes of system (6.1), which are approximately 1.01 and 1.02, we expect faster identification of unstable modes of system (6.2).

	Steps	Max state norm	Total cost
CEC	42 ± 14	14 ± 52	$1.5 \cdot 10^4 \pm 1.4 \cdot 10^5$
MinMax	40 ± 15	$58 \pm 1.3 \cdot 10^2$	$6.3 \cdot 10^4 \pm 2.2 \cdot 10^5$
RelaxedSdp	41 ± 14	40 ± 85	$5 \cdot 10^4 \pm 1.9 \cdot 10^5$
DeadBeat	14 ± 10	$1.7 \cdot 10^3 \pm 1.1 \cdot 10^4$	$3.7 \cdot 10^8 \pm 3.1 \cdot 10^9$

Table 6.3: Performance of different controllers if we apply them on system eq. (6.1) before we find a stabilizing controller.

6.3 Stabilizing region

Using the Bayesian initial belief we obtain that the true system matrices A_*, B_* lie in the set $\Theta = \{(A, B) | X^\top D_s X \preceq I, X^\top = (A\ B) - (A_s\ B_s)\}$. The goal is then to find a controller K with property $\rho(A + BK) < 1$ for every $(A, B) \in \Theta$. This problem is in general non convex and we do not know how to solve it efficiently. However, as we have seen in section 3.2, we can formulate a convex SDP (c.f. eq. (3.31)) with a guarantee that the associated controller will stabilize the underlying system. In this section we will try to illustrate for how large ellipsoid region Θ the SDP (3.31) can synthesize a controller which stabilizes every system in Θ.

For that we will move to the case when system matrices are one dimensional and leverage the fact that in one dimension the spectral radius is equal to the spectral norm. For every system $(\widehat{A}, \widehat{B}) \in [-3, 3] \times [-3, 3]$ we will search for the smallest matrix $D = \frac{1}{r^2} I$ for which SDP (3.31) can synthesize the controller. In other words we search for the largest radius r for which SDP (3.31) is feasible and returns a controller $K_g(\widehat{A}, \widehat{B})$ with a guarantee to stabilize every system (A, B) with $(A - \widehat{A})^2 + (B - \widehat{B})^2 \leq r^2$. On fig. 6.6 we plot for every system $\widehat{A}, \widehat{B}$ the largest r with such property. Note that r is trivially upper bounded by 1. To see that without loss of generality assume that for controller K, that stabilizes the largest ball around $(\widehat{A}\ \widehat{B})$, we have $0 \leq \widehat{A} + \widehat{B}K < 1$. If $r \geq 1$ than K stabilizes also the system $(A, B) = (\widehat{A} + 1, \widehat{B})$. But $1 \leq A + BK$.

Controller synthesized from SDP (3.31) has a guarantee to stabilize every

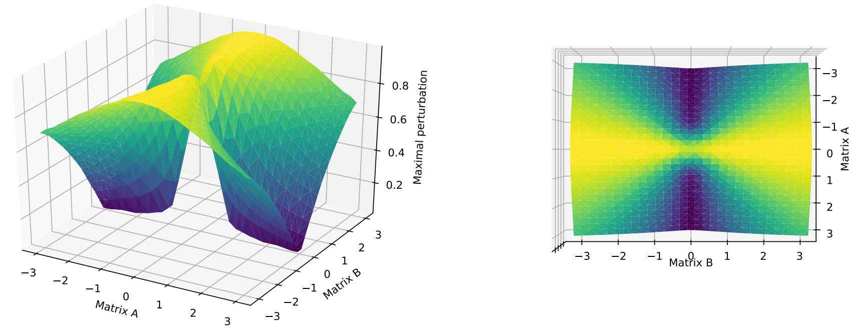


Figure 6.6: Maximal ball radius around estimates for which we have a guarantee to stabilize.

system in ellipsoid region associated with matrix D. However it can happen that in fact it stabilizes larger region. When the system matrices are one dimensional we can use SDP (3.46) to check what is actually the largest ball around the estimates which controller $K_g(\widehat{A}, \widehat{B})$ stabilizes. As we see on the fig. 6.7 the ball for which SDP (3.31) provides a guarantee that $K_g(\widehat{A}, \widehat{B})$ will stabilize it is in fact the largest ball that controller $K_g(\widehat{A}, \widehat{B})$ stabilizes. We can further use SDP (3.45) to search for the largest ball around estimates which are stabilized by any controller. Again we see that when systems are one dimensional the SDP (3.31) synthesize the controller which stabilizes the largest possible region. Since we computed the difference only to the accuracy 10^{-2} the plot obtained in fig. 6.7 is not identically zero.

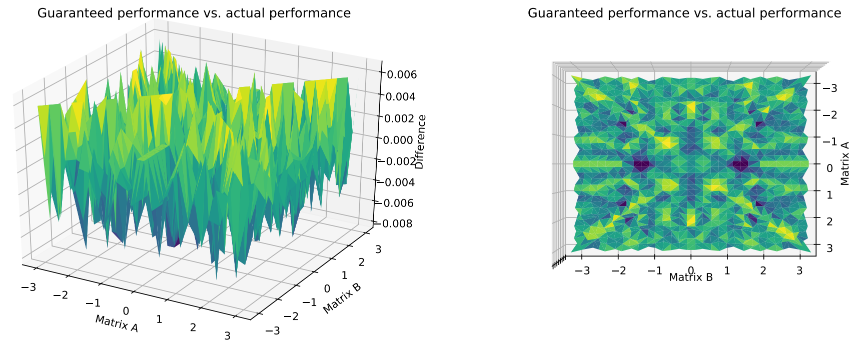


Figure 6.7: The area for which we have a guarantee that synthesized controller stabilizes is the same as the maximal area which the synthesized controller stabilizes.

We obtained that controller $K_g(\widehat{A}, \widehat{B})$ is the same as the controller obtained from SDP (3.45). Note that the fact, that we can efficiently compute the controller which stabilizes the ball with the largest radius around given system, crucially depends on the fact that system matrix A_* is one dimensional and its spectral radius is equal to its spectral norm. In higher dimension we can not compute such controller efficiently and there also exist controllers which stabilize larger ellipsoid than the one synthesized from SDP (3.31).

6.4 Comparison of CE and robust controller

As was argued in Simchowitz and Foster (2020) we also have a guarantee that CEC stabilizes some region around the estimates. Simchowitz and Foster (2020) use this fact to show that applying CEC the regret scales as $\mathcal{O}(\sqrt{d_u^2 d_x T})$. Since we only have upper bound but not lower bound on how large the confidence region should be for robust (CE) controller to stabilize the whole confidence region, we will try to illustrate the difference in the size of the largest ball which is stabilized either by robust or CE controller. First we will compare the maximal balls around the estimates they stabilize in one dimension. To compute CEC we also need to set matrices Q and R. We will compare robust controller with CEC, where we take $Q = 1, R = 1$, $Q = 1000, R = 1$ and $Q = 1, R = 1000$ respectively.

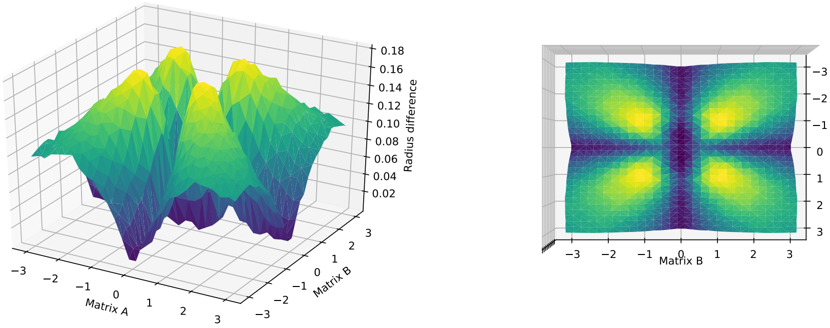


Figure 6.8: To compute CEC we used $Q = 1, R = 1$. Robust controller stabilizes larger ball around the estimates, however on some systems CEC stabilizes almost the same area as the robust controller.

First note that the absolute size of matrices Q and R does not matter in the computation of CE controller. Only relative size of matrices Q and R changes the CEC. As we can see on fig. 6.9 if we increase the relative size of state cost matrix Q compared to R CEC is starting to act as robust controller on most of the systems, however if we further increase the size of Q the difference between robust and CE controller is approximately the same as in the case when $Q = 1000$ and $R = 1$. On the other hand by increasing the

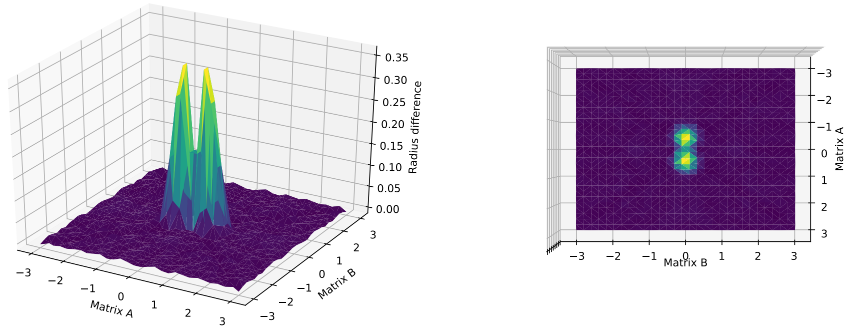


Figure 6.9: To compute CEC we used $Q = 1000, R = 1$. CEC stabilizes for many systems almost the same ball as robust controller.

relative size of the action matrix R, the CEC stabilizes significantly smaller region compared to the robust controller.

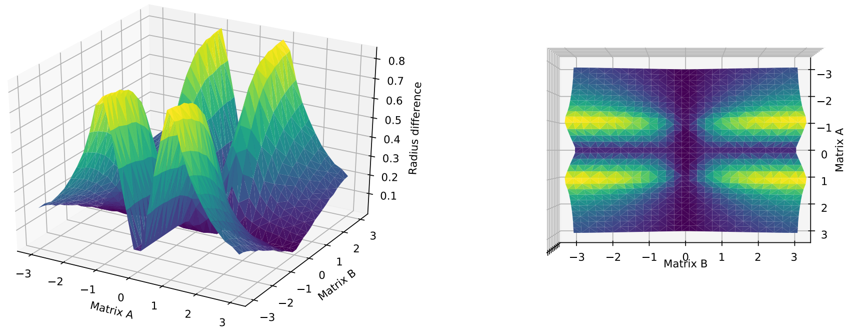


Figure 6.10: To compute CEC we used $Q = 1, R = 1000$. As we increase the cost for actions CEC controller for some estimates stabilizes significantly smaller region than robust controller.

In the following we will focus on the case $Q = 1, R = 1$ and will compare the cost of robust and CE controller for diferent systems $\widehat{A}, \widehat{B}$. For radius r for which CE controller stabilizes everything inside $\Theta_r = \{(A, B)|(A - \widehat{A})^2 + (B - \widehat{B})^2) \leq r^2\}$ we will find $A_w(r), B_w(r)$ on which CEC achieves the largest infinite horizon cost $J_{CE}(r, \widehat{A}, \widehat{B})$ among all systems in Θ_r and compute $J_{CE}(r, \widehat{A}, \widehat{B})$ using $A_w(r), B_w(r)$. At the same time for every radius r, for which SDP (3.31) stabilize every system inside Θ_r, denote $J_R(r, \widehat{A}, \widehat{B})$, which represent the largest cost robust controller achieves on systems in Θ_r. We selected 5 different systems uniformly at random $(\widehat{A}, \widehat{B}) \in [-3, 3] \times [-3, 3]$ and plotted their respective costs in fig. 6.11. We obtained that as radius goes to zero, performance of robust and CE controller on the worst system in r-ball neighborhood converges to the same - optimal cost. At the same time we also see that robust controller has smaller infinite horizon cost on the worst system compared to CE controller, which is not surprising since SDP (3.31) optimizes this objective.

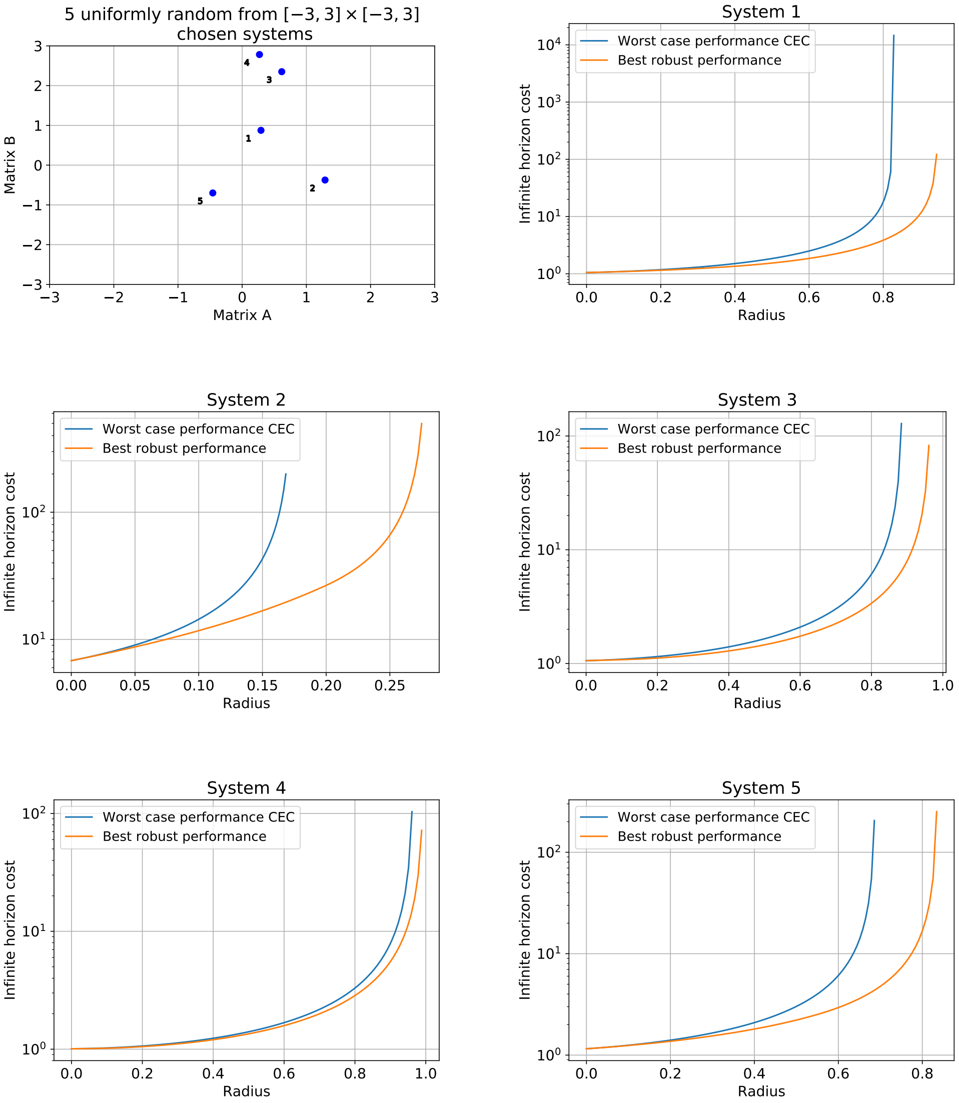


Figure 6.11: Comparison of robust and CE controller infinite horizon cost performance on the worst system in r-ball neighborhood. We selected uniformly at random 5 systems and tested performance on them. Robust controller stabilizes larger r-ball neighborhoods and achieves lower worst infinite horizon cost.

While fig. 6.11 shows the cost suffered by robust and CE controller on the worst case system inside confidence ball, fig. 6.12 reveals their behavior on all the systems inside the confidence ball. For the ball radius we took half of the maximal radius of the ball for which CEC stabilizes every system inside it. We then plot for every system inside this ball the difference between the cost suffered by CEC and the best robust controller for this ball.

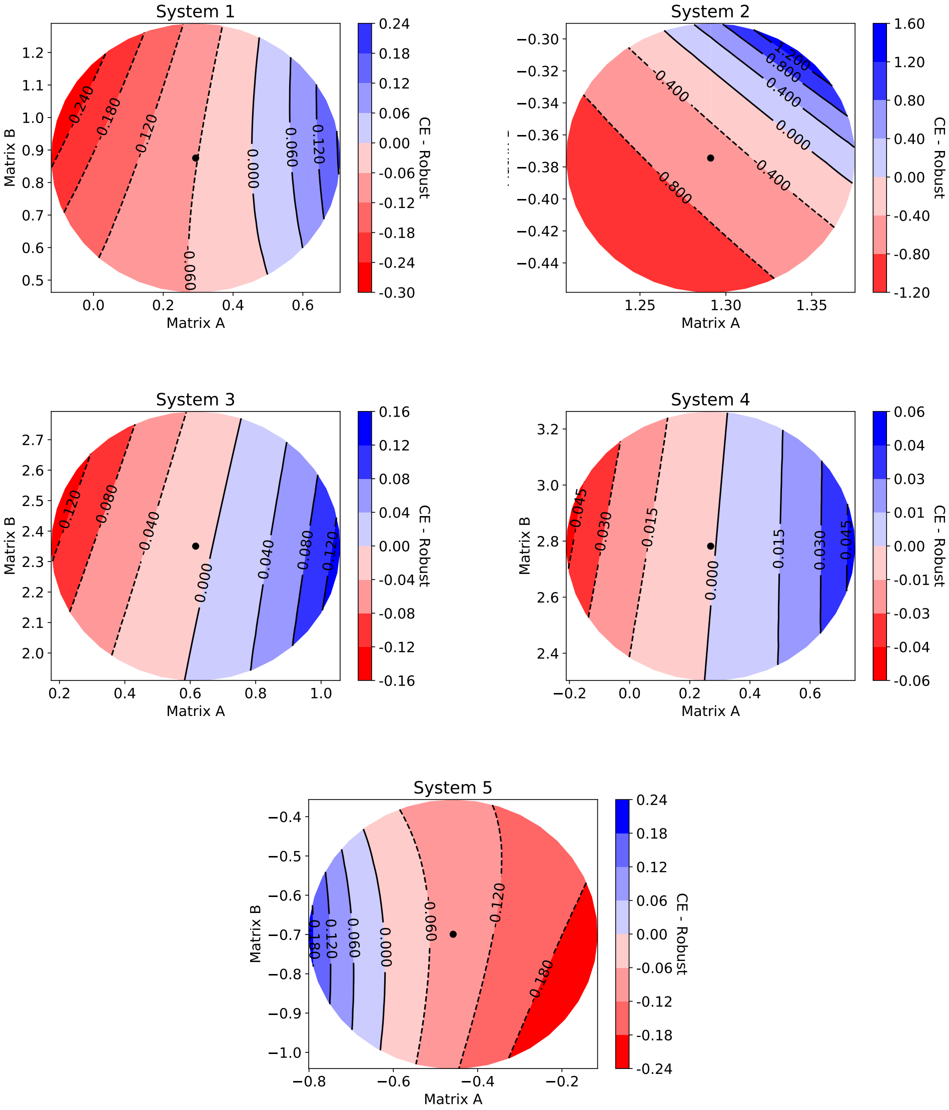


Figure 6.12: Difference between cost of CE and robust controller on the systems inside the confidence region.

Chapter 7

Discussion and outlook

The main topic in the thesis addresses the question how to find a stabilizing controller in a single trajectory with as little prior knowledge as possible. While we introduced an algorithm with a guarantee that will find, with high probability, a stabilizing controller in time depending only on system parameters, many other questions arose.

Firstly, running the algorithm the norm of the state can grow exponentially while we search for a stabilizing controller. Since this can be highly problematic for certain applications we changed the action inputs to the algorithm, in particular, instead of choosing zero mean Gaussian actions we choose actions as $u_i \sim \mathcal{N}(K_i x_i, \sigma_u^2 I)$, where K_i is a controller computed from the observed data. The question which arises is whether we can find a theoretical guarantee that for suitable nontrivial action selection we synthesize, with high probability, a stabilizing controller in finite time. Further, can we choose such actions that the cost suffered until we synthesize a stabilizing controller is not exponentially large? Or can we prove a lower bound on at least how much cost do we need to suffer to find a stabilizing controller? The answer to this question was already answered in the case when unknown matrix A_* is one dimensional and we know matrix B_*. Rantzer (2018) showed that by choosing the right controller the norm of states can be upper bounded in expectation as $\mathcal{O}\left(\frac{1}{s} + 1\right)$, where s is the number of taken steps. From their techniques it also follows that by choosing the proposed controller we find a stabilizing controller in finite time with high probability.

A minor step towards answering this questions in case when $d_x \geq 2$ is given by theorem 5.1 which shows that in the setting when we know matrix B_*, if we can find controllers K_i which bound the norm of states, then RLS estimate of A_* will be consistent and we will synthesize the controller in finite time.

The second question is based on the comparison of robust and CE controller. As we can see in section 6.4 robust controller stabilizes larger region around the estimates and achieves smaller worst infinite horizon cost. Simchowitz and Foster (2020) show that choosing CEC, when we have tight enough estimates, yields optimal $\Theta(\sqrt{d_u^2 d_x T})$ regret. Can we obtain the same regret with better other system parameters if we update optimistically robust controller synthesized from SDP (3.31) and use it to control the system? Does the performance of the robust compared to the CE controller change when we move to higher dimensions?

Finally, the most interesting question is whether we can formulate and prove finite termination of an algorithm which finds a stabilizing controller in partially observed setting, namely when we do not directly observe states x_i but rather $y_i = C_* x_i + v_i$, where C_* is unknown matrix and v_i unobserved noise. A good starting point would be to assume C_* is known or even $C_* = I$.

Appendix A

Supporting results

A.1 Elementary calculations

Lemma A.1 *For* $u = -(R + B_*^\top P_{t+1} B_*)^{-1} B_*^\top P_{t+1} A_* z$ *the expression* $u^\top R u + (A_* z + B_* u)^\top P_{t+1}(A_* z + B_* u)$ *equals to*

$$z^\top \left(A_*^\top P_{t+1} A_* - A_*^\top P_{t+1} B_* (R + B_*^\top P_{t+1} B_*)^{-1} B_*^\top P_{t+1} A_* \right) z.$$

Proof Denote by $K_t = -(R + B_*^\top P_{t+1} B_*)^{-1} B_*^\top P_{t+1} A_*$ and observe:

$$\begin{aligned} & u^\top R u + (A_* z + B_* u)^\top P_{t+1} (A_* z + B_* u) \\ =& z^\top \left(K_t^\top R K_t + (A_* + B_* K_t)^\top P_{t+1} (A_* + B_* K_t) \right) z. \end{aligned}$$

We further compute:

$$\begin{aligned} & K_t^\top R K_t + (A_* + B_* K_t)^\top P_{t+1} (A_* + B_* K_t) \\ &= A_*^\top P_{t+1} A_* + K_t^\top (R + B_*^\top P_{t+1} B_*) K_t + A_*^\top P_{t+1} B_* K_t + K_t^\top B_*^\top P_{t+1} A_* \\ &= A_*^\top P_{t+1} A_* - A_*^\top P_{t+1} B_* (R + B_*^\top P_{t+1} B_*)^{-1} B_*^\top P_{t+1} A_*. \end{aligned}$$

□

Lemma A.2 *We have:* $P_{t+1} - P_* = (A_* + B_* K_*)^\top (P_t - P_*)(A_* + B_* K_t)$

Proof Observe:

$$\begin{aligned} P_* &= Q + A_*^\top P_* A_* - A_*^\top P_* B_* (R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_* \\ &= Q + \left(A_* - B_* (R + B_*^\top P_* B_*)^{-1} B_*^\top P_* A_* \right)^\top P_* A_* \\ &= Q + (A_* + B_* K_*)^\top P_* A_*. \end{aligned}$$

With the same idea we obtain: $P_{t+1} = Q + A_*^\top P_t (A_* + B_* K_t)$. Hence we have:

$$\begin{aligned} P_{t+1} - P_* = & (A_* + B_* K_*)^\top (P_t - P_*)(A_* + B_* K_t) - K_*^\top B_*^\top P_t (A_* + B_* K_t) \\ & + (A_* + B_* K_*)^\top P_* B_* K_t \end{aligned}$$

To finish the proof we will show $K_*^\top B_*^\top P_t(A_* + B_*K_t) = (A_* + B_*K_*)^\top P_*B_*K_t$. Observe:

$$\begin{aligned}(A_* + B_*K_*)^\top P_*B_*K_t &= \left(A_*^\top P_*B_* + K_*^\top (R + B_*^\top P_*B_*) - K_*^\top R\right) K_t \\ &= -K_*^\top RK_t.\end{aligned}$$

The same idea shows $K_*^\top B_*^\top P_t(A_* + B_*K_t) = -K_*^\top RK_t$, which concludes the proof. □

Lemma A.3 *The following is equivalent:*

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \succeq 0 \iff \begin{pmatrix} A & -B \\ -B^\top & C \end{pmatrix} \succeq 0.$$

Proof By taking Schur complements we have:

$$\begin{aligned} &\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \succeq 0 \\ \iff\ & A - BC^{-1}B^\top \succeq 0 \\ \iff\ & A - (-B)C^{-1}(-B^\top) \succeq 0 \\ \iff\ & \begin{pmatrix} A & -B \\ -B^\top & C \end{pmatrix} \succeq 0. \end{aligned}$$

□

A.2 Dual problems

Here we derive the dual problem of SDP (2.20) and SDP (3.31). We first derive the dual problem of SDP (2.20). The Lagrangian for minimization problem eq. (2.20) is given by:

$$L(\Sigma, P) = \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \Sigma_{xx} - (A_*\ B_*)\Sigma(A_*\ B_*)^\top + \sigma_w^2 I \right\rangle,$$

where $P \in \mathbb{R}^{d_x \times d_x}$ is positive semi-definite matrix. For such P define:

$$g(P) = \inf_{\Sigma \succeq 0} L(\Sigma, P).$$

Further define the finite region of g as $F(g) = \{P | g(P) > -\infty\}$. Then the dual problem is defined as:

$$\begin{aligned} &\max_{P \succeq 0} g(P) \\ &\quad \text{s.t. } P \in F(g) \end{aligned}$$

Next we compute explicit formulas of g and $F(g)$.

$$\begin{aligned} g(P) &= \inf_{\Sigma} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \Sigma_{xx} - (A_* \; B_*)\Sigma(A_* \; B_*)^\top + \sigma_w^2 I \right\rangle \\ &= \langle P, \sigma_w^2 I \rangle + \inf_{\Sigma \succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} - \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + (A_* \; B_*)^\top P (A_* \; B_*), \Sigma \right\rangle \\ &= \begin{cases} \langle P, \sigma_w^2 I \rangle; & \begin{pmatrix} Q-P & 0 \\ 0 & R \end{pmatrix} + (A_* \; B_*)^\top P (A_* \; B_*) \succeq 0, \\ -\infty; & \text{otherwise.} \end{cases} \end{aligned}$$

Hence the dual is given by:

$$\begin{aligned} &\max_{P \succeq 0} \sigma_w^2 \, \|P\|_* \\ &\quad \text{s.t. } \begin{pmatrix} Q-P & 0 \\ 0 & R \end{pmatrix} + (A_* \; B_*)^\top P (A_* \; B_*) \succeq 0 \end{aligned} \tag{A.1}$$

It is interesting to note that the optimal solution P of the dual is equal to P_* associated with Riccati equations given in theorem 2.2.

Now we will derive the dual problem of SDP (3.31). The Lagrangian for SDP (3.31) is:

$$L(\Sigma, \lambda, P) = \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \begin{pmatrix} \Sigma_{xx} - (\widehat{A} \; \widehat{B})\Sigma(\widehat{A} \; \widehat{B})^\top - (\lambda + \sigma_w^2) I & (\widehat{A} \; \widehat{B})\Sigma \\ \Sigma(\widehat{A} \; \widehat{B})^\top & \lambda D - \Sigma \end{pmatrix} \right\rangle,$$

with $P \in \mathbb{R}^{(d+d_x)\times(d+d_x)}$ positive semi-definite matrix. Again for fixed P define

$$g(P) = \inf_{\Sigma \succeq 0, \lambda \geq 0} L(\Sigma, \lambda, P) \tag{A.2}$$

Further define finite region of P as $F(g) = \{P | g(P) > -\infty\}$. The dual problem of SDP (3.31) is then:

$$\begin{aligned} &\max_{P \succeq 0} g(P) \\ &\quad \text{s.t. } P \in F(g). \end{aligned}$$

We compute now function g and region $F(g)$. To ease the notation denote by $\theta = (\widehat{A} \; \widehat{B})$. With such a notation we have:

$$\begin{aligned} g(P) &= \inf_{\Sigma \succeq 0, \lambda \geq 0} L(\Sigma, \lambda, P) \\ &= \left\langle P, \begin{pmatrix} \sigma_w^2 I & 0 \\ 0 & 0 \end{pmatrix} \right\rangle + \inf_{\Sigma \succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \begin{pmatrix} \Sigma_{xx} - \theta\Sigma\theta^\top & \theta\Sigma \\ \Sigma\theta^\top & -\Sigma \end{pmatrix} \right\rangle \\ &+ \inf_{\lambda \geq 0} \lambda \left\langle P, \begin{pmatrix} I & 0 \\ 0 & -D \end{pmatrix} \right\rangle, \end{aligned}$$

where we notice that we can split the infimum over Σ and λ. We first compute infimum over Σ:

$$\begin{aligned}
&\inf_{\Sigma\succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \begin{pmatrix} \Sigma_{xx} - \theta\Sigma\theta^\top & \theta\Sigma \\ \Sigma\theta^\top & -\Sigma \end{pmatrix} \right\rangle \\
&= \inf_{\Sigma\succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}, \Sigma \right\rangle - \left\langle P, \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & 0 \end{pmatrix} \right\rangle + \left\langle P, \begin{pmatrix} \theta \\ -I \end{pmatrix} \Sigma \begin{pmatrix} \theta \\ -I \end{pmatrix}^\top \right\rangle \\
&= \inf_{\Sigma\succeq 0} \left\langle \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} - \begin{pmatrix} P_{xx} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \theta \\ -I \end{pmatrix}^\top P \begin{pmatrix} \theta \\ -I \end{pmatrix}, \Sigma \right\rangle \\
&= \begin{cases} 0; & \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} - \begin{pmatrix} P_{xx} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \theta \\ -I \end{pmatrix}^\top P \begin{pmatrix} \theta \\ -I \end{pmatrix} \succeq 0, \\ -\infty; & \text{otherwise.} \end{cases}
\end{aligned}$$

And then infimum over λ:

$$\inf_{\lambda\geq 0} \lambda \left\langle P, \begin{pmatrix} I & 0 \\ 0 & -D \end{pmatrix} \right\rangle = \begin{cases} 0; & \left\langle P, \begin{pmatrix} I & 0 \\ 0 & -D \end{pmatrix} \right\rangle \geq 0, \\ -\infty; & \text{otherwise.} \end{cases}$$

Now we can rewrite eq. (A.2) to obtain the Dual of SDP (3.31):

$$\begin{aligned}
\max_{P\succeq 0} &\left\langle P, \begin{pmatrix} \sigma_w^2 I & 0 \\ 0 & 0 \end{pmatrix} \right\rangle \\
\text{s.t. } &\left\langle P, \begin{pmatrix} I & 0 \\ 0 & -D \end{pmatrix} \right\rangle \geq 0 \\
&\begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} - \begin{pmatrix} P_{xx} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} (\widehat{A}\ \widehat{B}) \\ -I \end{pmatrix}^\top P \begin{pmatrix} (\widehat{A}\ \widehat{B}) \\ -I \end{pmatrix} \succeq 0
\end{aligned} \tag{A.3}$$

The Dual problem (A.3) is always feasible since $P = 0$ is a feasible solution. However its optimal value is unbounded as long as the primal SDP (3.31) is infeasible. When the primal becomes feasible, optimal value of SDP (A.3) is finite and strong duality applies - optimal value of the dual is equal to the optimal value of the primal problem. With very similar derivation as shown in this section we obtain that the dual problem of the Dual given by eq. (A.3) is the primal problem given by eq. (3.31).

A.3 Worst CE performing system

In this section we will derive $A_w(r), B_w(r)$ which we defined in section 6.4. Note that all the systems are one dimensional. In further computation denote by K the CEC. From the definition it follows:

$$\begin{aligned} A_w(r), B_w(r) = \underset{A,B}{\operatorname{argmin}} & (Q + K^2R)X \\ \text{s.t.} \quad & X = (A + BK)^2X + \sigma_w^2 \\ & (A - \widehat{A})^2 + (B - \widehat{B})^2 \leq r^2. \end{aligned} \tag{A.4}$$

Next by elementary transformations we obtain that we can equivalently to eq. (A.4) obtain $A_w(r), B_w(r)$ from:

$$\begin{aligned} A_w(r), B_w(r) = \underset{A,B}{\operatorname{argmin}} & (A + BK)^2 \\ \text{s.t.} \quad & (A - \widehat{A})^2 + (B - \widehat{B})^2 = r^2. \end{aligned} \tag{A.5}$$

To solve eq. (A.5) we write Lagrangian:

$$L(A, B, \lambda) = (A + BK)^2 - \lambda(r^2 - (A - \widehat{A})^2 + (B - \widehat{B})^2),$$

and solve a system of 3 equations with 3 unknowns:

$$\frac{\partial L}{\partial A} = 0, \quad \frac{\partial L}{\partial B} = 0, \quad \frac{\partial L}{\partial \lambda} = 0. \tag{A.6}$$

By solving eq. (A.6) we obtain:

$$(A_w(r), B_w(r)) = (\widehat{A} \pm \frac{r}{\sqrt{(K^2 + 1)}}, \widehat{B} \pm \frac{Kr}{\sqrt{K^2 + 1}}),$$

where we choose sign + if $\widehat{A} + \widehat{B}K \geq 0$ and sign $-$ otherwise. The corresponding worst case CE cost is then:

$$(Q + K^2R) \frac{\sigma_w^2}{1 - \left(\widehat{A} + \widehat{B}K \pm r\sqrt{K^2 + 1}\right)^2}.$$

Bibliography

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*.

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

Anderson, B. and Moore, J. (1979). *Optimal filtering*. Prentice-Hall information and system sciences series. Prentice-Hall.

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

Bart, H., ter Horst, S., Ran, A. C., and Woerdeman, H. J., editors (2018). *Operator Theory, Analysis and the State Space Approach*. Springer International Publishing.

Boyd, S. (2009). EE363 linear dynamical systems. [Online; accessed 29-July-2020].

Chen, X. and Hazan, E. (2020). Black-box control for linear dynamical systems. *arXiv preprint arXiv:2007.06650*.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. (2018). Online linear quadratic control. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1029–1038, Stockholmsmässan, Stockholm Sweden. PMLR.

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. *arXiv e-prints*, page arXiv:1902.06223.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2017). On the Sample Complexity of the Linear Quadratic Regulator. *arXiv e-prints*, page arXiv:1710.01688.

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.

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

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2018). Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64(8):3498–3505.

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.

Ibrahimi, M., Javanmard, A., and Roy, B. V. (2012). Efficient reinforcement learning for high dimensional linear quadratic systems. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2636–2644. Curran Associates, Inc.

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.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020a). Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020b). Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

Luo, Z.-Q., Sturm, J. F., and Zhang, S. (2004). Multivariate nonnegative quadratic mappings. *SIAM Journal on Optimization*, 14(4):1140–1162.

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.

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

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

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. (2018). Near optimal finite time identification of arbitrary linear dynamical systems. *arXiv e-prints*, page arXiv:1812.01251.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. (2019). Nonparametric finite time lti system identification.

Shirani Faradonbeh, M. K., Tewari, A., and Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica*, 96:342 – 353.

Simchowitz, M. (2020). Making non-stochastic control (almost) as easy as stochastic. *arXiv preprint arXiv:2006.05910*.

Simchowitz, M., Boczar, R., and Recht, B. (2019). Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802.

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

Simchowitz, M., Singh, K., and Hazan, E. (2020). Improper Learning for Non-Stochastic Control. *arXiv e-prints*, page arXiv:2001.09254.

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

Stoorvogel, A. A. (1992). *The H_∞ control problem: a state space approach*. Prentice-Hall, New York.

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.

Umenberger, J., Ferizbegovic, M., Schön, T. B., and Hjalmarsson, H. k. (2019). Robust exploration in linear quadratic reinforcement learning. 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 15336–15346. Curran Associates, Inc.

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.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

ONLINE LEARNING OF LINEAR-QUADRATIC REGULATORS

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
TREVEN	LENART

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 14.9.2020	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.