Sample Complexity Bounds for the Linear Quadratic Regulator

Stephen Tu

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-42
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-42.html

May 15, 2019

Sample Complexity Bounds for the Linear Quadratic Regulator

by

Stephen L. Tu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin Recht, Chair
Professor Peter Bartlett
Professor Francesco Borrelli

Spring 2019

Sample Complexity Bounds for the Linear Quadratic Regulator

Abstract

Sample Complexity Bounds for the Linear Quadratic Regulator

by

Stephen L. Tu

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Benjamin Recht, Chair

Reinforcement learning (RL) has demonstrated impressive performance in various domains such as video games, Go, robotic locomotion, and manipulation tasks. As we turn towards RL to power autonomous systems in the physical world, a natural question to ask is, how do we ensure that the behavior observed in the laboratory reflects the behavior that occurs when systems are deployed in the real world? How much data do we need to collect in order to learn how to control a system with a high degree of confidence?

This thesis takes a step towards answering these questions by establishing the Linear Quadratic Regulator (LQR) as a baseline for comparison of RL algorithms. LQR is a fundamental problem in optimal control theory for which the exact solution is efficiently computable with perfect knowledge of the underlying dynamics. This makes LQR well suited as a baseline for studying the sample complexity of RL algorithms which learn how to control from observing repeated interactions with the system.

The first part of this thesis focuses on *model-based* algorithms which estimate a model of the underlying system, and then build a controller based on the estimated dynamics. We show that the classic *certainty equivalence controller*, which discards confidence intervals surrounding the estimated dynamics, is efficient in regimes of low uncertainty. For regimes of moderate uncertainty, we propose a new model-based algorithm based on robust optimization, and show that it is also sample efficient.

The second part studies *model-free* algorithms which learn intermediate representations instead, or directly search for the parameters of the optimal controller. We first look at the classical least-squares policy iteration algorithm, and establish an upper bound on its sample complexity. We then use tools from asymptotic statistics to characterize the asymptotic behavior of both the certainty equivalence controller and the popular policy gradient method on a particular family of LQR instances, which allows us to directly compare the bounds. This comparison reveals that the model-free policy gradient method has polynomial in state/input dimension and horizon length worse sample complexity than the model-based certainty equivalence controller. Our experiments corroborate this finding and show that model-based algorithms are more sample efficient than model-free algorithms for LQR.

To my parents

Contents

List of Figures

Acknowledgments

This thesis is the product of the tremendous support that I have received from advisors, mentors, collaborators, friends, and family along the journey. When I was an undergraduate at UC Berkeley, Mike Franklin and Michael Armbrust showed me how fun and collaborative research really could be. This lead me to MIT, where I was a student in the database group advised by Sam Madden. Sam is truly a wonderful advisor and mentor, and even though this thesis has nothing to do with databases, many of the lessons I learned from working with Sam are reflected in this work. At MIT, I also had the wonderful privilege of collaborating with a lot of truly inspirational people. Frans Kaashoek and Nickolai Zeldovich made me feel right at home in PDOS. There was never a dull moment talking to Mike Stonebraker. Barbara Liskov and Eddie Kohler were the source of inspiration behind Silo. Both Kamalika Chaudhuri and Vinod Vaikuntanathan were kind enough to listen to my many ramblings while I futilely attempted to work on problems in differential privacy.

I am extremely grateful and lucky that Ben Recht took a chance on me when my research interests shifted towards machine learning and optimization. I do not know how I would have succeeded as an ML researcher if it were not for Ben. The five years I spent as a graduate student at UC Berkeley have been extremely rewarding. I want to acknowledge the tremendously important role that Ross Boczar, Sarah Dean, Horia Mania, Nikolai Matni, and Max Simchowitz played in this thesis as my closest collaborators at Berkeley– it is no exaggeration to say that this thesis would not exist without their collaboration. A special acknowledgment goes out to Andy Packard, who taught me everything I know about robust control, and who is an infinite source of inspiration. I would also like to acknowledge Francesco Borrelli and the students of the MPC lab for providing lots of valuable feedback in our joint group meetings, Michael Jordan and the students of SAIL for letting me selectively join their wonderful reading group, Mahdi Soltanolkotabi for his never ending patience as we worked on Procrustes Flow together (and for rolling with the name), Laurent Lessard for his clutch `#mathshop` answers on Slack and for being an awesome dinner companion at ICML in Sydney, Kevin Jamieson for instilling in me an appreciation for adaptive learning problems, Jason Lee for his ability to see the core idea of any paper in a matter of minutes, Alex Gittens, Becca Roelofs, Shivaram Venkataraman, and Ashia Wilson for entertaining my fascination with kernel methods and block coordinate descent, Xinghao Pan and Dimitris Papailiopoulos for involving me with the Cyclades project, Eric Jonas for allowing me to channel my inner Bayesian, and Evan Sparks for sharing his deep knowledge about both systems and the startup world. Finally, I would like to thank Murat Arcak, Peter Bartlett, Francesco Borrelli, and Andy Packard for agreeing to sit on my quals and dissertation committees.

The results on certainty equivalence control in Section 5.1 were directly inspired by Elad Hazan and Martin Wainwright who both asked whether or not it was possible to obtain a fast rate for LQR. Furthermore, the study of model-based and model-free methods in Section 6.2 was inspired by John Duchi and Dan Russo who both independently suggested asymptotic analysis and CLT based arguments. Finally, the policy iteration results of Section 6.1 are in collaboration with Karl Krauth.

In the summer of 2017 I had the wonderful experience of interning at Google Brain Robotics in New York City under the mentorship of Vikas Sindhwani. I cannot say enough wonderful things about Vikas both as a manager and as a collaborator. I would also like to especially thank my office mate over the summer, Xinyan Yan. His interest in reinforcement learning really kindled my own interest, which planted the seeds of this thesis.

Outside of research, graduate school is much more fun with other distractions. I would like to thank the wonderful community of the Recurse Center (RC). RC hosted me as a resident in the summer of 2015, and I have enjoyed being an active member of their community ever since. The Peoples Improv Theater in NYC and Endgames Improv in SF put on wonderful improv classes that I highly recommend as an escape from research. Thanks to all those who came out to see my shows.

The final thanks goes to my friends and especially my family. My parents have given nothing but the utmost support, even when it seemed like I had no clear direction.

Chapter 1

Introduction

Reinforcement learning (RL) has achieved impressive performance in the past decade on a wide variety of tasks across different domains such as video games [33, 81, 85], Go [103, 104], robotic locomotion [56, 66, 69, 111], autonomous racing [91, 125], and object manipulation tasks [59, 67, 68, 86]. Given these successes, it is natural to expect an increasing reliance on RL in the systems that we interact with on a daily basis. While widespread deployment of RL presents tremendous opportunities for our society as a whole, it is prudent to critically examine the potential consequences. In particular, when RL is deployed in the real world on a physical robotic system, how do we ensure that the system will act as it did in the laboratory? Put differently, how much data do we need to collect in order to learn how to control an autonomous system with a high degree of confidence?

In this thesis, we take a step towards answering these questions by establishing a useful baseline for comparison of core algorithms in RL. We turn to one of the most fundamental problems in optimal control theory, the Linear Quadratic Regulator (LQR). The LQR problem seeks to control a *linear* dynamical system, that is a dynamical system where the state evolution equation is described by a linear function of the current state and input, subject to a *quadratic* cost. We use LQR to understand the *sample complexity* of learning to control an unknown dynamical system. A celebrated result in control theory states that with perfect knowledge of the dynamics, an optimal control strategy can be exactly and efficiently computed. Therefore, LQR is a useful baseline for delineating the performance of methods which operate with imperfect knowledge of the dynamics.

We will study algorithms which broadly fall into two categories: (1) *model-based* approaches and (2) *model-free* approaches. While there is not a widely accepted technical definition of what it means for an algorithm to be model-based or model-free, we will use a common informal definition that describes model-based algorithms as those which, as an intermediate step, build an estimate of the transition dynamics which is then used to solve the control problem. On the other hand, model-free algorithms are defined as those which either directly search for the optimal controller, or learn other types of representations (such as value functions) as an intermediate step. It is important to note that this definition is not absolute; hybrid approaches that build models and learn other representations are in-

deed possible. This categorization into model-based and model-free is mostly for our own understanding, helping us to group algorithms together in order to understand better how they relate to each other.

One reoccurring theme we will see in this thesis is that learning the transition matrices that describe the underlying linear dynamics is one of the most statistically efficient ways to use the input/output data given to us, compared to learning other representations such as value and state-value functions. As a consequence, both in theory and practice, we will see that model-based algorithms tend to out-perform various model-free approaches for LQR. A natural question to ask as the reader is, what is the overall takeaway from this finding? We caution against coming to the naïve conclusion that model-based methods are always preferable to model-free methods. Indeed, it is hard for the author to imagine that this is true in full generality. Instead, the reader is asked to reflect upon why such a separation exists for LQR. What we will see is that the structure of the LQR problem lends itself very naturally to classical least-squares estimation of the linear dynamics, whereas estimating other representations is less natural and less sample efficient. Hence the real takeaway from this thesis is a lesson that we knew all along: take advantage of known structure when possible, and use the right tools for the problem at hand.

1.1 Markov Decision Processes and the Linear Quadratic Regulator

The Markov Decision Process (MDP) is the central object of study in RL. See Bertsekas [14] for an overview of RL and fundamental results. An infinite horizon average cost MDP is defined as a 4-tuple $(\mathcal{X}, \mathcal{U}, p, c)$ where $\mathcal{X}$ is the state space, $\mathcal{U}$ is the input space, $p(\cdot|x,u)$ for $x \in \mathcal{X}$, $u \in \mathcal{U}$ describes the probability distribution over the next state conditioned on the pair (x,u), and $c(x,u)$ denotes the stage-wise cost. The MDP task is to find a policy $\pi = \{u_t(\cdot)\}_{t=1}^{\infty}$ that minimizes the infinite horizon average cost:

$$J_\star = \inf_{\pi} \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} c(x_t,u_t)\right] \quad \text{s.t. } x_{t+1} \sim p(\cdot|x_t,u_t)\,. \tag{1.1.1}$$

Here, each function $u_t(\cdot)$ is allowed to depend on the history $(x_1, u_1, ..., x_{t-1}, u_{t-1}, x_t)$, but not on the future values. This type of function is called *causal*.

In general, without any more assumptions, solving (1.1.1) is intractable, even given perfect knowledge of the dynamics p and cost c. The classic assumption of RL is that the state space $\mathcal{X}$ and input space $\mathcal{U}$ is finite (this is often referred to as the "tabular setting"). Under this finiteness assumption, algorithms based on dynamic programming such as policy iteration or Q-learning can be used to solve (1.1.1). However, these algorithms typically scale polynomially in the size of $\mathcal{X}$ and $\mathcal{U}$, in both space and time complexity. While this is feasible for small problems, when $\mathcal{X}$ and $\mathcal{U}$ become large (e.g. arising from the discretization of a continuous space), then dynamic programming without any further structure becomes

intractable. Therefore, since we are primarily interested in the application of RL for continuous control problems, tabular methods are insufficient for our purposes.

One common assumption frequently made in the RL literature to move beyond a discrete state space is the *function approximation* setting (see e.g. Tsitsiklis and Van Roy [115] and the references within). The assumption is that we are given a finite set of basis functions $\{\phi_i(\cdot)\}$ such that the relevant quantities we are interested in can be well approximated in the span of the basis functions. For instance, if we are interested in learning a value function V^π, we would assume that $V^\pi \approx \sum_i w_i \phi_i$ for some set of coefficients $\{w_i\}$. It turns out that many tabular algorithms have corresponding variants in the function approximation setting. The function approximation setting is an important framework for RL that gives us a natural way to tackle continuous state spaces, but it is still a very general setting. In particular, minimal assumptions on the basis functions are assumed. Furthermore, given a particular problem at hand, it is often not clear what basis functions to pick, and quantifying the approximation error introduced by a particular set of basis function is a very non-trivial task.

In light of this discussion, we turn to the Linear Quadratic Regulator (LQR) as our main object of study. Let us now formally introduce the infinite horizon average cost LQR problem. Let S[1] be an $n \times n$ positive definite matrix and R be a $d \times d$ positive definite matrix. The LQR problem is to find the optimal policy π that minimizes:

$$J_\star = \inf_\pi \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t\right] \tag{1.1.2}$$

$$\text{s.t.} \quad x_{t+1} = Ax_t + Bu_t + w_t\,, \quad w_t \sim \mathcal{N}(0, W)\,. \tag{1.1.3}$$

Here, the sequence $\{w_t\}$ will be referred to as the *process noise*. It is understood that w_t is independent across time, i.e. w_i is independent from w_j for $i \neq j$. Comparing with (1.1.1), we see that the LQR problem (1.1.2)-(1.1.3) is a specific instance of an MDP with the state space $\mathcal{X} = \mathbb{R}^n$, the input space $\mathcal{U} = \mathbb{R}^d$, the transition probability $p(\cdot|x,u) = \mathcal{N}(Ax+Bu, W)$, and the stage wise cost $c(x,u) = x^\mathsf{T} S x + u^\mathsf{T} R u$.

The LQR problem is the most fundamental problem studied in the field of optimal control. See the book of Anderson and Moore [8] for an excellent introduction to linear quadratic control. We assume that the pair (A, B) is *stabilizable*, which means there exists a feedback matrix F such that $A+BF$ is a stable matrix (all eigenvalues have modulus strictly less than one). This assumption has several remarkable consequences. First, the optimal control law is a *stationary linear feedback* policy, i.e. $u_t = Kx_t$. Second, the feedback matrix K can be recovered by first solving for the unique[2] positive definite solution to the *discrete algebraic Riccati equation* (DARE):

$$V = A^\mathsf{T} V A - A^\mathsf{T} V B (B^\mathsf{T} V B + R)^{-1} B^\mathsf{T} V A + S\,, \tag{1.1.4}$$

[1] We depart from the usual convention of calling the cost matrix associated to the state as Q, in order to avoid confusion with the Q in Q-function.

[2] We made a simplifying assumption that both S and R are positive definite, which eliminates cases where the solution is not unique.

and setting $K = -(B^\mathsf{T} V B + R)^{-1} B^\mathsf{T} V A$. We will refer to the positive definite solution V of (1.1.4) as $V = \mathsf{dare}(A, B, S, R)$. The optimal cost $J_\star$ is then given by $\mathrm{tr}(VW)$.

Why is the LQR model a reasonable one to study from a practical engineering standpoint? Suppose for now that our dynamics are non-linear, that is:

$$x_{t+1} = f(x_t, u_t) \,,$$

where $f(\cdot)$ is no longer a linear function of x_t, u_t. It is clear that (1.1.3) no longer applies globally to our new dynamics. However, we can still apply (1.1.3) locally. There are many different methods of increasing complexity: here we briefly describe *Jacobian linearization*, which is one of the simplest schemes. Suppose that $(x_\star, 0)$ is an equilibrium point of f, meaning that $x_\star = f(x_\star, 0)$. Let us define the error $e_t := x_t - x_\star$, and suppose our goal is to send e_t to zero. Assuming that f is continuously differentiable and differentiating around this equilibrium point,

$$e_{t+1} = [D_x f(x_\star, 0)] e_t + [D_u f(x_\star, 0)] u_t + \text{H.O.T.} \,. \tag{1.1.5}$$

We expect this linear model to be a reasonable approximation to the true non-linear dynamics locally around the equilibrium point.

It is hopefully clear from this discussion what the advantages are of studying LQR for obtaining an understanding of RL for continuous control: the state and input spaces are naturally continuous, and given knowledge of the dynamics (A, B) and the cost matrices (S, R), we can efficiently compute the optimal solution. Therefore, we can completely isolate the effects of model uncertainty.

Problem statement. With this setup in place, we can more formally describe the core problem studied in this thesis:

> Find a nearly optimal solution to the LQR problem (1.1.2)-(1.1.3), where the cost matrices (S, R) are known, but the only access to the transition matrices (A, B) is via input/output data. Specifically, design an algorithm that chooses an input sequence $\{u_t\}$, observes the resulting trajectory $\{x_t\}$ of (1.1.3) induced by the chosen sequence $\{u_t\}$, and then returns a controller $\widehat{K}$ that stabilizes (A, B) and incurs average cost $J(\widehat{K})$ such that the sub-optimality gap $J(\widehat{K}) - J_\star$ is small. Here, the notation $J(\widehat{K})$ refers to the infinite horizon average cost induced by the controller $u_t = \widehat{K} x_t$.

We will be primarily interested in algorithms which come with *probably approximately correct* (PAC) style guarantees: after observing $T = T(\varepsilon, \delta)$ timesteps from the trajectory $\{x_t\}$, with probability at least $1 - \delta$ we have that $J(\widehat{K}) - J_\star \le \varepsilon$. The quantity $T(\varepsilon, \delta)$ will be referred to as the *sample complexity* of the algorithm.

1.2 Model-based Methods for LQR

The first step of model-based methods is to construct an estimate $(\widehat{A}, \widehat{B})$ of the true dynamics (A, B) from the trajectory data $\{x_t\}$ and input data $\{u_t\}$. While there are many ways to do this, we exploit the linear nature of the dynamics (1.1.3) and use ordinary least-squares (OLS):

$$(\widehat{A}, \widehat{B}) = \arg\min_{A,B} \frac{1}{2} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|^2 . \tag{1.2.1}$$

It is also possible to use a Tikhonov regularized least-squares estimator of the form:

$$(\widehat{A}, \widehat{B}) = \arg\min_{A,B} \frac{1}{2} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|^2 + \frac{\lambda}{2}(\|A\|_F^2 + \|B\|_F^2) . \tag{1.2.2}$$

Under the appropriate invertibility assumptions, the solution to (1.2.1) is given by:

$$(\widehat{A}, \widehat{B}) = \left(\sum_{t=0}^{T-1} x_{t+1} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\mathsf{T} \right) \left(\sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\mathsf{T} \right)^{-1} . \tag{1.2.3}$$

Of course, since $(\widehat{A}, \widehat{B}) \approx (A, B)$ only, an important question to answer is how to characterize the quality of the estimate $(\widehat{A}, \widehat{B})$. In general, describing the resulting confidence set is a non-trivial task, and we will see that the answer depends intimately on whether or not the matrix A is stable or not.

Another question that remains to be answered is how to best choose the input sequence $\{u_t\}$ to maximize the amount of information gained from the rollout. Indeed, there is an interesting experiment design question lurking here: given the past history $x_0, u_0, ..., x_{t-1}, u_{t-1}, x_t$, adaptively choose the next input u_t to play that would reveal the maximal amount of information. In this thesis, we leave the experiment design question to future work. Instead, we work with a passive input sequence where $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ and the u_t's are independent across time, or an feedback sequence of the form $u_t = Kx_t + \eta_t$ with $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2 I)$ and the η_t's are independent across time. Here, K is some stabilizing controller for (A, B) that is not necessarily optimal for LQR. We will see that both of these sequences provide rich enough excitation of the system that allows us to construct reasonable (not necessarily optimal, however) confidence intervals. In particular, these confidence intervals will be of the form:

$$C_t(\delta) := \{(A, B) : \|A - \widehat{A}\| \le \varepsilon_A(t, \delta) , \ \ \|B - \widehat{B}\| \le \varepsilon_B(t, \delta)\} . \tag{1.2.4}$$

where t here denotes the length of the trajectory observed and $\delta \in (0, 1)$ will be such that $\mathbb{P}((A, B) \in C_t(\delta)) \ge 1 - \delta$. Here, the probability is taken over the randomness of the process noise $\{w_t\}$ (c.f. (1.1.3)) and the randomness of the inputs $\{u_t\}$.

Armed with these confidence intervals, we now consider the question of learning a controller from the estimated model $(\widehat{A}, \widehat{B})$ and the interval $C_t(\delta)$. The most natural solution to this is known as the *certainty equivalence principle* [11]. The idea is to discard the interval $C_t(\delta)$, and use the controller $\widehat{K}$ given by:

$$\widehat{K} = -(\widehat{B}^\mathsf{T}\widehat{V}\widehat{B} + R)^{-1}\widehat{B}^\mathsf{T}\widehat{V}\widehat{A}\,, \tag{1.2.5}$$

$$\widehat{V} = \mathsf{dare}(\widehat{A}, \widehat{B}, S, R)\,. \tag{1.2.6}$$

We will also interchangeably refer to this controller as the *nominal* controller or the *plug-in* controller. When the interval $C_t(\delta)$ is very small (i.e. $\varepsilon_A \ll 1$ and $\varepsilon_B \ll 1$), then we expect that the nominal controller will not only stabilize (A, B) but also deliver cost $J(\widehat{K})$ that is quite competitive with $J_\star$. Quantifying when this happens, however, requires some work, and is one of the contributions of this thesis (Section 5.1). Furthermore, for moderate to large intervals, we do not expect nominal control to perform well because it does not take the uncertainty into account.

In regimes where the uncertainty in the model is moderate to large, we will look at a more sophisticated algorithm inspired from robust control. In particular, we will look at solving the following *robust* optimization procedure:

$$\inf_{\pi\in\Pi} \sup_{(A,B)\in C_t(\delta)} J(A, B, \pi)\,. \tag{1.2.7}$$

Here, the notation $J(A, B, \pi)$ denotes the average cost assuming the dynamics are described by (A, B) and the policy π is followed. We will take Π, the policy class we optimize over, to be the space of time-invariant linear feedback policies with memory. That is, the policy π is itself a linear dynamical system taking as input the sequence $\{x_t\}$ and outputting the control signal $\{u_t\}$. We note that as written, solving (1.2.7) is a non-trivial task. One of the contributions of this thesis (Section 5.2) is to show how to reasonably approximate (1.2.7) with tools from convex optimization.

1.3 Model-free Methods for LQR

We now describe methods which skip the model identification step and learn other representations for control. We first review two fundamental representations from RL, which are the value and state-value functions, the latter which is often referred to as the Q-function. Because we are dealing with infinite horizon average cost problems, the definition of these representations is more nuanced than the corresponding definitions in the finite horizon or discounted infinite horizon settings. Let K be a feedback policy which stabilizes (A, B), and let λ_K be the infinite horizon average cost associated to the policy K. We follow Tsitsiklis and Van Roy [116] and define the (relative) value function of K as:

$$V^K(x) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T}(x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t - \lambda_K)\,\middle|\, x_0 = x\right] \;\text{ s.t. }\; u_t = Kx_t\,. \tag{1.3.1}$$

The Bellman equation associated to (1.3.1) is:

$$\lambda_K + V^K(x) = c(x, Kx) + \mathbb{E}_{x' \sim p(\cdot|x,Kx)}[V^K(x')] \,. \tag{1.3.2}$$

From this, we see that $V^K(x) = x^\mathsf{T} V x$, where V solves the discrete Lyapunov equation

$$V = (A + BK)^\mathsf{T} V (A + BK) + S + K^\mathsf{T} R K \,. \tag{1.3.3}$$

We will denote this solution as $V = \mathsf{dlyap}(A + BK, S + K^\mathsf{T} R K)$.

Now, similar to (1.3.1), we define the relative Q-function of K as:

$$Q^K(x,u) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T} (x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t - \lambda_K) \,\middle|\, x_0 = x, u_0 = u\right] \text{ s.t. } u_t = K x_t \,. \tag{1.3.4}$$

The Bellman equation associated to (1.3.4) is:

$$\lambda_K + Q^K(x,u) = c(x,u) + \mathbb{E}_{x' \sim p(\cdot|x,u)}[Q^K(x', Kx')] \,. \tag{1.3.5}$$

Solving this equation, we obtain that $Q^K(x,u) = \begin{bmatrix} x \\ u \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} x \\ u \end{bmatrix}$ where Q is:

$$Q = \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A^\mathsf{T} \\ B^\mathsf{T} \end{bmatrix} V \begin{bmatrix} A & B \end{bmatrix} \,. \tag{1.3.6}$$

Equation 1.3.5 gives us a natural algorithm to estimate the parameter Q of the Q-function Q^K. We let $q = \mathrm{svec}(Q)$[3] and write $Q^K(x,u) = \phi(x,u)^\mathsf{T} q$, where $\phi(x,u) = \mathrm{svec}\left(\begin{bmatrix} x \\ u \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}^\mathsf{T}\right)$. Substituting this into (1.3.4) and using the fact that $\lambda_K = \left\langle Q, \begin{bmatrix} I \\ K \end{bmatrix} W \begin{bmatrix} I \\ K \end{bmatrix}^\mathsf{T} \right\rangle$, we obtain:

$$q^\mathsf{T}(\phi(x,u) - \mathbb{E}_{x' \sim p(\cdot|x,u)}[\phi(x', Kx')] + f) = c(x,u) \,, \tag{1.3.7}$$

where $f = \mathrm{svec}\left(\begin{bmatrix} I \\ K \end{bmatrix} W \begin{bmatrix} I \\ K \end{bmatrix}^\mathsf{T}\right)$. Therefore, given samples of the form $\{(x_t, u_t, x_{t+1})\}_{t=0}^{T-1}$ we can use the *errors-in-variables* approach [25] to construct (under the necessary invertibility assumptions) the following least-squares estimator for q:

$$\widehat{q} = \left(\sum_{t=0}^{T-1} \phi(x_t,u_t)(\phi(x_t,u_t) - \phi(x_{t+1}, Kx_{t+1}) + f)^\mathsf{T}\right)^{-1} \sum_{t=0}^{T-1} c(x_t,u_t)\phi(x_t,u_t) \,. \tag{1.3.8}$$

[3]Here, $\mathrm{svec} : \mathrm{Sym}_{n\times n} \to \mathbb{R}^{n(n+1)/2}$ is the linear operator mapping the space of $n \times n$ symmetric matrices (denoted $\mathrm{Sym}_{n\times n}$) to vectors while preserving the property that $\langle \mathrm{svec}(M_1), \mathrm{svec}(M_2)\rangle_{\mathbb{R}^{n(n+1)/2}} = \langle M_1, M_2\rangle_{\mathbb{R}^{n\times n}}$ for all symmetric M_1, M_2.

This estimator $\widehat{q}$ will be referred to as the *LSTD-Q estimator*. One of the contributions of this thesis (Section 6.1.2) will be to quantify the error $\|q - \widehat{q}\|$ incurred by LSTD-Q.

Given an algorithm that estimates the parameters of its associated Q-function for a policy K, we can use this algorithm as a sub-routine to construct an algorithm for policy optimization. One of the most classical approaches in RL is *approximate policy iteration* (PI), and it is a basic form of approximate dynamic programming. Algorithm 1 presents least-squares policy iteration (LSPI) from Lagoudakis and Parr [63], which is approximate PI combined with LSTD-Q for policy evaluation.

Algorithm 1 Least-Squares Policy Iteration (LSPI) for LQR

Require: Initial stabilizing controller K_0, exploration controller K_{play}, N number of policy iterations, T length of rollout for estimation, σ_η^2 exploration variance, μ lower eigenvalue bound.

1: Collect a trajectory $\mathcal{D} = \{(x_k, u_k, x_{k+1}\}_{k=1}^T$ using input $u_k = K_{\text{play}} x_k + \eta_k$, with $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2 I)$.
2: **for** $t = 0, ..., N-1$ **do**
3: $\quad \widehat{Q}_t = \mathsf{Proj}_\mu(\mathsf{LSTDQ}(\mathcal{D}, K_t))$.
4: $\quad K_{t+1} = G(\widehat{Q}_t)$ (c.f. (1.3.9)).
5: **end for**
6: return K_N.

In Algorithm 1, $\mathsf{Proj}_\mu(\cdot) = \arg\min_{X = X^\mathsf{T}: X \succeq \mu \cdot I} \|X - \cdot\|_F$ is the Euclidean projection onto the set of symmetric matrices lower bounded by $\mu \cdot I$. Furthermore, the map $G(\cdot)$ takes an $(n+d) \times (n+d)$ positive definite matrix and returns a $d \times n$ matrix:

$$G\left(\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^\mathsf{T} & Q_{22} \end{bmatrix}\right) = -Q_{22}^{-1} Q_{12}^\mathsf{T} \,. \tag{1.3.9}$$

One of the contributions of this thesis is to provide a sample complexity analysis of Algorithm 1 (Section 6.1).

We now turn our attention to an alternative style of model-free algorithm which is based on ideas from derivative-free optimization (DFO) instead of approximate dynamic programming. See the book by Spall [107] for an overview of derivative-free optimization. We will present an overview of two methods based on DFO that are closely related. The first is based on random perturbations. As before, define the function $J(K)$ as:

$$J(K) = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t\right] \text{ s.t. } u_t = K x_t \,. \tag{1.3.10}$$

The function $J(K)$ is finite when K stabilizes (A, B), and equals $+\infty$ otherwise. In the domain of stabilizing K, the function $J(K)$ is differentiable. Therefore, in principle we could optimize $J(K)$ via a local search method:

$$K_{t+1} = K_t - \eta_t \nabla J(K_t) \, .$$

However, in order to compute $\nabla J(K)$, we need access to the transition dynamics (A, B) (and if we knew (A, B) we could just compute the optimal controller directly). The trick to get around needing the dynamics is to smooth (convolve) the function $J(K)$, and construct a stochastic gradient estimate of the smoothed function. In particular, we fix a $\sigma > 0$ and define:

$$J_\sigma(K) = \mathbb{E}_\xi[J(K + \sigma\xi)] \, , \tag{1.3.11}$$

where each entry of ξ is drawn independently from a $\mathcal{N}(0, 1)$ distribution. A standard fact is that the gradient of $J_\sigma(K)$ can be expressed as:

$$\nabla J_\sigma(K) = \mathbb{E}_\xi \left[\frac{J(K + \sigma\xi) - J(K - \sigma\xi)}{2\sigma} \xi \right] \, . \tag{1.3.12}$$

Hence we can use the following stochastic gradient estimator $\widehat{g}$:

$$\widehat{g} = \frac{J(K + \sigma\xi) - J(K - \sigma\xi)}{2\sigma} \xi \, . \tag{1.3.13}$$

Notice how this formula does not involve computing any gradients of the original function $J(K)$, but only requires pointwise evaluation of $J(K)$ which we can obtain from rollouts. That is, we have side-stepped the issue of needing to know the model dynamics (A, B). We can now use $\widehat{g}$ as a stochastic gradient estimate of $\nabla J_\sigma(K)$ and plug it into our favorite stochastic optimization algorithm to optimize J_σ. Furthermore, for small σ we expect that $J_\sigma(K) \approx J(K)$. An analysis of the behavior of these type of DFO algorithms on LQR can be found in Malik et al. [73].

We now describe a different approach based on policy gradients (REINFORCE) [126]. Instead of directly perturbing the parameters of the controller K, we will perturb the actions u_t. Again, we fix a $\sigma > 0$, and this time we define:

$$J_\sigma(K) = \lim_{T \to \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t \right] \text{ s.t. } u_t = K x_t + \eta_t \, . \tag{1.3.14}$$

Here, we have $\eta_t \sim \mathcal{N}(0, \sigma^2 I)$ and the η_t's are independent across time. Let $\tau_{1:T} = (x_1, u_1, x_2, u_2, ..., x_T, u_T)$ denote a trajectory of length T. Then under appropriate regularity

conditions, with $c(\tau_{s:T}) := \sum_{t=s}^{T} c(x_t, u_t)$,

$$\begin{aligned}
\nabla J_\sigma(K) &= \nabla \lim_{T\to\infty} \int_{\tau_{1:T}} \frac{1}{T} c(\tau_{1:T}) p(\tau_{1:T})\, d\tau_{1:T} \\
&= \lim_{T\to\infty} \nabla \int_{\tau_{1:T}} \frac{1}{T} c(\tau_{1:T}) p(\tau_{1:T})\, d\tau_{1:T} \\
&= \lim_{T\to\infty} \int_{\tau_{1:T}} \frac{1}{T} c(\tau_{1:T}) \nabla p(\tau_{1:T})\, d\tau_{1:T} \\
&= \lim_{T\to\infty} \int_{\tau_{1:T}} \frac{1}{T} c(\tau_{1:T}) \nabla \log p(\tau_{1:T}) p(\tau_{1:T})\, d\tau_{1:T} \,.
\end{aligned}$$

Now we write:

$$\log p(\tau_{1:T}) = \sum_{t=0}^{T-1} \log p(x_{t+1}|x_t, u_t) + \sum_{t=1}^{T} \log p(u_t|x_t) \,,$$

and because the transition dynamics $p(x_{t+1}|x_t, u_t)$ do not depend on K, we have that the gradient $\nabla \log p(x_{t+1}|x_t, u_t) = 0$ and therefore:

$$\nabla \log p(\tau_{1:T}) = \sum_{t=1}^{T} \nabla \log p(u_t|x_t) = \sum_{t=1}^{T} \frac{1}{\sigma^2} \eta_t x_t^\mathsf{T} \,.$$

Above, the last equality follows since $p(u_t|x_t) = \mathcal{N}(Kx_t, \sigma^2 I)$. Therefore, we have:

$$\begin{aligned}
\nabla J_\sigma(K) &= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \frac{c(\tau_{1:T})}{\sigma^2} \eta_t x_t^\mathsf{T} \right] \\
&= \lim_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{1}{T} \frac{c(\tau_{1:T})}{\sigma^2} \eta_t x_t^\mathsf{T} \,\middle|\, x_1, \eta_1, ..., x_t \right] \right] \\
&= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \frac{c(\tau_{t:T})}{\sigma^2} \eta_t x_t^\mathsf{T} \right] \,.
\end{aligned}$$

We can therefore choose a large T and use the following stochastic gradient estimate $\widehat{g}$ which we compute from a rollout with $u_t = Kx_t + \eta_t$:

$$\widehat{g} = \frac{1}{T} \sum_{t=1}^{T} \frac{c(\tau_{t:T})}{\sigma^2} \eta_t x_t^\mathsf{T} \,.$$

As with DFO, we can plug the estimate $\widehat{g}$ into our favorite stochastic optimization algorithm. An upper bound analysis of the behavior of policy gradient like algorithms on LQR can be found in Fazel et al. [41]. In this thesis, we will study REINFORCE in particular using asymptotic analysis in order to compare the performance of REINFORCE to the model-based nominal control method (Section 6.2).

Chapter 2

Related Work

We survey the literature on obtaining approximate LQR solutions from imperfect knowledge of the dynamics, focusing on work which combines learning and control. We mostly focus on work that contains non-asymptotic results, since non-asymptotic results are the main focal point of this thesis. We divide the related work broadly into two categories: model-based and model-free methods. Within each of these categories, there is a further sub-division between offline (batch) settings and online (adaptive) settings. While this thesis will focus only on the offline setting, we discuss the online setting here for completeness.

2.1 Model-based Methods

We first discuss model-based methods in the offline setting. Fiechter [42] is the first to consider PAC style bounds for LQR with unknown dynamics. He studies the infinite horizon *discounted* LQR problem, and shows that the nominal controller, if it stabilizes the true system, achieves cost $J(\widehat{K})$ satisfying $J(\widehat{K}) - J_\star \leq \mathcal{O}(\varepsilon)$, where ε is the error of the estimated parameters. In Section 5.1, we will give sufficient conditions to ensure that the true system is stabilized by the nominal controller, and we will also show that the sub-optimality gap actually scales as $\mathcal{O}(\varepsilon^2)$ instead of $\mathcal{O}(\varepsilon)$. Turning to methods based on robust control, we note there is a rich literature in controls dealing with structured uncertainty such as μ-synthesis [89] or integral quadratic constraints [79]. The main drawback of traditional robust control approaches is that we are unaware of a way to quantify the performance degradation as a function of the size of the uncertainty set. Our approach in Section 5.2, which is based on a recent development in robust control called System Level Synthesis (SLS) [35], is to the best of our knowledge, the first robust control approach that comes with a guaranteed bound on the sub-optimality gap.

We now turn our attention to the online adaptive setting, which is inspired from the classical problem of adaptive control of LQR [11]. We will focus on the regret formulation introduced by Abbasi-Yadkori and Szepesvári [1]. Inspired by the classic "bet on the best" principle of Bittanti and Campi [18], Abbasi-Yadkori and Szepesvári [1] show how *opti-*

mism in the face of uncertainty (OFU) yields a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret algorithm for LQR. Ibrahimi et al. [50] extend this work to high dimensional systems with sparsity in the dynamics, and Faradonbeh et al. [38] remove some un-necessary technical assumptions. The main issue with the OFU algorithm proposed by Abbasi-Yadkori and Szepesvári [1] is computational: it is not clear if the OFU sub-routine can be efficiently solved, as it is a non-convex optimization problem. There have been several attempts to remedy this issue. First, several works have looked at Thompson sampling as an alternative to OFU [5, 6, 87]. The work of Abeille and Lazaric [5] shows a $\widetilde{\mathcal{O}}(T^{2/3})$ regret for Thompson sampling which was later improved by Abeille and Lazaric [6] to $\widetilde{\mathcal{O}}(T^{1/2})$, but the analysis only applies to scalar systems. On the other hand, Ouyang et al. [87] study a Bayesian regret formulation and show $\widetilde{\mathcal{O}}(T^{1/2})$ regret under strong technical assumptions. Dean et al. [32] show how to use SLS to achieve an efficient $\widetilde{\mathcal{O}}(T^{2/3})$ regret algorithm. Based on the techniques in Section 5.1, it is possible to show that the nominal controller coupled with a greedy exploration strategy is sufficient to achieve $\widetilde{\mathcal{O}}(T^{1/2})$ regret: this is described in more detail in Mania et al. [75]. Parallel to Mania et al. [75], Cohen et al. [30] give an efficient procedure that also achieves $\widetilde{\mathcal{O}}(T^{1/2})$ regret and is based on semidefinite programming.

2.2 Model-free Methods

We now survey the literature studying model-free algorithms on LQR. The Ph.D. thesis of Bradtke [24] studies the least-squares policy iteration algorithm applied to noiseless LQR and shows asymptotic consistency. For policy evaluation, Tu and Recht [117] give a non-asymptotic bound for least-squares temporal difference learning for infinite horizon discounted LQR. They empirically evaluate LSPI and observe that it has worse sample complexity than the model-based methods they compared to, but do not provide an analysis. In Section 6.1, we will provide the first non-asymptotic analysis for LSPI. For policy gradients, Fazel et al. [41] show that policy gradients converges to the optimal solution with polynomial (in the relevant quantities) sample complexity. However, Fazel et al. [41] focus on the case where the only noise in the system is in the initial state, and the rest of the state transitions are deterministic. Malik et al. [73] study derivative-free optimization for LQR and also show polynomial sample complexity. Their bounds suggest that having two samples in each evaluation allow one to obtain $\widetilde{\mathcal{O}}(1/\varepsilon)$ sample complexity versus $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ with only one sample. Mania et al. [74] empirically evaluate derivative-free optimization on LQR and show that it is competitive with LSPI. Vemula et al. [120] study the different regimes in which action space perturbation outperforms parameter space perturbation and vice-versa.

For the online setting, Abbasi-Yadkori et al. [4] show that a model-free algorithm based on follow the leader achieves $\widetilde{\mathcal{O}}(T^{2/3+\varepsilon})$ regret for any $T \geq C^{1/\varepsilon}$ where $C > 0$ is a constant that depends on the system. Based on the techniques in Section 6.1, this can be improved to $\widetilde{\mathcal{O}}(T^{2/3})$ using LSPI coupled with greedy exploration. It remains open whether a model-free algorithm based on approximate policy iteration can achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

2.3 Identification of Linear Systems

While not the direct focus of this thesis, we review results surrounding non-asymptotic identification of an unknown linear system from trajectory data. The asymptotic version of this question has classically been studied in the sub-field of control theory known as system identification [71]. Early non-asymptotic rates are given by Campi and Weyer [28] and Vidyasagar and Karandikar [123]. Goldenshluger [44] and Tu et al. [119] study the identification of a linear system from an input/output map (transfer function) perspective. Hardt et al. [46] show that gradient descent can learn the parameters of a linear system such that the resulting parameters deliver good predictive performance, under some technical assumptions on the A matrix. Hazan and Zhang [47] and Hazan et al. [48] propose a spectral filtering technique to learn unknown linear systems, with performance measured in a regret framework. One notable property of their bounds is that they apply to marginally stable systems when the spectral radius $\rho(A) = 1$; the bounds do not degrade as $\rho(A)$ approaches one.

By making the additional assumption that the state can be observed, one can prove stronger results regarding parameter estimation. Simchowitz et al. [105] show how to recover the A matrix of a linear system $x_{t+1} = Ax_t + w_t$, with the results applicable to the regime of marginal stability. This result is discussed in more detail in Chapter 3. Faradonbeh et al. [39] and Sarkar and Rakhlin [98] show how to obtain identification results for unstable systems under some technical assumptions. Rantzer [95] gives concentration bounds for the least-squares estimator in the scalar setting (which was sharpened by the scalar analysis in Simchowitz et al. [105]). Finally, there has been recent work in extending the parameter identification results to the case of partially observed linear systems [88, 99, 106, 114].

Chapter 3

Linear System Identification

In this chapter, we survey some basic non-asymptotic results for linear system identification. While not the primary focus of this thesis, the results here and the techniques behind them will play an important role for the remainder of the thesis.

We recall the setting from Chapter 1. We are interested in an unknown linear dynamical system:

$$x_{t+1} = Ax_t + Bu_t + w_t \,. \tag{3.0.1}$$

Here and for the remainder of this thesis, we will assume for simplicity that $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$ instead of the more general $\mathcal{N}(0, W)$. We will also assume that $x_0 = 0$. Suppose that we have access to N independent trajectories of (3.0.1), each of length T (that is, we perform N rollouts of length T, resetting the system after each rollout). Let us denote this data by $\{x_t^{(i)}\}$ with $1 \le i \le N$ and $1 \le t \le T+1$. Given this data, we consider the least-squares estimator:

$$(\widehat{A}, \widehat{B}) = \arg\min_{A,B} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \|x_{t+1}^{(i)} - Ax_t^{(i)} - Bu_t^{(i)}\|^2 \,. \tag{3.0.2}$$

The solution to (3.0.2) is given by:

$$(\widehat{A}, \widehat{B}) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{t+1}^{(i)} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix}^\mathsf{T} \right) \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix}^\mathsf{T} \right)^{-1} . \tag{3.0.3}$$

Here, we assume the empirical covariance matrix is invertible. It is not hard to see that the error of (3.0.3) is given by:

$$(\widehat{A} - A, \widehat{B} - B) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} w_t^{(i)} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix}^\mathsf{T} \right) \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix} \begin{bmatrix} x_t^{(i)} \\ u_t^{(i)} \end{bmatrix}^\mathsf{T} \right)^{-1} . \tag{3.0.4}$$

We consider the following question: how do we bound the error $\|\widehat{A} - A\|$ and $\|\widehat{B} - B\|$ as a function of N, T and various quantities relating to (A, B)? We note that this question is non-trivial because along the i-th trajectory, the covariates $\{x_t^{(i)}\}$ are correlated across time (the covariates are independent across trajectories, however). Therefore, the standard analysis of random design least-squares regression (see e.g. Hsu et al. [49]) does not apply.

Before we proceed, we discuss heuristically how we qualitatively expect the error to depend on (A, B). The more stable A is roughly means that the process noise entering the system dampens out quicker, whereas if A is unstable then the process noise drives the system state away from the origin exponentially fast. While the latter behavior is undesirable for regulation purposes, it is actually quite desirable from an estimation perspective, because it means that the signal to noise ratio is extremely high. This heuristic reasoning suggests that systems that are more "explosive" should yield better estimation rates than those that are very stable. We note this qualitative behavior is at odds with estimation bounds that depend on the *mixing time* of the system [61, 62, 127]. Indeed, the mixing time of a linear system degrades as the system tends towards instability, and does not exist for unstable systems. It is for this reason we do not utilize learning bounds based on mixing time arguments in the sequel.

3.1 A Simple Analysis Based on Independent Rollouts

We first consider a slight modification to the estimator (3.0.3) which yields a simple and very general analysis that applies to any (A, B). This modification is that, we will discard all the trajectory data except the very last state transition, exploiting the independence across trajectories. Mathematically,

$$(\widehat{A}, \widehat{B}) = \left(\sum_{i=1}^{N} x_{T+1}^{(i)} \begin{bmatrix} x_T^{(i)} \\ u_T^{(i)} \end{bmatrix}^\mathsf{T} \right) \left(\sum_{i=1}^{N} \begin{bmatrix} x_T^{(i)} \\ u_T^{(i)} \end{bmatrix} \begin{bmatrix} x_T^{(i)} \\ u_T^{(i)} \end{bmatrix}^\mathsf{T} \right)^{-1} . \tag{3.1.1}$$

This modification makes the analysis simple because now we have that $(x_T^{(i)}, u_T^{(i)})$ is independent across rollouts. Therefore, the existing tools we have for analyzing least-squares apply.

Proposition 3.1.1. *Define the matrices*

$$G_T = \begin{bmatrix} A^{T-1}B & A^{T-2}B & \dots & B \end{bmatrix} \quad \text{and} \quad F_T = \begin{bmatrix} A^{T-1} & A^{T-2} & \dots & I \end{bmatrix} . \tag{3.1.2}$$

Assume we collect data from the linear, time-invariant system initialized at $x_0 = 0$, using inputs $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $t = 1, ..., T$. Suppose that the process noise is $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$ and that

$$N \geq 8(n + d) + 16 \log(4/\delta) .$$

Then, with probability at least $1-\delta$*, the least squares estimator using only the final sample of each trajectory satisfies both the inequality*

$$\|\widehat{A}-A\| \leq \frac{16\sigma_w}{\sqrt{\lambda_{\min}(\sigma_u^2 G_T G_T^\mathsf{T} + \sigma_w^2 F_T F_T^\mathsf{T})}} \sqrt{\frac{(n+2d)\log(36/\delta)}{N}}\,, \tag{3.1.3}$$

and the inequality

$$\|\widehat{B}-B\| \leq \frac{16\sigma_w}{\sigma_u} \sqrt{\frac{(n+2d)\log(36/\delta)}{N}}\,. \tag{3.1.4}$$

Note that Proposition 3.1.1 yields an optimal dependence in terms of the number of parameters from a parameter counting perspective: (A, B) together have $n(n+d)$ parameters to learn and each measurement consists of n values. Moreover, this proposition further illustrates that not all linear systems are equally easy to estimate. The matrices $G_T G_T^\mathsf{T}$ and $F_T F_T^\mathsf{T}$ are finite time *controllability Gramians* for the control and noise inputs, respectively. These are standard objects in control: each eigenvalue/vector pair of such a Gramian characterizes how much input energy is required to move the system in that particular direction of the state-space. Therefore $\lambda_{\min}\left(\sigma_u^2 G_T G_T^\mathsf{T} + \sigma_w^2 F_T F_T^\mathsf{T}\right)$ quantifies the least controllable, and hence most difficult to excite and estimate, mode of the system. This property is captured nicely in our bound, which indicates that for systems for which all modes are easily excitable (i.e., all modes of the system amplify the applied inputs and disturbances), the identification task becomes easier.

The bounds (3.1.3) and (3.1.4) in Proposition 3.1.1 require knowledge of the true (A, B) in order to compute. Section 2 of Dean et al. [31] describes two alternative methods that avoid knowledge of (A, B), based on data-dependent concentration bounds and the bootstrap.

Before we present a proof of Proposition 3.1.1, we state some auxiliary results which will aid our analysis.

Lemma 3.1.2. *Fix a* $\delta \in (0,1)$ *and* $N \geq 2\log(1/\delta)$*. Let* $f_k \in \mathbb{R}^m$*,* $g_k \in \mathbb{R}^n$ *be independent random vectors* $f_k \sim \mathcal{N}(0, \Sigma_f)$ *and* $g_k \sim \mathcal{N}(0, \Sigma_g)$ *for* $1 \leq k \leq N$*. With probability at least* $1-\delta$*,*

$$\left\| \sum_{k=1}^{N} f_k g_k^\mathsf{T} \right\| \leq 4\|\Sigma_f\|^{1/2}\|\Sigma_g\|^{1/2}\sqrt{N(m+n)\log(9/\delta)}\,.$$

Proof. First, recall Bernstein's inequality. Let $X_1, ..., X_p$ be zero-mean independent r.v.s satisfying the Orlicz norm bound $\|X_i\|_{\psi_1} \leq K$ (see Section 2.7.1 of Vershynin [122] for an overview of Orlicz spaces). Then as long as $p \geq 2\log(1/\delta)$, with probability at least $1-\delta$,

$$\sum_{i=1}^{p} X_i \leq K\sqrt{2n\log(1/\delta)}\,.$$

Next, let Q be an $m \times n$ matrix. Let $u_1, ..., u_{M_\varepsilon}$ be a ε-net for the m-dimensional ℓ_2 ball, and similarly let $v_1, ..., v_{N_\varepsilon}$ be a ε covering for the n-dimensional ℓ_2 ball. For each $\|u\| = 1$ and $\|v\| = 1$, let u_i, v_j denote the elements in the respective nets such that $\|u - u_i\| \leq \varepsilon$ and $\|v - v_j\| \leq \varepsilon$. Then,

$$\begin{aligned} u^\mathsf{T} Q v &= (u - u_i + u_i)^\mathsf{T} Q v = (u - u_i)^\mathsf{T} Q v + u_i^\mathsf{T} Q (v - v_j + v_j) \\ &= (u - u_i)^\mathsf{T} Q v + u_i^\mathsf{T} Q (v - v_j) + u_i^\mathsf{T} Q v_j \,. \end{aligned}$$

Hence,

$$u^\mathsf{T} Q v \leq 2\varepsilon \|Q\| + u_i^\mathsf{T} Q v_j \leq 2\varepsilon \|Q\| + \max_{1 \leq i \leq M_\varepsilon, 1 \leq j \leq N_\varepsilon} u_i^\mathsf{T} Q v_j \,.$$

Since u, v are arbitrary on the sphere,

$$\|Q\| \leq \frac{1}{1 - 2\varepsilon} \max_{1 \leq i \leq M_\varepsilon, 1 \leq j \leq N_\varepsilon} u_i^\mathsf{T} Q v_j \,.$$

Now we study the problem at hand. Choose $\varepsilon = 1/4$. By a standard volume comparison argument, we have that $M_\varepsilon \leq 9^m$ and $N_\varepsilon \leq 9^n$, and that

$$\left\| \sum_{k=1}^N f_k g_k^\mathsf{T} \right\| \leq 2 \max_{1 \leq i \leq M_\varepsilon, 1 \leq j \leq N_\varepsilon} \sum_{k=1}^N (u_i^\mathsf{T} f_k)(g_k^\mathsf{T} v_j) \,.$$

Note that $u_i^\mathsf{T} f_k \sim \mathcal{N}(0, u_i^\mathsf{T} \Sigma_f u_i)$ and $g_k^\mathsf{T} v_j \sim \mathcal{N}(0, v_j^\mathsf{T} \Sigma_g v_j)$. By independence of f_k and g_k, $(u_i^\mathsf{T} f_k)(g_k^\mathsf{T} v_j)$ is a zero mean sub-Exponential random variable, and therefore $\|(u_i^\mathsf{T} f_k)(g_k^\mathsf{T} v_j)\|_{\psi_1} \leq \sqrt{2} \|\Sigma_f\|_2^{1/2} \|\Sigma_g\|_2^{1/2}$. Hence, for each pair u_i, v_j we have with probability at least $1 - \delta/9^{m+n}$,

$$\sum_{k=1}^N (u_i^\mathsf{T} f_k)(g_k^\mathsf{T} v_j) \leq 2 \|\Sigma_f\|^{1/2} \|\Sigma_g\|^{1/2} \sqrt{N(m+n)\log(9/\delta)} \,.$$

Taking a union bound over all pairs in the ε-net yields the claim. □

Lemma 3.1.2 shows that if X is $n_1 \times N$ with i.i.d. $\mathcal{N}(0,1)$ entries and Y is $N \times n_2$ with i.i.d. $\mathcal{N}(0,1)$ entries, and X and Y are independent, then with probability at least $1 - \delta$ we have

$$\|XY\| \leq 4\sqrt{N(n_1 + n_2)\log(9/\delta)} \,.$$

Next, we state a standard non-asymptotic bound on the minimum singular value of a standard Wishart matrix (see e.g. Corollary 5.35 of Vershynin [121]).

Lemma 3.1.3. *Let $X \in \mathbb{R}^{N \times n}$ have i.i.d. $\mathcal{N}(0,1)$ entries. With probability at least $1 - \delta$,*

$$\sqrt{\lambda_{\min}(X^\mathsf{T} X)} \geq \sqrt{N} - \sqrt{n} - \sqrt{2\log(1/\delta)} \,.$$

We combine the previous lemmas into a statement on the error of random design regression.

Proposition 3.1.4. *Let $z_1, ..., z_N \in \mathbb{R}^n$ be i.i.d. from $\mathcal{N}(0, \Sigma)$ with Σ invertible. Let $Z^\mathsf{T} := \begin{bmatrix} z_1 & ... & z_N \end{bmatrix}$. Let $W \in \mathbb{R}^{N \times p}$ with each entry i.i.d. $\mathcal{N}(0, \sigma_w^2)$ and independent of Z. Let $E := (Z^\mathsf{T} Z)^\dagger Z^\mathsf{T} W$, and suppose that*

$$N \geq 8n + 16 \log(2/\delta) \,. \tag{3.1.5}$$

For any fixed matrix Q, we have with probability at least $1 - \delta$,

$$\|QE\| \leq 16 \sigma_w \|Q\Sigma^{-1/2}\| \sqrt{\frac{(n+p)\log(18/\delta)}{N}} \,.$$

Proof. First, observe that Z is equal in distribution to $X\Sigma^{1/2}$, where $X \in \mathbb{R}^{N \times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries. By Lemma 3.1.3, with probability at least $1 - \delta/2$,

$$\sqrt{\lambda_{\min}(X^\mathsf{T} X)} \geq \sqrt{N} - \sqrt{n} - \sqrt{2\log(2/\delta)} \geq \sqrt{N}/2 \,.$$

The last inequality uses (3.1.5) combined with the inequality $(a+b)^2 \leq 2(a^2+b^2)$. Furthermore, by Lemma 3.1.2 and (3.1.5), with probability at least $1 - \delta/2$,

$$\|X^\mathsf{T} W\| \leq 4\sigma_w \sqrt{N(n+p)\log(18/\delta)} \,.$$

Let $\mathcal{E}$ denote the event which is the intersection of the two previous events. By a union bound, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. We continue the rest of the proof assuming the event $\mathcal{E}$ holds. Since $X^\mathsf{T} X$ is invertible,

$$QE = Q(Z^\mathsf{T} Z)^\dagger Z^\mathsf{T} W = Q(\Sigma^{1/2} X^\mathsf{T} X \Sigma^{1/2})^\dagger \Sigma^{1/2} X^\mathsf{T} W = Q\Sigma^{-1/2}(X^\mathsf{T} X)^{-1} X^\mathsf{T} W \,.$$

Taking operator norms on both sides,

$$\|QE\| \leq \|Q\Sigma^{-1/2}\| \|(X^\mathsf{T} X)^{-1}\| \|X^\mathsf{T} W\| = \|Q\Sigma^{-1/2}\| \frac{\|X^\mathsf{T} W\|}{\lambda_{\min}(X^\mathsf{T} X)} \,.$$

Combining the inequalities above,

$$\frac{\|X^\mathsf{T} W\|}{\lambda_{\min}(X^\mathsf{T} X)} \leq 16\sigma_w \sqrt{\frac{(n+p)\log(18/\delta)}{N}} \,.$$

The result now follows. □

We now have the necessary tools in place to prove Proposition 3.1.1.

Proof of Proposition 3.1.1. It is not hard to see that

$$\begin{bmatrix} x_T^{(i)} \\ u_T^{(i)} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_u^2 G_T G_T^\mathsf{T} + \sigma_w^2 F_T F_T^\mathsf{T} & 0 \\ 0 & \sigma_u^2 I \end{bmatrix}\right) . \tag{3.1.6}$$

Then applying Proposition 3.1.4 with $Q_A = \begin{bmatrix} I & 0 \end{bmatrix}$ so that $Q_A E$ extracts only the estimate for A, we conclude that with probability at least $1 - \delta/2$,

$$\|\widehat{A} - A\| \le \frac{16\sigma_w}{\sqrt{\lambda_{\min}(\sigma_u^2 G_T G_T^\mathsf{T} + \sigma_w^2 F_T F_T^\mathsf{T})}} \sqrt{\frac{(n+2d)\log(36/\delta)}{N}} , \tag{3.1.7}$$

as long as $N \ge 8(n+d) + 16\log(4/\delta)$. Now applying Proposition 3.1.4 under the same condition on N with $Q_B = \begin{bmatrix} 0 & I \end{bmatrix}$, we have with probability at least $1 - \delta/2$,

$$\|\widehat{B} - B\| \le \frac{16\sigma_w}{\sigma_u} \sqrt{\frac{(n+2d)\log(36/\delta)}{N}} . \tag{3.1.8}$$

The result follows by application of the union bound. □

3.2 Results for Stable Systems

We now present results for the estimation of (A, B) based on the estimator (3.0.3), without having to discard all but the last state transition. In this section, for simplicity we assume that the number of independent rollouts $N = 1$. It is not hard to generalize these results to handle $N > 1$. The key assumption in this section will be that the matrix A is stable, i.e. $\rho(A) \le 1$.

Proposition 3.2.1. *Fix a $\delta \in (0,1)$ and $k \ge 1$. Define the matrices:*

$$\begin{aligned} \Gamma_t &= \begin{bmatrix} \sigma_w^2 \sum_{k=0}^{t-1} (A^k)(A^k)^\mathsf{T} + \sigma_u^2 \sum_{k=0}^{t-2} (A^k) B B^\mathsf{T} (A^k)^\mathsf{T} & 0 \\ 0 & \sigma_u^2 I \end{bmatrix} , \\ \Gamma_{\mathrm{sb}} &= \Gamma_{\lceil k/2 \rceil} , \\ \overline{\Gamma} &= \frac{n+d}{\delta} \Gamma_T . \end{aligned}$$

Then as long as T satisfies:

$$T/k \ge c_0(\log(1/\delta) + (n+d) + \log\det(\overline{\Gamma}\Gamma_{\mathrm{sb}}^{-1})) , \tag{3.2.1}$$

with probability at least $1 - \delta$ we have:

$$\max\{\|\widehat{A} - A\|, \|\widehat{B} - B\|\} \le c_1 \sigma_w \sqrt{\frac{(n+d) + \log\det(\overline{\Gamma}\Gamma_{\mathrm{sb}}^{-1}) + \log(1/\delta)}{T \lambda_{\min}(\Gamma_{\mathrm{sb}})}} . \tag{3.2.2}$$

Here c_0, c_1 are universal constants.

Proof. Fix an $s \geq 0$ and $t \geq 1$. Let $\mathcal{F}_s = \sigma(w_0, ..., w_{s-1}, u_0, ..., u_s)$. Then we have that (x_s, u_s) is $\mathcal{F}_s$-measurable. Therefore,

$$\begin{bmatrix} x_{t+s} \\ u_{t+s} \end{bmatrix} | \mathcal{F}_s \stackrel{d}{=} \mathcal{N}\left(\begin{bmatrix} A^t x_s + A^{t-1} B u_s \\ 0 \end{bmatrix}, \Gamma_t \right) .$$

We can follow the argument of Proposition 3.1 from Simchowitz et al. [105] to conclude that for any unit vector v, then process $Z_t = \left\langle v, \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\rangle$ satisfies the block martingale small ball condition with parameters $(k, \Gamma_{\text{sb}}, p)$ given by $(k, \Gamma_{\lceil k/2 \rceil}, 3/20)$. Next, it is simple to verify with Markov's inequality that:

$$\mathbb{P}\left(\sum_{t=1}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^{\mathsf{T}} \not\preceq T \frac{n+d}{\delta} \Gamma_T \right) \leq \delta .$$

This means we can set $\overline{\Gamma} = \frac{n+d}{\delta} \Gamma_T$. The claim now follows applying Theorem 2.4 of Simchowitz et al. [105]. □

We can simplify Proposition 3.2.1 by choosing $k = 1$ and assuming that A is strictly stable, i.e. $\rho(A) < 1$. We fix a $\gamma \in (\rho(A), 1)$ and set $\tau = \sup\{\|A^k\| \gamma^{-k} : k = 0, 1, ...\}$. Because A is stable, we have $\Gamma_t \preceq \Gamma_\infty$, where Γ_∞ is given by

$$\Gamma_\infty = \begin{bmatrix} \mathsf{dlyap}(A^{\mathsf{T}}, \sigma_w^2 I + \sigma_u^2 B B^{\mathsf{T}}) & 0 \\ 0 & \sigma_u^2 I \end{bmatrix} .$$

Here, $\mathsf{dlyap}(A, M)$ for a stable matrix A and a symmetric matrix M is the solution P to the discrete Lyapunov equation $A^{\mathsf{T}} P A - P + M = 0$. It is also not hard to see that $\lambda_{\min}(\Gamma_{\text{sb}}) \geq \min\{\sigma_w^2, \sigma_u^2\}$. Using the inequality $\log\det(M) \leq n \log(\|M\|)$ for any positive definite M, as long as T satisfies:

$$T \geq c_0 \left(\log(1/\delta) + n + d + n \log\left(\frac{\tau^2}{1-\gamma^2} \left(1 + \frac{\sigma_u^2}{\sigma_w^2} \|B\|^2 \right) \right) \right) ,$$

we have that with probability at least $1 - \delta$:

$$\max\{\|\widehat{A} - A\|, \|\widehat{B} - B\|\} \leq c_1 \sigma_w \sqrt{\frac{n + d + n \log\left(\frac{\tau^2}{1-\gamma^2} \left(1 + \frac{\sigma_u^2}{\sigma_w^2} \|B\|^2 \right) \right) + \log(1/\delta)}{T \min\{\sigma_w^2, \sigma_u^2\}}} . \tag{3.2.3}$$

We note that we present the bound (3.2.3) because it is simple to interpret– sharper bounds are in general possible by exploiting the degree of freedom in choosing k, at the expense of a less interpretable result.

We now compare (3.2.3) to the bounds (3.1.3) and (3.1.4) from Proposition 3.1.1. First, both bounds are of the form $\widetilde{\mathcal{O}}(\sqrt{\frac{n+d}{T}})$, where the $\widetilde{\mathcal{O}}(\cdot)$ hides specific constants depending on

A. However, the analysis of Proposition 3.1.1 is able to separately treat $\|\widehat{A}-A\|$ and $\|\widehat{B}-B\|$ instead of Proposition 3.2.1, which combines the two. This gives (3.1.3) a dependence on the properties of A that is closer to what we expect intuitively compared to (3.2.3). It is an open question of how to decouple the estimation of A and B using the proof technique of Simchowitz et al. [105].

Chapter 4

Basic Robustness and Perturbation Results

In this chapter we cover basic robustness results which will play a fundamental part of the analysis to follow. We first start with a basic definition.

Definition 1. *Let L be a square matrix. Let $\tau \geq 1$ and $\rho \in (0,1)$. We say that L is (τ, ρ) stable if*

$$\|L^k\| \leq \tau \rho^k \ , \ \ k = 0, 1, 2, \ldots \ .$$

While stability of a matrix is an asymptotic notion, Definition 1 quantifies the degree of stability by characterizing the transient response of the powers of a matrix by the parameter τ. It is closely related to the notion of *strong stability* from Cohen et al. [29, 30].

The first question we will study is a fundamental one. Suppose that A is a (τ, ρ) stable matrix, and we perturb A by Δ. Can we find a $\tilde{\tau}, \tilde{\rho}$ such that $A + \Delta$ is $(\tilde{\tau}, \tilde{\rho})$ stable? We note that the question of whether $A + \Delta$ is stable is answered precisely by the structured singular value (SSV) [89]. Here, we will present an answer to our quantitative version of stability that is merely sufficient (as opposed to the exact characterization of SSV).

Proposition 4.0.1. *Let A be a (τ, ρ) stable matrix. Fix a $\gamma \in (\rho, 1)$. Suppose that Δ is a perturbation that satisfies:*

$$\|\Delta\| \leq \frac{\gamma - \rho}{\tau} \ .$$

Then we have that (a) $A + \Delta$ is a stable matrix with $\rho(A + \Delta) \leq \gamma$ and (b) $A + \Delta$ is (τ, γ) stable.

Proof. We start by proving (b). Fix an integer $k \geq 1$. Consider the expansion of $(A + \Delta)^k$ into 2^k terms. Label all these terms as $T_{i,j}$ for $i = 0, ..., k$ and $j = 1, ..., \binom{k}{i}$ where i denotes the degree of Δ in the term (hence there are $\binom{k}{i}$ terms with a degree of i for Δ). Using the

fact that $\|A^k\| \leq \tau\rho^k$ for all $k \geq 0$, we can bound $\|T_{i,j}\| \leq \tau^{i+1}\rho^{k-i}\|\Delta\|^i$. Hence by triangle inequality:

$$\begin{aligned}
\|(A+\Delta)^k\| &\leq \sum_{i=0}^{k}\sum_{j}\|T_{i,j}\| \\
&\leq \sum_{i=0}^{k}\binom{k}{i}\tau^{i+1}\rho^{k-i}\|\Delta\|^i \\
&= \tau\sum_{i=0}^{k}\binom{k}{i}(\tau\|\Delta\|)^i\rho^{k-i} \\
&= \tau(\tau\|\Delta\|+\rho)^k \\
&\leq \tau\gamma^k\,,
\end{aligned}$$

where the last inequality uses the assumption $\|\Delta\| \leq \frac{\gamma-\rho}{\tau}$. This gives the claim (b).

To derive the claim (a), we use the inequality that $\rho(A+\Delta) \leq \|(A+\Delta)^k\|^{1/k} \leq \tau^{1/k}\gamma$ for any $k \geq 1$. Since this holds for any $k \geq 1$, we can take the infimum over all $k \geq 1$ on the RHS, which yields the desired claim. □

Next, we introduce some notation that we will use heavily in the sequel. Let L be a stable matrix and M be a symmetric matrix. We write $P = \mathsf{dlyap}(L, M)$ to denote the unique solution to the discrete Lyapunov equation, i.e.

$$L^{\mathsf{T}}PL - P + M = 0\,.$$

We now present an upper bound on the norm of the discrete Lyapunov solution that uses the notion of (τ, ρ) stability.

Proposition 4.0.2. *Let A be a (τ, ρ) stable matrix, and let $|||\cdot|||$ be either the operator or Frobenius norm. We have that:*

$$|||\mathsf{dlyap}(A, M)||| \leq \frac{\tau^2}{1-\rho^2}|||M|||\,. \tag{4.0.1}$$

Proof. It is a well known fact that we can write $P = \sum_{k=0}^{\infty}(A^k)^{\mathsf{T}}M(A^k)$. Therefore the bound follows from triangle inequality and the (τ, ρ) stability assumption. □

Next, we look at the following question which arises in the context of policy iteration. Suppose that we have two controllers K, K_0 such that their associated value functions V, V_0 satisfy the inequality $V \preceq V_0$. How can we deduce (τ, ρ) stability bounds from V_0 alone?

Proposition 4.0.3. *Let K, K_0 be two stabilizing policies for (A, B). Let V, V_0 denote their respective value functions and suppose that $V \preceq V_0$. We have that for all $k \geq 0$:*

$$\|(A+BK)^k\| \leq \sqrt{\frac{\lambda_{\max}(V_0)}{\lambda_{\min}(S)}}(1-\lambda_{\min}(V_0^{-1}S))^{k/2}\,.$$

Proof. This proof is inspired by the proof of Lemma 5.1 of Abbasi-Yadkori et al. [4]. Since V is the value function for K, we have:

$$\begin{aligned} V &= (A+BK)^\mathsf{T} V (A+BK) + S + K^\mathsf{T} R K \\ &\succeq (A+BK)^\mathsf{T} V (A+BK) + S \,. \end{aligned}$$

Conjugating both sides by $V^{-1/2}$ and defining $H := V^{1/2}(A+BK)V^{-1/2}$,

$$\begin{aligned} I &\succeq V^{-1/2}(A+BK)^\mathsf{T} V (A+BK) V^{-1/2} + V^{-1/2} S V^{-1/2} \\ &= H^\mathsf{T} H + V^{-1/2} S V^{-1/2} \,. \end{aligned}$$

This implies that $\|H\|^2 = \|H^\mathsf{T} H\| \leq \|I - V^{-1/2} S V^{-1/2}\| = 1 - \lambda_{\min}(S^{1/2} V^{-1} S^{1/2}) \leq 1 - \lambda_{\min}(S^{1/2} V_0^{-1} S^{1/2})$. The last inequality holds since $V \preceq V_0$ iff $V^{-1} \succeq V_0^{-1}$, Now observe:

$$\|V^{1/2}(A+BK)^k V^{-1/2}\| = \|H^k\| \leq \|H\|^k \leq (1 - \lambda_{\min}(V_0^{-1} S))^{k/2}$$

Next, for M positive definite and N square, observe that:

$$\begin{aligned} \|MNM^{-1}\| &= \sqrt{\lambda_{\max}(MNM^{-2}N^\mathsf{T} M)} \\ &\geq \sqrt{\lambda_{\min}(M^{-2})\lambda_{\max}(MNN^\mathsf{T} M)} \\ &= \sqrt{\lambda_{\min}(M^{-2})\lambda_{\max}(N^\mathsf{T} M^2 N)} \\ &\geq \sqrt{\lambda_{\min}(M^{-2})\lambda_{\min}(M^2)\|N\|^2} \\ &= \frac{\|N\|}{\kappa(M)} \,. \end{aligned}$$

Therefore, we have shown that:

$$\|(A+BK)^k\| \leq \sqrt{\kappa(V)}(1 - \lambda_{\min}(V_0^{-1} S))^{k/2} \leq \sqrt{\frac{\lambda_{\max}(V_0)}{\lambda_{\min}(S)}}(1 - \lambda_{\min}(V_0^{-1} S))^{k/2} \,.$$

□

Next, we look at perturbations to discrete Lyapunov equations. A similar result to the following proposition can be found in Gahinet et al. [43].

Proposition 4.0.4. *Suppose that A_1, A_2 are stable matrices. Suppose furthermore that A_i is (τ, ρ) stable for $i = 1, 2$. Let Q_1, Q_2 be PSD matrices. Put $P_i = \mathsf{dlyap}(A_i, Q_i)$. We have that:*

$$\|P_1 - P_2\| \leq \frac{\tau^2}{1-\rho^2}\|Q_1 - Q_2\| + \frac{\tau^4}{(1-\rho^2)^2}\|A_1 - A_2\|(\|A_1\| + \|A_2\|)\|Q_2\| \,.$$

Proof. Let the linear operators F_1, F_2 be such that $P_i = F_i^{-1}(Q_i)$, i.e. $F_i(X) = X - A_i^\mathsf{T} X A_i$. Then:

$$\begin{aligned} P_1 - P_2 &= F_1^{-1}(Q_1) - F_2^{-1}(Q_2) \\ &= F_1^{-1}(Q_1 - Q_2) + F_1^{-1}(Q_2) - F_2^{-1}(Q_2) \\ &= F_1^{-1}(Q_1 - Q_2) + (F_1^{-1} - F_2^{-1})(Q_2) \,. \end{aligned}$$

Hence $\|P_1 - P_2\| \leq \|F_1^{-1}\|\|Q_1 - Q_2\| + \|F_1^{-1} - F_2^{-1}\|\|Q_2\|$. Now for any M satisfying $\|M\| \leq 1$

$$\|F_i^{-1}(M)\| = \left\| \sum_{k=0}^{\infty} (A_i^\mathsf{T})^k M A_i^k \right\| \leq \frac{\tau^2}{1-\rho^2} \,.$$

Next, we have that:

$$\|F_1^{-1} - F_2^{-1}\| = \|F_1^{-1}(F_2 - F_1)F_2^{-1}\| \leq \|F_1^{-1}\|\|F_2^{-1}\|\|F_1 - F_2\| \leq \frac{\tau^4}{(1-\rho^2)^2}\|F_1 - F_2\| \,.$$

Now for any M satisfying $\|M\| \leq 1$,

$$\begin{aligned} \|F_1(M) - F_2(M)\| &= \|A_2^\mathsf{T} M A_2 - A_1^\mathsf{T} M A_1\| \\ &= \|(A_2 - A_1)^\mathsf{T} M A_2 + A_1^\mathsf{T} M (A_2 - A_1)\| \\ &\leq \|A_1 - A_2\|(\|A_1\| + \|A_2\|) \,. \end{aligned}$$

The claim now follows. □

We now turn to the following question. Suppose that $\widehat{K}$ is a controller that stabilizes (A, B), but is not necessarily optimal for the LQR problem with parameters (A, B, S, R). On the other hand, suppose that K is the unique optimal LQR controller. The following lemma shows how to relate the cost sub-optimality gap $J(\widehat{K}) - J_\star$ to the error $\|\widehat{K} - K\|_F$.

Lemma 4.0.5 (Lemma 12, Fazel et al. [41])**.** *Let $\widehat{K}$ stabilize (A, B) and let K denote the optimal LQR controller for (A, B, S, R). Let $\Sigma(\widehat{K}) = \mathsf{dlyap}((A + B\widehat{K})^\mathsf{T}, W)$ denote the stationary covariance matrix of (A, B) in feedback with $\widehat{K}$, and let $V = \mathsf{dare}(A, B, S, R)$. We have that:*

$$J(\widehat{K}) - J_\star = \mathrm{tr}(\Sigma(\widehat{K})(\widehat{K} - K)^\mathsf{T}(R + B^\mathsf{T} V B)(\widehat{K} - K)) \,. \tag{4.0.2}$$

Proof. We give a slightly more direct proof than Fazel et al. [41]. Let $V(K) = \mathsf{dlyap}(L(K), S + K^\mathsf{T} R K)$ and let $\Sigma(K) = \mathsf{dlyap}(L(K)^\mathsf{T}, W)$ with $L(K) = A + BK$. We abbreviate $V = V(K)$ and $\widehat{V} = V(\widehat{K})$, and similarly with $\Sigma, \widehat{\Sigma}$ and $L, \widehat{L}$. Let $\Delta = \widehat{K} - K$. We have that:

$$\begin{aligned} J(\widehat{K}) - J_\star &= \mathrm{tr}(W\widehat{V}) - \mathrm{tr}(WV) \\ &= \mathrm{tr}(\widehat{\Sigma}(S + \widehat{K}^\mathsf{T} R \widehat{K})) - \mathrm{tr}((\widehat{\Sigma} - \widehat{L}\widehat{\Sigma}\widehat{L}^\mathsf{T})V) \\ &= \mathrm{tr}(\widehat{\Sigma}(S + \widehat{K}^\mathsf{T} R \widehat{K} + \widehat{L}^\mathsf{T} V \widehat{L} - V)) \,. \end{aligned}$$

Now we expand out:

$$\begin{aligned}
&S + \widehat{K}^\mathsf{T} R \widehat{K} + \widehat{L}^\mathsf{T} V \widehat{L} - V \\
&= S + \Delta^\mathsf{T} R \Delta + \Delta^\mathsf{T} R K + K^\mathsf{T} R \Delta + K^\mathsf{T} R K \\
&\quad + L^\mathsf{T} V L + L^\mathsf{T} V B \Delta + \Delta^\mathsf{T} B^\mathsf{T} V L + \Delta^\mathsf{T} B^\mathsf{T} V B \Delta - V \\
&= (S + K^\mathsf{T} R K + L^\mathsf{T} V L - V) + \Delta^\mathsf{T} R K + K^\mathsf{T} R \Delta + L^\mathsf{T} V B \Delta \\
&\quad + \Delta^\mathsf{T} B^\mathsf{T} V L + \Delta^\mathsf{T} (R + B^\mathsf{T} V B) \Delta \\
&= \Delta^\mathsf{T} (R K + B^\mathsf{T} V L) + (K^\mathsf{T} R + L^\mathsf{T} V B) \Delta + \Delta^\mathsf{T} (R + B^\mathsf{T} V B) \Delta \\
&= \Delta^\mathsf{T} (R + B^\mathsf{T} V B) \Delta \,.
\end{aligned}$$

The last equality holds because $RK + B^\mathsf{T} V L = (R + B^\mathsf{T} V B) K + B^\mathsf{T} V A = 0$ by the optimality of K. The claim now follows. □

We conclude this section with a simple perturbation result for the minimizers of strongly convex functions.

Proposition 4.0.6. *Let f_1, f_2 be two μ-strongly convex twice differentiable functions. Let $x_1 = \arg\min_x f_1(x)$ and $x_2 = \arg\min_x f_2(x)$. Suppose $\|\nabla f_1(x_2)\| \le \varepsilon$. Then $\|x_1 - x_2\| \le \frac{\varepsilon}{\mu}$.*

Proof. Taylor expanding ∇f_1, we have:

$$\nabla f_1(x_2) = \nabla f_1(x_1) + \nabla^2 f_1(\tilde{x})(x_2 - x_1) = \nabla^2 f_1(\tilde{x})(x_2 - x_1) \,.$$

for $\tilde{x} = t x_1 + (1-t) x_2$ with some $t \in [0, 1]$. Therefore:

$$\mu \|x_1 - x_2\| \le \|\nabla^2 f_1(\tilde{x})(x_2 - x_1)\| = \|\nabla f_1(x_2)\| \le \varepsilon \,.$$

□

Proposition 4.0.6 gives us a way to bound the difference between the resulting greedily induced controllers of two different Q-functions for LQR.

Proposition 4.0.7. *Let $M \succeq \mu I$ and $N \succeq \mu I$ be a positive definite matrices partitioned as $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22} \end{bmatrix}$ and similarly for N. Let $T(M) = -M_{22}^{-1} M_{12}^\mathsf{T}$. We have that:*

$$\|T(M) - T(N)\| \le \frac{(1 + \|T(N)\|)\|M - N\|}{\mu} \,.$$

Proof. Fix a unit norm x. Define $f(u) = (1/2) x^\mathsf{T} M_{11} x + (1/2) u^\mathsf{T} M_{22} u + x^\mathsf{T} M_{12} u$ and $g(u) = (1/2) x^\mathsf{T} N_{11} x + (1/2) u^\mathsf{T} N_{22} u + x^\mathsf{T} N_{12} u$. Let $u_\star = T(N) x$. We have that

$$\nabla f(u_\star) = \nabla f(u_\star) - \nabla g(u_\star) = (M_{22} - N_{22}) u_\star + (M_{12} - N_{12})^\mathsf{T} x \,.$$

Hence, $\|\nabla f(u_\star)\| \le \|M_{12} - N_{12}\| + \|M_{22} - N_{22}\| \|u_\star\|$. We can bound $\|u_\star\| = \|T(N) x\| \le \|T(N)\|$. The claim now follows using Proposition 4.0.6. □

Chapter 5

Model-based Methods for LQR

In this chapter, we present results on the performance bounds for model-based methods on LQR. The results in this chapter are based on Mania et al. [75] and Dean et al. [31]. We first present results for certainty equivalent control (also referred to nominal control), followed by results for a new synthesis method based on robust control.

We recall the basic problem setting. We assume (A, B) is a stabilizable system, and S, R are positive definite matrices. Our goal is to solve for an optimal controller for the infinite horizon LQR problem (1.1.2). We assume we are given estimates $\widehat{A}, \widehat{B}$ that satisfy the bounds $\|\widehat{A} - A\| \le \varepsilon_A$ and $\|\widehat{B} - B\| \le \varepsilon_B$. These bounds can, for instance, be probabilistic in nature, coming from the concentration inequalities of Chapter 3. As a reminder, we are interested in designing algorithms which take in as input $(\widehat{A}, \widehat{B}, \varepsilon_A, \varepsilon_B, S, R)$ and output a controller $\widehat{K}$ such that $J(\widehat{K}) - J_\star$ is well controlled.

5.1 Certainty Equivalence Control

Recall that the certainty equivalence controller discards the error bounds $\varepsilon_A, \varepsilon_B$ and outputs:

$$\widehat{K} = -(\widehat{B}^\mathsf{T}\widehat{V}\widehat{B} + R)^{-1}\widehat{B}^\mathsf{T}\widehat{V}\widehat{A}\,, \tag{5.1.1}$$

$$\widehat{V} = \mathsf{dare}(\widehat{A}, \widehat{B}, S, R)\,. \tag{5.1.2}$$

Here, $\widehat{V}$ is the unique positive definite solution to the discrete algebraic Riccati equation:

$$\widehat{V} = \widehat{A}^\mathsf{T}\widehat{V}\widehat{A} - \widehat{A}^\mathsf{T}\widehat{V}\widehat{B}(\widehat{B}^\mathsf{T}\widehat{V}\widehat{B} + R)^{-1}\widehat{V}^\mathsf{T}\widehat{B}\widehat{A} + S\,.$$

While this control scheme is simple and intuitive, it raises many questions. The first is, how do we ensure that the solution to (5.1.2) exists? We know that $\mathsf{dare}(A, B, S, R)$ must exist by the stabilizable assumption of (A, B), but how small do the errors $\varepsilon_A, \varepsilon_B$ need to be so that this transfers over to $\mathsf{dare}(\widehat{A}, \widehat{B}, S, R)$? The second question is, once we ensure that $\widehat{V}$ and $\widehat{K}$ are well-defined, how do we quantify the sub-optimality incurred with $J(\widehat{K}) - J_\star$?

5.1.1 A Meta Theorem

We first focus on the question of quantifying the sub-optimality incurred in terms of the distance of $\widehat{V} = \mathsf{dare}(\widehat{A}, \widehat{B}, S, R)$ to $V = \mathsf{dare}(A, B, S, R)$. We define the constant $\Gamma := 1 + \max\{\|A\|, \|B\|, \|V\|, \|K\|\}$.

Theorem 5.1.1. *Suppose that $d \le n$. Let $L := A + BK$ be the optimal closed loop matrix, and suppose that L is (τ, ρ) stable. Also, let $\varepsilon > 0$ such that $\|\widehat{A} - A\| \le \varepsilon$ and $\|\widehat{B} - B\| \le \varepsilon$ and assume $\|\widehat{V} - V\| \le f(\varepsilon)$ for some function f such that $f(\varepsilon) \ge \varepsilon$. Then, if (S, R) are both positive definite with $\lambda_{\min}(R) \ge 1$, the certainty equivalent controller $u_t = \widehat{K} x_t$ satisfies the sub-optimality gap*

$$J(\widehat{K}) - J_\star \le 200\, \sigma_w^2\, d\, \Gamma^9\, \frac{\tau^2}{1-\rho^2} f(\varepsilon)^2, \tag{5.1.3}$$

as long as $f(\varepsilon)$ is small enough so that the right hand side is smaller than σ_w^2.

The proof of Theorem 5.1.1 builds on tools from Chapter 4. We first relate the difference of $\|\widehat{K} - K\|$ to the difference of $\|\widehat{V} - V\|$.

Proposition 5.1.2. *Define $f_i(u; x) = \frac{1}{2} u^\mathsf{T} R u + \frac{1}{2}(A_i x + B_i u)^\mathsf{T} V_i (A_i x + B_i u)$ for $i = 1, 2$, with R, V_1, and V_2 positive definite matrices. Let K_i be the unique matrix such that $u_i := \arg\min_u f_i(u; x) = K_i x$ for any vector x. Also, denote $\Gamma_1 := 1 + \max\{\|A_1\|, \|B_1\|, \|V_1\|, \|K_1\|\}$. Suppose there exists ε such that $0 \le \varepsilon < 1$ and $\|A_1 - A_2\| \le \varepsilon$, $\|B_1 - B_2\| \le \varepsilon$, and $\|V_1 - V_2\| \le \varepsilon$. Then, we have*

$$\|K_1 - K_2\| \le \frac{7\varepsilon \Gamma_1^3}{\sigma_{\min}(R)}\,.$$

Proof. We first compute the gradient $\nabla f_i(u; x)$ with respect to u:

$$\nabla f_i(u; x) = (B_i^\mathsf{T} V_i B_i + R) u + B_i^\mathsf{T} V_i A_i x\,.$$

Now, we observe that:

$$\|B_1^\mathsf{T} V_1 B_1 - B_2^\mathsf{T} V_2 B_2\| \le 7\Gamma_1^2 \varepsilon \quad \text{and} \quad \|B_1^\mathsf{T} V_1 A_1 - B_2^\mathsf{T} V_2 A_2\| = 7\Gamma_1^2 \varepsilon.$$

Hence, for any vector x with $\|x\| \le 1$, we have

$$\|\nabla f_1(u; x) - \nabla f_2(u; x)\| \le 7\Gamma_1^2 \varepsilon (\|u\| + 1).$$

We can bound $\|u_1\| \le \|K_1\| \|x\| \le \|K_1\|$. Then, from Proposition 4.0.6 we obtain

$$\sigma_{\min}(R) \|(K_1 - K_2) x\| = \sigma_{\min}(R) \|u_1 - u_2\| \le 7\Gamma_1^3 \varepsilon.$$

□

The previous proposition allows us to upper bound $\|\widehat{K} - K\|$.

Proposition 5.1.3. *Let $\varepsilon > 0$ such that $\|\widehat{A} - A\| \le \varepsilon$ and $\|\widehat{B} - B\| \le \varepsilon$. Also, let $\|\widehat{V} - V\| \le f(\varepsilon)$ for some function f such that $f(\varepsilon) \ge \varepsilon$. Then, if $\lambda_{\min}(R) \ge 1$,*

$$\|\widehat{K} - K\| \le 7\Gamma^3 f(\varepsilon). \tag{5.1.4}$$

Now suppose that $L = A + BK$ is (τ, ρ) stable. Then, if $f(\varepsilon)$ is small enough so that the right hand side of (5.1.4) is smaller than $\frac{1-\gamma}{2\tau}$, we have $A + B\widehat{K}$ is $(\tau, (1+\gamma)/2)$ stable.

Proof. By our assumptions $\|\widehat{A} - A\|$, $\|\widehat{B} - B\|$, and $\|\widehat{V} - V\|$ are smaller than $f(\varepsilon)$, and $\sigma_{\min}(R) \ge 1$. Then, Proposition 5.1.2 ensures that

$$\|\widehat{K} - K\| \le 7\Gamma^3 f(\varepsilon).$$

Finally, when ε is small enough so that the right hand side of (5.1.4) is smaller or equal than $\frac{1-\rho}{2\tau}$, we can apply the perturbation result of Proposition 4.0.1 to guarantee that $\|(A + B\widehat{K})^k\| \le \tau \left(\frac{1+\rho}{2}\right)^k$ for all $k \ge 0$. □

Now, we have the necessary ingredients to complete the proof of Theorem 5.1.1.

Proof of Theorem 5.1.1. The second order perturbation result of Lemma 4.0.5 implies:

$$J(\widehat{K}) - J_\star \le \|\Sigma(\widehat{K})\|\|R + B^\mathsf{T} V B\|\|\widehat{K} - K\|_F^2 .$$

Proposition 5.1.3 states that $\widehat{K}$ stabilizes the system (A, B) when the estimation error is small enough. More precisely, under the assumptions of Theorem 5.1.1, we have $\widehat{L}$ is $(\tau, (1+\rho)/2)$ stable with $\widehat{L} = A + B\widehat{K}$. Therefore by Proposition 4.0.2,

$$\|\Sigma(\widehat{K})\| \le \frac{\sigma_w^2 \tau^2}{1 - \left(\frac{\rho+1}{2}\right)^2} \le \frac{4\sigma_w^2 \tau^2}{1 - \rho^2} .$$

Recalling that $\Gamma = 1 + \max\{\|A\|, \|B\|, \|V\|, \|K\|\}$, we have $\|R + B^\mathsf{T} V B\| \le \Gamma^3$. Then,

$$\begin{aligned} J(\widehat{K}) - J_\star &\le 4\sigma_w^2 \Gamma^3 \frac{\tau^2}{1-\rho^2} \|\widehat{K} - K\|_F^2 \\ &\le 4\sigma_w^2 \min\{n, d\} \Gamma^3 \frac{\tau^2}{1-\rho^2} \|\widehat{K} - K\|^2 \\ &\le 200 \sigma_w^2 d \Gamma^9 \frac{\tau^2}{1-\rho^2} f(\varepsilon)^2 . \end{aligned}$$

□

5.1.2 Riccati Perturbation

Theorem 5.1.1 reduces the problem of bounding $J(\widehat{K}) - J_\star$ to the problem of bounding the error $\|\widehat{V} - V\|$ of the solutions to the Riccati equation. In particular, if we can show that $\|\widehat{V} - V\| \leq L\varepsilon$ if $\varepsilon < b$, for some b and L, then Theorem 5.1.1 implies that $J(\widehat{K}) - J_\star = \mathcal{O}(\varepsilon^2)$. In other words, it suffices to show that the solutions to the discrete Riccati equation are locally Lipschitz with respect to the problem parameters.

However, we note that one cannot hope to find universal values b and L such that for any $0 < \varepsilon < b$ one has $\|\widehat{V} - V\| \leq L\varepsilon$ for arbitrary (A, B) and $(\widehat{A}, \widehat{B})$ with $\|\widehat{A} - A\| \leq \varepsilon$ and $\|\widehat{B} - B\| \leq \varepsilon$. To see this, consider the one dimensional linear system ($n = 1$) given by $A = 1$ and $B = \varepsilon$ and consider the estimated system $\widehat{A} = 1$ and $\widehat{B} = 0$. Then, the estimated system is ε close to the optimal system, but the estimated system is not stabilizable and hence $\widehat{V}$ is not finite. Even when $\widehat{B} = \varepsilon/2$, there is no universal L such that the desired inequality holds for all positive ε. Therefore, b and L must depend on the system parameters (A, B).

While there is a long line of work analyzing perturbations of Riccati equations, we are not aware of any result that offers explicit and easily interpretable b and L for a fixed system (A, B). See the book by Konstantinov et al. [58] for an overview of this literature. We present two new results for Riccati perturbation which offer interpretable bounds. The first one expands upon the operator-theoretic proof of Konstantinov et al. [57], and the second one is based on a new elementary approach. The proofs for these results are omitted here, and can be found in Mania et al. [75].

Proposition 5.1.4. *Let $L := A + BK$ be the optimal closed-loop matrix and suppose that L is (τ, ρ) stable. Let ε be such that $\|\widehat{A} - A\| \leq \varepsilon$ and $\|\widehat{B} - B\| \leq \varepsilon$. Put $\|\cdot\|_+ := \|\cdot\| + 1$. Suppose that $\lambda_{\min}(S) \geq 1$, $\lambda_{\min}(R) \geq 1$ and the system (A, B) is stabilizable. We have*

$$\|\widehat{V} - V\| \leq \mathcal{O}(1)\,\varepsilon\,\frac{\tau^2}{1-\rho^2}\|A\|_+^2\|V\|_+^2\|B\|_+\|R^{-1}\|_+,$$

as long as

$$\varepsilon \leq \mathcal{O}(1)\frac{(1-\rho^2)^2}{\tau^4}\|A\|_+^{-2}\|V\|_+^{-2}\|B\|_+^{-3}\|R^{-1}\|_+^{-2}\min\left\{\|L\|_+^{-2}, \|V\|_+^{-1}\right\}.$$

Before we present a new direct bound, we need a new definition. Recall that a linear system (A, B) is called *controllable* when the *controllability matrix*

$$\begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix}$$

has full row rank. Controllability is a fundamental concept in control theory; it states that there exists a sequence of inputs to the system (A, B) that moves it from any starting state to any final state in at most n steps. We now introduce a definition that quantifies how controllable a linear system is. We denote, for any integer $\ell \geq 1$, the matrix $\mathcal{C}_\ell :=$

$\begin{bmatrix} B & AB & \dots & A^{\ell-1}B \end{bmatrix}$ and call the system (ℓ, ν)-*controllable* if the n-th singular value of $\mathcal{C}_\ell$ is greater or equal than ν, i.e. $\sigma_{\min}(\mathcal{C}_\ell) = \sqrt{\lambda_{\min}\left(\mathcal{C}_\ell \mathcal{C}_\ell^\top\right)} \geq \nu$. Intuitively, the larger ν is, the less control effort is needed to move the system between two different states.

We now present a new direct approach, which uses (ℓ, ν)-controllability to give a bound which is sharper for some systems (A, B) then the one provided by Proposition 5.1.4. Recall that any controllable system is always (ℓ, ν)-controllable for some ℓ and ν. For any square matrix M, we define

$$\tau(M, \rho) := \sup\left\{\|M^k\|\rho^{-k} \colon k \geq 0\right\} . \tag{5.1.5}$$

Note that if $\rho \geq \rho(M)$, then $\tau(M, \rho)$ is guaranteed to be finite (this is a consequence of Gelfand's formula). This holds even if $\rho(M) \geq 1$.

Proposition 5.1.5. *Suppose that (A, B) is (ℓ, ν)-controllable. Let $\rho \geq \rho(A)$ and also let $\varepsilon \geq 0$ such that $\|\widehat{A} - A\| \leq \varepsilon$ and $\|\widehat{B} - B\| \leq \varepsilon$. Let $\beta := \max\{1, \varepsilon\tau(A, \rho) + \rho\}$. Suppose that $\lambda_{\min}(S) \geq 1$ and $\lambda_{\min}(R) \geq 1$. We have that:*

$$\|\widehat{V} - V\| \leq 32\, \varepsilon\, \ell^{\frac{5}{2}} \tau(A, \rho)^3 \beta^{2(\ell-1)} \left(1 + \frac{1}{\nu}\right) (1 + \|B\|)^2 \|V\| \frac{\max\{\|S\|, \|R\|\}}{\min\{\lambda_{\min}(S), \lambda_{\min}(R)\}},$$

as long as ε is small enough so that the right hand side is smaller or equal than one.

Proposition 5.1.5 requires an (ℓ, ν)-controllable system (A, B), whereas Proposition 5.1.4 only requires a stabilizable system, which is a milder assumption. However, Proposition 5.1.5 can offer a sharper guarantee. For example, consider the linear system with two dimensional states ($n = 2$) given by $A = 1.01 \cdot I_2$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix}$. The cost functions S and R are chosen to be the identity matrix I_2. This system (A, B) is readily checked to be $(1, \beta)$-controllable. It is also straightforward to verify that as β tends to zero, Proposition 5.1.4 gives a bound of $\|\widehat{V} - V\| = \mathcal{O}(\varepsilon/\beta^4)$, whereas Proposition 5.1.5 gives a sharper bound of $\|\widehat{V} - V\| = \mathcal{O}(\varepsilon/\beta^3)$.

5.1.3 Putting it Together

We now combine the meta theorem Theorem 5.1.1 with the Riccati perturbation result Proposition 5.1.5 and obtain the main sub-optimality bound for certainty equivalence control.

Theorem 5.1.6. *Suppose that $d \leq n$. Let ρ and γ be two real values such that $\rho(A) \leq \rho$ and $\rho(L) \leq \gamma < 1$, where $L := A + BK$ is the optimal closed-loop matrix. Also, let $\varepsilon > 0$ such that $\|\widehat{A} - A\| \leq \varepsilon$ and $\|\widehat{B} - B\| \leq \varepsilon$ and define $\beta := \max\{1, \varepsilon\tau(A, \rho) + \rho\}$. Suppose that $\lambda_{\min}(S) \geq 1$ and $\lambda_{\min}(R) \geq 1$. Suppose also that (A, B) is (ℓ, ν)-controllable. Then, the certainty equivalent controller $u_t = \widehat{K} x_t$ satisfies the sub-optimality gap*

$$J(\widehat{K}) - J_\star \leq \mathcal{O}(1)\, \sigma_w^2\, d\, \ell^5\, \Gamma^{15}\, \tau(A, \rho)^6 \beta^{4(\ell-1)}\, \frac{\tau(L, \gamma)^2}{1 - \gamma^2} \frac{\max\{\|S\|^2, \|R\|^2\}}{\min\left\{\lambda_{\min}(S)^2, \lambda_{\min}(R)^2\right\}} \left(1 + \frac{1}{\nu}\right)^2 \varepsilon^2 , \tag{5.1.6}$$

as long as ε is small enough so that the right hand side is smaller than σ_w^2. Here, $\mathcal{O}(1)$ denotes a universal constant.

The exact form of Equation 5.1.6, such as the polynomial dependence on ℓ, Γ, etc, can be improved at the expense of conciseness of the expression. In our proof we optimized for the latter. The factor $\max\{\|S\|^2, \|R\|^2\}/\min\{\lambda_{\min}(S)^2, \lambda_{\min}(R)^2\}$ is the squared condition number of the cost function, a natural quantity in the context of the optimization problem (1.1.2), which can be seen as an infinite dimensional quadratic program with a linear constraint. The term $\frac{\tau(L,\gamma)^2}{1-\gamma^2}$ quantifies the rate at which the optimal controller drives the state towards zero. Generally speaking, the less stable the optimal closed loop system is, the larger this term becomes.

An interesting trade-off arises between the factor $\ell^5\beta^{4(\ell-1)}$ (which arises from upper bounding perturbations of powers of A on a time interval of length ℓ) and the factor ν (the lower bound on $\sigma_{\min}(\mathcal{C}_\ell)$), which is increasing in ℓ. Hence, the parameter ℓ should be seen as a free-parameter that can be tuned to minimize the right hand side of (5.1.6). Now, we specialize Theorem 5.1.6 to a few cases.

Case: A is contractive, i.e. $\|A\| < 1$. In this case, we can choose $\rho = \|A\|$ and ε small enough so that $\varepsilon \le 1 - \|A\|$. Then, (5.1.6) simplifies to:

$$J(\widehat{K}) - J_\star \le \mathcal{O}(1)\, d\, \sigma_w^2\, \ell^5\, \Gamma^{15} \frac{\tau(L,\gamma)^2}{1-\gamma^2} \frac{\max\{\|S\|^2, \|R\|^2\}}{\min\{\lambda_{\min}(S)^2, \lambda_{\min}(R)^2\}} \left(1 + \frac{1}{\nu}\right)^2 \varepsilon^2 .$$

Case: B has rank n. In this case, we can choose $\ell = 1$. Then, (5.1.6) simplifies to:

$$J(\widehat{K}) - J_\star \le \mathcal{O}(1)\, d\, \sigma_w^2\, \Gamma^{15} \tau(A,\rho)^6 \frac{\tau(L,\gamma)^2}{1-\gamma^2} \frac{\max\{\|S\|^2, \|R\|^2\}}{\min\{\lambda_{\min}(S)^2, \lambda_{\min}(R)^2\}} \left(1 + \frac{1}{\nu}\right)^2 \varepsilon^2 .$$

5.2 Robust Control

In this section we discuss our approach based on *System Level Synthesis* (SLS), a recently developed approach to control design that relies on a particular parameterization of signals in a control system [78, 124]. We review the main SLS framework, highlighting the key constructions that we will use to solve the robust LQR problem. As we show in this and the following section, using the SLS framework, as opposed to traditional techniques from robust control, allows us to (a) compute robust controllers using semidefinite programming, and (b) provide sub-optimality guarantees in terms of the size of the uncertainties on our system estimates.

5.2.1 Useful Results from System Level Synthesis

The SLS framework focuses on the *system responses* of a closed-loop system. As a motivating example, consider linear dynamics under a fixed a static state-feedback control policy K,

i.e., let $u_k = Kx_k$. Then, the closed loop map from the disturbance process $\{w_0, w_1, \dots\}$ to the state x_k and control input u_k at time k is given by

$$\begin{array}{rcl} x_k & = & \sum_{t=1}^{k} (A+BK)^{k-t} w_{t-1} \,, \\ u_k & = & \sum_{t=1}^{k} K(A+BK)^{k-t} w_{t-1} \,. \end{array} \tag{5.2.1}$$

Letting $\Phi_x(k) := (A+BK)^{k-1}$ and $\Phi_u(k) := K(A+BK)^{k-1}$, we can rewrite (5.2.1) as

$$\begin{bmatrix} x_k \\ u_k \end{bmatrix} = \sum_{t=1}^{k} \begin{bmatrix} \Phi_x(k-t+1) \\ \Phi_u(k-t+1) \end{bmatrix} w_{t-1} \,, \tag{5.2.2}$$

where $\{\Phi_x(k), \Phi_u(k)\}$ are called the *closed-loop system response elements* induced by the static controller K.

Note that even when the control is a linear function of the state and its past history (i.e. a linear dynamic controller), the expression (5.2.2) is valid. Though we conventionally think of the control policy as a function mapping states to input, whenever such a mapping is linear, both the control input and the state can be written as linear functions of the disturbance signal w_t. With such an identification, the dynamics require that the $\{\Phi_x(k), \Phi_u(k)\}$ must obey the constraints

$$\Phi_x(k+1) = A\Phi_x(k) + B\Phi_u(k) \,, \;\; \Phi_x(1) = I \,, \;\; \forall k \geq 1 \,, \tag{5.2.3}$$

As we describe in more detail below in Theorem 5.2.1, these constraints are in fact both necessary and sufficient. Working with closed-loop system responses allows us to cast optimal control problems as optimization problems over elements $\{\Phi_x(k), \Phi_u(k)\}$, constrained to satisfy the affine equations (5.2.3). Comparing equations (5.2.1) and (5.2.2), we see that the former is non-convex in the controller K, whereas the latter is affine in the elements $\{\Phi_x(k), \Phi_u(k)\}$.

As we work with infinite horizon problems, it is notationally more convenient to work with *transfer function* representations of the above objects, which can be obtained by taking a z-transform of their time-domain representations. The frequency domain variable z can be informally thought of as the time-shift operator, i.e., $z\{x_k, x_{k+1}, \dots\} = \{x_{k+1}, x_{k+2}, \dots\}$, allowing for a compact representation of LTI dynamics. We use boldface letters to denote such transfer functions signals in the frequency domain, e.g., $\mathbf{\Phi}_x(z) = \sum_{k=1}^{\infty} \Phi_x(k) z^{-k}$. Then, the constraints (5.2.3) can be rewritten as

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I \,,$$

and the corresponding (not necessarily static) control law $\mathbf{u} = \mathbf{K}\mathbf{x}$ is given by $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$. The relevant frequency domain connections for LQR are illustrated in Section 5.2.6.

We formalize our discussion by introducing notation that is common in the controls literature. For a thorough introduction to the functional analysis commonly used in control

theory, see Chapters 2 and 3 of Zhou et al. [129]. Let $\mathbb{T}$ (resp. $\mathbb{D}$) denote the unit circle (resp. open unit disk) in the complex plane. The restriction of the Hardy spaces $\mathcal{H}_\infty(\mathbb{T})$ and $\mathcal{H}_2(\mathbb{T})$ to matrix-valued real-rational functions that are analytic on the complement of $\mathbb{D}$ will be referred to as $\mathcal{RH}_\infty$ and $\mathcal{RH}_2$, respectively. In controls parlance, this corresponds to (discrete-time) stable matrix-valued transfer functions. For these two function spaces, the $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norms simplify to

$$\|\mathbf{G}\|_{\mathcal{H}_\infty} = \sup_{z\in\mathbb{T}} \|G(z)\|_2\,, \quad \|\mathbf{G}\|_{\mathcal{H}_2} = \sqrt{\frac{1}{2\pi}\int_{\mathbb{T}} \|G(z)\|_F^2\, dz}\,. \tag{5.2.4}$$

Finally, the notation $\frac{1}{z}\mathcal{RH}_\infty$ refers to the set of transfer functions $\mathbf{G}$ such that $z\mathbf{G} \in \mathcal{RH}_\infty$. Equivalently, $\mathbf{G} \in \frac{1}{z}\mathcal{RH}_\infty$ if $\mathbf{G} \in \mathcal{RH}_\infty$ and $\mathbf{G}$ is strictly proper.

The most important transfer function for the LQR problem is the map from the state sequence to the control actions: the control policy. Consider an arbitrary transfer function $\mathbf{K}$ denoting the map from state to control action, $\mathbf{u} = \mathbf{K}\mathbf{x}$. Then the closed-loop transfer matrices from the process noise $\mathbf{w}$ to the state $\mathbf{x}$ and control action $\mathbf{u}$ satisfy

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{u}\end{bmatrix} = \begin{bmatrix}(zI - A - B\mathbf{K})^{-1}\\ \mathbf{K}(zI - A - B\mathbf{K})^{-1}\end{bmatrix}\mathbf{w}. \tag{5.2.5}$$

We then have the following theorem parameterizing the set of stable closed-loop transfer matrices, as described in (5.2.5), that are achievable by a given stabilizing controller $\mathbf{K}$.

Theorem 5.2.1 (State-Feedback Parameterization [124])**.** *The following are true:*

- *The affine subspace defined by*

$$\begin{bmatrix}zI - A & -B\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x\\ \mathbf{\Phi}_u\end{bmatrix} = I,\ \ \mathbf{\Phi}_x, \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty \tag{5.2.6}$$

  parameterizes all system responses (5.2.5) *from* $\mathbf{w}$ *to* $(\mathbf{x}, \mathbf{u})$*, achievable by an internally stabilizing state-feedback controller* $\mathbf{K}$*.*

- *For any transfer matrices* $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ *satisfying* (5.2.6)*, the controller* $\mathbf{K} = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}$ *is internally stabilizing and achieves the desired system response* (5.2.5)*.*

Note that in particular, $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\} = \{(zI-A-B\mathbf{K})^{-1}, \mathbf{K}(zI-A-B\mathbf{K})^{-1}\}$ as in (5.2.5) are elements of the affine space defined by (5.2.6) whenever $\mathbf{K}$ is a causal stabilizing controller.

We will also make extensive use of a robust variant of Theorem 5.2.1.

Theorem 5.2.2 (Robust Stability [78])**.** *Suppose that the transfer matrices* $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\} \in \frac{1}{z}\mathcal{RH}_\infty$ *satisfy*

$$\begin{bmatrix}zI - A & -B\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x\\ \mathbf{\Phi}_u\end{bmatrix} = I + \mathbf{\Delta}. \tag{5.2.7}$$

Then the controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ *stabilizes the system described by* (A, B) *if and only if* $(I + \mathbf{\Delta})^{-1} \in \mathcal{RH}_\infty$. *Furthermore, the resulting system response is given by*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} (I + \mathbf{\Delta})^{-1} \mathbf{w}. \tag{5.2.8}$$

Corollary 5.2.3. *Under the assumptions of Theorem 5.2.2, if* $\|\mathbf{\Delta}\| < 1$ *for any induced norm* $\|\cdot\|$, *then the controller* $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ *stabilizes the system described by* (A, B).

Proof. Follows immediately from the small gain theorem, see for example Section 9.2 in Zhou et al. [129]. □

5.2.2 Robust LQR Synthesis

We return to the problem setting where estimates $(\widehat{A}, \widehat{B})$ of a true system (A, B) satisfy

$$\|\Delta_A\| \leq \varepsilon_A, \quad \|\Delta_B\| \leq \varepsilon_B$$

where $\Delta_A := \widehat{A} - A$ and $\Delta_B := \widehat{B} - B$ and where we wish to minimize the LQR cost for the worst instantiation of the parametric uncertainty.

Before proceeding, we must formulate the LQR problem in terms of the system responses $\{\Phi_x(k), \Phi_u(k)\}$. It follows from Theorem 5.2.1 and the standard equivalence between infinite horizon LQR and $\mathcal{H}_2$ optimal control that, for a disturbance process distributed as $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$, the standard LQR problem (1.1.2) can be equivalently written as

$$\min_{\mathbf{\Phi}_x, \mathbf{\Phi}_u} \sigma_w^2 \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2}^2 \text{ s.t. equation (5.2.6).} \tag{5.2.9}$$

We provide a full derivation of this equivalence in Section 5.2.6. Going forward, we drop the σ_w^2 multiplier in the objective function as it affects neither the optimal controller nor the sub-optimality guarantees that we compute in Section 5.2.3.

We begin with a simple sufficient condition under which any controller $\mathbf{K}$ that stabilizes $(\widehat{A}, \widehat{B})$ also stabilizes the true system (A, B). To state the lemma, we introduce one additional piece of notation. For a matrix M, we let $\mathfrak{R}_M$ denote the resolvent

$$\mathfrak{R}_M := (zI - M)^{-1}. \tag{5.2.10}$$

We now can state our robustness lemma.

Lemma 5.2.4. *Let the controller* $\mathbf{K}$ *stabilize* $(\widehat{A}, \widehat{B})$ *and* $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ *be its corresponding system response (5.2.5) on system* $(\widehat{A}, \widehat{B})$. *Then if* $\mathbf{K}$ *stabilizes* (A, B), *it achieves the following LQR cost*

$$\sqrt{J(A, B, \mathbf{K})} := \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \left(I + \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right)^{-1} \right\|_{\mathcal{H}_2}. \tag{5.2.11}$$

Furthermore, letting

$$\hat{\mathbf{\Delta}} := \begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} = (\Delta_A + \Delta_B\mathbf{K})\mathfrak{R}_{\widehat{A}+\widehat{B}\mathbf{K}}\,. \tag{5.2.12}$$

a sufficient condition for $\mathbf{K}$ *to stabilize* (A, B) *is that* $\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} < 1$.

Proof. Follows immediately from Theorems 5.2.1, 5.2.2 and Corollary 5.2.3 by noting that for system responses $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ satisfying

$$\begin{bmatrix}zI - \widehat{A} & -\widehat{B}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} = I,$$

it holds that

$$\begin{bmatrix}zI - A & -B\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} = I + \hat{\mathbf{\Delta}}$$

for $\hat{\mathbf{\Delta}}$ as defined in equation (5.2.12). □

We can therefore recast the robust LQR problem (1.2.7) in the following equivalent form

$$\begin{aligned}&\min_{\mathbf{\Phi}_x, \mathbf{\Phi}_u} \sup_{\substack{\|\Delta_A\| \le \varepsilon_A \\ \|\Delta_B\| \le \varepsilon_B}} J(A, B, \mathbf{K}) \\ &\text{s.t.} \begin{bmatrix}zI - \widehat{A} & -\widehat{B}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} = I, \quad \mathbf{\Phi}_x, \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{R}\mathcal{H}_\infty\,.\end{aligned} \tag{5.2.13}$$

The resulting robust control problem is one subject to real-parametric uncertainty, a class of problems known to be computationally intractable [23]. Although effective computational heuristics (e.g., DK iteration [129]) exist, the performance of the resulting controller on the true system is difficult to characterize analytically in terms of the size of the perturbations.

To circumvent this issue, we take a slightly conservative approach and find an upper-bound to the cost $J(A, B, \mathbf{K})$ that is independent of the uncertainties Δ_A and Δ_B. First, note that if $\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} < 1$, we can write

$$\sqrt{J(A, B, \mathbf{K})} \le \|(I + \hat{\mathbf{\Delta}})^{-1}\|_{\mathcal{H}_\infty}\sqrt{J(\widehat{A}, \widehat{B}, \mathbf{K})} \le \frac{1}{1 - \|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty}}\sqrt{J(\widehat{A}, \widehat{B}, \mathbf{K})}\,. \tag{5.2.14}$$

Because $J(\widehat{A}, \widehat{B}, \mathbf{K})$ captures the performance of the controller $\mathbf{K}$ on the nominal system $(\widehat{A}, \widehat{B})$, it is not subject to any uncertainty. It therefore remains to compute a tractable bound for $\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty}$, which we do using the following fact.

Proposition 5.2.5. *For any* $\alpha \in (0, 1)$ *and* $\hat{\mathbf{\Delta}}$ *as defined in* (5.2.12)

$$\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} \le \left\|\begin{bmatrix}\frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u\end{bmatrix}\right\|_{\mathcal{H}_\infty} =: H_\alpha(\mathbf{\Phi}_x, \mathbf{\Phi}_u)\,. \tag{5.2.15}$$

Proof. Note that for any block matrix of the form $\begin{bmatrix} M_1 & M_2 \end{bmatrix}$, we have

$$\left\| \begin{bmatrix} M_1 & M_2 \end{bmatrix} \right\|_2 \leq \left(\|M_1\|_2^2 + \|M_2\|_2^2 \right)^{1/2} . \tag{5.2.16}$$

To verify this assertion, note that

$$\left\| \begin{bmatrix} M_1 & M_2 \end{bmatrix} \right\|^2 = \lambda_{\max}(M_1 M_1^* + M_2 M_2^*) \leq \lambda_{\max}(M_1 M_1^*) + \lambda_{\max}(M_2 M_2^*) = \|M_1\|^2 + \|M_2\|^2 .$$

With (5.2.16) in hand, we have

$$\begin{aligned} \left\| \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} &= \left\| \begin{bmatrix} \frac{\sqrt{\alpha}}{\varepsilon_A}\Delta_A & \frac{\sqrt{1-\alpha}}{\varepsilon_B}\Delta_B \end{bmatrix} \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \\ &\leq \left\| \begin{bmatrix} \frac{\sqrt{\alpha}}{\varepsilon_A}\Delta_A & \frac{\sqrt{1-\alpha}}{\varepsilon_B}\Delta_B \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} , \end{aligned}$$

completing the proof. □

The following corollary is then immediate.

Corollary 5.2.6. *Let the controller* $\mathbf{K}$ *and resulting system response* $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ *be as defined in Lemma 5.2.4. Then if* $H_\alpha(\mathbf{\Phi}_x, \mathbf{\Phi}_u) < 1$*, the controller* $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ *stabilizes the true system* (A, B)*.*

Applying Proposition 5.2.5 in conjunction with the bound (5.2.14), we arrive at the following upper bound to the cost function of the robust LQR problem (1.2.7), which is independent of the perturbations (Δ_A, Δ_B):

$$\sup_{\substack{\|\Delta_A\| \leq \varepsilon_A \\ \|\Delta_B\| \leq \varepsilon_B}} \sqrt{J(A, B, \mathbf{K})} \leq \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \frac{1}{1 - H_\alpha(\mathbf{\Phi}_x, \mathbf{\Phi}_u)} = \frac{\sqrt{J(\widehat{A}, \widehat{B}, \mathbf{K})}}{1 - H_\alpha(\mathbf{\Phi}_x, \mathbf{\Phi}_u)} . \tag{5.2.17}$$

The upper bound is only valid when $H_\alpha(\mathbf{\Phi}_x, \mathbf{\Phi}_u) < 1$, which guarantees the stability of the closed-loop system as in Corollary 5.2.6. We remark that Corollary 5.2.6 and the bound in (5.2.17) are of interest independent of the synthesis procedure for $\mathbf{K}$. In particular, they can be applied to the optimal LQR controller $\widehat{K}$ computed using the nominal system $(\widehat{A}, \widehat{B})$.

As the next lemma shows, the right hand side of Equation 5.2.17 can be efficiently optimized by an appropriate decomposition. The proof of the lemma is immediate.

Lemma 5.2.7. *For functions* $f : \mathcal{X} \to \mathbb{R}$ *and* $g : \mathcal{X} \to \mathbb{R}$ *and constraint set* $C \subseteq \mathcal{X}$*, consider*

$$\min_{x \in C} \frac{f(x)}{1 - g(x)} .$$

Assuming that $f(x) \geq 0$ and $0 \leq g(x) < 1$ for all $x \in C$, this optimization problem can be reformulated as an outer single-variable problem and an inner constrained optimization problem (the objective value of an optimization over the emptyset is defined to be infinity):

$$\min_{x \in C} \frac{f(x)}{1 - g(x)} = \min_{\gamma \in [0,1)} \tfrac{1}{1-\gamma} \min_{x \in C} \{ f(x) \mid g(x) \leq \gamma \}$$

Then combining Lemma 5.2.7 with the upper bound in (5.2.17) results in the following optimization problem:

$$\begin{aligned} \text{minimize}_{\gamma \in [0,1)} \; & \frac{1}{1-\gamma} \min_{\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u} \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\ & \text{s.t.} \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_u \end{bmatrix} = I, \quad \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}} \boldsymbol{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}} \boldsymbol{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \gamma \\ & \boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u \in \frac{1}{z} \mathcal{RH}_\infty. \end{aligned} \tag{5.2.18}$$

We note that this optimization objective is jointly quasi-convex in $(\gamma, \boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u)$. Hence, as a function of γ alone the objective is quasi-convex, and furthermore is smooth in the feasible domain. Therefore, the outer optimization with respect to γ can effectively be solved with methods like golden section search. We remark that the inner optimization is a convex problem, though an infinite dimensional one. We show in Section 5.2.4 that a simple finite impulse response truncation yields a finite dimensional problem with similar guarantees of robustness and performance.

We further remark that because $\gamma \in [0, 1)$, any feasible solution $(\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u)$ to optimization problem (5.2.18) generates a controller $\mathbf{K} = \boldsymbol{\Phi}_u \boldsymbol{\Phi}_x^{-1}$ satisfying the conditions of Corollary 5.2.6, and hence stabilizes the true system (A, B). Therefore, even if the solution is approximated, as long as it is feasible, it will be stabilizing. As we show in the next section, for sufficiently small estimation error bounds ε_A and ε_B, we can further bound the sub-optimality of the performance achieved by our robustly stabilizing controller relative to that achieved by the optimal LQR controller K.

5.2.3 Sub-optimality Guarantees

We now return to analyzing the Coarse-ID control problem. We upper bound the performance of the controller synthesized using the optimization (5.2.18) in terms of the size of the perturbations (Δ_A, Δ_B) and a measure of complexity of the LQR problem defined by A, B, S, and R. The following result is one of our main contributions.

Theorem 5.2.8. *Let $J_\star$ denote the minimal LQR cost achievable by any controller for the dynamical system with transition matrices (A, B), and let K denote the optimal controller.*

Let $(\widehat{A}, \widehat{B})$ *be estimates of the transition matrices such that* $\|\Delta_A\| \leq \varepsilon_A$, $\|\Delta_B\| \leq \varepsilon_B$. *Then, if* $\mathbf{K}$ *is synthesized via* (5.2.18) *with* $\alpha = 1/2$, *the relative error in the LQR cost is*

$$\frac{\sqrt{J(A, B, \mathbf{K})} - \sqrt{J_\star}}{\sqrt{J_\star}} \leq 5(\varepsilon_A + \varepsilon_B \|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} , \tag{5.2.19}$$

as long as $(\varepsilon_A + \varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} \leq 1/5$.

This result offers a guarantee on the performance of the SLS synthesized controller regardless of the estimation procedure used to estimate the transition matrices. Together with the estimation results in Chapter 3, Theorem 5.2.8 yields a sample complexity upper bound on the performance of the robust SLS controller $\mathbf{K}$ when (A, B) are not known. We make this guarantee precise in Corollary 5.2.10 below.

We now compare the guarantee offered by Theorem 5.2.8 to the certainty equivalence controller guarantee from Theorem 5.1.6. Theorem 5.2.8 states that as long as

$$\varepsilon \leq \frac{1}{5(1 + \|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}} ,$$

then the resulting controller satisfies:

$$\sqrt{J(\mathbf{K})} - \sqrt{J_\star} \leq 5(1 + \|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}\sqrt{J_\star}\,\varepsilon . \tag{5.2.20}$$

Equation 5.2.20 implies that:

$$J(\mathbf{K}) - J_\star \leq 10(1 + \|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} J_\star \varepsilon + \mathcal{O}(\varepsilon^2) . \tag{5.2.21}$$

In order to compare Equation 5.2.21 to Equation 5.1.6, we upper bound the quantity $\Psi_\star$ in terms of $\tau(L, \gamma)$ and γ. In particular, by a infinite series expansion of the inverse:

$$\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} = \sup_{z \in \mathbb{T}} \|(zI - L)^{-1}\| = \sup_{z \in \mathbb{T}} \left\| \sum_{k=0}^{\infty} L^k z^{-(k+1)} \right\| \leq \sum_{k=0}^{\infty} \|L^k\| \leq \frac{\tau(L, \gamma)}{1 - \gamma} .$$

We can also upper bound $J_\star = \sigma_w^2 \operatorname{tr}(V) \leq \sigma_w^2 n \Gamma$. Therefore, Equation 5.2.21 gives us that:

$$J(\mathbf{K}) - J_\star \leq \mathcal{O}(1) n \sigma_w^2 \Gamma^2 \frac{\tau(L, \gamma)}{1 - \gamma} \varepsilon + \mathcal{O}(\varepsilon^2) .$$

We see that the dependence on the parameters Γ and $\tau(L, \gamma)$ is significantly milder compared to Equation 5.1.6. Furthermore, this upper bound is valid for larger ε than the upper bound given in Theorem 5.1.6. Comparing these upper bound suggests that there is a price to pay for obtaining a fast rate, and that in regimes of moderate uncertainty (moderate size of ε), being robust to model uncertainty is important.

A similar trade-off between slow and fast rates arises in the setting of first-order convex stochastic optimization. The convergence rate $\mathcal{O}(1/\sqrt{T})$ of the stochastic gradient descent

method can be improved to $\mathcal{O}(1/T)$ under a strong convexity assumption. However, the performance of stochastic gradient descent, which can achieve a $\mathcal{O}(1/T)$ rate, is sensitive to poorly estimated problem parameters [84]. Similarly, in the case of LQR, the nominal controller achieves a fast rate, but it is much more sensitive to estimation error than the robust controller based on SLS.

The rest of the section is dedicated to proving Theorem 5.2.8. Recall that K is the optimal LQR static state feedback matrix for the true dynamics (A, B), and let $\mathbf{\Delta} := -\left[\Delta_A + \Delta_B K\right] \mathfrak{R}_{A+BK}$. We begin with a technical result.

Lemma 5.2.9. *Define* $\zeta := (\varepsilon_A + \varepsilon_B \|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}$, *and suppose that* $\zeta < (1+\sqrt{2})^{-1}$. *Then* $(\gamma_0, \tilde{\mathbf{\Phi}}_x, \tilde{\mathbf{\Phi}}_u)$ *is a feasible solution of* (5.2.18) *with* $\alpha = 1/2$, *where*

$$\gamma_0 = \frac{\sqrt{2}\zeta}{1-\zeta}, \quad \tilde{\mathbf{\Phi}}_x = \mathfrak{R}_{A+BK}(I+\mathbf{\Delta})^{-1}, \quad \tilde{\mathbf{\Phi}}_u = K\mathfrak{R}_{A+BK}(I+\mathbf{\Delta})^{-1}. \tag{5.2.22}$$

Proof. By construction $\tilde{\mathbf{\Phi}}_x, \tilde{\mathbf{\Phi}}_u \in \frac{1}{z}\mathcal{R}\mathcal{H}_\infty$. Therefore, we are left to check three conditions:

$$\gamma_0 < 1, \quad \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Phi}}_x \\ \tilde{\mathbf{\Phi}}_u \end{bmatrix} = I \text{ , and } \quad \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}} \tilde{\mathbf{\Phi}}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}} \tilde{\mathbf{\Phi}}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \le \frac{\sqrt{2}\zeta}{1-\zeta}. \tag{5.2.23}$$

The first two conditions follow by simple algebraic computations. Before we check the last condition, note that $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \le (\varepsilon_A + \varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} = \zeta < 1$. Now observe that,

$$\begin{aligned}
\left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}} \tilde{\mathbf{\Phi}}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}} \tilde{\mathbf{\Phi}}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} &= \sqrt{2} \left\| \begin{bmatrix} \varepsilon_A \mathfrak{R}_{A+BK} \\ \varepsilon_B K \mathfrak{R}_{A+BK} \end{bmatrix} (I+\mathbf{\Delta})^{-1} \right\|_{\mathcal{H}_\infty} \\
&\le \sqrt{2}\|(I+\mathbf{\Delta})^{-1}\|_{\mathcal{H}_\infty} \left\| \begin{bmatrix} \varepsilon_A \mathfrak{R}_{A+BK} \\ \varepsilon_B K \mathfrak{R}_{A+BK} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \\
&\le \frac{\sqrt{2}}{1-\|\mathbf{\Delta}\|_{\mathcal{H}_\infty}} \left\| \begin{bmatrix} \varepsilon_A I \\ \varepsilon_B K \end{bmatrix} \mathfrak{R}_{A+BK} \right\|_{\mathcal{H}_\infty} \\
&\le \frac{\sqrt{2}(\varepsilon_A + \varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}}{1-\|\mathbf{\Delta}\|_{\mathcal{H}_\infty}} \le \frac{\sqrt{2}\zeta}{1-\zeta}\,.
\end{aligned}$$

□

Proof of Theorem 5.2.8. Let $(\gamma_\star, \mathbf{\Phi}_x^\star, \mathbf{\Phi}_u^\star)$ be an optimal solution to problem (5.2.18) and let $\mathbf{K} = \mathbf{\Phi}_u^\star(\mathbf{\Phi}_x^\star)^{-1}$. We can then write

$$\sqrt{J(A,B,\mathbf{K})} \le \frac{1}{1-\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty}} \sqrt{J(\widehat{A},\widehat{B},\mathbf{K})} \le \frac{1}{1-\gamma_\star} \sqrt{J(\widehat{A},\widehat{B},\mathbf{K})}\,,$$

where the first inequality follows from the bound (5.2.14), and the second follows from the fact that $\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} \le \gamma_\star$ due to Proposition 5.2.5 and the constraint in optimization problem (5.2.18).

From Lemma 5.2.9 we know that $(\gamma_0, \tilde{\mathbf{\Phi}}_x, \tilde{\mathbf{\Phi}}_u)$ defined in equation (5.2.22) is also a feasible solution. Therefore, because $K = \tilde{\mathbf{\Phi}}_u \tilde{\mathbf{\Phi}}_x^{-1}$, we have by optimality,

$$\frac{1}{1-\gamma_\star}\sqrt{J(\widehat{A},\widehat{B},\mathbf{K})} \le \frac{1}{1-\gamma_0}\sqrt{J(\widehat{A},\widehat{B},K)} \le \frac{\sqrt{J(A,B,K)}}{(1-\gamma_0)(1-\|\mathbf{\Delta}\|_{\mathcal{H}_\infty})} = \frac{\sqrt{J_\star}}{(1-\gamma_0)(1-\|\mathbf{\Delta}\|_{\mathcal{H}_\infty})}\,,$$

where the second inequality follows by the argument used to derive (5.2.14) with the true and estimated transition matrices switched. Recall that $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} \le \zeta$ and that $\gamma_0 = \sqrt{2}\zeta/(1+\zeta)$. Therefore

$$\frac{\sqrt{J(A,B,\mathbf{K})} - \sqrt{J_\star}}{\sqrt{J_\star}} \le \frac{1}{1-(1+\sqrt{2})\zeta} - 1 = \frac{(1+\sqrt{2})\zeta}{1-(1+\sqrt{2})\zeta} \le 5\zeta\,,$$

where the last inequality follows because $\zeta < 1/5 < 1/(2+2\sqrt{2})$. The conclusion follows. □

With this sub-optimality result in hand, we are now ready to give an end-to-end performance guarantee for our procedure when the independent data estimation scheme is used.

Corollary 5.2.10. *Let $\lambda_G = \lambda_{\min}(\sigma_u^2 G_T G_T^\mathsf{T} + \sigma_w^2 F_T F_T^\mathsf{T})$, where F_T, G_T are defined in (3.1.2). Suppose the independent data estimation procedure described in (3.1.1) is used to produce estimates $(\widehat{A}, \widehat{B})$ and $\mathbf{K}$ is synthesized via (5.2.18) with $\alpha = 1/2$. Then there are universal constants C_0 and C_1 such that the relative error in the LQR cost satisfies*

$$\frac{\sqrt{J(A,B,\mathbf{K})} - \sqrt{J_\star}}{\sqrt{J_\star}} \le C_0 \sigma_w \|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} \left(\frac{1}{\sqrt{\lambda_G}} + \frac{\|K\|}{\sigma_u}\right) \sqrt{\frac{(n+d)\log(1/\delta)}{N}} \tag{5.2.24}$$

with probability $1-\delta$, as long as $N \ge C_1(n+d)\sigma_w^2 \|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}^2 (1/\lambda_G + \|K\|^2/\sigma_u^2)\log(1/\delta)$.

Proof. Recall from Proposition 3.1.1 that for the independent data estimation scheme, we have

$$\varepsilon_A \le \frac{16\sigma_w}{\sqrt{\lambda_G}}\sqrt{\frac{(n+2d)\log(32/\delta)}{N}}\,, \text{ and } \quad \varepsilon_B \le \frac{16\sigma_w}{\sigma_u}\sqrt{\frac{(n+2d)\log(32/\delta)}{N}}, \tag{5.2.25}$$

with probability $1-\delta$, as long as $N \ge 8(n+d) + 16\log(4/\delta)$.

To apply Theorem 5.2.8 we need $(\varepsilon_A + \varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty} < 1/5$, which will hold as long as $N \ge \mathcal{O}\left\{(n+d)\sigma_w^2\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}^2(1/\lambda_G + \|K\|^2/\sigma_u^2)\log(1/\delta)\right\}$. A direct plug in of (5.2.25) in (5.2.19) yields the conclusion. □

Corollary 5.2.10 states that the sub-optimality gap is:

$$\frac{\sqrt{J(A,B,\mathbf{K})} - \sqrt{J_\star}}{\sqrt{J_\star}} \le \mathcal{C}_{\mathrm{LQR}}\sqrt{\frac{(n+d)\log(1/\delta)}{N}}\,.$$

with the constant $\mathcal{C}_{\text{LQR}}$ defined as:

$$\mathcal{C}_{\text{LQR}} := C_0 \sigma_w \left(\frac{1}{\sqrt{\lambda_G}} + \frac{\|K\|_2}{\sigma_u} \right) \|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}.$$

Note that $\mathcal{C}_{\text{LQR}}$ decreases as the minimum eigenvalue of the sum of the input and noise controllability Gramians increases. This minimum eigenvalue tends to be larger for systems that amplify inputs in all directions of the state-space. $\mathcal{C}_{\text{LQR}}$ increases as function of the operator norm of the gain matrix K and the $\mathcal{H}_\infty$ norm of the transfer function from disturbance to state of the closed-loop system. These two terms tend to be larger for systems that are "harder to control." The dependence on Q and R is implicit in this definition since the optimal control matrix K is defined in terms of these two matrices. Note that when R is large in comparison to Q, the norm of the controller K tends to be smaller because large inputs are more costly. However, such a change in the size of the controller could cause an increase in the $\mathcal{H}_\infty$ norm of the closed-loop system. Thus, our upper bound suggests an odd balance. Stable and highly damped systems are easy to control but hard to estimate, whereas unstable systems are easy to estimate but hard to control. Our theorem suggests that achieving a small relative LQR cost requires for the system to be somewhere in the middle of these two extremes.

Finally, we remark that the above analysis holds more generally when we apply additional constraints to the controller in the synthesis problem (5.2.18). In this case, the sub-optimality bounds presented in Theorem 5.2.8 and Corollary 5.2.10 are true with respect to the minimal cost achievable by the constrained controller with access to the true dynamics. In particular, the bounds hold unchanged if the search is restricted to static controllers, i.e. $u_t = Kx_t$. This is true because the optimal controller is static and therefore feasible for the constrained synthesis problem.

As posed, the main optimization problem (5.2.18) is a semi-infinite program, and we are not aware of a way to solve this problem efficiently. We now turn to two alternative formulations that provide upper bounds to the optimal value and that can be solved in polynomial time.

5.2.4 Finite impulse response approximation

An elementary approach to reducing the aforementioned semi-infinite program to a finite dimensional one is to only optimize over the first L elements of the transfer functions $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$, effectively taking a finite impulse response (FIR) approximation. Since these are both stable maps, we expect the effects of such an approximation to be negligible as long as the optimization horizon L is chosen to be sufficiently large – in what follows, we show that this is indeed the case.

By restricting our optimization to FIR approximations of $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$, we can cast the $\mathcal{H}_2$ cost as a second order cone constraint. The only difficulty arises in posing the $\mathcal{H}_\infty$ constraint as a semidefinite program. Though there are several ways to cast $\mathcal{H}_\infty$ constraints

as linear matrix inequalities, we use the formulation in Theorem 5.8 of Dumitrescu's text to take advantage of the FIR structure in our problem [36]. We note that using Dumitrescu's formulation, the resulting problem is affine in α when γ is fixed, and hence we can solve for the optimal value of α. Then the resulting system response elements can be cast as a dynamic feedback controller using Theorem 2 of Anderson and Matni [9].

5.2.4.1 Sub-optimality guarantees

Here, we show that optimizing over FIR approximations incurs only a small degradation in performance relative to the solution to the infinite-horizon problem. In particular, this degradation in performance decays exponentially in the FIR horizon L, where the rate of decay is specified by the decay rate of the spectral elements of the optimal closed loop system response $\mathfrak{R}_{A+BK}$.

Before proceeding, we introduce additional concepts and notation needed to formalize guarantees in the FIR setting. A linear-time-invariant transfer function is stable if and only if it is exponentially stable, i.e., $\mathbf{\Phi} = \sum_{t=0}^{\infty} z^{-t}\Phi(t) \in \mathcal{RH}_\infty$ if and only if there exists positive values τ and $\rho \in [0, 1)$ such that for every spectral element $\Phi(t)$, $t \geq 0$, it holds that

$$\|\Phi(t)\| \leq \tau\rho^t. \tag{5.2.26}$$

In what follows, we pick $\tau_\star$ and $\rho_\star$ to be any such constants satisfying $\|\mathfrak{R}_{A+BK}(t)\| \leq \tau_\star\rho_\star^t$ for all $t \geq 0$.

We introduce a version of the optimization problem (5.2.13) with a finite number of decision variables:

$$\begin{aligned}
\text{minimize}_{\gamma\in[0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi}_x,\mathbf{\Phi}_u,V} & \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\
\text{s.t.} & \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I + \frac{1}{z^L}V, \\
& \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} + \|V\| \leq \gamma \\
& \mathbf{\Phi}_x = \sum_{t=1}^{L} \frac{1}{z^t}\Phi_x(t),\ \mathbf{\Phi}_u = \sum_{t=1}^{L} \frac{1}{z^t}\Phi_u(t).
\end{aligned} \tag{5.2.27}$$

In this optimization problem we search over finite response transfer functions $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$. Given a feasible solution $\mathbf{\Phi}_x$, $\mathbf{\Phi}_u$ of problem (5.2.27), we can implement the controller $\mathbf{K}_L = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}$ with an equivalent state-space representation (A_K, B_K, C_K, D_K) using the response elements $\{\Phi_x(k)\}_{k=1}^L$ and $\{\Phi_u(k)\}_{k=1}^L$ via Theorem 2 of Anderson and Matni [9].

The slack term V accounts for the error introduced by truncating the infinite response transfer functions of problem (5.2.13). Intuitively, if the truncated tail is sufficiently small,

then the effects of this approximation should be negligible on performance. The next result formalizes this intuition.

Theorem 5.2.11. *Set $\alpha = 1/2$ in (5.2.27) and let $\tau_\star > 0$ and $\rho_\star \in [0,1)$ be such that $\|\mathfrak{R}_{(A+BK)}(t)\| \le \tau_\star \rho_\star^t$ for all $t \ge 0$. Then, if $\mathbf{K}_L$ is synthesized via (5.2.27), the relative error in the LQR cost is*

$$\frac{\sqrt{J(A,B,\mathbf{K}_L)} - \sqrt{J_\star}}{\sqrt{J_\star}} \le 10(\varepsilon_A + \varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty},$$

as long as

$$\varepsilon_A + \varepsilon_B\|K\| \le \frac{1-\rho_\star}{10\tau_\star} \quad \text{and} \quad L \ge \frac{4\log\left(\frac{\tau_\star}{(\varepsilon_A+\varepsilon_B\|K\|)\|\mathfrak{R}_{A+BK}\|_{\mathcal{H}_\infty}}\right)}{1-\rho_\star}.$$

The proof of this result can be found in Dean et al. [31]. It is conceptually the same as that of the infinite horizon setting. The main difference is that care must be taken to ensure that the approximation horizon L is sufficiently large so as to ensure stability and performance of the resulting controller. From the theorem statement, we see that for such an appropriately chosen FIR approximation horizon L, our performance bound is the same, up to universal constants, to that achieved by the solution to the infinite horizon problem. Furthermore, the approximation horizon L only needs to grow logarithmically with respect to one over the estimation rate in order to preserve the same statistical rate as the controller produced by the infinite horizon problem. Finally, an end-to-end sample complexity result analogous to that stated in Corollary 5.2.10 can be easily obtained by simply substituting in the sample-complexity bounds on ε_A and ε_B specified in Proposition 3.1.1.

5.2.5 Static controller and a common Lyapunov approximation

As we have reiterated above, when the dynamics are known, the optimal LQR control law takes the form $u_t = Kx_t$ for properly chosen static gain matrix K. We can reparameterize the optimization problem (5.2.18) to restrict our attention to such static control policies:

$$\begin{aligned}\text{minimize}_{\gamma\in[0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi}_x,\mathbf{\Phi}_u,K} &\left\|\begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix}\right\|_{\mathcal{H}_2} \\ \text{s.t.} \begin{bmatrix} zI - \widehat{A} & -\widehat{B}\end{bmatrix}&\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} = I, \quad \left\|\begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}}\mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}\mathbf{\Phi}_u\end{bmatrix}\right\|_{\mathcal{H}_\infty} \le \gamma \\ &\mathbf{\Phi}_x, \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{R}\mathcal{H}_\infty, \quad K = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}.\end{aligned} \tag{5.2.28}$$

Under this reparameterization, the problem is no longer convex. Here we present a simple application of the *common Lyapunov relaxation* that allows us to find a controller K using semidefinite programming.

Note that the equality constraints imply:

$$I = \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} I \\ K \end{bmatrix} \mathbf{\Phi}_x = (zI - \widehat{A} - \widehat{B}K)\mathbf{\Phi}_x \,,$$

revealing that we must have

$$\mathbf{\Phi}_x = (zI - \widehat{A} - \widehat{B}K)^{-1} \quad \text{and} \quad \mathbf{\Phi}_u = K(zI - \widehat{A} - \widehat{B}K)^{-1} \,.$$

With these identifications, (5.2.28) can be reformulated as

$$\begin{aligned} \text{minimize}_{\gamma \in [0,1)} \; & \frac{1}{1-\gamma} \min_K \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}}K \end{bmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1} \right\|_{\mathcal{H}_2} \\ \text{s.t.} \; & \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}} \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}}K \end{bmatrix} (zI - \widehat{A} - \widehat{B}K)^{-1} \right\|_{\mathcal{H}_\infty} \leq \gamma \end{aligned} \tag{5.2.29}$$

Using standard techniques from the robust control literature, we can upper bound this problem via the semidefinite program

$$\begin{aligned} \text{minimize}_{X,Z,W,\alpha,\gamma} \quad & \tfrac{1}{(1-\gamma)^2} \left\{ \text{tr}(SW_{11}) + \text{tr}(RW_{22}) \right\} \\ \text{subject to} \quad & \begin{bmatrix} X & X & Z^\mathsf{T} \\ X & W_{11} & W_{12} \\ Z & W_{21} & W_{22} \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} X - I & (\widehat{A} + \widehat{B}K)X & 0 & 0 \\ X(\widehat{A} + \widehat{B}K)^\mathsf{T} & X & \varepsilon_A X & \varepsilon_B Z^\mathsf{T} \\ 0 & \varepsilon_A X & \alpha\gamma^2 I & 0 \\ 0 & \varepsilon_B Z & 0 & (1-\alpha)\gamma^2 I \end{bmatrix} \succeq 0 \,. \end{aligned} \tag{5.2.30}$$

Note that this optimization problem is affine in α when γ is fixed. Hence, in practice we can find the optimal value of α as well. A static controller can then be extracted from this optimization problem by setting $K = ZX^{-1}$. A full derivation of this relaxation can be found in Dean et al. [31]. Note that this compact SDP is simpler to solve than the truncated FIR approximation.

5.2.6 Derivation of the LQR cost as an $\mathcal{H}_2$ norm

In this section, we consider the transfer function description of the infinite horizon LQR optimal control problem. In particular, we show how it can be recast as an equivalent $\mathcal{H}_2$ optimal control problem in terms of the system response variables defined in Theorem 5.2.1.

Recall that stable and achievable system responses $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$, as characterized in equation (5.2.6), describe the closed-loop map from disturbance signal $\mathbf{w}$ to the state and control

action $(\mathbf{x}, \mathbf{u})$ achieved by the controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$, i.e., $\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \mathbf{w}$. Letting $\mathbf{\Phi}_x = \sum_{t=1}^{\infty} \Phi_x(t) z^{-t}$ and $\mathbf{\Phi}_u = \sum_{t=1}^{\infty} \Phi_u(t) z^{-t}$, we can then equivalently write for any $t \geq 1$

$$\begin{bmatrix} x_t \\ u_t \end{bmatrix} = \sum_{k=1}^{t} \begin{bmatrix} \Phi_x(k) \\ \Phi_u(k) \end{bmatrix} w_{t-k}. \tag{5.2.31}$$

For a disturbance process distributed as $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$, it follows from equation (5.2.31) that

$$\begin{aligned} \mathbb{E}\left[x_t^\mathsf{T} S x_t\right] &= \sigma_w^2 \sum_{k=1}^{t} \operatorname{tr}(\Phi_x(k)^\mathsf{T} S \Phi_x(k)) \,, \\ \mathbb{E}\left[u_t^\mathsf{T} R u_t\right] &= \sigma_w^2 \sum_{k=1}^{t} \operatorname{tr}(\Phi_u(k)^\mathsf{T} R \Phi_u(k)) \,. \end{aligned}$$

We can then write

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t\right] &= \sigma_w^2 \left[\sum_{t=1}^{\infty} \operatorname{tr}(\Phi_x(t)^\mathsf{T} S \Phi_x(t)) + \operatorname{tr}(\Phi_u(t)^\mathsf{T} R \Phi_u(t))\right] \\ &= \sigma_w^2 \sum_{t=1}^{\infty} \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \Phi_x(t) \\ \Phi_u(t) \end{bmatrix} \right\|_F^2 \\ &= \frac{\sigma_w^2}{2\pi} \int_{\mathbb{T}} \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_F^2 dz \\ &= \sigma_w^2 \left\| \begin{bmatrix} S^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2}^2, \end{aligned}$$

where the second to last equality is due to Parseval's Theorem.

Chapter 6

Model-free Methods for LQR

In this chapter, we study model-free methods for LQR. We first turn to a non-asymptotic analysis of the classic least-squares policy iteration (LSPI) algorithm from Lagoudakis and Parr [63] applied to LQR. One issue with the non-asymptotic analysis is that we will produce an upper bound on the sample complexity, which is not directly comparable to the upper bounds established in Chapter 5. We partially resolve this issue in the later part of this chapter by turning to an asymptotic analysis of various model-based and model-free algorithms, which allows us to directly compare the sample complexity of model-based and model-free algorithms on LQR. The LSPI analysis in this chapter is new material, whereas the later part of the thesis is based on Tu and Recht [118].

Notation. For a positive scalar $x \geq 0$, we let $x_+ = \max\{1, x\}$. As before, we let $\mathrm{svec}(M) \in \mathbb{R}^{n(n+1)/2}$ denote the vectorized version of the upper triangular part of a symmetric matrix M so that $\|M\|_F^2 = \langle \mathrm{svec}(M), \mathrm{svec}(M) \rangle$. Finally, $\mathrm{smat}(\cdot)$ denotes the inverse of $\mathrm{svec}(\cdot)$, so that $\mathrm{smat}(\mathrm{svec}(M)) = M$.

6.1 Least-squares Policy Iteration for LQR

6.1.1 Related Work

We first discuss related RL results for the general function approximation setting. Antos et al. [10] and Lazaric et al. [64] analyze variants of LSPI for discounted MDPs where the state space is a compact set, the action space finite, and the feature vectors and rewards are uniformly bounded. Furthermore, Lazaric et al. [64] study a version of LSPI where LSTD is applied to learn the *value* function of the current policy, and the policy is greedily updated via an update operator that requires access to the underlying dynamics (and is therefore not implementable). Farahmand et al. [40] extend the results of Lazaric et al. [64] to when the function spaces considered are reproducing kernel Hilbert spaces. Zou et al. [130] give a finite-time analysis of both Q-learning and SARSA, combining the asymptotic analysis

of Melo et al. [80] with the finite-time analysis of TD-learning from Bhandari et al. [16]. We note that checking the required assumptions to apply the results of Zou et al. [130] is non-trivial (c.f. Section 3.1, [80]). We also note that we are un-aware of any non-asymptotic analysis of LSPI in the *average cost* setting considered in this paper, which is substantially more difficult as the Bellman operator is no longer a contraction.

6.1.2 Least-squares temporal difference learning for Q-functions

The first component towards an understanding of approximate PI is to understand least-squares temporal difference learning (LSTD-Q) for Q-functions, which is the fundamental building block of LSPI. We briefly recap the LSTD-Q algorithm from the discussion in Section 1.3.

Given a policy K_{eval} which stabilizes (A, B), the goal of LSTD-Q is to estimate the parameters of the Q-function associated to K_{eval}. Bellman's equation for infinite-horizon average cost MDPs (c.f. Bertsekas [14]) states that the (relative) Q-function associated to a policy π satisfies the following fixed-point equation:

$$\lambda + Q(x,u) = c(x,u) + \mathbb{E}_{x' \sim p(\cdot|x,u)}[Q(x', \pi(x'))] \,. \tag{6.1.1}$$

Here, $\lambda \in \mathbb{R}$ is a free parameter chosen so that the fixed-point equation holds. LSTD-Q operates under the *linear architecture* assumption, which states that the Q-function can be described as $Q(x,u) = q^{\mathsf{T}}\phi(x,u)$, for a known (possibly non-linear) feature map $\phi(x,u)$. It is well known that LQR satisfies this assumption, since we have:

$$\begin{aligned}
Q(x,u) &= \text{svec}(Q)^{\mathsf{T}} \text{svec}\left(\begin{bmatrix} x \\ u \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}^{\mathsf{T}} \right) \,, \\
Q &= \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A^{\mathsf{T}} \\ B^{\mathsf{T}} \end{bmatrix} V \begin{bmatrix} A & B \end{bmatrix} \,, \\
V &= \mathsf{dlyap}(A + BK_{\text{eval}}, S + K_{\text{eval}}^{\mathsf{T}} R K_{\text{eval}}) \,, \\
\lambda &= \left\langle Q, \sigma_w^2 \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^{\mathsf{T}} \right\rangle \,.
\end{aligned}$$

Here, we slightly abuse notation and let Q denote the function and also the matrix parameterizing the Q function. Now suppose that a trajectory $\{(x_t, u_t, x_{t+1})\}_{t=1}^{T}$ is collected. Note that LSTD-Q is an *off-policy* method (unlike the closely related LSTD estimator for value functions), and therefore the inputs u_t can come from any sequence that provides sufficient excitation for learning. In particular, it does *not* have to come from the policy K_{eval}. In this paper, we will consider inputs of the form:

$$u_t = K_{\text{play}} x_t + \eta_t \,, \;\; \eta_t \sim \mathcal{N}(0, \sigma_\eta^2 I) \,, \tag{6.1.2}$$

where K_{play} is a stabilizing controller for (A, B). Once again we emphasize that $K_{\text{play}} \neq K_{\text{eval}}$ in general. The injected noise η_t is needed in order to provide sufficient excitation for learning.

In order to describe the LSTD-Q estimator, we define the following quantities which play a key role throughout the paper:

$$\phi_t := \phi(x_t, u_t)\ ,\ \ \psi_t := \phi(x_t, K_{\text{eval}} x_t)\ ,$$
$$f := \text{svec}\left(\sigma_w^2 \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^{\mathsf{T}}\right)\ ,\ \ c_t := x_t^{\mathsf{T}} S x_t + u_t^{\mathsf{T}} R u_t\ .$$

The LSTD-Q estimator estimates q via:

$$\widehat{q} := \left(\sum_{t=1}^{T} \phi_t(\phi_t - \psi_{t+1} + f)^{\mathsf{T}}\right)^{\dagger} \sum_{t=1}^{T} \phi_t c_t\ . \tag{6.1.3}$$

Here, $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. Our first result establishes a non-asymptotic bound on the quality of the estimator $\widehat{q}$, measured in terms of $\|\widehat{q} - q\|$.

Theorem 6.1.1. *Fix a $\delta \in (0,1)$. Let policies K_{play} and K_{eval} stabilize (A, B), and assume that both $A + BK_{\text{play}}$ and $A + BK_{\text{eval}}$ are (τ, ρ)-stable. Let the initial state $x_0 \sim \mathcal{N}(0, \Sigma_0)$ and consider the inputs $u_t = K_{\text{play}} x_t + \eta_t$ with $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2 I)$. For simplicity, assume that $\sigma_\eta \le \sigma_w$. Let P_∞ denote the steady-state covariance of the trajectory $\{x_t\}$:*

$$P_\infty = \mathsf{dlyap}((A + BK_{\text{play}})^{\mathsf{T}}, \sigma_w^2 I + \sigma_\eta^2 BB^{\mathsf{T}})\ . \tag{6.1.4}$$

Define the proxy variance $\overline{\sigma}^2$ by:

$$\overline{\sigma}^2 := \tau^2 \rho^4 \|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2 \|B\|^2\ . \tag{6.1.5}$$

Suppose that T satisfies:

$$T \ge \widetilde{\mathcal{O}}(1) \max\left\{(n+d)^2, \frac{\tau^4}{\rho^4(1-\rho^2)^2} \frac{(n+d)^4}{\sigma_\eta^4} \sigma_w^2 \overline{\sigma}^2 \|K_{\text{play}}\|_+^4 \|K_{\text{eval}}\|_+^8 (\|A\|^4 + \|B\|^4)_+\right\}\ . \tag{6.1.6}$$

Then we have with probability at least $1 - \delta$,

$$\|\widehat{q} - q\| \le \widetilde{\mathcal{O}}(1) \frac{\tau^2}{\rho^2(1-\rho^2)} \frac{(n+d)}{\sigma_\eta^2 \sqrt{T}} \sigma_w \overline{\sigma} \|K_{\text{play}}\|_+^2 \|K_{\text{eval}}\|_+^4 (\|A\|^2 + \|B\|^2)_+ \|Q^{K_{\text{eval}}}\|_F\ . \tag{6.1.7}$$

Here the $\widetilde{\mathcal{O}}(1)$ hides $\text{polylog}(n, \tau, \|\Sigma_0\|, \|P_\infty\|, \|K_{\text{play}}\|, T/\delta, 1/\sigma_\eta)$ *factors.*

Theorem 6.1.1 states that:

$$T \le \widetilde{\mathcal{O}}\left((n+d)^4, \frac{1}{\sigma_\eta^4} \frac{(n+d)^3}{\varepsilon^2}\right)$$

timesteps are sufficient to achieve error $\|\widehat{q} - q\| \leq \varepsilon$. Several remarks are in order. First, while the $(n+d)^4$ burn-in is sub-optimal, the $(n+d)^3/\varepsilon^2$ dependence is likely sharp as suggested by the asymptotic results in the later part of this chapter. We leave improving the polynomial dependence of the burn-in period to future work.

Before we turn to the proof of Theorem 6.1.1, we remark that it rests on top of several recent advances. First, we build off the work of Abbasi-Yadkori et al. [4] to derive a new basic inequality for LSTD-Q which serves as a starting point for the analysis. Next, we combine the small-ball techniques of Simchowitz et al. [105] with the self-normalized martingale inequalities of Abbasi-Yadkori et al. [2]. While an analysis of LSTD-Q is presented in Abbasi-Yadkori et al. [4] (which builds on the analysis for LSTD from Tu and Recht [117]), a direct application of their result yields a $1/\sigma_\eta^8$ dependence; the use of self-normalized inequalities is necessary in order to reduce this dependence to $1/\sigma_\eta^4$.

We now turn to the proof. Because of stability, we have that P_t converges to a limit $P_\infty = \mathsf{dlyap}((A + BK_{\text{play}})^\mathsf{T}, \sigma_w^2 I + \sigma_\eta^2 BB^\mathsf{T})$, where P_t is:

$$P_t := \sum_{k=0}^{t-1} (A + BK_{\text{play}})^k (\sigma_w^2 I + \sigma_\eta BB^\mathsf{T})((A + BK_{\text{play}})^\mathsf{T})^k \,.$$

The covariance of x_t for $t \geq 1$ is:

$$\text{Cov}(x_t) = \Sigma_t := P_t + (A + BK_{\text{play}})^t \Sigma_0 ((A + BK_{\text{play}})^\mathsf{T})^t \,.$$

We define the following data matrices:

$$\Phi = \begin{bmatrix} -\phi_1^\mathsf{T} - \\ \vdots \\ -\phi_T^\mathsf{T} - \end{bmatrix}, \quad \Psi_+ = \begin{bmatrix} -\psi_2^\mathsf{T} - \\ \vdots \\ -\psi_{T+1}^\mathsf{T} - \end{bmatrix}, \quad c = (c_1, ..., c_T)^\mathsf{T}, \quad F = \begin{bmatrix} -f^\mathsf{T} - \\ \vdots \\ -f^\mathsf{T} - \end{bmatrix}.$$

With this notation, the LSTD-Q estimator is:

$$\widehat{q} = \left(\Phi^\mathsf{T}(\Phi - \Psi_+ + F)\right)^\dagger \Phi^\mathsf{T} c \,.$$

Next, let Ξ be the matrix:

$$\Xi = \begin{bmatrix} -\mathbb{E}[\phi(x_2, K_{\text{eval}} x_2) | x_1, u_1]^\mathsf{T} - \\ \vdots \\ -\mathbb{E}[\phi(x_{T+1}, K_{\text{eval}} x_{T+1}) | x_T, u_T]^\mathsf{T} - \end{bmatrix}.$$

For what follows, we let the notation $\otimes_s$ denote the *symmetric* Kronecker product. See Schäcke [100] for more details. The following lemma, based on Lemma 4.1 of Abbasi-Yadkori et al. [4], gives us a starting point for analysis. Recall that $q = \text{svec}(Q)$ and Q is the matrix which parameterizes the Q-function for K_{eval}.

Lemma 6.1.2 (Lemma 4.1, [4]). *Let* $L := \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}$. *Suppose that* Φ *has full column rank, and that*

$$\frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I - L \otimes_s L)} \leq 1/2\,.$$

Then we have:

$$\|\widehat{q} - q\| \leq 2\frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)q\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I - L \otimes_s L)}\,. \tag{6.1.8}$$

Proof. By the Bellman equation (6.1.1), we have the identity:

$$\Phi q = c + (\Xi - F)q$$

By the definition of $\widehat{q}$, we have the identity:

$$\Phi\widehat{q} = P_\Phi(c + (\Psi_+ - F)\widehat{q})\,,$$

where $P_\Phi = \Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}$ is the orthogonal projector onto the columns of Φ. Combining these two identities gives us:

$$P_\Phi(\Phi - \Xi + F)(q - \widehat{q}) = P_\Phi(\Xi - \Psi_+)\widehat{q}\,.$$

Next, the i-th row of $\Phi - \Xi + F$ is:

$$\begin{aligned}
&\text{svec}\left(\begin{bmatrix} x_i \\ u_i \end{bmatrix}\begin{bmatrix} x_i \\ u_i \end{bmatrix}^\mathsf{T} - \mathbb{E}\left[\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}\tilde{x}\tilde{x}^\mathsf{T}\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T}\,\middle|\, x_i, u_i\right] + \sigma_w^2\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T}\right) \\
&= \text{svec}\left(\begin{bmatrix} x_i \\ u_i \end{bmatrix}\begin{bmatrix} x_i \\ u_i \end{bmatrix}^\mathsf{T} - L\begin{bmatrix} x_i \\ u_i \end{bmatrix}\begin{bmatrix} x_i \\ u_i \end{bmatrix}^\mathsf{T} L^\mathsf{T}\right) \\
&= (I - L \otimes_s L)\phi(x_i, u_i)\,.
\end{aligned}$$

Therefore, $\Phi - \Xi + F = \Phi(I - L \otimes_s L)^\mathsf{T}$. Combining with the above identity:

$$\Phi(I - L \otimes_s L)^\mathsf{T}(q - \widehat{q}) = P_\Phi(\Xi - \Psi_+)\widehat{q}\,.$$

Because Φ has full column rank, this identity implies that:

$$(I - L \otimes_s L)^\mathsf{T}(q - \widehat{q}) = (\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}(\Xi - \Psi_+)\widehat{q}\,.$$

Using the inequalities:

$$\begin{aligned}
&\|(I - L \otimes_s L)^\mathsf{T}(q - \widehat{q})\| \geq \sigma_{\min}((I - L \otimes_s L))\|q - \widehat{q}\|\,, \\
&(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}(\Xi - \Psi_+)\widehat{q} \leq \frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\widehat{q}\|}{\lambda_{\min}((\Phi^\mathsf{T}\Phi)^{-1/2})} = \frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\widehat{q}\|}{\sigma_{\min}(\Phi)}\,,
\end{aligned}$$

we obtain:

$$\|q - \widehat{q}\| \leq \frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\widehat{q}\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I - L \otimes_s L)}.$$

Next, let $\Delta = q - \widehat{q}$. By triangle inequality:

$$\|\Delta\| \leq \frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\|\|\Delta\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I - L \otimes_s L)} + \frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)q\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I - L \otimes_s L)}.$$

The claim now follows. □

In order to apply Lemma 6.1.2, we first bound the minimum singular value $\sigma_{\min}(\Phi)$. We do this using the small-ball argument of Simchowitz et al. [105].

Proposition 6.1.3. *Given an arbitrary vector $y \in \mathcal{S}^{n+d-1}$, define the process $Z_t := \langle \phi_t, y\rangle$, the filtration $\mathcal{F}_t := \sigma(\{u_i, w_{i-1}\}_{i=0}^t)$, and the matrix $C := \begin{bmatrix} I & 0 \\ K_{\text{play}} & I \end{bmatrix}\begin{bmatrix} \sigma_w I & 0 \\ 0 & \sigma_\eta I \end{bmatrix}$. Then $(Z_t)_{t\geq 1}$ satisfies the $(1, \sigma_{\min}^2(C), 1/324)$ block martingale small-ball (BMSB) condition from Definition 2.1 of Simchowitz et al. [105]. That is, almost surely, we have:*

$$\mathbb{P}(|Z_{t+1}| \geq \sigma_{\min}^2(C)|\mathcal{F}_t) \geq 1/324.$$

Proof. Let $Y := \text{smat}(y)$ and $\mu_t := Ax_t + Bu_t$. We have that:

$$\begin{bmatrix} x_{t+1} \\ u_{t+1} \end{bmatrix} = \begin{bmatrix} I \\ K_{\text{play}} \end{bmatrix}\mu_t + \begin{bmatrix} I & 0 \\ K_{\text{play}} & I \end{bmatrix}\begin{bmatrix} w_t \\ \eta_{t+1} \end{bmatrix}.$$

Therefore:

$$\begin{aligned}\langle \phi_{t+1}, y\rangle &= \begin{bmatrix} x_{t+1} \\ u_{t+1} \end{bmatrix}^\mathsf{T} Y \begin{bmatrix} x_{t+1} \\ u_{t+1} \end{bmatrix} \\ &= \left(\begin{bmatrix} I \\ K_{\text{play}} \end{bmatrix}\mu_t + \begin{bmatrix} I & 0 \\ K_{\text{play}} & I \end{bmatrix}\begin{bmatrix} w_t \\ \eta_{t+1} \end{bmatrix}\right)^\mathsf{T} Y \left(\begin{bmatrix} I \\ K_{\text{play}} \end{bmatrix}\mu_t + \begin{bmatrix} I & 0 \\ K_{\text{play}} & I \end{bmatrix}\begin{bmatrix} w_t \\ \eta_{t+1} \end{bmatrix}\right),\end{aligned}$$

which is clearly a Gaussian polynomial of degree 2 given $\mathcal{F}_t$. Hence by Gaussian hypercontractivity results (see e.g. [20]), we have that almost surely:

$$\mathbb{E}[|Z_{t+1}|^4|\mathcal{F}_t] \leq 81\mathbb{E}[|Z_{t+1}|^2|\mathcal{F}_t]^2.$$

Hence we can invoke the Paley-Zygmund inequality to conclude that for any $\theta \in (0, 1)$, almost surely we have:

$$\mathbb{P}(|Z_{t+1}| \geq \sqrt{\theta\mathbb{E}[|Z_{t+1}|^2|\mathcal{F}_t]}|\mathcal{F}_t) \geq (1-\theta)^2\frac{\mathbb{E}[|Z_{t+1}|^2|\mathcal{F}_t]^2}{\mathbb{E}[|Z_{t+1}|^4|\mathcal{F}_t]} \geq \frac{(1-\theta)^2}{81}.$$

We now state an useful proposition.

Proposition 6.1.4. *Let μ, C, Y be fixed and $g \sim \mathcal{N}(0, I)$. We have that:*

$$\mathbb{E}[((\mu + Cg)^\mathsf{T} Y(\mu + Cg))^2] \geq 2\|C^\mathsf{T} YC\|_F^2 .$$

Proof. Let $Z := (\mu + Cg)^\mathsf{T} Y(\mu + Cg)$. We know that $\mathbb{E}[Z^2] \geq \mathbb{E}[(Z - \mathbb{E}[Z])^2]$. A quick computation yields that $\mathbb{E}[Z] = \mu^\mathsf{T} Y\mu + \mathrm{tr}(C^\mathsf{T} YC)$. Hence

$$Z - \mathbb{E}[Z] = g^\mathsf{T} C^\mathsf{T} YCg - \mathrm{tr}(C^\mathsf{T} YC) + 2\mu^\mathsf{T} YCg .$$

Therefore,

$$\mathbb{E}[(Z - \mathbb{E}[Z])^2] \geq \mathbb{E}[(g^\mathsf{T} C^\mathsf{T} YCg - \mathrm{tr}(C^\mathsf{T} YC))^2] = 2\|C^\mathsf{T} YC\|_F^2 .$$

□

Invoking Proposition 6.1.4 and using basic properties of the Kronecker product, we have that:

$$\mathbb{E}[Z_{t+1}^2 | \mathcal{F}_t] \geq 2\|C^\mathsf{T} YC\|_F^2 = 2\|(C^\mathsf{T} \otimes C^\mathsf{T})y\|^2 \geq 2\sigma_{\min}^2(C^\mathsf{T} \otimes C^\mathsf{T}) = 2\sigma_{\min}^4(C) .$$

The claim now follows by setting $\theta = 1/2$. □

With the BMSB bound in place, we can now utilize Proposition 2.5 of Simchowitz et al. [105] to obtain the following lower bound on the minimum singular value $\sigma_{\min}(\Phi)$.

Proposition 6.1.5. *Fix $\delta \in (0, 1)$. Suppose that $\sigma_\eta \leq \sigma_w$, and that T exceeds:*

$$324^2 \cdot 8 \left((n + d)^2 \log \left(1 + \frac{20736\sqrt{3}}{\sqrt{\delta}} \frac{(1 + \|K_{\mathrm{play}}\|^2)^2(\tau^2\rho^2 n\|\Sigma_0\| + \mathrm{tr}(P_\infty))}{\sigma_\eta^2} \right) + \log(2/\delta) \right) . \tag{6.1.9}$$

Suppose also that $A + BK_{\mathrm{play}}$ is (τ, ρ)-stable. Then we have with probability at least $1 - \delta$,

$$\sigma_{\min}(\Phi) \geq \frac{\sigma_\eta^2}{1296\sqrt{8}} \frac{1}{1 + \|K_{\mathrm{play}}\|^2} \sqrt{T} .$$

We also have with probability at least $1 - \delta$,

$$\|\Phi^\mathsf{T} \Phi\| \leq \frac{12T}{\delta}(1 + \|K_{\mathrm{play}}\|^2)^2(\tau^2\rho^2 n\|\Sigma_0\| + \mathrm{tr}(P_\infty))^2 .$$

Proof. We first compute a crude upper bound on $\|\Phi\|$ using Markov's inequality:

$$\mathbb{P}(\|\Phi\|^2 \geq t^2) = \frac{\mathbb{E}[\lambda_{\max}(\Phi^\mathsf{T} \Phi)]}{t^2} \leq \frac{\mathrm{tr}(\mathbb{E}[\Phi^\mathsf{T} \Phi])}{t^2} .$$

Now we upper bound $\mathbb{E}[\|\phi_t\|^2]$. Letting $z_t = (x_t, u_t)$, we have that $\mathbb{E}[\|\phi_t\|^2] = \mathbb{E}[\|z_t\|^4] \leq 3(\mathbb{E}[\|z_t\|^2])^2$. We bound $\mathbb{E}[\|z_t\|^2] \leq (1 + \|K_{\text{play}}\|^2)\operatorname{tr}(\Sigma_t) + \sigma_\eta^2 d$, and therefore:

$$\begin{aligned}\sqrt{\mathbb{E}[\|\phi_t\|^2]} &\leq \sqrt{3}((1 + \|K_{\text{play}}\|^2)\operatorname{tr}(\Sigma_t) + \sigma_\eta^2 d) \\ &\leq \sqrt{3}((1 + \|K_{\text{play}}\|^2)(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty)) + \sigma_\eta^2 d) \\ &\leq 2\sqrt{3}(1 + \|K_{\text{play}}\|^2)(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty))\,.\end{aligned}$$

Above, the last inequality holds because $\sigma_\eta^2 d \leq \sigma_w^2 n \leq \operatorname{tr}(P_\infty)$. Therefore, we have from Markov's inequality:

$$\mathbb{P}\left(\|\Phi\| \geq \frac{\sqrt{T}}{\sqrt{\delta}} 2\sqrt{3}(1 + \|K_{\text{play}}\|^2)(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty))\right) \leq \delta\,.$$

Fix an $\varepsilon > 0$, and let $\mathcal{N}(\varepsilon)$ denote an ε-net of the unit sphere $\mathcal{S}^{(n+d)(n+d+1)/2-1}$. Next, by Proposition 2.5 of Simchowitz et al. [105] and a union bound over $\mathcal{N}(\varepsilon)$:

$$\mathbb{P}\left(\min_{v\in\mathcal{N}(\varepsilon)}\|\Phi v\| \geq \frac{\sigma_{\min}^2(C)}{324\sqrt{8}}\sqrt{T}\right) \geq 1 - (1 + 2/\varepsilon)^{(n+d)^2} e^{-\frac{T}{324^2\cdot 8}}\,.$$

Now set

$$\varepsilon = \frac{\sqrt{\delta}}{5184\sqrt{3}} \frac{\sigma_{\min}^2(C)}{(1 + \|K_{\text{play}}\|^2)(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty))}\,,$$

and observe that as long as T exceeds:

$$324^2 \cdot 8\left((n+d)^2 \log\left(1 + \frac{10368\sqrt{3}}{\sqrt{\delta}} \frac{(1 + \|K_{\text{play}}\|^2)(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty))}{\sigma_{\min}^2(C)}\right) + \log(2/\delta)\right)\,,$$

we have that $\mathbb{P}\left(\min_{v\in\mathcal{N}(\varepsilon)}\|\Phi v\| \geq \frac{\sigma_{\min}^2(C)}{324\sqrt{8}}\sqrt{T}\right) \geq 1 - \delta/2$. To conclude, observe that:

$$\sigma_{\min}(\Phi) = \inf_{\|v\|=1}\|\Phi v\| \geq \min_{v\in\mathcal{N}(\varepsilon)}\|\Phi v\| - \|\Phi\|\varepsilon\,,$$

and union bound over the two events. To conclude the proof, note that Lemma F.6 in Dean et al. [32] yields that $\sigma_{\min}^2(C) \geq \frac{\sigma_\eta^2}{2}\frac{1}{1+\|K_{\text{play}}\|^2}$ since $\sigma_\eta \leq \sigma_w$. □

We now turn our attention to upper bounding the self-normalized martingale terms:

$$\|(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}(\Xi - \Psi_+)\| \text{ and } \|(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}(\Xi - \Psi_+)q\|\,.$$

Our main tool here will be the self-normalized tail bounds of Abbasi-Yadkori et al. [2].

Lemma 6.1.6 (Corollary 1, [2]). *Let $\{\mathcal{F}_t\}$ be a filtration. Let $\{x_t\}$ be a $\mathbb{R}^{d_1}$ process that is adapted to $\{\mathcal{F}_t\}$ and let $\{w_t\}$ be a $\mathbb{R}^{d_2}$ martingale difference sequence that is adapted to $\{\mathcal{F}_t\}$. Let V be a fixed positive definite $d_1 \times d_1$ matrix and define:*

$$\bar{V}_t = V + \sum_{s=1}^{t} x_s x_s^\mathsf{T} \,, \quad S_t = \sum_{s=1}^{t} x_s w_{s+1}^\mathsf{T} \,.$$

(a) Suppose for any fixed unit $h \in \mathbb{R}^{d_2}$ we have that $\langle w_t, h \rangle$ is conditionally R-sub-Gaussian, that is:

$$\forall \lambda \in \mathbb{R}, t \geq 1 \,, \;\; \mathbb{E}[e^{\lambda \langle w_{t+1}, h \rangle} | \mathcal{F}_t] \leq e^{\frac{\lambda^2 R^2}{2}} \,.$$

We have that with probability at least $1-\delta$, for all $t \geq 1$,

$$\|\bar{V}_t^{-1/2} S_t\|^2 \leq 8R^2 \left(d_2 \log 5 + \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \right) \,.$$

(b) Now suppose that $\bar{\delta}$ satisfies the condition:

$$\sum_{s=2}^{T+1} \mathbb{P}(\|w_s\| > R) \leq \bar{\delta} \,.$$

Then with probability at least $1 - \delta - \bar{\delta}$, for all $1 \leq t \leq T$,

$$\|\bar{V}_t^{-1/2} S_t\|^2 \leq 32R^2 \left(d_2 \log 5 + \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \right) \,.$$

Proof. Fix a unit $h \in \mathbb{R}^{d_2}$. By Corollary 1 of Abbasi-Yadkori et al. [2], we have with probability at least $1-\delta$,

$$\|\bar{V}_t^{-1/2} S_t h\|^2 \leq 2R^2 \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \,, \;\; 1 \leq t \leq T \,.$$

A standard covering argument yields that:

$$\|\bar{V}_t^{-1/2} S_t\|^2 \leq 4 \max_{h \in \mathcal{N}(1/2)} \|\bar{V}_t^{-1/2} S_t h\|^2 \,.$$

Union bounding over $\mathcal{N}(1/2)$, we obtain that:

$$\begin{aligned} \|\bar{V}_t^{-1/2} S_t\|^2 &\leq 8R^2 \log \left(5^{d_2} \frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \\ &= 8R^2 \left(d_2 \log 5 + \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \right) \,. \end{aligned}$$

This yields (a).

For (b), we use a simple stopping time argument to handle truncation. Define the stopping time $\tau := \inf\{t \geq 1 : \|w_t\| > R\}$ and the truncated process $\tilde{w}_t := w_t \mathbf{1}_{\tau \geq t}$. Because τ is a stopping time, this truncated process $\{\tilde{w}_t\}$ remains a martingale difference sequence. Define $Z_t = \sum_{s=1}^{t} x_s \tilde{w}_{s+1}^{\mathsf{T}}$. For any $\ell > 0$ we observe that:

$$\begin{aligned}
&\mathbb{P}(\exists 1 \leq t \leq T : \|\bar{V}_t^{-1/2} S_t\| > \ell) \\
&\leq \mathbb{P}(\{\exists 1 \leq t \leq T : \|\bar{V}_t^{-1/2} S_t\| > \ell\} \cap \{\tau > T+1\}) + \mathbb{P}(\tau \leq T+1) \\
&= \mathbb{P}(\{\exists 1 \leq t \leq T : \|\bar{V}_t^{-1/2} Z_t\| > \ell\} \cap \{\tau > T+1\}) + \mathbb{P}(\tau \leq T+1) \\
&\leq \mathbb{P}(\exists t \geq 1 : \|\bar{V}_t^{-1/2} Z_t\| > \ell) + \mathbb{P}(\tau \leq T+1) \\
&\leq \mathbb{P}(\exists t \geq 1 : \|\bar{V}_t^{-1/2} Z_t\| > \ell) + \sum_{s=2}^{T+1} \mathbb{P}(\|w_s\| > R) \\
&\leq \mathbb{P}(\exists t \geq 1 : \|\bar{V}_t^{-1/2} Z_t\| > \ell) + \bar{\delta} \,.
\end{aligned}$$

Now set $\ell = 32R^2 \left(d_2 \log 5 + \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right) \right)$ and using the fact that a R bounded random variable is $2R$-sub-Gaussian, the claim now follows by another application of Corollary 1 from [2]. □

With Lemma 6.1.6 in place, we are ready to bound the martingale difference terms.

Proposition 6.1.7. *Suppose the hypothesis of Proposition 6.1.5 hold. With probability at least* $1 - \delta$,

$$\begin{aligned}
\|(\Phi^{\mathsf{T}}\Phi)^{-1/2}\Phi^{\mathsf{T}}(\Xi - \Psi_+)q\| &\leq (n+d)\sigma_w \sqrt{\tau^2\rho^4\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}(1 + \|K_{\text{eval}}\|^2)\|Q\|_F \\
&\quad \times \text{polylog}(n, \tau, \|\Sigma_0\|, \|P_\infty\|, \|K_{\text{play}}\|, T/\delta, 1/\sigma_\eta) \,, \\
\|(\Phi^{\mathsf{T}}\Phi)^{-1/2}\Phi^{\mathsf{T}}(\Xi - \Psi_+)\| &\leq (n+d)^2\sigma_w \sqrt{\tau^2\rho^4\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}(1 + \|K_{\text{eval}}\|^2) \\
&\quad \times \text{polylog}(n, \tau, \|\Sigma_0\|, \|P_\infty\|, \|K_{\text{play}}\|, T/\delta, 1/\sigma_\eta) \,.
\end{aligned}$$

Proof. For the proof, constants c, c_i will denote universal constants. Define two matrices:

$$\begin{aligned}
V_1 &:= c_1 \frac{\sigma_\eta^4}{(1 + \|K_{\text{play}}\|^2)^2} T \cdot I \,, \\
V_2 &:= c_2 \frac{T}{\delta}(1 + \|K_{\text{play}}\|^2)^2(\tau^2\rho^2 n\|\Sigma_0\| + \text{tr}(P_\infty))^2 \cdot I \,.
\end{aligned}$$

By Proposition 6.1.5, with probability at least $1 - \delta/2$, we have that:

$$V_1 \preceq \Phi^{\mathsf{T}}\Phi \preceq V_2 \,.$$

Call this event $\mathcal{E}_1$.

Next, we have:

$$\begin{aligned}
&\mathbb{E}[x_{t+1}x_{t+1}^\mathsf{T}|x_t,u_t] - x_{t+1}x_{t+1}^\mathsf{T} \\
&= \mathbb{E}[(Ax_t+Bu_t+w_t)(Ax_t+Bu_t+w_t)^\mathsf{T}|x_t,u_t] - (Ax_t+Bu_t+w_t)(Ax_t+Bu_t+w_t)^\mathsf{T} \\
&= (Ax_t+Bu_t)(Ax_t+Bu_t)^\mathsf{T} + \sigma_w^2 I \\
&\quad - (Ax_t+Bu_t)(Ax_t+Bu_t)^\mathsf{T} - (Ax_t+Bu_t)w_t^\mathsf{T} - w_t(Ax_t+Bu_t)^\mathsf{T} - w_t w_t^\mathsf{T} \\
&= \sigma_w^2 I - w_t w_t^\mathsf{T} - (Ax_t+Bu_t)w_t^\mathsf{T} - w_t(Ax_t+Bu_t)^\mathsf{T} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\psi_{t+1}|x_t,u_t] - \psi_{t+1} \\
&= \operatorname{svec}\left(\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} (\sigma_w^2 I - w_t w_t^\mathsf{T} - (Ax_t+Bu_t)w_t^\mathsf{T} - w_t(Ax_t+Bu_t)^\mathsf{T}) \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T}\right) .
\end{aligned}$$

Taking the inner product of this term with q,

$$\begin{aligned}
&(\mathbb{E}[\psi_{t+1}|x_t,u_t] - \psi_{t+1})^\mathsf{T} q \\
&= \operatorname{tr}\left((\sigma_w^2 I - w_t w_t^\mathsf{T} - (Ax_t+Bu_t)w_t^\mathsf{T} - w_t(Ax_t+Bu_t)^\mathsf{T}) \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}\right) \\
&= \operatorname{tr}\left((\sigma_w^2 I - w_t w_t^\mathsf{T}) \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}\right) - 2w_t^\mathsf{T} \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} (Ax_t+Bu_t) .
\end{aligned}$$

By the Hanson-Wright inequality (see e.g. Rudelson and Vershynin [97]), with probability at least $1-\delta/T$,

$$\left|\operatorname{tr}\left((\sigma_w^2 I - w_t w_t^\mathsf{T}) \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}\right)\right| \le c_1 \sigma_w^2 (1+\|K_{\text{eval}}\|^2)\|Q\|_F \log(T/\delta) .$$

Now, let $L_{\text{play}} := A + BK_{\text{play}}$. By Proposition 4.7 in Tu and Recht [117], with probability at least $1-\delta/T$,

$$\begin{aligned}
&\left|w_t^\mathsf{T} \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T} Q \begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} (Ax_t+Bu_t)\right| \\
&\le c_1 \sigma_w (1+\|K_{\text{eval}}\|^2)\sqrt{\|L_{\text{play}}^{t+1}\Sigma_0 (L_{\text{play}}^{t+1})^\mathsf{T}\| + \|L_{\text{play}} P_t L_{\text{play}}^\mathsf{T}\| + \sigma_\eta^2\|B\|^2}\|Q\|_F \log(T/\delta) \\
&\le c_1 \sigma_w (1+\|K_{\text{eval}}\|^2)\sqrt{\tau^2\rho^{2(t+1)}\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}\|Q\|_F \log(T/\delta) ,
\end{aligned}$$

where the inequality above comes from $P_t \preceq P_\infty$ and $L_{\text{play}} P_\infty L_{\text{play}}^\mathsf{T} \preceq P_\infty$. Therefore, we have:

$$\begin{aligned}
&|(\mathbb{E}[\psi_{t+1}|x_t, u_t] - \psi_{t+1})^\mathsf{T} v| \\
&\le c_2(\sigma_w^2 + \sigma_w\sqrt{\tau^2\rho^{2(t+1)}\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2})(1 + \|K_{\text{eval}}\|^2)\|Q\|_F \log(T/\delta) \\
&\le c_3\sigma_w\sqrt{\tau^2\rho^{2(t+1)}\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}(1 + \|K_{\text{eval}}\|^2)\|Q\|_F \log(T/\delta)\,.
\end{aligned}$$

The last inequality holds because $P_\infty \succeq \sigma_w^2 I$ and hence $\sigma_w \le \|P_\infty\|^{1/2}$. Therefore we can set

$$R = c_3\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}(1 + \|K_{\text{eval}}\|^2)\|Q\|_F \log(T/\delta)\,,$$

and invoke Lemma 6.1.6 to conclude that with probability at least $1 - \delta/2$,

$$\|(V_1 + \Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)v\| \le c_4(n+d)R + c_5 R\sqrt{\log(\det((V_1 + \Phi^\mathsf{T}\Phi)V_1^{-1})^{1/2}/\delta)}\,.$$

Call this event $\mathcal{E}_2$.

For the remainder of the proof we work on $\mathcal{E}_1 \cap \mathcal{E}_2$, which has probability at least $1 - \delta$. Since $\Phi^\mathsf{T}\Phi \succeq V_1$, we have that $(\Phi^\mathsf{T}\Phi)^{-1} \le 2(V_1 + \Phi^\mathsf{T}\Phi)^{-1}$. Therefore, by another application of Lemma 6.1.6:

$$\begin{aligned}
&\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\| \\
&\le \sqrt{2}\|(V_1 + \Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi - \Psi_+)\| \\
&\le c_6(n+d)R + c_7 R\sqrt{\log(\det((V_1 + \Phi^\mathsf{T}\Phi)V_1^{-1})^{1/2}/\delta)} \\
&\le c_6(n+d)R + c_7 R\sqrt{\log(\det((V_1 + V_2)V_1^{-1})^{1/2}/\delta)} \\
&\le c_6(n+d)R + c_8 R(n+d)\sqrt{\log\left(\frac{(1 + \|K_{\text{play}}\|^2)^4}{\delta}\frac{(\tau^2\rho^2 n\|\Sigma_0\| + \operatorname{tr}(P_\infty))^2}{\sigma_\eta^4}\right)} \\
&\le c(n+d)R\,\text{polylog}(n, \tau, \|\Sigma_0\|, \|P_\infty\|, \|K_{\text{play}}\|, 1/\delta, 1/\sigma_\eta)\,.
\end{aligned}$$

Next, we bound:

$$\begin{aligned}
&\|\mathbb{E}[\psi_{t+1}|x_t, u_t] - \psi_{t+1}\| \\
&\le \left\|\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}(\sigma_w^2 I - w_t w_t^\mathsf{T})\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T}\right\|_F + \left\|\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix} w_t(Ax_t + Bu_t)^\mathsf{T}\begin{bmatrix} I \\ K_{\text{eval}} \end{bmatrix}^\mathsf{T}\right\|_F \\
&\le (1 + \|K_{\text{eval}}\|^2)(\|\sigma_w^2 I - w_t w_t^\mathsf{T}\|_F + \|w_t(Ax_t + Bu_t)^\mathsf{T}\|_F)\,.
\end{aligned}$$

Now, by standard Gaussian concentration results, with probability $1 - \delta/T$,

$$\|\sigma_w^2 I - w_t w_t^\mathsf{T}\|_F \le c\sigma_w^2(n + \log(T/\delta))\,,$$

and also

$$\begin{aligned}
&\|w_t(Ax_t + Bu_t)^\mathsf{T}\|_F \\
&\le c\sigma_w(\sqrt{n} + \sqrt{\log(T/\delta)})(\sqrt{\mathrm{tr}(L_{\mathrm{play}}^{t+1}\Sigma_0(L_{\mathrm{play}}^{t+1})^\mathsf{T}) + \mathrm{tr}(L_{\mathrm{play}}P_tL_{\mathrm{play}}^\mathsf{T}) + \sigma_\eta^2\|B\|_F^2} \\
&\quad + \sqrt{\|L_{\mathrm{play}}^{t+1}\Sigma_0(L_{\mathrm{play}}^{t+1})^\mathsf{T}\| + \|L_{\mathrm{play}}P_tL_{\mathrm{play}}^\mathsf{T}\| + \sigma_\eta^2\|B\|^2}\sqrt{\log(T/\delta)}) \\
&\le c\sigma_w(n+d)\sqrt{\tau^2\rho^4\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|}\log(T/\delta)\,.
\end{aligned}$$

Therefore, with probability $1 - \delta/T$,

$$\begin{aligned}
&\|\mathbb{E}[\psi_{t+1}|x_t, u_t] - \psi_{t+1}\| \\
&\le c(1 + \|K_{\mathrm{eval}}\|^2)(n+d)\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2}\log(T/\delta)\,.
\end{aligned}$$

□

We are now in a position to prove Theorem 6.1.1. We first observe that we can lower bound $\sigma_{\min}(I - L \otimes_s L)$ using the (τ, ρ)-stability of $A + BK_{\mathrm{eval}}$. This is because for $k \ge 1$,

$$\begin{aligned}
\|L^k\| &= \left\| \begin{bmatrix} I \\ K_{\mathrm{eval}} \end{bmatrix} (A + BK_{\mathrm{eval}})^{k-1} \begin{bmatrix} A & B \end{bmatrix} \right\| \\
&\le 2\|K_{\mathrm{eval}}\|_+ \|\begin{bmatrix} A & B \end{bmatrix}\| \tau\rho^{k-1} \\
&\le \frac{2\|K_{\mathrm{eval}}\|_+ \max\{1, \sqrt{\|A\|^2 + \|B\|^2}\}}{\rho}\tau \cdot \rho^k\,.
\end{aligned}$$

Hence we see that L is $(\frac{2\|K_{\mathrm{eval}}\|_+ \max\{1,\sqrt{\|A\|^2+\|B\|^2}\}}{\rho}\tau, \rho)$-stable. Next, we know that $\sigma_{\min}(I - L \otimes_s L) = \frac{1}{\|(I-L\otimes_s L)^{-1}\|}$. Therefore, for any unit norm v,

$$\begin{aligned}
\|(I - L \otimes_s L)^{-1}v\| &= \|(I - L \otimes_s L)^{-1}\mathrm{svec}(\mathrm{smat}(v))\| = \|\mathsf{dlyap}(L^\mathsf{T}, \mathrm{smat}(v))\|_F \\
&\le \frac{4\|K_{\mathrm{eval}}\|_+^2(\|A\|^2 + \|B\|^2)_+\tau^2}{\rho^2(1-\rho^2)}\,.
\end{aligned}$$

Here, the last inequality uses Proposition 4.0.2. Hence we have the bound:

$$\sigma_{\min}(I - L \otimes_s L) \ge \frac{\rho^2(1-\rho^2)}{4\|K_{\mathrm{eval}}\|_+^2(\|A\|^2 + \|B\|^2)_+\tau^2}\,.$$

By Proposition 6.1.5, as long as $T \ge \widetilde{\mathcal{O}}(1)(n+d)^2$ with probability at least $1 - \delta/2$:

$$\sigma_{\min}(\Phi) \ge c\frac{\sigma_\eta^2}{\|K_{\mathrm{play}}\|_+^2}\sqrt{T}\,.$$

By Proposition 6.1.7, with probability at least $1-\delta/2$:

$$\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi-\Psi_+)q\| \le (n+d)\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2}\|K_{\mathrm{eval}}\|_+^2\|Q^{K_{\mathrm{eval}}}\|_F\widetilde{\mathcal{O}}(1)\,,$$
$$\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi-\Psi_+)\| \le (n+d)^2\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2}\|K_{\mathrm{eval}}\|_+^2\widetilde{\mathcal{O}}(1)\,.$$

We first check the condition

$$\frac{\|(\Phi^\mathsf{T}\Phi)^{-1/2}\Phi^\mathsf{T}(\Xi-\Psi_+)\|}{\sigma_{\min}(\Phi)\sigma_{\min}(I-L\otimes_s L)} \le 1/2\,,$$

from Lemma 6.1.2. A sufficient condition is that T satisfies:

$$\begin{aligned}
T &\ge \widetilde{\mathcal{O}}(1)\frac{\|K_{\mathrm{play}}\|_+^4}{\sigma_\eta^4}\cdot(n+d)^4\sigma_w^2(\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2)\\
&\quad\times\|K_{\mathrm{eval}}\|_+^4\cdot\frac{\|K_{\mathrm{eval}}\|_+^4(\|A\|^2+\|B\|^2)_+^2\tau^4}{\rho^4(1-\rho^2)^2}\\
&=\widetilde{\mathcal{O}}(1)\frac{\tau^4}{\rho^4(1-\rho^2)^2}\frac{(n+d)^4}{\sigma_\eta^4}\sigma_w^2(\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2)\\
&\quad\times\|K_{\mathrm{play}}\|_+^4\|K_{\mathrm{eval}}\|_+^8(\|A\|^4+\|B\|^4)_+\,.
\end{aligned}$$

Once this condition on T is satisfied, then we have that the error $\|\widehat{q}-q\|$ is bounded by:

$$\begin{aligned}
&\widetilde{\mathcal{O}}(1)\frac{\|K_{\mathrm{play}}\|_+^2}{\sigma_\eta^2\sqrt{T}}\cdot(n+d)\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2}\\
&\quad\times\|K_{\mathrm{eval}}\|_+^2\|Q^{K_{\mathrm{eval}}}\|_F\cdot\frac{\|K_{\mathrm{eval}}\|_+^2(\|A\|^2+\|B\|^2)_+\tau^2}{\rho^2(1-\rho^2)}\\
&=\widetilde{\mathcal{O}}(1)\frac{\tau^2}{\rho^2(1-\rho^2)}\frac{(n+d)}{\sigma_\eta^2\sqrt{T}}\sigma_w\sqrt{\tau^2\rho^4\|\Sigma_0\|+\|P_\infty\|+\sigma_\eta^2\|B\|^2}\\
&\quad\times\|K_{\mathrm{play}}\|_+^2\|K_{\mathrm{eval}}\|_+^4(\|A\|^2+\|B\|^2)_+\|Q^{K_{\mathrm{eval}}}\|_F\,.
\end{aligned}$$

Theorem 6.1.1 now follows from Lemma 6.1.2.

6.1.3 Exact Policy Iteration for LQR

Exact policy iteration works as follows. We start with a stabilizing controller K_0 for (A,B), and let V_0 denote its associated value function. We then apply the following recursions for $t=0,1,2,\ldots$:

$$K_{t+1} = -(R+B^\mathsf{T}V_tB)^{-1}B^\mathsf{T}V_tA\,, \tag{6.1.10}$$
$$V_{t+1} = \mathsf{dlyap}(A+BK_{t+1}, S+K_{t+1}^\mathsf{T}RK_{t+1})\,. \tag{6.1.11}$$

Note that this recurrence is related to, but different from, that of *value iteration*, which starts from a PSD V_0 and recurses:

$$V_{t+1} = A^\mathsf{T} V_t A - A^\mathsf{T} V_t B (R + B^\mathsf{T} V_t B)^{-1} B^\mathsf{T} V_t A + S \,.$$

While the behavior of value iteration for LQR is well understood (see e.g. Lincoln and Rantzer [70] or Kailath et al. [55]), the behavior of policy iteration is less studied. Fazel et al. [41] show that policy iteration is equivalent to the Gauss-Newton method on the objective $J(K)$ with a specific step-size, and give a simple analysis which shows linear convergence to the optimal controller.

In this section, we present an analysis of the behavior of exact policy iteration that builds on top of the fixed-point theory from Lee and Lim [65]. A key component of our analysis is the following invariant metric δ_∞ on positive definite matrices:

$$\delta_\infty(A, B) := \|\log(A^{-1/2} B A^{-1/2})\| \,.$$

Various properties of δ_∞ are reviewed at the end of this section. We use the fixed-point analysis because it lends itself nicely to handling error in the updates, which we exploit in the next section on approximate policy iteration.

Our analysis proceeds as follows. First, we note by the matrix inversion lemma:

$$S + A^\mathsf{T}(B R^{-1} B^\mathsf{T} + V^{-1})^{-1} A = S + A^\mathsf{T} V A - A^\mathsf{T} V B (R + B^\mathsf{T} V B)^{-1} B^\mathsf{T} V A =: F(V) \,.$$

Let $V_\star$ be the unique positive definite solution to $V = F(V)$. For any positive definite V we have by Lemma 6.1.10:

$$\delta_\infty(F(V), V_\star) \leq \frac{\alpha}{\lambda_{\min}(S) + \alpha} \delta_\infty(V, V_\star) \,, \tag{6.1.12}$$

with $\alpha = \max\{\lambda_{\max}(A^\mathsf{T} V A), \lambda_{\max}(A^\mathsf{T} V_\star A)\}$. Indeed, (6.1.12) gives us another method to analyze value iteration, since it shows that the Riccati operator $F(V)$ is contractive in the δ_∞ metric. Our next result combines this contraction property with the policy iteration analysis of Bertsekas [15].

Proposition 6.1.8 (Policy Iteration for LQR)**.** *Suppose that S, R are positive definite and there exists a unique positive definite solution to the discrete algebraic Riccati equation (DARE). Let K_0 be a stabilizing policy for (A, B) and let $V_0 = \mathsf{dlyap}(A + BK_0, S + K_0^\mathsf{T} R K_0)$. Consider the following sequence of updates for $t = 0, 1, 2, \ldots$:*

$$\begin{aligned} K_{t+1} &= -(R + B^\mathsf{T} V_t B)^{-1} B^\mathsf{T} V_t A \,, \\ V_{t+1} &= \mathsf{dlyap}(A + B K_{t+1}, S + K_{t+1}^\mathsf{T} R K_{t+1}) \,. \end{aligned}$$

The following statements hold:

(i) K_t stabilizes (A, B) for all $t = 0, 1, 2, \ldots$,

(ii) $V_\star \preceq V_{t+1} \preceq V_t$ *for all* $t = 0, 1, 2, ...,$

(iii) $\delta_\infty(V_{t+1}, V_\star) \leq \rho \cdot \delta_\infty(V_t, V_\star)$ *for all* $t = 0, 1, 2, ...,$ *with* $\rho := \frac{\lambda_{\max}(A^\mathsf{T} V_0 A)}{\lambda_{\min}(S) + \lambda_{\max}(A^\mathsf{T} V_0 A)}$. *Consequently,* $\delta_\infty(V_t, V_\star) \leq \rho^t \cdot \delta_\infty(V_0, V_\star)$ *for* $t = 0, 1, 2, ...$

Proof. We first prove (i) and (ii) using the argument of Proposition 1.3 from Bertsekas [15].

Let $c(x, u) = x^\mathsf{T} S x + u^\mathsf{T} R u$, $f(x, u) = Ax + Bu$, and $V^K(x_1) = \sum_{t=1}^\infty c(x_t, u_t)$ with $x_{t+1} = f(x_t, u_t)$ and $u_t = Kx_t$. Let $V_t = V^{K_t}$. With these definitions, we have that for all x:

$$K_{t+1}x = \arg\min_u c(x, u) + V_t(f(x, u)) \, .$$

Therefore,

$$\begin{aligned}
V_t(x) &= c(x, K_t x) + V_t(f(x, K_t x)) \\
&\geq c(x, K_{t+1} x) + V_t(f(x, K_{t+1} x)) \\
&= c(x, K_{t+1} x) + c(f(x, K_{t+1} x), K_t f(x, K_{t+1} x)) + V_t(f(f(x, K_{t+1} x), K_t f(x, K_{t+1} x))) \\
&\geq c(x, K_{t+1} x) + c(f(x, K_{t+1} x), K_{t+1} f(x, K_{t+1} x)) + V_t(f(f(x, K_{t+1} x), K_{t+1} f(x, K_{t+1} x))) \\
&\vdots \\
&\geq V_{t+1}(x) \, .
\end{aligned}$$

This proves (i) and (ii).

Now, observe that by partial minimization of a strongly convex quadratic:

$$\begin{aligned}
c(x, K_{t+1} x) + V_t(f(x, K_{t+1} x)) &= \min_u c(x, u) + V_t(f(x, u)) \\
&= x^\mathsf{T}(S + A^\mathsf{T} V_t A - A^\mathsf{T} V_t B (R + B^\mathsf{T} V_t B)^{-1} B^\mathsf{T} V_t A) x \\
&= x^\mathsf{T} F(V_t) x \, .
\end{aligned}$$

Combined with the above inequalities, this shows that $V_{t+1} \preceq F(V_t) \preceq V_t$. Therefore, by (6.1.12) and Proposition 6.1.12,

$$\begin{aligned}
\delta_\infty(V_{t+1}, V_\star) &\leq \delta_\infty(F(V_t), V_\star) \\
&= \delta_\infty(F(V_t), F(V_\star)) \\
&\leq \frac{\alpha_t}{\lambda_{\min}(Q) + \alpha_t} \delta_\infty(V_t, V_\star) \, ,
\end{aligned}$$

where $\alpha_t = \max\{\lambda_{\max}(A^\mathsf{T} V_t A), \lambda_{\max}(A^\mathsf{T} V_\star A)\} = \lambda_{\max}(A^\mathsf{T} V_t A)$, since $V_\star \preceq V_t$. But since $V_t \preceq V_0$, we can upper bound $\alpha_t \leq \lambda_{\max}(A^\mathsf{T} V_0 A)$. This proves (iii). □

6.1.3.1 Properties of the Invariant Metric

Here we review relevant properties of the invariant metric $\delta_\infty(A, B) = \|\log(A^{-1/2} B A^{-1/2})\|$ over positive definite matrices.

Lemma 6.1.9 (c.f. [65]). *Suppose that A is positive semidefinite and X, Y are positive definite. Also suppose that M is invertible. We have:*

(i) $\delta_\infty(X, Y) = \delta_\infty(X^{-1}, Y^{-1}) = \delta_\infty(MXM^\mathsf{T}, MYM^\mathsf{T})$.

(ii) $\delta_\infty(A + X, A + Y) \leq \frac{\alpha}{\alpha+\beta}\delta_\infty(X, Y)$, *where* $\alpha = \max\{\lambda_{\max}(X), \lambda_{\max}(Y)\}$ *and* $\beta = \lambda_{\min}(A)$.

Lemma 6.1.10 (c.f. Theorem 4.4, [65]). *Consider the map $f(X) = A + M(B + X^{-1})^{-1}M^\mathsf{T}$, where A, B are PSD and X is positive definite. Suppose that X, Y are two positive definite matrices and A is invertible. We have:*

$$\delta_\infty(f(X), f(Y)) \leq \frac{\max\{\lambda_1(MXM^\mathsf{T}), \lambda_1(MYM^\mathsf{T})\}}{\lambda_{\min}(A) + \max\{\lambda_1(MXM^\mathsf{T}), \lambda_1(MYM^\mathsf{T})\}}\delta_\infty(X, Y)\,.$$

Proof. We first assume that M is invertible. Using the properties of δ_∞ from Lemma 6.1.9, we have:

$$\begin{aligned}
\delta_\infty(f(X), f(Y)) &= \delta_\infty(A + M(B + X^{-1})^{-1}M^\mathsf{T}, A + M(B + Y^{-1})^{-1}M^\mathsf{T}) \\
&\leq \frac{\alpha}{\lambda_{\min}(A) + \alpha}\delta_\infty(M(B + X^{-1})^{-1}M^\mathsf{T}, M(B + Y^{-1})^{-1}M^\mathsf{T}) \\
&= \frac{\alpha}{\lambda_{\min}(A) + \alpha}\delta_\infty((B + X^{-1})^{-1}, (B + Y^{-1})^{-1}) \\
&= \frac{\alpha}{\lambda_{\min}(A) + \alpha}\delta_\infty(B + X^{-1}, B + Y^{-1}) \\
&\leq \frac{\alpha}{\lambda_{\min}(A) + \alpha}\delta_\infty(X^{-1}, Y^{-1}) \\
&= \frac{\alpha}{\lambda_{\min}(A) + \alpha}\delta_\infty(X, Y)\,,
\end{aligned}$$

where $\alpha = \max\{\lambda_{\max}(M(B+X^{-1})^{-1}M^\mathsf{T}), \lambda_{\max}(M(B+X^{-1})^{-1}M^\mathsf{T})\}$. Now, we observe that:

$$B + X^{-1} \succeq X^{-1} \iff (B + X^{-1})^{-1} \preceq X\,.$$

This means that $M(B+X^{-1})^{-1}M^\mathsf{T} \preceq MXM^\mathsf{T}$ and similarly $M(B+Y^{-1})^{-1}M^\mathsf{T} \preceq MYM^\mathsf{T}$. This proves the claim when M is invertible. When M is not invertible, we use a standard limiting argument, for which the details are omitted. □

Proposition 6.1.11. *Suppose that A, B are positive definite matrices satisfying $A \succeq \mu I$, $B \succeq \mu I$. We have that:*

$$\delta_\infty(A, B) \leq \frac{\|A - B\|}{\mu}\,.$$

Proof. We have that:

$$\|A^{-1/2}BA^{-1/2}\| = \|A^{-1/2}(B-A)A^{-1/2} + I\| \leq 1 + \frac{\|B-A\|}{\mu} .$$

Taking log on both sides and using $\log(1+x) \leq x$ for $x \geq 0$ yields the claim. □

Proposition 6.1.12. *Suppose that $B \preceq A_1 \preceq A_2$ are all positive definite matrices. We have that:*

$$\delta_\infty(A_1, B) \leq \delta_\infty(A_2, B) .$$

Proof. The chain of orderings implies that:

$$I \preceq B^{-1/2}A_1B^{-1/2} \preceq B^{-1/2}A_2B^{-1/2} .$$

Therefore:

$$\delta_\infty(A_1, B) = \log \lambda_{\max}(B^{-1/2}A_1B^{-1/2}) \leq \log \lambda_{\max}(B^{-1/2}A_2B^{-1/2}) = \delta_\infty(A_2, B) .$$

Each step requires careful justification. The first equality holds because $I \preceq B^{-1/2}A_1B^{-1/2}$ and the second inequality uses the monotonicity of the scalar function $x \mapsto \log x$ on $\mathbb{R}_+$ in addition to $B^{-1/2}A_1B^{-1/2} \preceq B^{-1/2}A_2B^{-1/2}$. □

Proposition 6.1.13. *Suppose that A, B are positive definite matrices with $B \succeq A$. We have that:*

$$\|A - B\| \leq \|A\|(\exp(\delta_\infty(A, B)) - 1) .$$

Furthermore, if $\delta_\infty(A, B) \leq 1$ we have:

$$\|A - B\| \leq e\|A\|\delta_\infty(A, B) .$$

Proof. The assumption that $B \succeq A$ implies that $A^{-1/2}BA^{-1/2} \succeq I$ and that $\|A - B\| = \lambda_{\max}(B - A)$. Now observe that:

$$\begin{aligned}
\|A - B\| &= \lambda_{\max}(B - A) \\
&= \lambda_{\max}(A^{1/2}(A^{-1/2}BA^{-1/2} - I)A^{1/2}) \\
&\leq \|A\|\lambda_{\max}(A^{-1/2}BA^{-1/2} - I) \\
&= \|A\|(\lambda_{\max}(A^{-1/2}BA^{-1/2}) - 1) \\
&= \|A\|(\exp(\delta_\infty(A, B)) - 1) .
\end{aligned}$$

This yields the first claim. The second follows from the crude bound that $e^x \leq 1 + ex$ for $x \in (0, 1)$. □

6.1.4 Approximate Policy Iteration for LQR

We now turn to the analysis of approximate policy iteration. We present an analysis of a modified version of Algorithm 1, which is described in Algorithm 2. The main difference between Algorithm 1 and Algorithm 2 is that Algorithm 2 does not reuse trajectory data for each invocation of LSTD-Q unlike Algorithm 1. This makes each invocation of LSTD-Q in Algorithm 2 independent from the previous invocations, which is much more amenable to analysis. We leave open the question of analyzing data reuse in the context of LSPI. Algorithm 2 also uses the initial policy K_0 as the exploration policy throughout the algorithm unlike Algorithm 1 which delineates between the generated policies $\{K_t\}$ and the evaluation policy K_{eval}. This simplification is not fundamental to the analysis.

Algorithm 2 Least-Squares Policy Iteration (LSPI) for LQR

Require: Initial stabilizing controller K_0, N number of policy iterations, T length of rollout for estimation, σ_η^2 exploration variance, μ lower eigenvalue bound.

1: **for** $t = 0, ..., N-1$ **do**
2: Collect a trajectory $\mathcal{D}_t = \{(x_k^{(t)}, u_k^{(t)}, x_{k+1}^{(t)}\}_{k=1}^T$ using input $u_k^{(t)} = K_0 x_k^{(t)} + \eta_k^{(t)}$, with $\eta_k^{(t)} \sim \mathcal{N}(0, \sigma_\eta^2 I)$.
3: $\widehat{Q}_t = \mathsf{Proj}_\mu(\mathsf{LSTDQ}(\mathcal{D}_t, K_t))$.
4: $K_{t+1} = G(\widehat{Q}_t)$ (c.f. (1.3.9)).
5: **end for**
6: return K_N.

Before we state our main finite-sample guarantee for Algorithm 2, we review the notion of a (relative) value-function. The infinite horizon average-cost Bellman equation states that the (relative) value function V associated to a policy π satisfies the fixed-point equation:

$$\lambda + V(x) = c(x, \pi(x)) + \mathbb{E}_{x' \sim p(\cdot|x,\pi(x))}[V(x')] \,. \tag{6.1.13}$$

For a stabilizing policy K, it is well known that for LQR the value function $V(x) = x^\mathsf{T} V x$ with

$$V = \mathsf{dlyap}(A + BK, S + K^\mathsf{T} R K) \,, \quad \lambda = \langle \sigma_w^2 I, V \rangle \,.$$

Once again as we did for Q-functions, we slightly abuse notation and let V denote the value function and the matrix that parameterizes the value function. Our main result of Algorithm 1 appears in the following theorem. For simplicity, we will assume that $\|S\| \geq 1$ and $\|R\| \geq 1$.

Theorem 6.1.14. *Fix a $\delta \in (0,1)$. Let the initial policy K_0 to Algorithm 2 stabilize (A,B). Suppose the initial state $x_0 \sim \mathcal{N}(0,\Sigma_0)$ and that the excitation noise satisfies $\sigma_\eta \le \sigma_w$. Recall that the steady-state covariance of the trajectory $\{x_t\}$ is*

$$P_\infty = \mathsf{dlyap}((A+BK_0)^\mathsf{T}, \sigma_w^2 I + \sigma_\eta^2 BB^\mathsf{T}) \,.$$

Let V_0 denote the value function associated to the initial policy K_0, and $V_\star$ denote the value function associated to the optimal policy $K_\star$ for the LQR problem (1.1.2). Define the variables μ, L as:

$$\begin{aligned} \mu &:= \min\{\lambda_{\min}(S), \lambda_{\min}(R)\} \,, \\ L &:= \max\{\|S\|, \|R\|\} + 2(\|A\|^2 + \|B\|^2 + 1)\|V_0\|_+ \,. \end{aligned}$$

Fix an $\varepsilon > 0$ that satisfies:

$$\varepsilon \le 5\left(\frac{L}{\mu}\right)^2 \min\left\{1, \frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{e}, \frac{\|V_\star\|^2}{8\mu^2\log(\|V_0\|/\lambda_{\min}(V_\star))}\right\} \,. \tag{6.1.14}$$

Suppose we run Algorithm 1 for $N := N_0 + 1$ policy improvement iterations where

$$N_0 := \left\lceil (1 + L/\mu)\log\left(\frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{\varepsilon}\right)\right\rceil \,, \tag{6.1.15}$$

and we set the rollout length T to satisfy:

$$\begin{aligned} T \ge \widetilde{\mathcal{O}}(1)\max\Bigg\{ & (n+d)^2, \\ & \frac{L^2}{(1-\mu/L)^2}\left(\frac{L}{\mu}\right)^{17}\frac{(n+d)^4}{\sigma_\eta^4}\sigma_w^2(\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2), \\ & \frac{1}{\varepsilon^2}\frac{L^4}{(1-\mu/L)^2}\left(\frac{L}{\mu}\right)^{42}\frac{(n+d)^3}{\sigma_\eta^4}\sigma_w^2(\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2\|B\|^2)\Bigg\} \,. \end{aligned} \tag{6.1.16}$$

Then with probability $1-\delta$, we have that each policy K_t for $t = 1, ..., N$ stabilizes (A,B) and furthermore:

$$\|K_N - K_\star\| \le \varepsilon \,.$$

Here the $\widetilde{\mathcal{O}}(1)$ hides $\mathrm{polylog}(n, \tau, \|\Sigma_0\|, \|P_\infty\|, L/\mu, T/\delta, N_0, 1/\sigma_\eta)$ *factors.*

Theorem 6.1.14 states roughly that $T \cdot N \le \widetilde{\mathcal{O}}(\frac{(n+d)^3}{\varepsilon^2}\log(1/\varepsilon))$ samples are sufficient for LSPI to recover a controller K that is within ε of the optimal $K_\star$. That is, only $\log(1/\varepsilon)$ iterations of policy improvement are necessary, and furthermore more outer iterations of policy improvement do not necessary help due to the inherent statistical noise of estimating the Q-function for every policy K_t. We note that the polynomial factors in (L/μ) is by no

means optimal and was made quite conservative in order to simplify the presentation of the bound. A sharper bound can be recovered from our analysis techniques at the expense of keeping track of extra terms.

It is worth taking a moment to compare Theorem 6.1.14 to classical results in the RL literature regarding approximate policy iteration. For example, a well known result reported in Theorem 3.1 of Lagoudakis and Parr [63] states that if LSTD-Q is able to return Q-function estimates with error L_∞ bounded by ε, then we have that:

$$\limsup_{t\to\infty} \|\widehat{Q}_t - Q_\star\|_\infty \leq \frac{2\gamma\varepsilon}{(1-\gamma)^2}\,.$$

Here, $Q_\star$ is the optimal Q-function and γ is the discount factor of the MDP. Theorem 6.1.14 is qualitatively similar to this result in that we show roughly that ε error in the Q-function estimate translates to ε error in the estimated policy. However, there are several fundamental differences. First, our analysis does not rely on discounting to show contraction of the Bellman operator. Instead, we use the (τ, ρ) stability of closed loop system to achieve this effect. Second, our analysis does not rely on L_∞ bounds on the estimated Q-function, which are generally not possible to achieve with LQR since the Q-function is a quadratic function and the states and inputs are not uniformly bounded. And finally, our analysis gives a non-asymptotic result.

The proof of Theorem 6.1.14 combines the estimation guarantee of Theorem 6.1.1 with a new analysis of policy iteration for LQR, which we believe is of independent interest. Our new policy iteration analysis combines the work of Bertsekas [15] on policy iteration in infinite horizon average cost MDPs with the contraction theory of Lee and Lim [65] for non-linear matrix equations.

Algorithm 3 Approximate Policy Iteration for LQR (offline)

Require: Initial stabilizing controller K_0, N number of policy iterations, T length of rollout for estimation, σ_η^2 exploration variance.

1: **for** $t = 0, ..., N-1$ **do**
2: Collect a trajectory $\mathcal{D}_t = \{(x_k^{(t)}, u_k^{(t)}, x_{k+1}^{(t)})\}_{k=1}^T$ using input $u_k^{(t)} = K_0 x_k^{(t)} + \eta_k^{(t)}$, with $\eta_k^{(t)} \sim \mathcal{N}(0, \sigma_\eta^2 I)$.
3: $\widehat{Q}_t = \mathsf{EstimateQ}(\mathcal{D}_t, K_t)$.
4: $K_{t+1} = G(\widehat{Q}_t)$.
5: **end for**
6: return K_N.

Before analyzing Algorithm 2, we analyze a slightly more general algorithm described in Algorithm 3. In Algorithm 3, the procedure EstimateQ takes as input an off-policy trajectory $\mathcal{D}_t$ and a policy K_t, and returns an estimate $\widehat{Q}_t$ of the true Q function Q_t. We will analyze

Algorithm 3 first assuming that the procedure EstimateQ delivers an estimate with a certain level of accuracy. We will then use the results of Section 6.1.2 to specialize EstimateQ to LSTD-Q. We define the sequence of variables for our analysis:

(i) Q_t is true state-value function for K_t.

(ii) V_t is true value function for K_t.

(iii) $\overline{K}_{t+1} = G(Q_t)$.

(iv) $\overline{V}_t$ is true value function for $\overline{K}_t$.

The following proposition is our main result regarding Algorithm 3.

Proposition 6.1.15. *Consider the sequence of updates defined by Algorithm 3. Suppose we start with a stabilizing K_0 and let V_0 denote its value function. Fix an $\varepsilon > 0$. Define the following variables:*

$$\begin{aligned}
\mu &:= \min\{\lambda_{\min}(S), \lambda_{\min}(R)\}\,, \\
Q_{\max} &:= \max\{\|S\|, \|R\|\} + 2(\|A\|^2 + \|B\|^2)\|V_0\|\,, \\
\gamma &:= \frac{2\|A\|^2\|V_0\|}{\mu + 2\|A\|^2\|V_0\|}\,, \\
N_0 &:= \lceil \frac{1}{1-\gamma}\log(2\delta_\infty(V_0, V_\star)/\varepsilon)\rceil\,, \\
\tau &:= \sqrt{\frac{2\|V_0\|}{\mu}}\,, \\
\rho &:= \sqrt{1 - 1/\tau^2}\,, \\
\overline{\rho} &:= \mathsf{Avg}(\rho, 1)\,.
\end{aligned}$$

Let $N_1 \geq N_0$. Suppose the estimates $\widehat{Q}_t$ output by EstimateQ *satisfy, for all $0 \leq t \leq N_1 - 1$, $\widehat{Q}_t \succeq \mu I$ and furthermore,*

$$\|\widehat{Q}_t - Q_t\| \leq \min\left\{\frac{\|V_0\|}{N_1}, \varepsilon\mu(1-\gamma)\right\}\left(\frac{\mu}{28}\frac{(1-\overline{\rho}^2)^2}{\tau^5}\frac{1}{\|B\|_+\max\{\|S\|, \|R\|\}}\frac{\mu^3}{Q_{\max}^3}\right)\,.$$

Then we have for any N satisfying $N_0 \leq N \leq N_1$ the bound $\delta_\infty(V_N, V_\star) \leq \varepsilon$. We also have that for all $0 \leq t \leq N_1$, $A + BK_t$ is $(\tau, \overline{\rho})$-stable and $\|K_t\| \leq 2Q_{\max}/\mu$.

Proof. We first start by observing that if V, V_0 are value functions satisfying $V \preceq V_0$, then their state-value functions also satisfy $Q \preceq Q_0$. This is because

$$\begin{aligned}
Q &= \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A^\mathsf{T} \\ B^\mathsf{T} \end{bmatrix} V \begin{bmatrix} A & B \end{bmatrix} \\
&\preceq \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A^\mathsf{T} \\ B^\mathsf{T} \end{bmatrix} V_0 \begin{bmatrix} A & B \end{bmatrix} = Q_0\,.
\end{aligned}$$

From this we also see that any state-value function satisfies $Q \succeq \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix}$.

The proof proceeds as follows. We observe that since $\overline{V}_{t+1} \preceq V_t$ (Proposition 6.1.8-(ii)):

$$V_t = V_t - \overline{V}_t + \overline{V}_t - V_{t-1} + V_{t-1} \preceq V_t - \overline{V}_t + V_{t-1} \,.$$

Therefore, by triangle inequality we have $\|V_t\| \leq \|V_t - \overline{V}_t\| + \|V_{t-1}\|$. Supposing for now that we can ensure for all $1 \leq t \leq N_1$:

$$\|V_t - \overline{V}_t\| \leq \frac{\|V_0\|}{N} \,, \tag{6.1.17}$$

unrolling the recursion for $\|V_t\|$ for N_1 steps ensures that $\|V_t\| \leq 2\|V_0\|$ for all $0 \leq t \leq N_1$. Furthermore,

$$\begin{aligned} \|Q_t\| &\leq \max\{\|S\|, \|R\|\} + \|\begin{bmatrix} A & B \end{bmatrix}\|^2 \|V_t\| \\ &\leq \max\{\|S\|, \|R\|\} + 2(\|A\|^2 + \|B\|^2)\|V_0\| \\ &= Q_{\max} \,. \end{aligned}$$

for all $0 \leq t \leq N_1$.

Now, by triangle inequality and Proposition 6.1.8-(iii), for all $0 \leq t \leq N_1 - 1$,

$$\begin{aligned} \delta_\infty(V_{t+1}, V_\star) &\leq \delta_\infty(V_{t+1}, \overline{V}_{t+1}) + \delta_\infty(\overline{V}_{t+1}, V_\star) \\ &\leq \delta_\infty(V_{t+1}, \overline{V}_{t+1}) + \gamma \cdot \delta_\infty(V_t, V_\star) \\ &\leq \frac{\|V_{t+1} - \overline{V}_{t+1}\|}{\mu} + \gamma \cdot \delta_\infty(V_t, V_\star) \,, \end{aligned} \tag{6.1.18}$$

where $\gamma = \frac{2\|A\|^2\|V_0\|}{\mu + 2\|A\|^2\|V_0\|}$, and the last inequality uses Proposition 6.1.11 combined with the fact that $V_{t+1} \succeq \mu I$ and $\overline{V}_{t+1} \succeq \mu I$.

We now focus on bounding $\|V_{t+1} - \overline{V}_{t+1}\|$. To do this, we first bound $\|K_{t+1} - \overline{K}_{t+1}\|$, and then use the Lyapunov perturbation result from Chapter 4. First, observe the simple bounds:

$$\begin{aligned} \|\overline{K}_{t+1}\| = \|T(Q_t)\| &\leq \frac{\|Q_t\|}{\mu} \leq \frac{Q_{\max}}{\mu} \,, \\ \|K_{t+1}\| = \|T(\widehat{Q}_t)\| &\leq \frac{\|\widehat{Q}_t\|}{\mu} \leq \frac{\Delta + Q_{\max}}{\mu} \leq \frac{2Q_{\max}}{\mu} \,. \end{aligned}$$

where the second bound uses the assumption that the estimates $\widehat{Q}_t$ satisfy $\widehat{Q}_t \succeq \mu I$ and $\|\widehat{Q}_t - Q_t\| \leq \Delta$ with

$$\Delta \leq Q_{\max} \,. \tag{6.1.19}$$

Now, by Proposition 4.0.7 we have:

$$\begin{aligned}\|K_{t+1} - \overline{K}_{t+1}\| &= \|G(\widehat{Q}_t) - G(Q_t)\| \\ &\leq \frac{(1 + \|\overline{K}_{t+1}\|)\|\widehat{Q}_t - Q_t\|}{\mu} \\ &\leq \frac{(1 + Q_{\max}/\mu)\Delta}{\mu} \\ &\leq \frac{2Q_{\max}}{\mu^2}\Delta\,.\end{aligned}$$

Above, the last inequality holds since $Q_{\max} \geq \mu$ by definition.

By Proposition 4.0.3, because $\overline{V}_{t+1} \preceq V_t$, we know that $\overline{K}_{t+1}$ satisfies for all $k \geq 0$:

$$\begin{aligned}\|(A + B\overline{K}_{t+1})^k\| &\leq \sqrt{\frac{\|V_t\|}{\lambda_{\min}(S)}} \cdot \sqrt{1 - \lambda_{\min}(V_t^{-1}S)}^k \\ &\leq \sqrt{\frac{2\|V_0\|}{\mu}}\sqrt{1 - \frac{\mu}{2\|V_0\|}}^k = \tau \cdot \rho^k\,.\end{aligned}$$

Let us now assume that Δ satisfies:

$$\frac{2Q_{\max}}{\mu^2} \cdot \Delta \leq \frac{1-\rho}{2\tau\|B\|}\,. \tag{6.1.20}$$

Then by Proposition 4.0.1, we know that $\|(A + BK_{t+1})^k\| \leq \tau \cdot \overline{\rho}^k$. Hence, we have that $A + BK_{t+1}$ is $(\tau, \overline{\rho})$-stable.

Next, by the Lyapunov perturbation result of Proposition 4.0.4,

$$\begin{aligned}&\|V_{t+1} - \overline{V}_{t+1}\| \\ &= \|\mathsf{dlyap}(A + BK_{t+1}, S + K_{t+1}^\mathsf{T} R K_{t+1}) - \mathsf{dlyap}(A + B\overline{K}_{t+1}, S + \overline{K}_{t+1}^\mathsf{T} R \overline{K}_{t+1})\| \\ &\leq \frac{\tau^2}{1-\overline{\rho}^2}\|K_{t+1}^\mathsf{T} R K_{t+1} - \overline{K}_{t+1}^\mathsf{T} R \overline{K}_{t+1}\| \\ &\quad + \frac{\tau^4}{(1-\overline{\rho}^2)^2}\|B(K_{t+1} - \overline{K}_{t+1})\|(\|A + BK_{t+1}\| + \|A + B\overline{K}_{t+1}\|)\|S + \overline{K}_{t+1}^\mathsf{T} R \overline{K}_{t+1}\|\,.\end{aligned}$$

We bound:

$$\begin{aligned}\|K_{t+1}^\mathsf{T} R K_{t+1} - \overline{K}_{t+1}^\mathsf{T} R \overline{K}_{t+1}\| &\leq \|R\|\|K_{t+1} - \overline{K}_{t+1}\|(\|K_{t+1}\| + \|\overline{K}_{t+1}\|) \\ &\leq \frac{6\|R\|Q_{\max}^2}{\mu^3}\Delta\,, \\ \|B(K_{t+1} - \overline{K}_{t+1})\| &\leq \frac{2\|B\|Q_{\max}}{\mu^2}\Delta\,, \\ \max\{\|A + BK_{t+1}\|, \|A + B\overline{K}_{t+1}\|\} &\leq \tau\,, \\ \|S + \overline{K}_{t+1}^\mathsf{T} R \overline{K}_{t+1}\| &\leq \|S\| + \frac{\|R\|Q_{\max}^2}{\mu^2}\,.\end{aligned}$$

Therefore,

$$\begin{aligned}\|V_{t+1} - \overline{V}_{t+1}\| &\leq \frac{\tau^2}{1-\overline{\rho}^2}\frac{6\|R\|Q_{\max}^2}{\mu^3}\Delta + 8\frac{\tau^5}{(1-\overline{\rho}^2)^2}\|B\|\max\{\|S\|,\|R\|\}\frac{Q_{\max}^3}{\mu^4}\Delta \\ &= \frac{1}{\mu}\left(\frac{\tau^2}{1-\overline{\rho}^2}\frac{6\|R\|Q_{\max}^2}{\mu^2} + 8\frac{\tau^5}{(1-\overline{\rho}^2)^2}\|B\|\max\{\|S\|,\|R\|\}\frac{Q_{\max}^3}{\mu^3}\right)\Delta \\ &\leq \frac{14}{\mu}\frac{\tau^5}{(1-\overline{\rho}^2)^2}\|B\|_+\max\{\|S\|,\|R\|\}\frac{Q_{\max}^3}{\mu^3}\Delta\,.\end{aligned}$$

Now suppose that Δ satisfies:

$$\begin{aligned}\Delta &\leq \frac{1}{2}\varepsilon\mu(1-\gamma)\left(\frac{\mu}{14}\frac{(1-\overline{\rho}^2)^2}{\tau^5}\frac{1}{\|B\|_+\max\{\|S\|,\|R\|\}}\frac{\mu^3}{Q_{\max}^3}\right) \\ &= \frac{\varepsilon}{28}\mu^2(1-\gamma)\frac{(1-\overline{\rho}^2)^2}{\tau^5}\frac{1}{\|B\|_+\max\{\|S\|,\|R\|\}}\frac{\mu^3}{Q_{\max}^3}\,,\end{aligned} \tag{6.1.21}$$

we have for all $t \leq N_1 - 1$ from (6.1.18):

$$\delta_\infty(V_{t+1}, V_\star) \leq (1-\gamma)\varepsilon/2 + \gamma\cdot\delta_\infty(V_t, V_\star)\,.$$

Unrolling this recursion, we have that for any $N \leq N_1$:

$$\delta_\infty(V_N, V_\star) \leq \gamma^N\cdot\delta_\infty(V_0, V_\star) + \varepsilon/2\,.$$

Now observe that for any $N \geq N_0 := \lceil \frac{1}{1-\gamma}\log(2\delta_\infty(V_0, V_\star)/\varepsilon)\rceil$, we obtain:

$$\delta_\infty(V_N, V_\star) \leq \varepsilon\,.$$

The claim now follows by combining our four requirements on Δ given in (6.1.19), (6.1.17), (6.1.20), and (6.1.21). □

We now proceed to make several simplifications to Proposition 6.1.15 in order to make the result more presentable. These simplifications come with the tradeoff of introducing extra conservatism into the bounds.

Our first simplification of Proposition 6.1.15 is the following corollary.

Corollary 6.1.16. *Consider the sequence of updates defined by Algorithm 3. Suppose we start with a stabilizing K_0 and let V_0 denote its value function. Define the following variables:*

$$\begin{aligned}\mu &:= \min\{\lambda_{\min}(S), \lambda_{\min}(R)\}\,, \\ L &:= \max\{\|S\|,\|R\|\} + 2(\|A\|^2 + \|B\|^2 + 1)\|V_0\|_+\,, \\ N_0 &:= \lceil (1 + L/\mu)\log(2\delta_\infty(V_0, V_\star)/\varepsilon)\rceil\,.\end{aligned}$$

Fix an $N_1 \geq N_0$ and suppose that

$$\varepsilon \leq \frac{1}{\mu}\left(1+\frac{L}{\mu}\right)\frac{\|V_0\|}{N_1}\,. \tag{6.1.22}$$

Suppose the estimates $\widehat{Q}_t$ output by EstimateQ *satisfy, for all $0 \leq t \leq N_1 - 1$, $\widehat{Q}_t \succeq \mu I$ and furthermore,*

$$\|\widehat{Q}_t - Q_t\| \leq \frac{\varepsilon}{448}\frac{\mu}{\mu+L}\left(\frac{\mu}{L}\right)^{19/2}\,.$$

Then we have for any $N_0 \leq N \leq N_1$ that $\delta_\infty(V_N, V_\star) \leq \varepsilon$. We also have that for any $0 \leq t \leq N_1$, that $A + BK_t$ is $(\sqrt{L/\mu}, \mathsf{Avg}(\sqrt{1-\mu/L}, 1))$-stable and $\|K_t\| \leq 2L/\mu$.

Proof. First, observe that the map $x \mapsto \frac{x}{\mu+x}$ is increasing, and therefore $\gamma \leq \frac{L}{\mu+L}$ which implies that $1-\gamma \geq \frac{\mu}{\mu+L}$. Therefore if $\varepsilon \leq \frac{1}{\mu}\left(1+\frac{L}{\mu}\right)\frac{\|V_0\|}{N_1}$ holds, then we can bound:

$$\min\left\{\frac{\|V_0\|}{N_1}, \varepsilon\mu(1-\gamma)\right\} \geq \varepsilon\mu\left(\frac{\mu}{\mu+L}\right)\,.$$

Next, observe that

$$1-\overline{\rho}^2 = (1+\overline{\rho})(1-\overline{\rho}) = (1+1/2+\rho/2)(1/2-\rho/2) \geq (1+\rho)(1-\rho)/4 = (1-\rho^2)/4\,.$$

Therefore,

$$(1-\overline{\rho}^2)^2 \geq (1-(1-\mu/L))^2/16 = (1/16)(\mu/L)^2\,.$$

We also have that $\tau \leq \sqrt{\frac{L}{\mu}}$. This means we can bound:

$$\frac{\mu}{28}\frac{(1-\overline{\rho}^2)^2}{\tau^5}\frac{1}{\|B\|_+\max\{\|S\|,\|R\|\}}\frac{\mu^3}{Q_{\max}^3} \geq \frac{\mu}{28\cdot 16}(\mu/L)^{5/2+2}\frac{\mu^3}{L^5} = \frac{1}{448L}\left(\frac{\mu}{L}\right)^{17/2}\,.$$

Therefore,

$$\min\left\{\frac{\|V_0\|}{N_1}, \varepsilon\mu(1-\gamma)\right\}\frac{\mu}{28}\frac{(1-\overline{\rho}^2)^2}{\tau^5}\frac{1}{\|B\|_+\max\{\|S\|,\|R\|\}}\frac{\mu^3}{Q_{\max}^3} \geq \frac{\varepsilon}{448}\left(\frac{\mu}{\mu+L}\right)\left(\frac{\mu}{L}\right)^{19/2}\,.$$

The claim now follows from Proposition 6.1.15. □

Corollary 6.1.16 gives a guarantee in terms of $\delta_\infty(V_N, V_\star) \leq \varepsilon$. By Proposition 6.1.13, this implies a bound on the error of the value functions $\|V_N - V_\star\| \leq \mathcal{O}(\varepsilon)$ for $\varepsilon \leq 1$. In the next corollary, we show we can also control the error $\|K_N - K_\star\| \leq \mathcal{O}(\varepsilon)$.

Corollary 6.1.17. *Consider the sequence of updates defined by Algorithm 3. Suppose we start with a stabilizing K_0 and let V_0 denote its value function. Define the following variables:*

$$\begin{aligned}\mu &:= \min\{\lambda_{\min}(S), \lambda_{\min}(R)\}\,,\\ L &:= \max\{\|S\|, \|R\|\} + 2(\|A\|^2 + \|B\|^2 + 1)\|V_0\|_+\,,\\ N_0 &:= \left\lceil (1 + L/\mu)\log\left(\frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{\varepsilon}\right)\right\rceil\,.\end{aligned}$$

Suppose that $\varepsilon > 0$ satisfies:

$$\varepsilon \le \min\left\{1, \frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{e}, \frac{\|V_\star\|^2}{8\mu^2\log(\|V_0\|/\lambda_{\min}(V_\star))}\right\}\,.$$

Suppose we run Algorithm 3 for $N := N_0 + 1$ iterations. Suppose the estimates $\widehat{Q}_t$ output by EstimateQ *satisfy, for all $0 \le t \le N_0$, $\widehat{Q}_t \succeq \mu I$ and furthermore,*

$$\|\widehat{Q}_t - Q_t\| \le \frac{\varepsilon}{448}\frac{\mu}{\mu + L}\left(\frac{\mu}{L}\right)^{19/2}\,. \tag{6.1.23}$$

We have that:

$$\|K_N - K_\star\| \le 5\left(\frac{L}{\mu}\right)^2\varepsilon$$

and that $A + BK_t$ is $(\sqrt{L/\mu}, \mathsf{Avg}(\sqrt{1 - \mu/L}, 1))$-stable and $\|K_t\| \le 2L/\mu$ for all $0 \le t \le N$.

Proof. We set $N_1 = N_0 + 1$. From this, we compute:

$$\begin{aligned}
\|K_{N_1} - K_\star\| &= \|G(\widehat{Q}_{N_0}) - G(Q_\star)\| \\
&\overset{(a)}{\leq} \frac{(1 + \|G(Q_\star)\|)}{\mu} \|\widehat{Q}_{N_0} - Q_\star\| \\
&\leq \frac{(1 + \|G(Q_\star)\|)}{\mu} (\|\widehat{Q}_{N_0} - Q_{N_0}\| + \|Q_{N_0} - Q_\star\|) \\
&= \frac{(1 + \|G(Q_\star)\|)}{\mu} \left(\|\widehat{Q}_{N_0} - Q_{N_0}\| + \left\| \begin{bmatrix} A^\mathsf{T} \\ B^\mathsf{T} \end{bmatrix} (V_{N_0} - V_\star) \begin{bmatrix} A & B \end{bmatrix} \right\| \right) \\
&\leq \frac{(1 + \|G(Q_\star)\|)}{\mu} (\|\widehat{Q}_{N_0} - Q_{N_0}\| + \|\begin{bmatrix} A & B \end{bmatrix}\|^2 \|V_{N_0} - V_\star\|) \\
&\overset{(b)}{\leq} \frac{(1 + \|G(Q_\star)\|)}{\mu} \left(\frac{\varepsilon}{448} \frac{\mu}{\mu + L} \left(\frac{\mu}{L}\right)^{19/2} + \|\begin{bmatrix} A & B \end{bmatrix}\|^2 \|V_{N_0} - V_\star\| \right) \\
&\overset{(c)}{\leq} \frac{(1 + \|G(Q_\star)\|)}{\mu} \left(\frac{\varepsilon}{448} \frac{\mu}{\mu + L} \left(\frac{\mu}{L}\right)^{19/2} + e(\|A\|^2 + \|B\|^2) \|V_\star\| \varepsilon \right) \\
&\leq \frac{2L}{\mu^2} \left(\frac{1}{448} \frac{\mu}{\mu + L} \left(\frac{\mu}{L}\right)^{19/2} + 2L \right) \varepsilon \\
&= \left(\frac{1}{224} \frac{1}{\mu + L} \left(\frac{\mu}{L}\right)^{17/2} + 4 \left(\frac{L}{\mu}\right)^2 \right) \varepsilon \\
&\leq 5 \left(\frac{L}{\mu}\right)^2 \varepsilon \,.
\end{aligned}$$

Above, (a) follows from Proposition 4.0.7, (b) follows from the bound on $\|\widehat{Q}_{N_0} - Q_{N_0}\|$ from Corollary 6.1.16, and (c) follows from Proposition 6.1.13 and the fact that $\delta_\infty(V_{N_0}, V_\star) \leq \varepsilon$ from Corollary 6.1.16.

Next, we observe that since $V_0 \succeq V_\star$:

$$\delta_\infty(V_0, V_\star) = \log(\|V_\star^{-1/2} V_0 V_\star^{-1/2}\|) \leq \log(\|V_0\| / \lambda_{\min}(V_\star)) \,.$$

Hence we can upper bound N_0 from Corollary 6.1.16 by:

$$N_0 = 2(1 + L/\mu) \log(2 \log(\|P_0\| / \lambda_{\min}(V_\star))/\varepsilon) \,.$$

From (6.1.22), the requirement on ε is that:

$$\varepsilon \leq \min \left\{ \frac{\|V_0\|}{2\mu} \frac{1}{\log\left(\frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{\varepsilon}\right)}, 1 \right\} \,.$$

We will show with Proposition 6.1.18 that a sufficient condition is that:

$$\varepsilon \leq \min \left\{ 1, \frac{2\log(\|V_0\|/\lambda_{\min}(V_\star))}{e}, \frac{\|V_\star\|^2}{8\mu^2 \log(\|V_0\|/\lambda_{\min}(V_\star))} \right\} \,.$$

□

With Corollary 6.1.17 in place, we are now ready to prove Theorem 6.1.14.

Proof of Theorem 6.1.14. Let $L_0 := A + BK_0$ and let (τ, ρ) be such that L_0 is (τ, ρ)-stable. We know we can pick $\tau = \sqrt{L/\mu}$ and $\rho = \sqrt{1 - \mu/L}$. The covariance Σ_t of x_t satisfies:

$$\Sigma_t = L_0^t \Sigma_0 (L_0^t)^\mathsf{T} + P_t \preceq \tau^2 \rho^{2t} \|\Sigma_0\| I + P_\infty \,.$$

Hence for either $t = 0$ or $t \geq \log(\tau)/(1-\rho)$, $\|\Sigma_t\| \leq \|\Sigma_0\| + \|P_\infty\|$. Therefore, if the trajectory length $T \geq \log(\tau)/(1-\rho)$, then the operator norm of the initial covariance for every invocation of LSTD-Q can be bounded by $\|\Sigma_0\| + \|P_\infty\|$, and therefore the proxy variance (6.1.5) can be bounded by:

$$\begin{aligned}\overline{\sigma}^2 &\leq \tau^2 \rho^4 \|\Sigma_0\| + (1 + \tau^2 \rho^4) \|P_\infty\| + \sigma_\eta^2 \|B\|^2 \\ &\leq 2(L/\mu)(\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2 \|B\|^2) \,.\end{aligned}$$

By Corollary 6.1.17, when condition (6.1.23) holds, we have that $A + BK_t$ is $(\tau, \mathsf{Avg}(\rho, 1))$ stable, $\|K_t\| \leq 2L/\mu$, and $\|Q_t\| \leq L$ for all $0 \leq t \leq N_0 + 1$. We now define $\overline{\varepsilon} := 5(L/\mu)^2 \varepsilon$. If we can ensure that

$$\|\widehat{Q}_t - Q_t\| \leq \frac{1}{2240} \left(\frac{\mu}{\mu + L}\right) \left(\frac{\mu}{L}\right)^{23/2} \overline{\varepsilon} \,, \tag{6.1.24}$$

then if

$$\overline{\varepsilon} \leq 5 \left(\frac{L}{\mu}\right)^2 \min \left\{1, \frac{2 \log(\|V_0\|/\lambda_{\min}(V_\star))}{e}, \frac{\|V_\star\|^2}{8\mu^2 \log(\|V_0\|/\lambda_{\min}(V_\star))}\right\} \,,$$

then by Corollary 6.1.17 we ensure that $\|K_N - K\| \leq \overline{\varepsilon}$. By Theorem 6.1.1, (6.1.24) can be ensured by first observing that $Q_t \succeq \mu I$ and therefore for any symmetric $\widehat{Q}$ we have:

$$\|\mathsf{Proj}_\mu(\widehat{Q}) - Q_t\| \leq \|\mathsf{Proj}_\mu(\widehat{Q}) - Q_t\|_F \leq \|\widehat{Q} - Q_t\|_F \,.$$

Above, the last inequality holds because $\mathsf{Proj}_\mu(\cdot)$ is the Euclidean projection operator associated with $\|\cdot\|_F$ onto the convex set $\{Q : Q \succeq \mu I \,,\ Q = Q^\mathsf{T}\}$. Now combining (6.1.7) and (6.1.6) and using the bound $\frac{\tau^2}{\rho^2(1-\rho^2)} \leq \frac{(L/\mu)^2}{1-\mu/L}$:

$$\begin{aligned}T \geq \widetilde{\mathcal{O}}(1) \max \Bigg\{ & (n+d)^2, \\ & \frac{L^2}{(1-\mu/L)^2} \left(\frac{L}{\mu}\right)^{17} \frac{(n+d)^4}{\sigma_\eta^4} \sigma_w^2 (\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2 \|B\|^2), \\ & \frac{1}{\overline{\varepsilon}^2} \frac{L^4}{(1-\mu/L)^2} \left(\frac{L}{\mu}\right)^{42} \frac{(n+d)^3}{\sigma_\eta^4} \sigma_w^2 (\|\Sigma_0\| + \|P_\infty\| + \sigma_\eta^2 \|B\|^2) \Bigg\} \,.\end{aligned}$$

Theorem 6.1.14 now follows. □

The following proposition allows us to solve explicitly for an upper bound on ε.

Proposition 6.1.18. *Let $C > 0$. Then for any $\varepsilon \in (0, \min\{1/e, C^2\})$, we have the following inequality holds:*

$$\varepsilon \log(1/\varepsilon) \leq C \, .$$

As a corollary, let $M > 0$, then for $\varepsilon \in (0, \min\{M/e, C^2/M\})$ we have that:

$$\varepsilon \log(M/\varepsilon) \leq C \, .$$

Proof. Let $f(\varepsilon) := \varepsilon \log(1/\varepsilon)$. We have that $\lim_{\varepsilon \to 0^+} f(\varepsilon) = 0$ and that $f'(\varepsilon) = \log(1/\varepsilon) - 1$. Hence f is increasing on the interval $\varepsilon \in [0, 1/e]$, and $f(1/e) = 1/e$. Therefore, if $C \geq 1/e$ then $f(\varepsilon) \leq C$ for any $\varepsilon \in (0, 1/e)$.

Now suppose that $C < 1/e$. One can verify that the function $g(x) := 1/x + 2 \log x$ satisfies $g(x) \geq 0$ for all $x > 0$. Therefore:

$$\begin{aligned} g(C) \geq 0 &\iff 1/C + 2 \log C \geq 0 \\ &\iff 1/C \geq \log(1/C^2) \\ &\iff C \geq C^2 \log(1/C^2) \\ &\iff f(C^2) \leq C \, . \end{aligned}$$

Since $C < 1/e$ we have $C^2 \leq C$ and therefore $f(\varepsilon) \leq f(C^2) \leq C$ for all $\varepsilon \in (0, C^2)$. This proves the first part.

To see the second part, use the variable substitution $\varepsilon \leftarrow \varepsilon/M$, $C \leftarrow C/M$. □

6.2 Asymptotic Analysis of Model-based and Model-free Methods for LQR

We focus our asymptotic analysis on two tasks for LQR. The first task is *policy evaluation*, which given a stabilizing policy K for (A, B), estimates the value function V associated with K. The second task is *policy optimization*, which finds the optimal LQR controller.

6.2.1 Related Work

We give a brief overview of known model-based and model-free results in the tabular MDP setting.

The best known regret bound in the model-based case is $\widetilde{\mathcal{O}}(\sqrt{H^2SAT})$ from Azar et al. [13], which matches the known lower bound of $\Omega(\sqrt{H^2SAT})$ from [51, 53] up to log factors. On the other hand, the best known regret bound in the model-free case is $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ from the UCB-style Q-learning algorithm of Jin et al. [53], which is worse than the model-based case by a factor of the horizon length H. Interestingly, there is no gap in terms of the number

of states S and actions A. It is open whether or not the gap in H for regret is fundamental or can be closed.

We now turn to the PAC setting. We first look at the "simulator" setting, where an oracle exists that allows one to query the state transition from any state/action pair at every timestep. For infinite horizon discounted MDPs, Azar et al. [12] show that model-based policy iteration can find a ε optimal policy with $\widetilde{\mathcal{O}}(SA/((1-\gamma)^3\varepsilon^2))$ samples as long as $\varepsilon \leq \mathcal{O}(1/\sqrt{(1-\gamma)S})$, which matches the minimax lower bound given in the same work. Sidford et al. [102] show that a model-free variance reduced value iteration algorithm also achieves $\widetilde{\mathcal{O}}(SA/((1-\gamma)^3\varepsilon^2))$ sample complexity even beyond the small ε regime of the model-based method. Therefore, in this setting, there is no gap between model-based and model-free methods in the small ε regime, and the best upper bounds currently suggest that model-free methods actually outperform model-based methods in the moderate ε regime by a factor of $1/(1-\gamma)^2$ in sample complexity. It is still open if this gap can be resolved in the moderate ε regime.

For tabular MDPs, a popular definition in the literature of a model-free method is one where the space complexity is sub-linear in the amount of storage needed for a model-based method. For example, for a finite horizon MDP, one can define a model-free method as having space complexity $o(S^2AH)$, where S is the number of states, A is the number of actions, and H is the horizon length (see e.g. Jin et al. [53] and Strehl et al. [109]). We note that this definition does not generalize to the continuous setting; using LQR as an example, storing the (A, B) matrices for the model requires $\mathcal{O}(n(n+d))$ space, whereas storing the Q-function requires $\mathcal{O}((n+d)^2)$ space. Sun et al. [110] present a new information-theoretic definition of model-free algorithms. Under their definition, they construct a family of factored MDPs with horizon length H where any model-free algorithm incurs sample complexity $\Omega(2^H)$, whereas there exists a model-based algorithm that has sample complexity polynomial in H and other relevant quantities.

6.2.2 Policy Evaluation

Given a controller $K \in \mathbb{R}^{d\times n}$ that stabilizes (A, B), the policy evaluation task is to compute the (relative) value function $V^K(x)$:

$$V^K(x) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - \lambda_K) \,\middle|\, x_0 = x\right] \;,\; u_t = Kx_t \;. \tag{6.2.1}$$

Above, λ_K is the infinite horizon average cost. It is well-known that $V^K(x)$ can be written as:

$$V^K(x) = \sigma_w^2 x^\mathsf{T} V_\star x \;,\; V_\star = \mathsf{dlyap}(A+BK, Q+K^\mathsf{T} R K) \;. \tag{6.2.2}$$

From the Lyapunov equation, it is clear that given (A, B), the solution to policy evaluation task is readily computable. W study algorithms which only have input/output access to

(A, B). Specifically, we study *on-policy* algorithms that operate on a *single* trajectory, where the input u_t is determined by $u_t = Kx_t$. The variable that controls the amount of information available to the algorithm is T, the trajectory length. The trajectory will be denoted as $\{x_t\}_{t=0}^T$. We are interested in the asymptotic behavior of algorithms as $T \to \infty$.

Model-based algorithm. In light of Equation (6.2.2), the plugin estimator is a very natural model-based algorithm to use. Let $L_\star := A + BK$ denote the true closed-loop matrix. The plugin estimator uses the trajectory $\{x_t\}_{t=0}^T$ to estimate $L_\star$ via least-squares; call this $\widehat{L}(T)$. The estimator then returns $\widehat{V}_{\mathrm{plug}}(T)$ by using $\widehat{L}(T)$ in-place of $L_\star$ in (6.2.2). Algorithm 4 describes this estimator.

Algorithm 4 Model-based algorithm for policy evaluation.

Require: Policy $\pi(x) = Kx$, rollout length T, regularization $\lambda > 0$, thresholds $\zeta \in (0, 1)$ and $\psi > 0$.

1: Collect trajectory $\{x_t\}_{t=0}^T$ using the feedback $u_t = \pi(x_t) = Kx_t$.

2: Estimate the closed-loop matrix via least-squares:

$$\widehat{L}(T) = \left(\sum_{t=0}^{T-1} x_{t+1}x_t^\mathsf{T}\right)\left(\sum_{t=0}^{T-1} x_t x_t^\mathsf{T} + \lambda I_n\right)^{-1}.$$

3: **if** $\rho(\widehat{L}(T)) > \zeta$ or $\|\widehat{L}(T)\| > \psi$ **then**

4: Set $\widehat{V}_{\mathrm{plug}}(T) = 0$.

5: **else**

6: Set $\widehat{V}_{\mathrm{plug}}(T) = \mathsf{dlyap}(\widehat{L}(T), Q + K^\mathsf{T}RK)$.

7: **end if**

8: return $\widehat{V}_{\mathrm{plug}}(T)$.

Model-free algorithm. By observing that $V^K(x) = \sigma_w^2 x^\mathsf{T} V_\star x = \sigma_w^2 \langle \mathrm{svec}(V_\star), \mathrm{svec}(xx^\mathsf{T})\rangle$, one can apply Least-Squares Temporal Difference Learning (LSTD) [22, 25] with the feature map $\phi(x) := \mathrm{svec}(xx^\mathsf{T})$ to estimate $V_\star$. This is a classical algorithm in RL related to the LSTD-Q algorithm studied in Section 6.1.2; the pseudocode is given in Algorithm 5.

We now proceed to compare the risk of Algorithm 4 versus Algorithm 5. Our notion of risk will be the expected squared error of the estimator: $\mathbb{E}[\|\widehat{V} - V_\star\|_F^2]$. Our first result gives an upper bound on the asymptotic risk of the model-based plugin Algorithm 4.

Theorem 6.2.1. *Let K stabilize (A, B). Define $L_\star$ to be the closed-loop matrix $A + BK$ and let $\rho(L_\star) \in (0, 1)$ denote its stability radius. Recall that $V_\star$ is the solution to the discrete-time Lyapunov equation (6.2.2) that parameterizes the value function $V^K(x)$. We have that*

Algorithm 5 Model-free algorithm for policy evaluation (LSTD) [25].

Require: Policy $\pi(x) = Kx$, rollout length T.

1: Collect trajectory $\{x_t\}_{t=0}^{T}$ using the feedback $u_t = \pi(x_t) = Kx_t$.

2: Estimate $\lambda_t \approx \sigma_w^2 \operatorname{tr}(V_\star)$ from $\{x_t\}_{t=0}^{T}$.

3: Compute (recall that $\phi(x) = \operatorname{svec}(xx^\mathsf{T})$):

$$\widehat{w}_{\mathrm{lstd}}(T) = \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1} (c_t - \lambda_t)\phi(x_t)\right) ,$$

4: Set $\widehat{V}_{\mathrm{lstd}}(T) = \operatorname{smat}(\widehat{w}_{\mathrm{lstd}}(T))$.

5: return $\widehat{V}_{\mathrm{lstd}}(T)$.

Algorithm 4 with thresholds (ζ, ψ) satisfying $\zeta \in (\rho(L_\star), 1)$ and $\psi \in (\|L_\star\|, \infty)$ and any fixed regularization parameter $\lambda > 0$ has the asymptotic risk upper bound:

$$\lim_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|_F^2] \le 4 \operatorname{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(L_\star^\mathsf{T} V_\star^2 L_\star \otimes_s \sigma_w^2 P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) .$$

Here, $P_\infty = \mathsf{dlyap}(L_\star^\mathsf{T}, \sigma_w^2 I_n)$ is the stationary covariance matrix of the closed-loop system $x_{t+1} = L_\star x_t + w_t$ and $\otimes_s$ denotes the symmetric Kronecker product.

We make a few quick remarks regarding Theorem 6.2.1. First, while the risk bound is presented as an upper bound, the exact asymptotic risk can be recovered from the proof. Second, the thresholds (ζ, ψ) and regularization parameter λ do not affect the final asymptotic bound, but do possibly affect both higher order terms and the rate of convergence to the limiting risk. We include these thresholds as they simplify the proof. In practice, we find that thresholding or regularization is generally not needed, with the caveat that if the estimate $\widehat{L}(T)$ is not stable then the solution to the discrete Lyapunov equation is not guaranteed to exist (and when it exists is not guaranteed to be positive semidefinite). Finally, we remark that a non-asymptotic high probability upper bound for the risk of Algorithm 4 can be easily derived by combining the single trajectory learning results of Simchowitz et al. [105] with the Lyapunov perturbation results of Chapter 4.

We now turn our attention to the model-free LSTD algorithm. Our next result gives a lower bound on the asymptotic risk of Algorithm 5.

Theorem 6.2.2. *Let K stabilize (A, B). Define $L_\star$ to be the closed-loop matrix $A + BK$. Recall that $V_\star$ is the solution to the discrete-time Lyapunov equation (6.2.2) that parameterizes the value function $V^K(x)$. We have that Algorithm 5 with the cost estimates λ_t set to the true cost $\lambda_\star := \sigma_w^2 \operatorname{tr}(V_\star)$ satisfies the asymptotic risk lower bound:*

$$\begin{aligned}\liminf_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{V}_{\mathrm{lstd}}(T) - V_\star\|_F^2] &\ge 4\mathcal{R}_{\mathsf{plug}} \\ &+ 8\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} V_\star^2 L_\star \rangle \operatorname{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}})\end{aligned}$$

Here, $\mathcal{R}_{\mathsf{plug}} := \lim_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|_F^2]$ *is the asymptotic risk of the plugin estimator,* $P_\infty = \mathsf{dlyap}(L_\star^\mathsf{T}, \sigma_w^2 I_n)$ *is the stationary covariance matrix of the closed loop system* $x_{t+1} = L_\star x_t + w_t$*, and* $\otimes_s$ *denotes the symmetric Kronecker product.*

Theorem 6.2.2 shows that the asymptotic risk of the model-free method always exceeds that of the model-based plugin method. We remark that we prove the theorem under an idealized setting where the infinite horizon cost estimate λ_t is set to the true cost $\lambda_\star$. In practice, the true cost is not known and must instead be estimated from the data at hand. However, for the purposes of our comparison this is not an issue because using the true cost $\lambda_\star$ over an estimator of $\lambda_\star$ only reduces the variance of the risk.

To get a sense of how much excess risk is incurred by the model-free method over the model-based method, consider the following family of instances, defined for $\rho \in (0,1)$ and $1 \le d \le n$:

$$\mathscr{F}(\rho, d, K) := \{(A,B) : A + BK = \tau P_E + \gamma I_n\,,\ (\tau,\gamma) \in (0,1)\,,\ \tau+\gamma \le \rho\,,\ \dim(E) \le d\}\,. \tag{6.2.3}$$

With this family, one can show with elementary computations that under the simplifying assumptions that $Q + K^\mathsf{T} R K = I_n$ and $d \asymp n$, Theorem 6.2.1 and Theorem 6.2.2 state that:

$$\lim_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|_F^2] \le \mathcal{O}\left(\frac{\rho^2 n^2}{(1-\rho^2)^3}\right)\,,$$
$$\liminf_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{V}_{\mathrm{lstd}}(T) - V_\star\|_F^2] \ge \Omega\left(\frac{\rho^2 n^3}{(1-\rho^2)^3}\right)\,.$$

That is, for $\mathscr{F}(\rho, d, K)$, the plugin risk is a factor of state-dimension n less than the LSTD risk. Moreover, the non-asymptotic result for LSTD-Q from Section 6.1.2 can be modified to give a bound $\|\widehat{V}_{\mathrm{lstd}}(T) - V_\star\|_F^2 \le \widetilde{\mathcal{O}}(n^3/T)$ w.h.p., which matches the asymptotic bound of Theorem 6.2.2 in terms of n up to logarithmic factors.

Our final result for policy evaluation is a minimax lower bound on the risk of any estimator over $\mathscr{F}(\rho, d, K)$.

Theorem 6.2.3. *Fix a* $\rho \in (0,1)$ *and suppose that* K *satisfies* $S + K^\mathsf{T} R K = I_n$*. Suppose that* n *is greater than an absolute constant and* $T \gtrsim n(1-\rho^2)/\rho^2$*. We have that:*

$$\inf_{\widehat{V}} \sup_{(A,B)\in\mathscr{F}(\rho,\frac{n}{4},K)} \mathbb{E}[\|\widehat{V} - V_\star\|_F^2] \gtrsim \frac{\rho^2 n^2}{(1-\rho^2)^3 T}\,,$$

where the infimum is taken over all estimators $\widehat{V}$ *taking input* $\{x_t\}_{t=0}^T$*.*

Theorem 6.2.3 states that the rate achieved by the model-based Algorithm 6 over the family $\mathscr{F}(\rho, d, K)$ cannot be improved beyond constant factors, at least asymptotically; its dependence on both the state dimension n and stability radius ρ is optimal.

6.2.3 Policy Optimization

Given a finite horizon length T, the policy optimization task we study in this section is to solve the *finite horizon* optimal control problem:

$$J_\star := \min_{u_t(\cdot)} \mathbb{E}\left[\sum_{t=0}^{T-1}(x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t) + x_T^\mathsf{T} S x_T\right], \quad x_{t+1} = Ax_t + Bu_t + w_t\,. \tag{6.2.4}$$

We will focus on a special case of this problem when there is no penalty on the input: $S = I_n$, $R = 0$, and range$(A) \subseteq$ range(B). In this situation, the cost function reduces to $\mathbb{E}[\sum_{t=0}^{T}\|x_t\|_2^2]$ and the optimal solution simply chooses a u_t that cancels out the state x_t; that is $u_t = K_\star x_t$ with $K_\star := -B^\dagger A$. We work with this simple class of instances so that we can ensure that policy gradient converges to the optimal solution; in general this is not guaranteed.

We consider a slightly different input/output oracle model in this setting than we did in Section 6.2.2. The horizon length T is now considered fixed, and N rounds are played. At each round $i = 1, ..., N$, the algorithm chooses a feedback matrix $K_i \in \mathbb{R}^{d\times n}$. The algorithm then observes the trajectory $\{x_t^{(i)}\}_{t=0}^T$ by playing the control input $u_t^{(i)} = K_i x_t^{(i)} + \eta_t^{(i)}$, where $\eta_t^{(i)} \sim \mathcal{N}(0, \sigma_\eta^2 I_d)$ is i.i.d. noise used for the policy. This process then repeats for N total rounds. After the N rounds, the algorithm is asked to output a $\widehat{K}(N)$ and is assigned the risk $\mathbb{E}[J(\widehat{K}(N)) - J_\star]$, where $J(\widehat{K}(N))$ denotes playing the feedback $u_t = \widehat{K}(N)x_t$ on the true system (A, B). We will study the behavior of algorithms when $N \to \infty$ (and T is held fixed).

Model-based algorithm. Under this oracle model, a natural model-based algorithm is to first use random open-loop feedback (i.e. $K_i = 0$) to observe N independent trajectories (each of length T), and then use the trajectory data to fit the state transition matrices (A, B); call this estimate $(\widehat{A}(N), \widehat{B}(N))$. After fitting the dynamics, the algorithm then returns the estimate of $K_\star$ by solving the finite horizon problem (6.2.4) with $(\widehat{A}(N), \widehat{B}(N))$ taking the place of (A, B). In general, however, the assumption that range$(\widehat{A}(N)) \subseteq$ range$(\widehat{B}(N))$ will not hold, and hence the optimal solution to (6.2.4) will not be time-invariant. Moreover, solving for the best time-invariant static feedback for the finite horizon problem in general is not tractable. In light of this, to provide the fairest comparison to the model-free policy gradient method, we use the time-invariant static feedback that arises from infinite horizon solution given by the discrete algebraic Riccati equation as a proxy. We note that under our range inclusion assumption, the infinite horizon solution is a consistent estimator of the optimal feedback. The pseudo-code for this model-based algorithm is described in Algorithm 6.

[1] A sufficient condition for the existence of a unique positive definite solution to the discrete algebraic Riccati equation when $R = 0$ is that (A, B) is stabilizable and B has full column rank (Lemma 6.2.19).

Algorithm 6 Model-based algorithm for policy optimization.

Require: Horizon length T, rollouts N, regularization λ, thresholds $\varrho \in (0,1), \zeta, \psi, \gamma$.

1: Collect trajectories $\{\{(x_t^{(i)}, u_t^{(i)})\}_{t=0}^{T}\}_{i=1}^{N}$ using the feedback $K_i = 0$ (open-loop).

2: Estimate the dynamics matrices (A, B) via regularized least-squares:

$$\widehat{\Theta}(N) = \left(\sum_{i=1}^{N}\sum_{t=0}^{T-1} x_{t+1}(z_t^{(i)})^{\mathsf{T}}\right)\left(\sum_{i=1}^{N}\sum_{t=0}^{T-1} z_t^{(i)}(z_t^{(i)})^{\mathsf{T}} + \lambda I_{n+d}\right)^{-1}, \quad z_t^{(i)} := (x_t^{(i)}, u_t^{(i)}) .$$

3: Set $(\widehat{A}, \widehat{B}) = \widehat{\Theta}(N)$.

4: **if** $\rho(\widehat{A}) > \varrho$ or $\|\widehat{A}\| > \zeta$ or $\|\widehat{B}\| > \psi$ or $\sigma_d(\widehat{B}) < \gamma$ **then**

5: Set $\widehat{K}_{\text{plug}}(N) = 0$.

6: **else**

7: Set $\widehat{V} = \mathsf{dare}(\widehat{A}, \widehat{B}, I_n, 0)$ as the positive definite solution to[1]:

$$V = \widehat{A}^{\mathsf{T}} V \widehat{A} - \widehat{A}^{\mathsf{T}} V \widehat{B}(\widehat{B}^{\mathsf{T}} V \widehat{B})^{-1} \widehat{B}^{\mathsf{T}} V \widehat{A} + I_n .$$

8: Set $\widehat{K}_{\text{plug}}(N) = -(\widehat{B}^{\mathsf{T}}\widehat{V}\widehat{B})^{-1}\widehat{B}^{\mathsf{T}}\widehat{V}\widehat{A}$.

9: **end if**

10: return $\widehat{K}_{\text{plug}}(N)$.

Model-free algorithm. We study a model-free algorithm based on policy gradients (see e.g. [92, 126]). Here, we choose to parameterize the policy as a time-invariant linear feedback. The algorithm is described in Algorithm 7.

Algorithm 7 Model-free algorithm for policy optimization (REINFORCE) [92, 126].

Require: Horizon length T, rollouts N, baseline functions $\{\Psi_t(\cdot;\cdot)\}$, step-sizes $\{\alpha_i\}$, initial K_1, threshold ζ.

1: **for** $i = 1, ..., N$ **do**

2: Collect trajectory $\mathcal{T}^{(i)} := \{(x_t^{(i)}, u_t^{(i)})\}_{t=0}^{T}$ using feedback K_i.

3: Compute policy gradient g_i as: $g_i = \frac{1}{\sigma_\eta^2}\sum_{t=0}^{T-1} \eta_t^{(i)}(x_t^{(i)})^{\mathsf{T}}\Psi_t(\mathcal{T}^{(i)}; K_i)$.

4: Take policy gradient step: $K_{i+1} = \mathsf{Proj}_{\|\cdot\|\le\zeta}(K_i - \alpha_i g_i)$.

5: **end for**

6: Set $\widehat{K}_{\text{pg}}(N) = K_N$.

7: return $\widehat{K}_{\text{pg}}(N)$.

In general for problems with a continuous action space, when applying policy gradient one has many degrees of freedom in choosing how to represent the policy π. Some of these degrees of freedom include whether or not the policy should be time-invariant and how much of the history before time t should be used to compute the action at time t. More broadly,

the question is what function class should be used to model the policy. Ideally, one chooses a function class which is both capable of expressing the optimal solution and is easy to optimize over.

Another issue that significantly impacts the performance of policy gradient in practice is choosing a baseline which effectively reduces the variance of the policy gradient estimate. What makes computing a baseline challenging is that good baselines (such as value or advantage functions) require knowledge of the unknown MDP transition dynamics in order to compute. Therefore, one has to estimate the baseline from the empirical trajectories, adding another layer of complexity to the policy gradient algorithm.

In general, these issues are still an active area of research in RL and present many hurdles to a general theory for policy optimization. However, by restriction our attention to LQR, we can sidestep these issues which enables our analysis. In particular, by studying problems with no penalty on the input and where the state can be canceled at every step, we know that the optimal control is a static time-invariant linear feedback. Therefore, we can restrict our policy representation to static linear feedback controllers without introducing any approximation error. Furthermore, it turns out that the specific assumptions on (A, B) that we impose imply that the optimization landscape satisfies a standard notion of restricted strong convexity. This allows us to study policy gradient by leveraging the existing theory on the asymptotic distribution of stochastic gradient descent for strongly convex objectives. Finally, we can compute many of the standard baselines used in closed form, which further enables our analysis.

We note that in the literature, the model-based method is often called *nominal control* or the *certainty equivalence principle*. As noted in Section 5.2, one issue with this approach is that on an infinite horizon, there is no guarantee of robust stability with nominal control. However, as we are dealing with only finite horizon problems, the notion of stability is irrelevant.

Our first result for policy optimization gives the asymptotic risk of the model-based Algorithm 6.

Theorem 6.2.4. *Let (A, B) be such that A is stable, $\mathrm{range}(A) \subseteq \mathrm{range}(B)$, and B has full column rank. We have that the model-based plugin Algorithm 6 with thresholds $(\varrho, \zeta, \psi, \gamma)$ such that $\varrho \in (\rho(A), 1)$, $\zeta \in (\|A\|, \infty)$, $\psi \in (\|B\|, \infty)$, and $\gamma \in (0, \sigma_d(B))$ satisfies the asymptotic risk bound:*

$$\lim_{N\to\infty} N \cdot \mathbb{E}[J(\widehat{K}_{\mathrm{plug}}(N)) - J_\star] = \mathcal{O}(d(\mathrm{tr}(P_\infty^{-1}) + \|K_\star\|_F^2)) + o_T(1)\,.$$

Here, $P_\infty = \mathsf{dlyap}(A, \sigma_\eta^2 BB^\mathsf{T} + \sigma_w^2 I_n)$ is the steady-state convariance of the system driven with control input $u_t \sim \mathcal{N}(0, \sigma_\eta^2 I_d)$, $K_\star$ is the optimal controller, and $\mathcal{O}(\cdot)$ hides constants depending only on $\sigma_w^2, \sigma_\eta^2$.

We can interpret Theorem 6.2.4 by upper bounding $P_\infty^{-1} \preceq \sigma_w^{-2} I_n$. In this case if $\|K_\star\|_F^2 \le \mathcal{O}(n)$, then this result states that the asymptotic risk scales as $\mathcal{O}(nd/N)$. Similar to Theorem 6.2.1, Theorem 6.2.4 requires the setting of thresholds $(\varrho, \zeta, \psi, \gamma)$. These

thresholds serve two purposes. First, they ensure the existence of a unique positive definite solution to the discrete algebraic Riccati solution with the input penalty $R = 0$ (the details of this are worked out in Section 6.2.7.2). Second, they simplify various technical aspects of the proof related to uniform integrability. In practice, such strong thresholds are not needed, and we leave either removing them or relaxing their requirements to future work.

Next, we look at the model-free case. As mentioned previously, baselines are very influential on the behavior of policy gradient. In our analysis, we consider three different baselines:

$$\Psi_t(\mathcal{T}; K) = \sum_{\ell=t+1}^{T} \|x_\ell\|_2^2 \,, \qquad (\textbf{Simple} \text{ baseline } b_t(x_t; K) = \|x_t\|_2^2.)$$

$$\Psi_t(\mathcal{T}; K) = \sum_{\ell=t}^{T} \|x_\ell\|_2^2 - V_t^K(x_t) \,, \quad (\textbf{Value function} \text{ baseline } b_t(x_t; K) = V_t^K(x_t).)$$

$$\Psi_t(\mathcal{T}; K) = A_t^K(x_t, u_t) \,. \qquad (\textbf{Advantage} \text{ baseline } A_t^K(x_t, u_t) = Q_t^K(x_t, u_t) - V_t^K(x_t).)$$

Above, the simple baseline should be interpreted as having effectively no baseline; it turns out to simplify the variance calculations. On the other hand, the value function baseline V_t^K is a very popular heuristic used in practice [92]. Typically one has to actually estimate the value function for a given policy, since computing it requires knowledge of the model dynamics. In our analysis however, we simply assume the true value function is known. While this is an unrealistic assumption in practice, we note that this assumption substantially reduce the variance of policy gradient, and hence only serves to reduce the asymptotic risk. The last baseline we consider is to use the advantage function A_t^K. Using advantage functions has been shown to be quite effective in practice [101]. It has the same issue as the value function baseline in that it needs to be estimated from the data; once again in our analysis we simply assume we have access to the true advantage function.

Our main result for model-free policy optimization is the following asymptotic risk lower bound on Algorithm 7.

Theorem 6.2.5. *Let (A, B) be such that A is stable,* $\mathrm{range}(A) \subseteq \mathrm{range}(B)$*, and B has full column rank. Consider Algorithm 7 with $K_1 = 0_{d\times n}$, step-sizes $\alpha_i = [2(T-1)\sigma_w^2\sigma_d(B)^2 \cdot i]^{-1}$, and threshold $\zeta \in (\|K_\star\|, \infty)$. We have that the risk is lower bounded by:*

$$\liminf_{N\to\infty} N \cdot \mathbb{E}[J(\widehat{K}_{\mathrm{pg}}(N)) - J_\star] \geq \frac{1}{\sigma_d(B)^2(1+\|B\|^2)} \times \begin{cases} \Omega(T^2 d(n + \|B\|_F^2)^3) + o_T(T^2) & \textit{(Simple baseline)} \\ \Omega(Td(n+\|B\|_F^2)(n + \|B^\mathsf{T}B\|_F^2)) + o_T(T) & \textit{(Value function baseline)} \\ \Omega(d(n+\|B\|_F^2)\|B^\mathsf{T}B\|_F^2) & \textit{(Advantage baseline)} \end{cases}.$$

Here, $\Omega(\cdot)$ hides constants depending only on $\sigma_w^2, \sigma_\eta^2$.

In order to interpret Theorem 6.2.5, we consider a restricted family of instances (A, B). For a $\rho \in (0, 1)$ and $1 \le d \le n$, we define the family $\mathscr{G}(\rho, d)$ over (A, B) as:

$$\mathscr{G}(\rho, d) := \{(\rho U_\star U_\star^\mathsf{T}, \rho U_\star) : U_\star \in \mathbb{R}^{n \times d} \ , \ U_\star^\mathsf{T} U_\star = I_d\} \ .$$

This is a simple family where the A matrix is stable and contractive, and furthermore we have range(A) = range(B). The optimal feedback is $K_\star = -U_\star^\mathsf{T}$ for each of these instances.

Theorem 6.2.5 states that for instances from $\mathscr{G}(\rho, d)$, the simple baseline has risk $\Omega(T^2 \cdot dn^3/N)$, the value function baseline has risk $\Omega(T \cdot dn^2/N)$, and the advantage baseline has risk $\Omega(d^2 n/N)$. On the other hand, Theorem 6.2.4 states that the model-based risk is upper bounded by $\mathcal{O}(nd/N)$, which is less than the lower bound for all baselines considered in Theorem 6.2.5. For the simple and value function baselines, we see that the sample complexity of the model-free policy gradient method is several factors of n and T more than the model-based method. The extra factors of the horizon length appear due to the large variance of the policy gradient estimator without the variance reduction effects of the advantage baseline. The advantage baseline performs the best, only one factor of d more than the model-based method.

We note that we prove Theorem 6.2.5 with a specific choice of step size α_i. This step size corresponds to the standard $1/(mt)$ step sizes commonly found in proofs for SGD on strongly convex functions (see e.g. Rakhlin et al. [94]), where m is the strong convexity parameter. We leave to future work extending our results to support Polyak-Ruppert averaging, which would yield asymptotic results that are more robust to specific step size choices.

Finally, we turn to our information-theoretic lower bound for any (possibly adaptive) method over the family $\mathscr{G}(\rho, d)$.

Theorem 6.2.6. *Fix a $d \le n/2$ and suppose $d(n-d)$ is greater than an absolute constant. Consider the family $\mathscr{G}(\rho, d)$ as describe above. Fix a time horizon T and number of rollouts N. The risk over any algorithm $\mathcal{A}$ which plays (possibly adaptive) feedbacks of the form $u_t = K_i x_t + \eta_t$ with $\|K_i\| \le 1$ and $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2 I_d)$ is lower bounded by:*

$$\inf_{\mathcal{A}} \sup_{\substack{\rho \in (0,1/4), \\ (A,B) \in \mathscr{G}(d,\rho)}} \mathbb{E}[J(\mathcal{A}) - J_\star] \gtrsim \frac{\sigma_w^4}{\sigma_w^2 + \sigma_\eta^2} \frac{d(n-d)}{N} \ .$$

Observe that this bound is $\Omega(nd/N)$. Therefore, Theorem 6.2.6 tells us that asymptotically, the model-based method in Algorithm 6 is optimal in terms of its dependence on the state and input dimensions n and d over the family $\mathscr{G}(\rho, d)$.

6.2.4 Asymptotic Toolbox

Our analysis relies heavily on computing limiting distributions for the various estimators we study. A crucial fact we use is that if the matrix $L_\star$ is stable, then the Markov chain $\{x_t\}$ given by $x_{t+1} = L_\star x_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ is geometrically ergodic. This allows us

to apply well known limit theorems for ergodic Markov chains. In what follows, we let $\xrightarrow{\text{a.s.}}$ denote almost sure convergence and $\overset{D}{\rightsquigarrow}$ denote convergence in distribution.

Our main limit theorem is the following CLT for ergodic Markov chains.

Theorem 6.2.7 (Corollary 2 of Jones [54])**.** *Suppose that $\{x_t\}_{t=0}^{\infty} \subseteq X$ is a geometrically ergodic (Harris) Markov chain with stationary distribution π. Let $f : X \to \mathbb{R}$ be a Borel function. Suppose that $\mathbb{E}_\pi[|f|^{2+\delta}] < \infty$ for some $\delta > 0$. Then for any initial distribution, we have:*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}_\pi[f(x)]\right) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_f^2)\,,$$

where

$$\sigma_f^2 := \mathrm{Var}_\pi(f(x_0)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(f(x_0), f(x_i))\,.$$

We first state a well-known result that concerns the least-squares estimator of a stable dynamical system. In the scalar case, this result dates back to Mann and Wald [76].

Lemma 6.2.8. *Let $x_{t+1} = L_\star x_t + w_t$ be a dynamical system with $L_\star$ stable and $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$. Given a trajectory $\{x_t\}_{t=0}^{T}$, let $\widehat{L}(T)$ denote the least-squares estimator of $L_\star$ with regularization $\lambda \geq 0$:*

$$\widehat{L}(T) = \arg\min_{L \in \mathbb{R}^{n\times n}} \frac{1}{2}\sum_{t=0}^{T-1} \|x_{t+1} - Lx_t\|_2^2 + \frac{\lambda}{2}\|L\|_F^2\,.$$

Let P_∞ denote the stationary covariance matrix of the process $\{x_t\}_{t=0}^{\infty}$, i.e. $L_\star P_\infty L_\star^{\mathsf{T}} - P_\infty + \sigma_w^2 I_n = 0$. We have that $\widehat{L}(T) \xrightarrow{\text{a.s.}} L_\star$ and furthermore:

$$\sqrt{T}\mathrm{vec}(\widehat{L}(T) - L_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_w^2(P_\infty^{-1} \otimes I_n))\,.$$

Proof. Let $X \in \mathbb{R}^{T\times n}$ be the data matrix with rows $(x_0, ..., x_{T-1})$ and $W \in \mathbb{R}^{T\times n}$ be the noise matrix with rows $(w_0, ..., w_{T-1})$. We write:

$$\widehat{L}(T) - L_\star = -\lambda L_\star(X^{\mathsf{T}}X + \lambda I_n)^{-1} + W^{\mathsf{T}}X(X^{\mathsf{T}}X + \lambda I_n)^{-1}\,.$$

Using the fact that $\mathrm{vec}(AXB) = (B^{\mathsf{T}} \otimes A)\mathrm{vec}(X)$,

$$\sqrt{T}\mathrm{vec}(\widehat{L}(T) - L_\star) = -\sqrt{T}\mathrm{vec}(\lambda L_\star(X^{\mathsf{T}}X + \lambda I_n)^{-1}) + ((T^{-1}X^{\mathsf{T}}X)^{-1} \otimes I_n)\mathrm{vec}(T^{-1/2}W^{\mathsf{T}}X)\,.$$

It is well-known that $\{x_t\}$ is geometrically ergodic (see e.g. Mokkadem [82]), and therefore the augmented Markov chain $\{(x_t, w_t)\}$ is geometrically ergodic as well. By Theorem 6.2.7 combined with the Cramér-Wold theorem we conclude:

$$\mathrm{vec}(T^{-1/2}W^{\mathsf{T}}X) = T^{-1/2}\sum_{t=1}^{T} \mathrm{vec}(w_t x_t^{\mathsf{T}}) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \mathbb{E}_{x\sim\nu_\infty, w}[\mathrm{vec}(wx^{\mathsf{T}})\mathrm{vec}(wx^{\mathsf{T}})^{\mathsf{T}}])\,.$$

Above, we let ν_∞ denote the stationary distribution of $\{x_t\}$. We note that the cross-correlation terms disappear in the asymptotic covariance due to the martingale property of $\sum_{t=0}^{T-1} w_t x_t^\mathsf{T}$. We now use the identity $\mathrm{vec}(wx^\mathsf{T}) = (x \otimes I_n)w$ and compute

$$\begin{aligned}\mathbb{E}_{x\sim\nu_\infty,w}[\mathrm{vec}(wx^\mathsf{T})\mathrm{vec}(wx^\mathsf{T})^\mathsf{T}] &= \mathbb{E}_{x\sim\nu_\infty,w}[(x\otimes I_n)ww^\mathsf{T}(x^\mathsf{T}\otimes I_n)] \\ &= \sigma_w^2\mathbb{E}_{x\sim\nu_\infty}[(x\otimes I_n)(x^\mathsf{T}\otimes I_n)] \\ &= \sigma_w^2\mathbb{E}_{x\sim\nu_\infty}[(xx^\mathsf{T}\otimes I_n)] \\ &= \sigma_w^2(P_\infty\otimes I_n)\,.\end{aligned}$$

We have that $T^{-1}X^\mathsf{T}X \xrightarrow{\text{a.s.}} P_\infty$ by the ergodic theorem. Therefore by the continuous mapping theorem followed by Slutsky's theorem, we have that

$$((T^{-1}X^\mathsf{T}X)^{-1}\otimes I_n)\mathrm{vec}(T^{-1/2}W^\mathsf{T}X) \overset{D}{\rightsquigarrow} \mathcal{N}(0,\sigma_w^2(P_\infty^{-1}\otimes I_n))\,.$$

On the other hand, we have:

$$\sqrt{T}\mathrm{vec}(\lambda L_\star(X^\mathsf{T}X+\lambda I_n)^{-1}) = \frac{1}{\sqrt{T}}\mathrm{vec}(\lambda L_\star(T^{-1}X^\mathsf{T}X+T^{-1}\lambda I_n)^{-1}) \xrightarrow{\text{a.s.}} 0\,.$$

The claim now follows by another application of Slutsky's theorem. □

We now consider a slightly altered process where the system is no longer autonomous, and instead will be driven by white noise.

Lemma 6.2.9. *Let $x_{t+1} = Ax_t + Bu_t + w_t$ be a stable dynamical system driven by $u_t \sim \mathcal{N}(0,\sigma_\eta^2 I_d)$ and $w_t \sim \mathcal{N}(0,\sigma_w^2 I_n)$. Consider a least-squares estimator $\widehat{\Theta}$ of $\Theta_\star := (A,B) \in \mathbb{R}^{n\times(n+d)}$ based off of N independent trajectories of length T, i.e. $\{\{z_t^{(i)} := (x_t^{(i)}, u_t^{(i)})\}_{t=0}^T\}_{i=1}^N$,*

$$\widehat{\Theta}(N) = \arg\min_{(A,B)\in\mathbb{R}^{n\times(n+d)}} \frac{1}{2}\sum_{i=1}^N\sum_{t=0}^{T-1}\|x_{t+1}^{(i)} - Ax_t^{(i)} - Bu_t^{(i)}\|_2^2 + \frac{\lambda}{2}\|\begin{bmatrix}A & B\end{bmatrix}\|_F^2\,.$$

Let P_∞ denote the stationary covariance of the process $\{x_t\}_{t=0}^\infty$, i.e. P_∞ solves

$$AP_\infty A^\mathsf{T} - P_\infty + \sigma_\eta^2 BB^\mathsf{T} + \sigma_w^2 I_n = 0\,.$$

We have that $\widehat{\Theta}(N) \xrightarrow{\text{a.s.}} \Theta_\star$ and furthermore:

$$\sqrt{N}\mathrm{vec}(\widehat{\Theta}(N)-\Theta_\star) \overset{D}{\rightsquigarrow} \mathcal{N}\left(0, \frac{\sigma_w^2}{T}\begin{bmatrix}P_\infty^{-1} & 0\\ 0 & (1/\sigma_\eta^2)I_d\end{bmatrix}\otimes I_n + o(1/T)\right)\,.$$

Proof. Let $Z^{(i)} \in \mathbb{R}^{T\times(n+d)}$ be a data matrix with the rows $(z_0^{(i)}, ..., z_{T-1}^{(i)})$, and let $W^{(i)} \in \mathbb{R}^{T\times n}$ be the noise matrix with the rows $(w_0^{(i)}, ..., w_{T-1}^{(i)})$. With this notation we write:

$$\begin{aligned}
\widehat{\Theta}(N) - \Theta_\star &= \left(\sum_{i=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1} z_{t+1}^{(i)}(z_t^{(i)})^\mathsf{T}\right)\left(\sum_{i=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1} z_t^{(i)}(z_t^{(i)})^\mathsf{T} + \lambda I_{n+d}\right)^{-1} - \Theta_\star \\
&= \Theta_\star\left(\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)}\right)\left(\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \lambda I_{n+d}\right)^{-1} - \Theta_\star \\
&\quad + \left(\sum_{i=1}^{N}\frac{1}{T}(W^{(i)})^\mathsf{T}Z^{(i)}\right)\left(\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&= -\lambda\Theta_\star\left(\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&\quad + \left(\sum_{i=1}^{N}\frac{1}{T}(W^{(i)})^\mathsf{T}Z^{(i)}\right)\left(\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&=: G_1(N) + G_2(N)\,.
\end{aligned}$$

Taking vec of $G_2(N)$:

$$\mathrm{vec}(G_2(N)) = \left(\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \frac{\lambda}{N}I_{n+d}\right)^{-1} \otimes I_n\right)\mathrm{vec}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1} w_t^{(i)}(z_t^{(i)})^\mathsf{T}\right)\,.$$

Now we write $\mathrm{vec}(w_t z_t^\mathsf{T}) = (z_t \otimes I_n)w_t$ and hence

$$\begin{aligned}
\mathbb{E}\left[\mathrm{vec}\left(\frac{1}{T}\sum_{t=0}^{T-1} w_t z_t^\mathsf{T}\right)\mathrm{vec}\left(\frac{1}{T}\sum_{t=0}^{T-1} w_t z_t^\mathsf{T}\right)^\mathsf{T}\right] &= \frac{1}{T^2}\sum_{t_1,t_2=0}^{T-1}\mathbb{E}[(z_{t_1}\otimes I_n)w_{t_1}w_{t_2}^\mathsf{T}(z_{t_2}^\mathsf{T}\otimes I_n)] \\
&= \frac{\sigma_w^2}{T^2}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}]\otimes I_n\,.
\end{aligned}$$

We have that:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}(Z^{(i)})^\mathsf{T}Z^{(i)} + \frac{\lambda}{N}I_{n+d} \xrightarrow{\text{a.s.}} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}]\,.$$

Hence by the central limit theorem combined with the continuous mapping theorem and

Slutsky's theorem,

$$\begin{aligned}
&\sqrt{N}\mathrm{vec}(G_1(N)) \xrightarrow{\text{a.s.}} 0\,,\\
&\sqrt{N}\mathrm{vec}(G_2(N)) \overset{D}{\rightsquigarrow} \mathcal{N}\left(0, \frac{\sigma_w^2}{T}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}]\right]^{-1} \otimes I_n\right)\\
&\qquad\qquad = \mathcal{N}\left(0, \frac{\sigma_w^2}{T}\begin{bmatrix}[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[x_t x_t^\mathsf{T}]]^{-1} & 0\\ 0 & (1/\sigma_\eta^2) I_d\end{bmatrix} \otimes I_n\right)\,.
\end{aligned}$$

To finish the proof, we note that $\mathbb{E}[x_t x_t^\mathsf{T}] = \sum_{\ell=0}^{t-1} A^\ell M (A^\ell)^\mathsf{T} := P_t$ with $M := \sigma_\eta^2 BB^\mathsf{T} + \sigma_w^2 I_n$ and $P_0 = 0$ (since $x_0 = 0$). Since A is stable, there exists a $\rho \in (0,1)$ and $C > 0$ such that $\|A^k\| \le C\rho^k$ for all $k \ge 0$. Hence,

$$\|P_\infty - P_t\| = \left\|\sum_{\ell=t}^{\infty} A^\ell M (A^\ell)^\mathsf{T}\right\| \le C^2\|M\|\sum_{\ell=t}^{\infty}\rho^{2\ell} = C^2\|M\|\frac{\rho^{2t}}{1-\rho^2}\,.$$

Therefore,

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=0}^{T-1} P_t - P_\infty\right\| &= \left\|\frac{1}{T}\sum_{t=1}^{T-1}(P_t - P_\infty) + \frac{1}{T}P_\infty\right\|\\
&\le \frac{1}{T}\sum_{t=1}^{T-1}\|P_\infty - P_t\| + \frac{1}{T}\|P_\infty\|\\
&\le \frac{C^2\|M\|}{T(1-\rho^2)}\sum_{t=1}^{T-1}\rho^{2t} + \frac{1}{T}\|P_\infty\|\\
&\le \frac{C^2\|M\|}{T(1-\rho^2)^2} + \frac{1}{T}\|P_\infty\| = \mathcal{O}(1/T)\,.
\end{aligned}$$

Therefore, $[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[x_t x_t^\mathsf{T}]]^{-1} = P_\infty^{-1} + \mathcal{O}(1/T)$ from which the claim follows. □

Next, we consider the asymptotic distribution of Least-Squares Temporal Difference Learning for LQR.

Lemma 6.2.10. *Let $x_{t+1} = Ax_t + Bu_t + w_t$ be a linear system driven by $u_t = Kx_t$ and $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$. Suppose the closed-loop matrix $A + BK$ is stable. Let ν_∞ denote the stationary distribution of the Markov chain $\{x_t\}_{t=0}^\infty$. Define the two matrices A_∞, B_∞, the*

mapping $\psi(x)$, and the vector $w_\star$ as

$$\begin{aligned} A_\infty &:= \mathop{\mathbb{E}}_{\substack{x\sim\nu_\infty,\\ x'\sim p(\cdot|x,\pi(x))}} [\phi(x)(\phi(x)-\phi(x'))^\mathsf{T}] \,, \\ B_\infty &:= \mathop{\mathbb{E}}_{\substack{x\sim\nu_\infty,\\ x'\sim p(\cdot|x,\pi(x))}} [((\phi(x')-\psi(x))^\mathsf{T} w_\star)^2 \phi(x)\phi(x)^\mathsf{T}] \,, \\ \psi(x) &:= \mathop{\mathbb{E}}_{x'\sim p(\cdot|x,\pi(x))} [\phi(x')] \,, \\ w_\star &:= \mathrm{svec}(V_\star) \,. \end{aligned}$$

Let $\widehat{w}_{\mathrm{lstd}}(T)$ denote the LSTD estimator given by:

$$\widehat{w}_{\mathrm{lstd}}(T) = \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1} (c_t-\lambda_t)\phi(x_t)\right) \,.$$

Suppose that LSTD is run with the true $\lambda_t = \lambda_\star := \sigma_w^2 \operatorname{tr}(V_\star)$ and that the matrix A_∞ is invertible. We have that $\widehat{w}_{\mathrm{lstd}}(T) \xrightarrow{\text{a.s.}} w_\star$ and furthermore:

$$\sqrt{T}(\widehat{w}_{\mathrm{lstd}}(T) - w_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}}) \,.$$

Proof. Let $c_t = x_t^\mathsf{T}(S + K^\mathsf{T} R K)x_t$. From Bellman's equation, we have $c_t - \lambda_\star = (\phi(x_t) - \psi(x_t))^\mathsf{T} w_\star$. We write:

$$\begin{aligned} \widehat{w}_{\mathrm{lstd}}(T) - w_\star &= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1} (c_t-\lambda_\star)\phi(x_t)\right) - w_\star \\ &= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\psi(x_t))^\mathsf{T}\right) w_\star - w_\star \\ &= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_{t+1})-\psi(x_t))^\mathsf{T} w_\star\right) \\ &= \left(\frac{1}{T}\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\frac{1}{T}\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_{t+1})-\psi(x_t))^\mathsf{T} w_\star\right) \,. \end{aligned}$$

We now proceed by considering the Markov chain $\{z_t := (x_t, w_t)\}$. Observe that x_{t+1} is z_t-measurable, and furthermore the stationary distribution of this chain is $\nu_\infty \times \mathcal{N}(0, \sigma_w^2 I_n)$. From this we conclude two things. First, we conclude by the ergodic theorem that the term inside the inverse converges a.s. to A_∞ and hence the inverse converges a.s. to A_∞^{-1} by the continuous mapping theorem. Next, Theorem 6.2.7 combined with the Cramér-Wold theorem allows us to conclude that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \phi(x_t)(\phi(x_{t+1})-\psi(x_t))^\mathsf{T} w_\star \overset{D}{\rightsquigarrow} \mathcal{N}(0, B_\infty) \,.$$

The final claim now follows by Slutsky's theorem. □

As a corollary to Lemma 6.2.10, we work out the formulas for A_∞ and B_∞ and a useful lower bound.

Corollary 6.2.11. *In the setting of Lemma 6.2.10, with $L_\star = A + BK$, we have that the matrix A_∞ is invertible, and:*

$$\begin{aligned}
A_\infty &= (P_\infty \otimes_s P_\infty) - (P_\infty L_\star^\mathsf{T} \otimes_s P_\infty L_\star^\mathsf{T}) \,, \\
B_\infty &= (\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} V_\star^2 L_\star \rangle + 2\sigma_w^4 \|V_\star\|_F^2)(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\
&\quad + 2\sigma_w^2 (\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L_\star^\mathsf{T} V_\star^2 L_\star P_\infty)^\mathsf{T} + \mathrm{svec}(P_\infty L_\star^\mathsf{T} V_\star^2 L_\star P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\
&\quad + 8\sigma_w^2 (P_\infty L_\star^\mathsf{T} V_\star^2 L_\star P_\infty \otimes_s P_\infty) \,.
\end{aligned}$$

Furthermore, we can lower bound the matrix $A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}}$ by:

$$\begin{aligned}
A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}} \succeq\ & 8\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} V_\star^2 L_\star \rangle (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (P_\infty^{-1} \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}} \\
& + 16\sigma_w^2 (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (L_\star^\mathsf{T} V_\star^2 L_\star \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}} \,. \qquad (6.2.5)
\end{aligned}$$

The proof of Corollary 6.2.11 is involved and deferred to the end of this section. Next, we state a standard lemma which we will use to convert convergence in distribution guarantees to guarantees regarding the convergence of risk.

Lemma 6.2.12. *Suppose that $\{X_n\}$ is a sequence of random vectors and $X_n \overset{D}{\rightsquigarrow} X$. Suppose that f is a non-negative continuous real-valued function such that $\mathbb{E}[f(X)] < \infty$. We have that:*

$$\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)] \,.$$

If additionally we have $\sup_{n\geq 1} \mathbb{E}[f(X_n)^{1+\varepsilon}] < \infty$ holds for some $\varepsilon > 0$, then the limit $\lim_{n\to\infty} \mathbb{E}[f(X_n)]$ exists and

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \,.$$

Proof. Both facts are standard consequences of weak convergence of probability measures; see e.g. Chapter 5 of Billingsley [17] for more details. □

The next claim uniformly controls the p-th moments of the regularized least-squares estimate when T is large enough. This technical result will allow us to invoke Lemma 6.2.12 to obtain convergence in L^p.

Lemma 6.2.13. *Let $x_{t+1} = L_\star x_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ and $L_\star$ stable. Fix a regularization parameter $\lambda > 0$ and let $\widehat{L}(T)$ denote the LS estimator:*

$$\widehat{L}(T) = \arg\min_{L\in\mathbb{R}^{n\times n}} \frac{1}{2} \sum_{t=0}^{T-1} \|x_{t+1} - L x_t\|_2^2 + \frac{\lambda}{2} \|L\|_F^2 \,.$$

Fix a finite $p \geq 1$. Let $C_{L_\star,\lambda,n}$ and $C_{L_\star,\lambda,n,p}$ denote constants that depend only on $L_\star, \lambda, n$ (resp. $L_\star, \lambda, n, p$) and not on T, δ. Fix a $\delta \in (0,1)$. With probability at least $1-\delta$, as long as $T \geq C_{L_\star,\lambda,n} \log(1/\delta)$ we have:

$$\|\widehat{L}(T) - L_\star\| \leq C'_{L_\star,\lambda,n} \sqrt{\frac{\log(1/\delta)}{T}} \, .$$

Furthermore, as long as $T \geq C_{L_\star,\lambda,n,p}$, then:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \leq C'_{L_\star,\lambda,n,p} \frac{1}{T^{p/2}} \, .$$

Proof. Recall in the notation of the proof of Lemma 6.2.8,

$$\widehat{L}(T) - L_\star = -\lambda L_\star (X^\mathsf{T} X + \lambda I_n)^{-1} + W^\mathsf{T} X (X^\mathsf{T} X + \lambda I_n)^{-1} \, .$$

Now let us suppose that we are on an event where $X^\mathsf{T} X$ is invertible. Let $X = U \Sigma V^\mathsf{T}$ denote the compact SVD of X. We have:

$$\begin{aligned} \|\widehat{L}(T) - L_\star\| &\leq \lambda \frac{\|L_\star\|}{\lambda_{\min}(X^\mathsf{T} X + \lambda I_n)} + \|W^\mathsf{T} X (X^\mathsf{T} X + \lambda I_n)^{-1}\| \\ &\overset{(a)}{\leq} \lambda \frac{\|L_\star\|}{\lambda_{\min}(X^\mathsf{T} X + \lambda I_n)} + \|W^\mathsf{T} X (X^\mathsf{T} X)^{-1}\| \, . \end{aligned}$$

The inequality (a) holds due to the following. First observe that $(X^\mathsf{T} X + \lambda I_n)^{-2} \preceq (X^\mathsf{T} X)^{-2}$. Therefore with $M = W^\mathsf{T} X$, conjugating both sides by M, we have $M(X^\mathsf{T} X + \lambda I_n)^{-2} M^\mathsf{T} \preceq M (X^\mathsf{T} X)^{-2} M^\mathsf{T}$. Hence,

$$\begin{aligned} \|M(X^\mathsf{T} X + \lambda I_n)^{-1}\| &= \sqrt{\lambda_{\max}(M(X^\mathsf{T} X + \lambda I_n)^{-2} M^\mathsf{T})} \\ &\leq \sqrt{\lambda_{\max}(M(X^\mathsf{T} X)^{-2} M^\mathsf{T})} \\ &= \|M(X^\mathsf{T} X)^{-1}\| \, . \end{aligned}$$

By Theorem 2.4 of Simchowitz et al. [105] for $T \geq C_{L_\star,n} \log(1/\delta)$, there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ such that on $\mathcal{E}$ we have:

$$\|\widehat{L}_{\mathrm{ols}}(T) - L_\star\| \leq C'_{L_\star,n} \sqrt{\log(1/\delta)/T} \, , \;\; X^\mathsf{T} X \succeq C''_{L_\star,n} T \cdot I_n \, .$$

Hence on this event we have $\|\widehat{L}(T) - L_\star\| \leq C'_{L_\star,n,\lambda} \sqrt{\log(1/\delta)/T}$.

For the remainder of the proof, $O(\cdot)$ will hide constants that depend on $L_\star, n, p, \lambda$ but not on T or δ. We bound the p-th moment as follows. We decompose:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] = \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}}] + \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] \, .$$

On $\mathcal{E}$ we have by the inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ for non-negative a, b,

$$\|\widehat{L}(T) - L_\star\|^p \leq 2^{p-1}(O(\lambda^p/T^p) + O((\log(1/\delta)/T)^{p/2})) \, .$$

On the other hand, we always have:

$$\|\widehat{L}(T) - L_\star\|^p \le 2^{p-1}(\|L_\star\|^p + (\|W^\mathsf{T} X\|/\lambda)^p) \,.$$

Hence:

$$\begin{aligned}\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] &\le 2^{p-1}\|L_\star\|^p \mathbb{P}(\mathcal{E}^c) + \frac{2^{p-1}}{\lambda^p}\mathbb{E}[\|W^\mathsf{T} X\|^p \mathbf{1}_{\mathcal{E}^c}] \\ &\le 2^{p-1}\|L_\star\|^p \delta + \frac{2^{p-1}}{\lambda^p}\sqrt{\mathbb{E}[\|W^\mathsf{T} X\|^{2p}]\delta} \,.\end{aligned}$$

We will now compute a very crude bound on $\mathbb{E}[\|W^\mathsf{T} X\|^{2p}]$ which will suffice. For non-negative a_t, we have $(a_1 + ... + a_T)^{2p} \le T^{2p-1}(\sum_{t=1}^T a_i^{2p})$ by Hölder's inequality. Hence

$$\begin{aligned}\mathbb{E}[\|W^\mathsf{T} X\|^{2p}] &= \mathbb{E}\left[\left\|\sum_{t=0}^{T-1} w_i x_i^\mathsf{T}\right\|^{2p}\right] \\ &\le T^{2p-1}\mathbb{E}\left[\sum_{t=1}^{T}\|w_t\|^{2p}\|x_t\|^{2p}\right] \\ &= T^{2p-1}\mathbb{E}[\|w_1\|^{2p}]\sum_{t=1}^{T}\mathbb{E}[\|x_t\|^{2p}] \\ &\le T^{2p}\mathbb{E}[\|w_1\|^{2p}]\|P_\infty\|^p \mathbb{E}_{g\sim\mathcal{N}(0,I)}[\|g\|^{2p}] \\ &= \mathcal{O}(T^{2p}) \,.\end{aligned}$$

Above, P_∞ denotes the covariance of the stationary distribution of $\{x_t\}$. Continuing from above:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] = 2^{p-1}\|L_\star\|^p \delta + \frac{2^{p-1}}{\lambda^p}\sqrt{\mathcal{O}(T^{2p})\delta} \,.$$

We now set $\delta = \mathcal{O}(1/T^{3p})$ so that the term above is $\mathcal{O}(1/T^{p/2})$. Doing this we obtain that for T sufficiently large (as a function of only $L_\star, p, \lambda$),

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \le \mathcal{O}(1/T^{p/2}) \,.$$

□

The next result is the analogue of Lemma 6.2.13 for the non-autonomous system driven by white noise.

Lemma 6.2.14. *Let $x_{t+1} = Ax_t + Bu_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$, $u_t \sim \mathcal{N}(0, \sigma_\eta^2 I_d)$, and A stable. Fix a regularization parameter $\lambda > 0$ and let $\widehat{\Theta}(N)$ denote the LS estimator:*

$$\widehat{\Theta}(N) = \arg\min_{(A,B)\in\mathbb{R}^{n\times(n+d)}} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\|x_{t+1}^{(i)} - Ax_t^{(i)} - Bu_t^{(i)}\|^2 + \frac{\lambda}{2}\|\begin{bmatrix}A & B\end{bmatrix}\|_F^2 \,.$$

Fix a finite $p \geq 1$. Let $C_{\Theta_\star,T,\lambda,n,d}$ and $C_{\Theta_\star,T,\lambda,n,d,p}$ denote constants that depend only on $\Theta_\star, T, \lambda, n, d$ (resp. $\Theta_\star, T, \lambda, n, d, p$) and not on N, δ. Fix a $\delta \in (0,1)$. With probability at least $1-\delta$, as long as $N \geq C_{\Theta_\star,T,\lambda,n,d} \log(1/\delta)$ we have:

$$\|\widehat{\Theta}(N) - \Theta_\star\| \leq C'_{\Theta_\star,T,\lambda,n,d} \sqrt{\frac{\log(1/\delta)}{N}} \, .$$

Furthermore, as long as $N \geq C_{\Theta_\star,T,\lambda,n,d,p}$, then:

$$\mathbb{E}[\|\widehat{\Theta}(N) - \Theta_\star\|^p] \leq C'_{\Theta_\star,T,\lambda,n,d,p} \frac{1}{N^{p/2}} \, .$$

Proof. The proof is nearly identical to that of Lemma 6.2.13, except we use the concentration result of Proposition 3.1.1 instead of Proposition 3.2.1 to establish concentration over multiple independent rollouts. We omit the details as they very closely mimic that of Lemma 6.2.13.

We note that in doing this we obtain a sub-optimal dependence on the horizon length T. This can be remedied by a more careful argument combining Proposition 3.1.1 with Proposition 3.2.1. However, as in our limit theorems only N the rollout length is being sent to infinity (e.g. T is considered a constant), a sub-optimal bound in T will suffice for our purpose. □

Our final asymptotic result deals with the performance of stochastic gradient descent (SGD) with projection. This will be our key ingredient in analyzing policy gradient (Algorithm 7). While the asymptotic performance of SGD (and more generally stochastic approximation) is well-established (see e.g. Kushner and Yin [60]), we consider a slight modification where the iterates are projected back into a compact convex set at every iteration. As long as the optimal solution is not on the boundary of the projection set, then one intuitively does not expect the asymptotic distribution to be affected by this projection, since eventually as SGD converges towards the optimal solution the projection step will effectively be inactive. Our result here makes this intuition rigorous. It follows by combining the asymptotic analysis of Toulis and Airoldi [113] with the high probability bounds for SGD from Rakhlin et al. [94].

To state the result, we need a few definitions. First, we say a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies *restricted strong convexity* (RSC) on a compact convex set $\Theta \subseteq \mathbb{R}^d$ if it has a unique minimizer $\theta_\star \in \text{int}(\Theta)$ and for some $m > 0$, we have $\langle \nabla F(\theta), \theta - \theta_\star \rangle \geq m\|\theta - \theta_\star\|^2$ for all $\theta \in \Theta$. We denote this by $\mathsf{RSC}(m, \Theta)$.

Lemma 6.2.15. *Let $F \in \mathcal{C}^3(\Theta)$ and suppose F satisfies $\mathsf{RSC}(m, \Theta)$. Let $\theta_\star \in \Theta$ denote the unique minimizer of F in Θ. Suppose we have a stochastic gradient oracle $g(\theta; \xi)$ such that g is continuous in both θ, ξ and $\nabla F(\theta) = \mathbb{E}_\xi[g(\theta; \xi)]$ for some distribution over ξ. Suppose*

that for some $G_1, G_2, L > 0$, for all $p \in [1, 4]$ and $\delta \in (0, 1)$, we have that

$$\sup_{\theta \in \Theta} \mathbb{E}_\xi[\|g(\theta; \xi)\|^p] \leq G_1^p \,, \tag{6.2.6}$$

$$\mathbb{P}_\xi \left(\sup_{\theta \in \Theta} \|g(\theta; \xi)\| > G_2 \operatorname{polylog}(1/\delta) \right) \leq \delta \,, \tag{6.2.7}$$

$$\mathbb{E}_\xi[\|g(\theta; \xi) - g(\theta_\star; \xi)\|^2] \leq L\|\theta - \theta_\star\|^2 \;\; \forall \theta \in \Theta \,. \tag{6.2.8}$$

Given an sequence $\{\xi_t\}_{t=1}^\infty$ drawn i.i.d. from the law of ξ, consider the sequence of iterates $\{\theta_t\}_{t=1}^\infty$ starting with $\theta_1 \in \Theta$ and defined as:

$$\theta_{t+1} = \mathsf{Proj}_\Theta(\theta_t - \alpha_t g(\theta_t; \xi_t)) \,, \;\; \alpha_t = \frac{1}{mt} \,.$$

We have that:

$$\lim_{T \to \infty} mT \cdot \operatorname{Var}(\theta_T) = \Xi \,, \tag{6.2.9}$$

where $\Xi = \mathsf{lyap}(\frac{m}{2} I_d - \nabla^2 F(\theta_\star), \mathbb{E}_\xi[g(\theta_\star; \xi) g(\theta_\star; \xi)^\mathsf{T}])$ solves the continuous-time Lyapunov equation:

$$\left(\frac{m}{2} I_d - \nabla^2 F(\theta_\star)\right) \Xi + \Xi \left(\frac{m}{2} I_d - \nabla^2 F(\theta_\star)\right) + \mathbb{E}_\xi[g(\theta_\star; \xi) g(\theta_\star; \xi)^\mathsf{T}] = 0 \,. \tag{6.2.10}$$

We also have that for any $G \in \mathcal{C}^3(\Theta)$ with $\nabla G(\theta_\star) = 0$ and $\nabla^2 G(\theta_\star) \succ 0$,

$$\liminf_{T \to \infty} T \cdot \mathbb{E}[G(\theta_T) - G(\theta_\star)] \geq \frac{1}{2m} \operatorname{tr}(\nabla^2 G(\theta_\star) \cdot \Xi) \,. \tag{6.2.11}$$

We defer the proof of this lemma to Section 6.2.5 of the Appendix. We quickly comment on how the last inequality can be used. Taking trace of both sides from Equation 6.2.10, we obtain:

$$\operatorname{tr}(\Xi \cdot (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) = \frac{1}{2} \mathbb{E}_\xi[\|g(\theta_\star; \xi)\|^2] \,.$$

We now upper bound the LHS as:

$$\begin{aligned}
&\operatorname{tr}(\Xi \cdot (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) \\
&= \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star)^{1/2} \cdot \nabla^2 G(\theta_\star)^{-1/2} (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d) \nabla^2 G(\theta_\star)^{-1/2} \cdot \nabla^2 G(\theta_\star)^{1/2}) \\
&\leq \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star)) \lambda_{\max}(\nabla^2 G(\theta_\star)^{-1/2} (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d) \nabla^2 G(\theta_\star)^{-1/2}) \\
&= \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star)) \lambda_{\max}(\nabla^2 G(\theta_\star)^{-1} (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) \,.
\end{aligned}$$

Combining the last two equations we obtain that:

$$\begin{aligned}\liminf_{T\to\infty} T\cdot\mathbb{E}[G(\theta_T)-G(\theta_\star)] &\geq \frac{1}{2m}\operatorname{tr}(\Xi\cdot\nabla^2 G(\theta_\star))\\ &\geq \frac{1}{4m\lambda_{\max}(\nabla^2 G(\theta_\star)^{-1}(\nabla^2 F(\theta_\star)-\frac{m}{2}I_d))}\mathbb{E}_\xi[\|g(\theta_\star;\xi)\|^2]\,.\end{aligned} \tag{6.2.12}$$

We will use this last estimate in our analysis.

6.2.4.1 Deferred Proof of Corollary 6.2.11

Proof. In the proof we write $\Sigma=\sigma_w^2 I_n$. First, we note that a quick computation shows that $\psi(x)=\operatorname{svec}(Lxx^\mathsf{T}L^\mathsf{T}+\Sigma)$.

Matrix A_∞. We have

$$\begin{aligned}\phi(x)-\phi(x') &= \operatorname{svec}(xx^\mathsf{T}-(Lx+w)(Lx+w)^\mathsf{T})\\ &= \operatorname{svec}(xx^\mathsf{T}-Lxx^\mathsf{T}L^\mathsf{T}-Lxw^\mathsf{T}-wx^\mathsf{T}L^\mathsf{T}-ww^\mathsf{T})\,.\end{aligned}$$

Hence, conditioning on x and iterating expectations, we have

$$A_\infty=\mathbb{E}_{x\sim\nu_\infty}[\phi(x)\operatorname{svec}(xx^\mathsf{T}-Lxx^\mathsf{T}L^\mathsf{T}-\Sigma)^\mathsf{T}]\,.$$

Now let m,n be two test vectors and $M=\operatorname{smat}(m), N=\operatorname{smat}(n)$. We have that,

$$\begin{aligned}m^\mathsf{T}A_\infty n &= \mathbb{E}_{x\sim\nu_\infty}[x^\mathsf{T}Mx\langle xx^\mathsf{T}-Lxx^\mathsf{T}L^\mathsf{T}-\Sigma,N\rangle]\\ &= \mathbb{E}_{x\sim\nu_\infty}[x^\mathsf{T}Mx(x^\mathsf{T}(N-L^\mathsf{T}NL)x-\langle\Sigma,N\rangle)]\\ &= \mathbb{E}_{x\sim\nu_\infty}[x^\mathsf{T}Mxx^\mathsf{T}(N-L^\mathsf{T}NL)x]-\langle\Sigma,N\rangle\mathbb{E}_{x\sim\nu_\infty}[x^\mathsf{T}Mx]\\ &= \mathbb{E}_g[g^\mathsf{T}P_\infty^{1/2}MP_\infty^{1/2}gg^\mathsf{T}P_\infty^{1/2}(N-L^\mathsf{T}NL)P_\infty^{1/2}g]-\langle\Sigma,N\rangle\langle M,P_\infty\rangle\\ &= 2\langle P_\infty^{1/2}MP_\infty^{1/2},P_\infty^{1/2}(N-L^\mathsf{T}NL)P_\infty^{1/2}\rangle+\langle M,P_\infty\rangle\langle N-L^\mathsf{T}NL,P_\infty\rangle\\ &\quad -\langle\Sigma,N\rangle\langle M,P_\infty\rangle\\ &= 2\langle P_\infty^{1/2}MP_\infty^{1/2},P_\infty^{1/2}(N-L^\mathsf{T}NL)P_\infty^{1/2}\rangle\,,\end{aligned}$$

where the last identity follows since $LP_\infty L^\mathsf{T}-P_\infty+\Sigma=0$. We therefore have:

$$\begin{aligned}A_\infty &= (P_\infty\otimes_s P_\infty)-(P_\infty L^\mathsf{T}\otimes_s P_\infty L^\mathsf{T})\\ &= (P_\infty\otimes_s P_\infty)(I-L^\mathsf{T}\otimes_s L^\mathsf{T})\,.\end{aligned}$$

Note that this writes A_∞ as the product of two invertible matrices and hence A_∞ is invertible.

Matrix B_∞. We have

$$\begin{aligned}\langle \phi(x') - \psi(x), w_\star \rangle &= \mathrm{svec}(Lxw^\mathsf{T} + wx^\mathsf{T}L^\mathsf{T} + ww^\mathsf{T} - \Sigma)^\mathsf{T} w_\star \\ &= 2x^\mathsf{T}L^\mathsf{T}V_\star w + \langle ww^\mathsf{T} - \Sigma, V_\star \rangle \,.\end{aligned}$$

Hence,

$$\begin{aligned}\langle \phi(x') - \psi(x), w_\star \rangle^2 &= 4(x^\mathsf{T}L^\mathsf{T}V_\star w)^2 + \langle ww^\mathsf{T} - \Sigma, V_\star \rangle^2 + 4x^\mathsf{T}L^\mathsf{T}V_\star w \langle ww^\mathsf{T} - \Sigma, V_\star \rangle \\ &=: T_1 + T_2 + T_3 \,.\end{aligned}$$

Now we have that $m^\mathsf{T}B_\infty n$ is

$$m^\mathsf{T}B_\infty n = \mathbb{E}[T_1 x^\mathsf{T}Mxx^\mathsf{T}Nx] + \mathbb{E}[T_2 x^\mathsf{T}Mxx^\mathsf{T}Nx] + \mathbb{E}[T_3 x^\mathsf{T}Mxx^\mathsf{T}Nx] \,. \tag{6.2.13}$$

First, we have

$$\begin{aligned}\mathbb{E}[T_1 x^\mathsf{T}Mxx^\mathsf{T}Nx] &= 4\mathbb{E}[(x^\mathsf{T}L^\mathsf{T}V_\star w)^2 x^\mathsf{T}Mxx^\mathsf{T}Nx] \\ &= 4\mathbb{E}[x^\mathsf{T}L^\mathsf{T}V_\star ww^\mathsf{T}V_\star Lxx^\mathsf{T}Mxx^\mathsf{T}Nx] \\ &= 4\mathbb{E}[x^\mathsf{T}L^\mathsf{T}V_\star \Sigma V_\star Lxx^\mathsf{T}Mxx^\mathsf{T}Nx] \\ &= 4\mathbb{E}_g[g^\mathsf{T}(P_\infty^{1/2}L^\mathsf{T}V_\star \Sigma V_\star L P_\infty^{1/2})gg^\mathsf{T}(P_\infty^{1/2}MP_\infty^{1/2})gg^\mathsf{T}(P_\infty^{1/2}NP_\infty^{1/2})g]\end{aligned}$$

Now we state a result from Magnus to compute the expectation of the product of three quadratic forms of Gaussians.

Lemma 6.2.16 (See e.g. Magnus [72])**.** *Let $g \sim \mathcal{N}(0, I)$ and A_1, A_2, A_3 be symmetric matrices. Then,*

$$\begin{aligned}\mathbb{E}[g^\mathsf{T}A_1gg^\mathsf{T}A_2gg^\mathsf{T}A_3g] &= \mathrm{tr}(A_1)\,\mathrm{tr}(A_2)\,\mathrm{tr}(A_3) \\ &\quad + 2(\mathrm{tr}(A_1)\,\mathrm{tr}(A_2A_3) + \mathrm{tr}(A_2)\,\mathrm{tr}(A_1A_3) + \mathrm{tr}(A_3)\,\mathrm{tr}(A_1A_2)) \\ &\quad + 8\,\mathrm{tr}(A_1A_2A_3) \,.\end{aligned}$$

Now by setting

$$\begin{aligned}A_1 &= P_\infty^{1/2}L^\mathsf{T}V_\star \Sigma V_\star L P_\infty^{1/2} \,, \\ A_2 &= P_\infty^{1/2}MP_\infty^{1/2} \,, \\ A_3 &= P_\infty^{1/2}NP_\infty^{1/2} \,,\end{aligned}$$

we can compute the expectation $\mathbb{E}[T_1 x^\mathsf{T}Mxx^\mathsf{T}Nx]$ using Lemma 6.2.16. In particular,

$$\begin{aligned}\mathrm{tr}(A_1)\,\mathrm{tr}(A_2)\,\mathrm{tr}(A_3) &= \langle P_\infty, L^\mathsf{T}V_\star \Sigma V_\star L \rangle m^\mathsf{T}\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}n \,, \\ \mathrm{tr}(A_1)\,\mathrm{tr}(A_2A_3) &= \langle P_\infty, L^\mathsf{T}V_\star \Sigma V_\star L \rangle m^\mathsf{T}(P_\infty \otimes_s P_\infty)n \,, \\ \mathrm{tr}(A_2)\,\mathrm{tr}(A_1A_3) &= m^\mathsf{T}\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\mathsf{T}V_\star \Sigma V_\star L P_\infty)^\mathsf{T}n \,, \\ \mathrm{tr}(A_3)\,\mathrm{tr}(A_1A_2) &= m^\mathsf{T}\mathrm{svec}(P_\infty L^\mathsf{T}V_\star \Sigma V_\star L P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}n \,, \\ \mathrm{tr}(A_1A_2A_3) &= m^\mathsf{T}(P_\infty L^\mathsf{T}V_\star \Sigma V_\star L P_\infty \otimes_s P_\infty)n \,.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}[g^\mathsf{T} A_1 g g^\mathsf{T} A_2 g g^\mathsf{T} A_3 g] = m^\mathsf{T}(&\langle P_\infty, L^\mathsf{T} V_\star \Sigma V_\star L\rangle(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\ &+ 2\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty)^\mathsf{T} + 2\mathrm{svec}(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T} \\ &+ 8(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty \otimes_s P_\infty))n\,.\end{aligned}$$

Next, we compute

$$\begin{aligned}\mathbb{E}[T_2 x^\mathsf{T} M x x^\mathsf{T} N x] &= \mathbb{E}[\langle w w^\mathsf{T} - \Sigma, V_\star\rangle^2 x^\mathsf{T} M x x^\mathsf{T} N x] \\ &= \mathbb{E}[\langle w w^\mathsf{T} - \Sigma, V_\star\rangle^2]\mathbb{E}[x^\mathsf{T} M x x^\mathsf{T} N x]\,.\end{aligned}$$

First, we have

$$\begin{aligned}\mathbb{E}[\langle w w^\mathsf{T} - \Sigma, V_\star\rangle^2] &= \mathbb{E}[(w^\mathsf{T} V_\star w)^2] - 2\langle \Sigma, V_\star\rangle \mathbb{E}[w^\mathsf{T} V_\star w] + \langle \Sigma, V_\star\rangle^2 \\ &= 2\|\Sigma^{1/2} V_\star \Sigma^{1/2}\|_F^2 + \langle V_\star, \Sigma\rangle^2 - 2\langle \Sigma, V_\star\rangle^2 + \langle V_\star, \Sigma\rangle^2 \\ &= 2\|\Sigma^{1/2} V_\star \Sigma^{1/2}\|_F^2\,.\end{aligned}$$

On the other hand,

$$\mathbb{E}[x^\mathsf{T} M x x^\mathsf{T} N x] = 2\langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} N P_\infty^{1/2}\rangle + \langle M, P_\infty\rangle\langle N, P_\infty\rangle\,.$$

Combining these calculations,

$$\begin{aligned}\mathbb{E}[T_2 x^\mathsf{T} M x x^\mathsf{T} N x] &= 2\|\Sigma^{1/2} V_\star \Sigma^{1/2}\|_F^2 (2\langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} N P_\infty^{1/2}\rangle + \langle M, P_\infty\rangle\langle N, P_\infty\rangle) \\ &= 2\|\Sigma^{1/2} V_\star \Sigma^{1/2}\|_F^2 m^\mathsf{T}(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T})n\end{aligned}$$

Finally, we have $\mathbb{E}[T_3 x^\mathsf{T} M x x^\mathsf{T} N x] = 0$, which is easy to see because it involves odd powers of w. This gives us that B_∞ is:

$$\begin{aligned}B_\infty = (&\langle P_\infty, L^\mathsf{T} V_\star \Sigma V_\star L\rangle + 2\|\Sigma^{1/2} V_\star \Sigma^{1/2}\|_F^2)(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\ &+ 2\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty)^\mathsf{T} + 2\mathrm{svec}(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T} \\ &+ 8(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty \otimes_s P_\infty)\,.\end{aligned}$$

This completes the proof of the formulas for A_∞ and B_∞.

To obtain the lower bound, we need the following lemma which gives a useful lower bound to Lemma 6.2.16.

Lemma 6.2.17. *Let A_1 be positive semi-definite and let A_2 be symmetric. Let $g \sim \mathcal{N}(0, I)$. We have that:*

$$\mathbb{E}[g^\mathsf{T} A_1 g (g^\mathsf{T} A_2 g)^2] \geq 2\,\mathrm{tr}(A_1)\,\mathrm{tr}(A_2^2) + 4\,\mathrm{tr}(A_1 A_2^2)\,.$$

Proof. Suppose that $A_1 \neq 0$, otherwise the bound holds vacuously. From Lemma 6.2.16,

$$\mathbb{E}[g^\mathsf{T} A_1 g (g^\mathsf{T} A_2 g)^2] = \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 2\operatorname{tr}(A_1)\operatorname{tr}(A_2^2) + 4\operatorname{tr}(A_2)\operatorname{tr}(A_1 A_2) + 8\operatorname{tr}(A_1 A_2^2)\,.$$

Since A_1 is PSD and non-zero, this means that $\operatorname{tr}(A_1) > 0$. We proceed as follows:

$$\begin{aligned}
4|\operatorname{tr}(A_2)\operatorname{tr}(A_1 A_2)| &= 2|\operatorname{tr}(A_2)\operatorname{tr}(A_1)^{1/2}|\left|2\frac{\operatorname{tr}(A_1 A_2)}{\operatorname{tr}(A_1)^{1/2}}\right| \\
&\overset{(a)}{\leq} \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\operatorname{tr}(A_1 A_2)^2}{\operatorname{tr}(A_1)} \\
&= \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\operatorname{tr}(A_1^{1/2} A_1^{1/2} A_2)^2}{\operatorname{tr}(A_1)} \\
&\overset{(b)}{\leq} \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\|A_1^{1/2}\|_F^2 \|A_1^{1/2} A_2\|_F^2}{\operatorname{tr}(A_1)} \\
&= \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\operatorname{tr}(A_1 A_2^2)\,,
\end{aligned}$$

where in (a) we used Young's inequality and in (b) we used Cauchy-Schwarz. The claim now follows. □

We now start from the decomposition (6.2.13) for B_∞, with $m = n$ and noting that $\mathbb{E}[T_2(x^\mathsf{T} M x)^2] \geq 0$ and $\mathbb{E}[T_3(x^\mathsf{T} M x)^3] = 0$:

$$\begin{aligned}
m^\mathsf{T} B_\infty m &\geq \mathbb{E}[T_1(x^\mathsf{T} M x)^2] \\
&\overset{(a)}{\geq} 8\langle P_\infty, L^\mathsf{T} V_\star \Sigma V_\star L\rangle m^\mathsf{T}(P_\infty \otimes_s P_\infty)m + 16 m^\mathsf{T}(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty \otimes_s P_\infty)m\,.
\end{aligned}$$

Above in (a) we applied the lower bound from Lemma 6.2.17. Hence since m is arbitrary,

$$B_\infty \succeq 8\langle P_\infty, L^\mathsf{T} V_\star \Sigma V_\star L\rangle(P_\infty \otimes_s P_\infty) + 16(P_\infty L^\mathsf{T} V_\star \Sigma V_\star L P_\infty \otimes_s P_\infty)\,.$$

We also have that $A_\infty = (P_\infty \otimes_s P_\infty)(I - L^\mathsf{T} \otimes L^\mathsf{T})$, and hence $A_\infty^{-1} = (I - L^\mathsf{T} \otimes L^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})$. Therefore,

$$\begin{aligned}
A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}} \succeq\ & 8\langle P_\infty, L^\mathsf{T} V_\star \Sigma V_\star L\rangle (I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-\mathsf{T}} \\
&+ 16(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-1}(L^\mathsf{T} V_\star \Sigma V_\star L \otimes_s P_\infty^{-1})(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-\mathsf{T}}\,.
\end{aligned}$$

□

6.2.5 Asymptotic Analysis of Projected SGD

We now state a high probability bound for SGD. This is a straightforward modification of Lemma 6 from Rakhlin et al. [94] (modifications are needed to deal with the lack of almost surely bounded gradients), and hence we omit the proof.

Lemma 6.2.18 (Lemma 6, Rakhlin et al. [94]). *Let the assumptions of Lemma 6.2.15 hold. Define two constants:*

$$M := \sup_{\theta \in \Theta} \|\theta\| \,, \;\; G_3 := \sup_{\theta \in \Theta} \|\nabla F(\theta)\| \,.$$

Note that since Θ is compact, both M and G_3 are finite. Fix a $T \geq 4$ and $\delta \in (0, 1/e)$. We have that with probability at least $1 - \delta$, for all $t \leq T$,

$$\|\theta_t - \theta_\star\|^2 \lesssim \frac{\text{polylog}(T/\delta)}{t} \left(\frac{G_1^2 + G_2^2}{m^2} + \frac{M(G_2 + G_3)}{m} \right) \,.$$

We are now in a position to analyze the asymptotic variance of SGD with projection. As mentioned previously, our argument follows closely that of Toulis and Airoldi [113]. For the remainder of the proof, $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ will hide all constants except those depending on t and δ. Introduce the notation:

$$\begin{aligned} \tilde{\theta}_{t+1} &= \theta_t - \alpha_t g(\theta_t; \xi_t) \,, \\ \theta_{t+1} &= \mathsf{Proj}_\Theta(\tilde{\theta}_{t+1}) \,. \end{aligned}$$

Let $\mathcal{E}_t := \{\tilde{\theta}_t = \theta_t\}$ be the event that the projection step is inactive at time t. Recall that we assumed that $\theta_\star$ is in the interior of Θ. This means there exists a radius $R > 0$ such that $\{\theta : \|\theta - \theta_\star\| \leq R\} \subseteq \Theta$. Therefore, the event $\{\|\tilde{\theta}_t - \theta_\star\| \leq R\} \subseteq \mathcal{E}_t$. We now decompose,

$$\begin{aligned} \text{Var}(\theta_{t+1}) &= \text{Var}(\theta_{t+1} - \tilde{\theta}_{t+1} + \tilde{\theta}_{t+1}) \\ &= \text{Var}(\tilde{\theta}_{t+1}) + \text{Var}(\theta_{t+1} - \tilde{\theta}_{t+1}) + \text{Cov}(\theta_{t+1} - \tilde{\theta}_{t+1}, \tilde{\theta}_{t+1}) + \text{Cov}(\tilde{\theta}_{t+1}, \theta_{t+1} - \tilde{\theta}_{t+1}) \,. \end{aligned}$$

We have that,

$$\theta_{t+1} - \tilde{\theta}_{t+1} = (\theta_{t+1} - \tilde{\theta}_{t+1}) \mathbf{1}_{\mathcal{E}_{t+1}^c} \,.$$

Hence,

$$\begin{aligned} \|\text{Var}(\theta_{t+1} - \tilde{\theta}_{t+1})\| &\leq \mathbb{E}[\|\tilde{\theta}_{t+1} \mathbf{1}_{\mathcal{E}_{t+1}^c} - \theta_{t+1} \mathbf{1}_{\mathcal{E}_{t+1}^c}\|^2] \\ &\leq 2(\mathbb{E}[\|\tilde{\theta}_{t+1}\|^2 \mathbf{1}_{\mathcal{E}_{t+1}^c}] + \mathbb{E}[\|\theta_{t+1}\|^2 \mathbf{1}_{\mathcal{E}_{t+1}^c}]) \\ &\leq 2(\sqrt{\mathbb{E}[\|\tilde{\theta}_{t+1}\|^4] \mathbb{E}[\mathbf{1}_{\mathcal{E}_{t+1}^c}]} + M^2 \mathbb{E}[\mathbf{1}_{\mathcal{E}_{t+1}^c}]) \,. \end{aligned}$$

We can bound $\mathbb{E}[\|\tilde{\theta}_{t+1}\|^4]$ by a constant for all t using our assumption (6.2.6). On the other hand,

$$\mathbb{E}[\mathbf{1}_{\mathcal{E}_{t+1}^c}] \leq \mathbb{P}(\|\tilde{\theta}_{t+1} - \theta_\star\| > R) \,.$$

By triangle inequality,

$$\|\tilde{\theta}_{t+1} - \theta_\star\| \leq \|\theta_t - \theta_\star\| + \alpha_t \|g_t\| \,.$$

By Lemma 6.2.18 and the concentration bound on $\|g_t\|$ from our assumption (6.2.7), with probability at least $1-\delta$,

$$\|\tilde{\theta}_{t+1} - \theta_\star\| \le \mathcal{O}(\mathrm{polylog}(t/\delta)/\sqrt{t}) \,.$$

Hence for t large enough, $\mathbb{E}[\mathbf{1}_{\mathcal{E}_{t+1}^c}] \le \mathcal{O}(\exp(-t^\alpha))$ for some $\alpha > 0$. This shows that $\|\mathrm{Var}(\theta_{t+1} - \tilde{\theta}_{t+1})\| \le \mathcal{O}(\exp(-t^\alpha))$. Similar arguments show that

$$\max\{\|\mathrm{Cov}(\theta_{t+1} - \tilde{\theta}_{t+1}, \tilde{\theta}_{t+1})\|, \|\mathrm{Cov}(\tilde{\theta}_{t+1}, \theta_{t+1} - \tilde{\theta}_{t+1})\|\} \le \mathcal{O}(\exp(-t^\alpha)) \,.$$

Hence:

$$\mathrm{Var}(\theta_{t+1}) = \mathrm{Var}(\tilde{\theta}_{t+1}) + \mathcal{O}(\exp(-t^\alpha)) \,.$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\theta_{t+1}) &= \mathrm{Var}(\tilde{\theta}_{t+1}) + \mathcal{O}(\exp(-t^\alpha)) \\
&= \mathrm{Var}(\theta_t - \alpha_t g(\theta_t;\xi_t)) + \mathcal{O}(\exp(-t^\alpha)) \\
&= \mathrm{Var}(\theta_t) + \alpha_t^2 \mathrm{Var}(g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(\theta_t, g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(g(\theta_t;\xi_t), \theta_t) \quad (6.2.14) \\
&\quad + \mathcal{O}(\exp(-t^\alpha)) \\
&= \mathrm{Var}(\theta_t) + \alpha_t^2 \mathrm{Var}(g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(\theta_t, \nabla F(\theta_t)) - \alpha_t \mathrm{Cov}(\nabla F(\theta_t), \theta_t) \quad (6.2.15) \\
&\quad + \mathcal{O}(\exp(-t^\alpha)) \,.
\end{aligned}$$

Now we write:

$$\begin{aligned}
\mathrm{Var}(g(\theta_t;\xi_t)) &= \mathrm{Var}(g(\theta_\star;\xi_t) + (g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))) \\
&= \mathrm{Var}(g(\theta_\star;\xi_t)) + \mathrm{Var}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)) \\
&\quad + \mathrm{Cov}(g(\theta_\star;\xi_t), g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)) + \mathrm{Cov}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t), g(\theta_\star;\xi_t)) \,.
\end{aligned}$$

We have by our assumption (6.2.8),

$$\begin{aligned}
\|\mathrm{Var}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))\| &\le \mathbb{E}[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|^2] \\
&= \mathbb{E}_{\theta_t}\mathbb{E}_\xi[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|^2] \\
&\le L\mathbb{E}[\|\theta_t - \theta_\star\|^2] \,.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\mathrm{Cov}(g(\theta_\star;\xi_t), g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))\| &\le 2\mathbb{E}[\|g(\theta_\star;\xi_t)\|\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|] \\
&\le 2\sqrt{\mathbb{E}[\|g(\theta_\star;\xi_t)\|^2]\mathbb{E}[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|^2]} \\
&\le 2\sqrt{LG_1^2\mathbb{E}[\|\theta_t - \theta_\star\|^2]} \,.
\end{aligned}$$

The same bound also holds for $\|\mathrm{Cov}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t), g(\theta_\star;\xi_t))\|$. Since we know that $\mathbb{E}[\|\theta_t - \theta_\star\|^2] \le \mathcal{O}(1/t)$, this shows that:

$$\mathrm{Var}(g(\theta_t;\xi_t)) = \mathrm{Var}(g(\theta_\star;\xi)) + o_t(1)\,.$$

Next, by a Taylor expansion of $\nabla F(\theta_t)$ around $\theta_\star$, we have that:

$$\nabla F(\theta_t) = \nabla^2 F(\theta_\star)(\theta_t - \theta_\star) + \mathsf{Rem}(\theta_t - \theta_\star)\,,$$

where $\|\mathsf{Rem}(\theta_t - \theta_\star)\| \le \mathcal{O}(\|\theta_t - \theta_\star\|^2)$. Therefore, utilizing the fact that adding a non-random vector does not change the covariance,

$$\begin{aligned}
\mathrm{Cov}(\theta_t, \nabla F(\theta_t)) &= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)(\theta_t - \theta_\star) + \mathsf{Rem}(\theta_t - \theta_\star)) \\
&= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)(\theta_t - \theta_\star)) + \mathrm{Cov}(\theta_t, \mathsf{Rem}(\theta_t - \theta_\star)) \\
&= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)\theta_t) + \mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star)) \\
&= \mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))\,.
\end{aligned}$$

We now bound $\mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))$ as:

$$\|\mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))\| \le \mathcal{O}(\mathbb{E}[\|\theta_t - \theta_\star\|^3]) \le \mathcal{O}(\mathrm{polylog}(t)/t^{3/2})\,.$$

Above, the last inequality comes from the high probability bound given in Lemma 6.2.18. Observing that $\mathrm{Cov}(\theta_t, \nabla F(\theta_t))^\mathsf{T} = \mathrm{Cov}(\nabla F(\theta_t), \theta_t)$, combining our calculations and continuing from Equation (6.2.15),

$$\begin{aligned}
\mathrm{Var}(\theta_{t+1}) = \mathrm{Var}(\theta_t) + \alpha_t^2(\mathrm{Var}(g(\theta_\star;\xi)) + o_t(1)) - \alpha_t(\mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \nabla^2 F(\theta_\star)\mathrm{Var}(\theta_t)) \\
+ \alpha_t\mathcal{O}(\mathrm{polylog}(t)/t^{3/2}) + \mathcal{O}(\exp(-t^\alpha))\,.
\end{aligned}$$

We now make two observations. Recall that $\alpha_t = 1/(mt)$. Hence we have $\mathcal{O}(\exp(-t^\alpha)) = \alpha_t^2\mathcal{O}(t^2\exp(-t^\alpha)) = \alpha_t^2 o_t(1)$. Similarly, $\alpha_t\mathcal{O}(\mathrm{polylog}(t)/t^{3/2}) = \alpha_t^2\mathcal{O}(\mathrm{polylog}(t)/t^{1/2}) = \alpha_t^2 o_t(1)$. Therefore,

$$\mathrm{Var}(\theta_{t+1}) = \mathrm{Var}(\theta_t) - \alpha_t(\mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \nabla^2 F(\theta_\star)\mathrm{Var}(\theta_t)) + \alpha_t^2(\mathrm{Var}(g(\theta_\star;\xi)) + o_t(1))\,.$$

This matrix recursion can be solved by Corollary C.1 of Toulis and Airoldi [113], yielding (6.2.9).

To complete the proof, by a Taylor expansion we have:

$$T\cdot\mathbb{E}[F(\theta_T) - F(\theta_\star)] = \frac{T}{2}\,\mathrm{tr}(\nabla^2 F(\theta_\star)\mathbb{E}[(\theta_T - \theta_\star)(\theta_T - \theta_\star)^\mathsf{T}]) + \frac{T}{6}\mathbb{E}[\nabla^3 f(\hat\theta)(\theta_T - \theta_\star)^{\otimes 3}]\,.$$

As above, we can bound $|\mathbb{E}[\nabla^3 f(\hat\theta)(\Theta_T - \theta_\star)^{\otimes 3}]| \le \mathcal{O}(\mathbb{E}[\|\theta_T - \theta_\star\|^3]) \le \mathcal{O}(\mathrm{polylog}(T)/T^{3/2})$, and hence $T\cdot|\mathbb{E}[\nabla^3 f(\hat\theta)(\Theta_T - \theta_\star)^{\otimes 3}]| \to 0$. On the other hand, letting $\mu_T := \mathbb{E}[\theta_T]$, by a bias-variance decomposition,

$$\begin{aligned}
\mathbb{E}[(\theta_T - \theta_\star)(\theta_T - \theta_\star)^\mathsf{T}] &= \mathbb{E}[(\theta_T - \mu_T)(\theta_T - \mu_T)^\mathsf{T}] + (\mu_T - \theta_\star)(\mu_T - \theta_\star)^\mathsf{T} \\
&\succeq \mathbb{E}[(\theta_T - \mu_T)(\theta_T - \mu_T)^\mathsf{T}] = \mathrm{Var}(\theta_T)\,.
\end{aligned}$$

Therefore,

$$T \cdot \mathbb{E}[F(\theta_T) - F(\theta_\star)] \geq \frac{1}{2m} \operatorname{tr}(\nabla^2 F(\theta_\star)(mT)\operatorname{Var}(\theta_T)) - \frac{T}{6}|\mathbb{E}[\nabla^3 f(\hat{\theta})(\theta_T - \theta_\star)^{\otimes 3}]| \,.$$

Taking limits on both sides yields (6.2.11). This concludes the proof of Lemma 6.2.15.

6.2.6 Analysis of Policy Evaluation Methods

In this section, recall that S, R, K are fixed, and furthermore define $M := S + K^\mathsf{T} R K$.

6.2.6.1 Proof of Theorem 6.2.1

The strategy is as follows. Recall that Lemma 6.2.8 gives us the asymptotic distribution of the (regularized) least-squares estimator $\widehat{L}(T)$ of the true closed-loop matrix $L_\star$. For a stable matrix L, let $V(L) = \mathsf{dlyap}(L, M)$. Since the map $L \mapsto V(L)$ is differentiable, using the delta method we can recover the asymptotic distribution of $\sqrt{T}\operatorname{svec}(V(\widehat{L}(T)) - V_\star)$. Upper bounding the trace of the covariance matrix for this asymptotic distribution then yields Theorem 6.2.1.

Let $[DV(L)]$ denote the Fréchet derivative of the map $V(\cdot)$ evaluated at L, and let $[DV(L)](X)$ denote the action of the linear operator $[DV(L)]$ on X. By a straightforward application of the implicit function theorem, we have that:

$$[DV(L_\star)](X) = \mathsf{dlyap}(L_\star, X^\mathsf{T} V_\star L_\star + L_\star^\mathsf{T} V_\star X) \,.$$

Before we proceed, we introduce some notation surrounding Kronecker products. Let Γ denote the matrix such that $(A \otimes_s B) = \frac{1}{2}\Gamma^\mathsf{T}(A \otimes B + B \otimes A)\Gamma$ for any square matrices A, B. It is a fact that $\Gamma \operatorname{vec}(S) = \operatorname{svec}(S)$ for any symmetric matrix S. Also let Π be the orthonormal matrix such that $\Pi \operatorname{vec}(X) = \operatorname{vec}(X^\mathsf{T})$ for all square matrices X. It is not hard to verify that $\Pi^\mathsf{T}(A \otimes B)\Pi = (B \otimes A)$, a fact we will use later. With this notation, we proceed as follows:

$$\begin{aligned}
\operatorname{svec}([DV(L_\star)](X)) &= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} \operatorname{svec}(X^\mathsf{T} V_\star L_\star + L_\star^\mathsf{T} V_\star X) \\
&= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} \Gamma \operatorname{vec}(X^\mathsf{T} V_\star L_\star + L_\star^\mathsf{T} V_\star X) \\
&= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} \Gamma((L_\star^\mathsf{T} V_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} V_\star)) \operatorname{vec}(X) \,.
\end{aligned}$$

Applying Lemma 6.2.8 in conjunction with the delta method, we obtain:

$$\sqrt{T}\operatorname{svec}(V(\widehat{L}(T)) - V_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_w^2 (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} \Sigma (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \,,$$

where,

$$\begin{aligned}
\Sigma &:= \Gamma[((L_\star^\mathsf{T} V_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} V_\star))(P_\infty^{-1} \otimes I_n)((L_\star^\mathsf{T} V_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} V_\star))^\mathsf{T}]\Gamma^\mathsf{T} \\
&\overset{(a)}{\preceq} 2\Gamma[(L_\star^\mathsf{T} V_\star \otimes I_n)\Pi(P_\infty^{-1} \otimes I_n)\Pi^\mathsf{T}(V_\star L_\star \otimes I_n) + (I_n \otimes L_\star^\mathsf{T} V_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes V_\star L_\star)]\Gamma^\mathsf{T} \\
&= 2\Gamma[(L_\star^\mathsf{T} V_\star \otimes I_n)(I_n \otimes P_\infty^{-1})(V_\star L_\star \otimes I_n) + (I_n \otimes L_\star^\mathsf{T} V_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes V_\star L_\star)]\Gamma^\mathsf{T} \\
&= 2\Gamma[(L_\star^\mathsf{T} V_\star^2 L_\star \otimes P_\infty^{-1}) + (P_\infty^{-1} \otimes L_\star^\mathsf{T} V_\star^2 L_\star)]\Gamma^\mathsf{T} \\
&= 4(L_\star^\mathsf{T} V_\star^2 L_\star \otimes_s P_\infty^{-1}) \,.
\end{aligned}$$

In (a), we used the inequality for any matrices X, Y and positive definite matrices F, G, (see e.g. Chapter 3, page 94 of Zhang [128]):

$$(X+Y)(F+G)^{-1}(X+Y)^\mathsf{T} \preceq XF^{-1}X^\mathsf{T} + YG^{-1}Y^\mathsf{T} \,.$$

Suppose that the sequence $\{\|Z_T\|_F^2\}$ is uniformly integrable, where $Z_T := \sqrt{T}\mathrm{svec}(V(\widehat{L}(T)) - V_\star)$. Then:

$$\lim_{T\to\infty} T \cdot \mathbb{E}[\|V(\widehat{L}(T)) - V_\star\|_F^2] \le 4\,\mathrm{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(L_\star^\mathsf{T} V_\star^2 L_\star \otimes_s \sigma_w^2 P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \,,$$

which is the desired bound on the asymptotic risk.

We now show that the sequence $\{\|Z_T\|_F^2\}$ is uniformly integrable. Fix a finite $p \ge 1$. Since $L_\star$ is stable and $\zeta \in (\rho(L_\star), 1)$, there exists a $C_\star$ such that $\|L_\star^k\| \le C_\star \zeta^k$ for all $k \ge 0$. For the rest of the proof, $\mathcal{O}(\cdot), \Omega(\cdot)$ will hide constants that depend on $L_\star, C_\star, n, p, \lambda, \zeta, \psi$, but not on T. Set $\delta_T = \mathcal{O}(1/T^{p/2})$ and let T be large enough so that there exists an event $\mathcal{E}_{\mathsf{Bdd}}$ promised by Lemma 6.2.13 such that $\mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}) \ge 1 - \delta_T$ and on $\mathcal{E}_{\mathsf{Bdd}}$ we have $\|\widehat{L}(T) - L_\star\| \le \mathcal{O}(\sqrt{\log(1/\delta_T)/T})$. Let T also be large enough so that on $\mathcal{E}_{\mathsf{Bdd}}$, we have $\|\widehat{L}(T) - L_\star\| \le \min((\gamma - \rho_\star)/C_\star, \psi - \|L_\star\|)$. With this setting, we have that on $\mathcal{E}_{\mathsf{Bdd}}$, for any $\alpha \in (0,1)$,

$$\begin{aligned}
\tilde{L}(\alpha) &:= \alpha\widehat{L}(T) + (1-\alpha)L_\star \\
&\in \left\{ L \in \mathbb{R}^{n\times n} : \rho(L) \le \zeta \,,\ \|L\| \le \min\left(\|L_\star\| + \frac{\gamma - \rho_\star}{C_\star}, \psi\right)\right\} =: \mathcal{G} \,.
\end{aligned}$$

Therefore on $\mathcal{E}_{\mathsf{Bdd}}$, for some $\alpha \in (0,1)$,

$$\begin{aligned}
\|V(\widehat{L}(T)) - V_\star\| &= \|[DV(\tilde{L}(\alpha))](\widehat{L}(T) - L_\star)\| \\
&\le \sup_{\tilde{L}\in\mathcal{G}} \|[DV(\tilde{L})]\| \|\widehat{L}(T) - L_\star\| =: C_{\mathrm{lip}} \|\widehat{L}(T) - L_\star\| \,.
\end{aligned}$$

Here the norm $\|[H]\| := \sup_{\|X\|\le 1}\|[H](X)\|$. We have that C_{lip} is finite since $\mathcal{G}$ is a compact set. Next, define the set $\mathcal{G}_{\mathsf{Alg}}$ as:

$$\mathcal{G}_{\mathsf{Alg}} := \{L \in \mathbb{R}^{n\times n} : \rho(L) \le \zeta \,,\ \|L\| \le \psi\} \,,$$

and define the event $\mathcal{E}_{\mathsf{Alg}}$ as $\mathcal{E}_{\mathsf{Alg}} := \{\widehat{L}(T) \in \mathcal{G}_{\mathsf{Alg}}\}$. Consider the decomposition:

$$\begin{aligned}\mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p] &= \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c}] \\ &\le \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\ &\quad + \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] \,.\end{aligned}$$

In what follows we will assume that T is sufficiently large.

On $\mathcal{E}_{\mathsf{Bdd}}$. On this event, since we have $\mathcal{E}_{\mathsf{Bdd}} \subseteq \mathcal{E}_{\mathsf{Alg}}$, we can bound by Lemma 6.2.13:

$$\mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] = \mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}} \cap \mathcal{E}_{\mathsf{Alg}}}] \le C_{\mathrm{lip}}^p \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \le \mathcal{O}(1/T^{p/2}) \,.$$

On $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}$. On this event, we use the fact that $\mathcal{G}_{\mathsf{Alg}}$ is compact to bound:

$$\begin{aligned}\mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] &\le \sup_{\widehat{L} \in \mathcal{G}_{\mathsf{Alg}}} \|\mathsf{dlyap}(\widehat{L}, S + K^\mathsf{T} R K) - V_\star\|^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}) \\ &\le \sup_{\widehat{L} \in \mathcal{G}_{\mathsf{Alg}}} \|\mathsf{dlyap}(\widehat{L}, S + K^\mathsf{T} R K) - V_\star\|^p \delta_T \\ &\le \mathcal{O}(1/T^{p/2}) \,.\end{aligned}$$

On $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c$. On this event, we simply have:

$$\mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] = \|V_\star\|^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c) \le \|V_\star\|^p \delta_T \le \mathcal{O}(1/T^{p/2}) \,.$$

Putting it together. Combining these bounds we obtain that $\mathbb{E}[\|\widehat{V}_{\mathrm{plug}}(T) - V_\star\|^p] \le \mathcal{O}(1/T^{p/2})$. Recall that $Z_T = \mathrm{svec}(V(\widehat{L}(T)) - V_\star)$. We have that for any finite $\gamma > 0$ and $T \ge \Omega(1)$:

$$\begin{aligned}\mathbb{E}[\|Z_T\|_F^{2+\gamma}] &= T^{(2+\gamma)/2} \mathbb{E}[\|V(\widehat{L}(T)) - V_\star\|_F^{2+\gamma}] \\ &\le n^{(2+\gamma)/2} T^{(2+\gamma)/2} \mathbb{E}[\|V(\widehat{L}(T)) - V_\star\|^{2+\gamma}] \\ &\le n^{(2+\gamma)/2} T^{(2+\gamma)/2} \mathcal{O}(1/T^{(2+\gamma)/2}) \\ &\le n^{(2+\gamma)/2} \mathcal{O}(1) \,.\end{aligned}$$

On the other hand, when $T \le \mathcal{O}(1)$ it is easy to see that $\mathbb{E}[\|Z_T\|_F^{2+\gamma}]$ is finite. Hence we have $\sup_{T \ge 1} \mathbb{E}[\|Z_T\|_F^{2+\gamma}] < \infty$ which shows the desired uniformly integrable condition. This concludes the proof of Theorem 6.2.1.

6.2.6.2 Proof of Theorem 6.2.2

Lemma 6.2.10 (specifically (6.2.5)) combined with Lemma 6.2.12 tells us that:

$$\begin{aligned}\liminf_{T\to\infty} T\cdot\mathbb{E}[\|\widehat{V}_{\mathrm{lstd}}(T)-V_\star\|_F^2] &\geq \mathrm{tr}(A_\infty^{-1}B_\infty A_\infty^{-\mathsf{T}})\\ &\geq 8\sigma_w^2\,\mathrm{tr}(\langle P_\infty, L_\star^{\mathsf{T}}V_\star^2 L_\star\rangle(I-L_\star^{\mathsf{T}}\otimes_s L_\star^{\mathsf{T}})^{-1}(P_\infty^{-1}\otimes_s P_\infty^{-1})(I-L_\star^{\mathsf{T}}\otimes_s L_\star^{\mathsf{T}})^{-\mathsf{T}})\\ &\quad + 16\sigma_w^2\,\mathrm{tr}((I-L_\star^{\mathsf{T}}\otimes_s L_\star^{\mathsf{T}})^{-1}(L_\star^{\mathsf{T}}V_\star^2 L_\star\otimes_s P_\infty^{-1})(I-L_\star^{\mathsf{T}}\otimes_s L_\star^{\mathsf{T}})^{-\mathsf{T}})\,.\end{aligned}$$

The claim now follows by using the risk bound from Theorem 6.2.1.

6.2.6.3 Proof of Theorem 6.2.3

Let $E_1,...,E_N$ be d-dimensional subspaces of $\mathbb{R}^n$ with $d\leq n/2$ such that $\|P_{E_i}-P_{E_j}\|_F\gtrsim\sqrt{d}$. By Proposition 8 of Pajor [90], we can take $N\geq e^{n(n-d)}$. Now consider instances A_i with $A_i=\tau P_{E_i}+\gamma I_n$ for a $\tau,\gamma\in(0,1)$ to be determined. We will set $\tau+\gamma=\rho$ so that each A_i is contractive (i.e. $\|A_i\|<1$) and hence stable. This means implicitly that we will require $\tau<\rho$. Let $\mathbb{P}_i$ denote the distribution over $(x_1,...,x_T)$ induced by instance A_i. We have that:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_i,\mathbb{P}_j)&=\sum_{t=1}^{T}\mathbb{E}_{x_t\sim\mathbb{P}_i}[\mathrm{KL}(\mathcal{N}(A_ix_t,\sigma^2I),\mathcal{N}(A_jx_t,\sigma^2I))]\\ &=\frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim\mathbb{P}_i}[\|(A_i-A_j)x_t\|_2^2]\\ &\leq\frac{\|A_i-A_j\|^2}{2\sigma^2}\sum_{t=1}^{T}\mathrm{tr}(\mathbb{E}_{x_t\sim\mathbb{P}_i}[x_tx_t^{\mathsf{T}}])\\ &\leq\frac{\tau^2}{\sigma^2}T\,\mathrm{tr}(P_\infty)\\ &=\tau^2T\left(\frac{d}{1-\rho^2}+\frac{n-d}{1-\gamma^2}\right)\\ &\leq\tau^2T\frac{n}{1-\rho^2}\,.\end{aligned}$$

Now if we choose $n(n-d)\geq 4\log 2$ and $T\gtrsim n(1-\rho^2)/\rho^2$, we can set $\tau^2\asymp\frac{n(1-\rho^2)}{T}$ and obtain that $\frac{I(V;X)+\log 2}{\log|V|}\leq 1/2$.

On the other hand, let $P_i=\mathsf{dlyap}(A_i,I_n)$. We have that for any integer $k\geq 0$:

$$\begin{aligned}(\tau P_{E_i}+\gamma I_n)^k-(\tau P_{E_j}+\gamma I_n)^k&=\sum_{\ell=0}^{k}\binom{k}{\ell}\gamma^{k-\ell}\tau^\ell(P_{E_i}^\ell-P_{E_j}^\ell)\\ &=k\gamma^{k-1}\tau(P_{E_i}-P_{E_j})+\sum_{\ell=2}^{k}\binom{k}{\ell}\gamma^{k-\ell}\tau^\ell(P_{E_i}^\ell-P_{E_j}^\ell)\,.\end{aligned}$$

Hence,

$$\begin{aligned} P_i - P_j &= \sum_{k=1}^{\infty}((A_i^k)^\mathsf{T} A_i^k - (A_j^k)^\mathsf{T} A_j^k) \\ &= \sum_{k=1}^{\infty}(A_i^{2k} - A_j^{2k}) \\ &= \left(\sum_{k=1}^{\infty} 2k\gamma^{2k-1}\tau + \sum_{k=2}^{\infty}\sum_{\ell=2}^{2k}\binom{2k}{\ell}\gamma^{2k-\ell}\tau^\ell\right)(P_{E_i} - P_{E_j}) \\ &= \left(\frac{2\gamma\tau}{(1-\gamma^2)^2} + \sum_{k=2}^{\infty}\sum_{\ell=2}^{k}\binom{k}{\ell}\gamma^{k-\ell}\tau^\ell\right)(P_{E_i} - P_{E_j})\,. \end{aligned}$$

Therefore,

$$\|P_i - P_j\|_F \geq \frac{2\gamma\tau}{(1-\gamma^2)^2}\|P_{E_i} - P_{E_j}\|_F \gtrsim \frac{\gamma\tau}{(1-\gamma^2)^2}\sqrt{d}\,.$$

The claim now follows by Fano's inequality and setting $d = n/4$.

6.2.7 Analysis of Policy Optimization Methods

6.2.7.1 Preliminary Calculations

Given (A, B) with $\text{range}(A) \subseteq \text{range}(B)$ and $\text{rank}(B) = d$, let $J_W(K)$ for a $K \in \mathbb{R}^{d\times n}$ denote the following cost:

$$J_W(K) := \mathbb{E}\left[\sum_{t=1}^{T}\|x_t\|^2\right]\ ,\ \ x_{t+1} = Ax_t + Bu_t + w_t\ ,\ \ u_t = Kx_t\ ,\ \ w_t \sim \mathcal{N}(0, W)\,.$$

Here we assume $T \geq 2$ and W is positive definite. We write $J(K) = J_{\sigma_w^2 I_n}(K)$ as shorthand. Under this feedback law, we have $x_t \sim \mathcal{N}(0, \sum_{\ell=0}^{t-1} L(K)^\ell W(L(K)^\ell)^\mathsf{T})$ with $L(K) := A + BK$. Letting L be shorthand for $L(K)$, the cost can be written as:

$$J_W(K) = \sum_{t=1}^{T}\sum_{\ell=0}^{t-1}\text{tr}(L^\ell W(L^\ell)^\mathsf{T}) = T\,\text{tr}(W) + \sum_{t=1}^{T}\sum_{\ell=1}^{t-1}\text{tr}(L^\ell W(L^\ell)^\mathsf{T})\,.$$

Let $K_\star$ denote the minimizer of $J_W(K)$; under our assumptions we have that $K_\star = -B^\dagger A$. Furthermore, because of the range condition we can write $A = BB^\dagger A$. Therefore, $L(K) = B(B^\dagger A + K)$. While the function $J_W(K)$ is not convex, it has many nice properties. First,

$J_W(K)$ satisfies a *quadratic growth condition*:

$$\begin{aligned} J_W(K) - J_W(K_\star) &\geq (T-1)\,\mathrm{tr}(LWL^\mathsf{T}) \\ &= (T-1)\,\mathrm{tr}(B(B^\dagger A + K)W(B^\dagger A + K)^\mathsf{T} B^\mathsf{T}) \\ &= (T-1)\mathrm{vec}(B^\dagger A + K)^\mathsf{T}(W \otimes B^\mathsf{T}B)\mathrm{vec}(B^\dagger A + K) \\ &\geq (T-1)\lambda_{\min}(W)\sigma_{\min}(B)^2\|K - K_\star\|_F^2 \,. \end{aligned} \tag{6.2.16}$$

Next, we will see $J_W(K)$ satisfies restricted strong convexity. To do this, we first compute the gradient $\nabla J_W(K)$. Consider the function $M \mapsto M^\ell$ for any integer $\ell \geq 2$. We have that the derivatives are:

$$[DM^\ell](\Delta) = \sum_{k=0}^{\ell-1} M^k \Delta M^{\ell-k-1}\,, \;\; [D(M^\ell)^\mathsf{T}](\Delta) = \sum_{k=0}^{\ell-1} (M^{\ell-k-1})^\mathsf{T}\Delta^\mathsf{T}(M^k)^\mathsf{T}\,.$$

By the chain rule,

$$[DL(K)^\ell](\Delta) = \sum_{k=0}^{\ell-1} L(K)^k B\Delta L(K)^{\ell-k-1}\,.$$

Hence by the chain rule again,

$$\begin{aligned} [D\,\mathrm{tr}(L(K)^\ell W(L^\ell)^\mathsf{T})](\Delta) &= \mathrm{tr}\left(L^\ell W \sum_{k=0}^{\ell-1}(L^{\ell-k-1})^\mathsf{T}\Delta^\mathsf{T}B^\mathsf{T}(L^k)^\mathsf{T}\right) \\ &\quad + \mathrm{tr}\left(\sum_{k=0}^{\ell-1} L^k B\Delta L^{\ell-k-1} W(L^\ell)^\mathsf{T}\right) \\ &= 2\left\langle \sum_{k=0}^{\ell-1} B^\mathsf{T}(L^k)^\mathsf{T}L^\ell W(L^{\ell-k-1})^\mathsf{T}, \Delta \right\rangle\,. \end{aligned}$$

We have shown that:

$$\nabla_K\,\mathrm{tr}(L(K)^\ell W(L(K)^\ell)^\mathsf{T}) = 2\sum_{k=0}^{\ell-1} B^\mathsf{T}(L^k)^\mathsf{T}L^\ell W(L^{\ell-k-1})^\mathsf{T}\,.$$

Therefore we can compute the gradient of $J_W(K)$ as:

$$\nabla J_W(K) = 2(T-1)B^\mathsf{T}LW + 2\sum_{\ell=2}^{T-1}\sum_{k=0}^{\ell-1}(T-\ell)B^\mathsf{T}(L^k)^\mathsf{T}L^\ell W(L^{\ell-k-1})^\mathsf{T}\,.$$

Now observe that $L(K) = B(K - K_\star)$ and therefore:

$$\begin{aligned}\langle \nabla J_W(K), K - K_\star \rangle &= \operatorname{tr}(\nabla J_W(K)(K - K_\star)^\mathsf{T}) \\ &= 2(T-1)\operatorname{tr}(LWL^\mathsf{T}) + 2\sum_{\ell=2}^{T-1}\sum_{k=0}^{\ell-1}(T-\ell)\operatorname{tr}(L^\ell W (L^\ell)^\mathsf{T}) \\ &\overset{(a)}{\geq} 2(T-1)\operatorname{tr}(LWL^\mathsf{T}) \\ &\geq 2(T-1)\lambda_{\min}(W)\sigma_{\min}(B)^2\|K - K_\star\|_F^2 \,.\end{aligned}$$

Above, (a) follows since $\operatorname{tr}(AB) \geq 0$ for positive semi-definite matrices A, B. This condition proves that $K = K_\star$ is the unique stationary point, and establishes the restricted strong convexity $\mathsf{RSC}(m, \mathbb{R}^{d\times n})$ condition for $J_W(K)$ with constant $m = 2(T-1)\lambda_{\min}(W)\sigma_{\min}(B)^2$.

Finally, we show that the Hessian of $J_W(K)$ evaluated at $K_\star$ is positive definite. Fix a test matrix $H \in \mathbb{R}^{d\times n}$, and define the function $g(t) := \langle H, \nabla J_W(K_\star + tH)\rangle$. By standard properties of the directional derivative, we have that $\mathrm{Hess} J_W(K_\star)[H,H] = g'(0)$. Observing that $L(K_\star + tH) = t \cdot BH$, we have that:

$$\begin{aligned}g(t) &= 2(T-1)t\operatorname{tr}(WH^\mathsf{T}B^\mathsf{T}BH) \\ &\quad + 2\sum_{\ell=2}^{T-1}\sum_{k=0}^{\ell-1}(T-\ell)t^{2\ell-1}\operatorname{tr}(H^\mathsf{T}B^\mathsf{T}(H^\mathsf{T}B^\mathsf{T})^k(BH)^\ell W(H^\mathsf{T}B^\mathsf{T})^{\ell-k-1}) \,,\end{aligned}$$

from which we conclude:

$$\mathrm{Hess} J_W(K_\star)[H,H] = 2(T-1)\operatorname{tr}(WH^\mathsf{T}B^\mathsf{T}BH) = 2(T-1)\mathrm{vec}(H)^\mathsf{T}(W \otimes B^\mathsf{T}B)\mathrm{vec}(H) \,.$$

6.2.7.2 Proof of Theorem 6.2.4

Recal that the pair (A, B) is stabilizable if there exists a feedback matrix K such that $\rho(A + BK) < 1$. We first state a result which gives a sufficient condition for the existence of a unique positive definite solution to the discrete algebraic Riccati equation.

Lemma 6.2.19 (Theorem 2, Molinari [83])**.** *Suppose that $V \succ 0$, (A, B) is stabilizable, and B has full column rank. Then there exists a unique positive definite solution V to the DARE:*

$$V = A^\mathsf{T}VA - A^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA + S \,. \tag{6.2.17}$$

This V satisfies the lower bound $V \succeq S$, and if A is contractive (i.e. $\|A\| < 1$) satisfies the upper bound $\|V\| \leq \frac{\|V\|}{1-\|A\|^2}$.

Proof. Define the map $\Psi(z; A) := B^\mathsf{T}(z^{-1}I_n - A)^{-\mathsf{T}}V(zI_n - A)^{-1}B$. Let K be such that $A + BK$ is stable. We observe that for $|z| = 1$, we have that:

$$\Psi(z; A + BK) = B^*(zI_n - (A + BK))^{-*}V(zI_n - (A + BK))^{-1}B \succ 0 \,.$$

This is because $V \succ 0$, $B^*B \succ 0$, and the matrix $zI_n - (A + BK)$ does not drop rank since $A+BK$ has no eigenvalues on the unit circle. Therefore by Theorem 2 of Molinari [83], there exists a unique symmetric solution V that satisfies (6.2.17) with the additional constraint that $B^\mathsf{T}VB \succ 0$ and that $\rho(A_c) < 1$ with $A_c := A - B(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA$. But (6.2.17) means that:

$$\begin{aligned} A_c^\mathsf{T}VA_c &= (A - B(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA)^\mathsf{T}V(A - B(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA) \\ &= A^\mathsf{T}VA - A^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA - A^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA \\ &\quad + A^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA \\ &= A^\mathsf{T}VA - A^\mathsf{T}VB(B^\mathsf{T}VB)^{-1}B^\mathsf{T}VA \\ &= V - S\,. \end{aligned}$$

Hence, we have $A_c^\mathsf{T}VA_c - V + S = 0$, and since A_c is stable by Lyapunov theory we know that $V \succeq S$. Furthermore, since $V \succeq 0$, (6.2.17) implies that $V \preceq A^\mathsf{T}VA + S$ from which the upper bound on $\|V\|$ follows under the contractivity assumptions. □

Next, we state a result which gives the derivative of the discrete algebraic Riccati equation.

Lemma 6.2.20 (Section A.2 of Abeille and Lazaric [5])**.** *Let (S, R) be positive semidefinite matrices. Suppose that (A, B) are such that there exists a unique positive definite solution $V(A, B)$ to $\mathsf{dare}(A, B, S, R)$. For a perturbation $\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix} \in \mathbb{R}^{n\times(n+d)}$, we have that the Fréchet derivative $[D_{(A,B)}V(A, B)]$ evaluated at the perturbation $\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}$ is given by:*

$$[D_{(A,B)}V(A, B)](\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}) = \mathsf{dlyap}\left(A_c, A_c^\mathsf{T}V\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}\begin{bmatrix}I_n \\ K\end{bmatrix} + \begin{bmatrix}I_n \\ K\end{bmatrix}^\mathsf{T}\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}^\mathsf{T}VA_c\right),$$

where $V = V(A, B)$, $K = -(B^\mathsf{T}VB + R)^{-1}B^\mathsf{T}VA$, and $A_c = A + BK$.

With these two lemmas, we are ready to proceed. We differentiate the map $h(A, B) := -(B^\mathsf{T}V(A, B)B + R)^{-1}B^\mathsf{T}V(A, B)A$. By the chain rule:

$$\begin{aligned} [D_{(A,B)}h(A, B)](\Delta) &= -(B^\mathsf{T}VB + R)^{-1}(B^\mathsf{T}V\Delta_A + \Delta_B^\mathsf{T}VA + B^\mathsf{T}[D_{(A,B)}V](\Delta)A) \\ &\quad + (B^\mathsf{T}VB + R)^{-1}(\Delta_B^\mathsf{T}VB + B^\mathsf{T}V\Delta_B + B^\mathsf{T}[D_{(A,B)}V](\Delta)B)(B^\mathsf{T}VB + R)^{-1}B^\mathsf{T}VA\,. \end{aligned}$$

We now evaluate this derivative at:

$$A = A\,, B = B\,, V = I_n\,, R = 0\,.$$

Note that $V(A, B) = I_n$ and also by Lemma 6.2.20, we have that $[D_{(A,B)}V(A, B)] = 0$, since $A_c = 0$. Therefore the derivative $[D_{(A,B)}h(A, B)](\Delta)$ simplifies to:

$$\begin{aligned} [D_{(A,B)}h(A, B)](\Delta) &= -(B^\mathsf{T}B)^{-1}(B^\mathsf{T}\Delta_A + \Delta_B^\mathsf{T}A) + (B^\mathsf{T}B)^{-1}(\Delta_B^\mathsf{T}B + B^\mathsf{T}\Delta_B)B^\dagger A \\ &= -B^\dagger\Delta_A + B^\dagger\Delta_B B^\dagger A\,. \end{aligned}$$

Hence we have:

$$\mathrm{vec}([D_{(A,B)}h(A,B)](\Delta)) = \begin{bmatrix} -(I_n \otimes B^\dagger) & (B^\dagger A)^\mathsf{T} \otimes B^\dagger \end{bmatrix} \mathrm{vec}(\Delta)\,.$$

Now using the assumption that A is stable, from Lemma 6.2.9 we have that by the delta method:

$$\sqrt{N}\mathrm{vec}(h(\widehat{A}(N),\widehat{B}(N)) - K_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0,\Psi) =: \varphi\,,$$

where

$$\Psi := \frac{\sigma_w^2}{T} \begin{bmatrix} -(I_n \otimes B^\dagger) & (B^\dagger A)^\mathsf{T} \otimes B^\dagger \end{bmatrix} \left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_\eta^2) I_d \end{bmatrix} \otimes I_n \right) \begin{bmatrix} -(I_n \otimes (B^\dagger)^\mathsf{T}) \\ B^\dagger A \otimes (B^\dagger)^\mathsf{T} \end{bmatrix} + o(1/T)\,.$$

We now make use of the second order delta method. Recall that the Hessian of J at $K_\star$ is $\mathrm{Hess}J(K_\star)[H,H] = 2(T-1)\sigma_w^2\langle H, B^\mathsf{T}BH\rangle$. If $\sqrt{N}\mathrm{vec}(\widehat{K}(N) - K_\star) \overset{D}{\rightsquigarrow} \varphi$, then by the second order delta method:

$$N \cdot (J(\widehat{K}(N)) - J_\star) \overset{D}{\rightsquigarrow} (T-1)\sigma_w^2 \varphi^\mathsf{T}(I_n \otimes B^\mathsf{T}B)\varphi\,.$$

Next we make an intermediate calculation:

$$\begin{aligned} &\begin{bmatrix} -(I_n \otimes (B^\dagger)^\mathsf{T}) \\ B^\dagger A \otimes (B^\dagger)^\mathsf{T} \end{bmatrix} (I_n \otimes B^\mathsf{T}B) \begin{bmatrix} -(I_n \otimes B^\dagger) & (B^\dagger A)^\mathsf{T} \otimes B^\dagger \end{bmatrix} \\ &= \begin{bmatrix} I_n \otimes BB^\dagger & -((B^\dagger A)^\mathsf{T} \otimes BB^\dagger) \\ -(B^\dagger A \otimes BB^\dagger) & B^\dagger A A^\mathsf{T}(B^\dagger)^\mathsf{T} \otimes BB^\dagger \end{bmatrix} \\ &= \begin{bmatrix} I_n & -(B^\dagger A)^\mathsf{T} \\ -B^\dagger A & B^\dagger A A^\mathsf{T}(B^\dagger)^\mathsf{T} \end{bmatrix} \otimes BB^\dagger\,. \end{aligned}$$

Let $Z_N := N \cdot (J(\widehat{K}(N)) - J_\star)$. To conclude the proof, we show that the sequence $\{Z_N\}$ is uniformly integrable. Once we have the uniform integrability in place, then by Lemma 6.2.12:

$$\begin{aligned} &\lim_{N\to\infty} N \cdot (J(\widehat{K}(N)) - J_\star) \\ &\quad = \sigma_w^4 \frac{T-1}{T} \mathrm{tr}\left(\left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_\eta^2) I_d \end{bmatrix} \otimes I_n \right) \left(\begin{bmatrix} I_n & -(B^\dagger A)^\mathsf{T} \\ -B^\dagger A & B^\dagger A A^\mathsf{T}(B^\dagger)^\mathsf{T} \end{bmatrix} \otimes BB^\dagger \right) \right) + o_T(1) \\ &\quad = \sigma_w^4 \frac{T-1}{T} \mathrm{tr}\left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_\eta^2) I_d \end{bmatrix} \begin{bmatrix} I_n & -(B^\dagger A)^\mathsf{T} \\ -B^\dagger A & B^\dagger A A^\mathsf{T}(B^\dagger)^\mathsf{T} \end{bmatrix} \right) \mathrm{tr}(BB^\dagger) + o_T(1) \\ &\quad = \sigma_w^4 \frac{T-1}{T} \left(\mathrm{tr}(P_\infty^{-1}) + \frac{\|B^\dagger A\|_F^2}{\sigma_\eta^2} \right) d + o_T(1)\,. \end{aligned}$$

To conclude the proof, let $C_\star, \rho_\star$ be such that $\|A^k\| \le C_\star \rho_\star^k$ with $\rho_\star \in (0,1)$: these constants exist because A is stable. Now define the events:

$$\begin{aligned} \mathcal{E}_{\mathsf{Alg}} &:= \{\rho(\widehat{A}(N)) \le \varrho\,,\ \|\widehat{A}(N)\| \le \zeta\,,\ \|\widehat{B}(N)\| \le \psi\,,\ \sigma_d(\widehat{B}(N)) \ge \gamma\}\,, \\ \mathcal{E}_{\mathsf{Bdd}} &:= \{\|\widehat{A}(N) - A\| \le \frac{1-\rho_\star}{2C_\star}\,,\ \|\widehat{B}(N) - B\| \le \sigma_d(B)/2\}\,. \end{aligned}$$

Fix a finite $p \geq 1$. We write:

$$\begin{aligned}
\mathbb{E}[Z_N^p] &= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c}] \\
&= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] \\
&= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p (J(0) - J_\star)^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c) \\
&\leq N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p (J(0) - J_\star)^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \, .
\end{aligned}$$

We now consider what happens on these three events. For the remainder of the proof, we let C denote a constant that depends on $n, d, p, C_\star, \rho_\star, \varrho, \zeta, \psi, \gamma, A, B, T, \varepsilon, \sigma_w^2, \sigma_\eta^2$ but not on N, whose value can change from line to line.

On the event $\mathcal{E}_{\mathsf{Bdd}}$. By a Taylor expansion we write:

$$h(\widehat{A}(N), \widehat{B}(N)) - h(A, B) = [D_{(A,B)} h(\tilde{A}, \tilde{B})] \left(\begin{bmatrix} \widehat{A}(N) - A & \widehat{B}(N) - B \end{bmatrix} \right) \, ,$$

where $\tilde{A} = tA + (1-t)\widehat{A}(N)$ and $\tilde{B} = tB + (1-t)\widehat{B}(N)$ for some $t \in [0,1]$. Observe that on $\mathcal{E}_{\mathsf{Bdd}}$, we have that

$$\tilde{A}, \tilde{B} \in \mathcal{G} := \left\{ (A,B) : \|A\| \leq \|A\| + \frac{1-\rho_\star}{2C_\star} \, , \|B\| \leq \|B\| + \sigma_d(B)/2 \, , \sigma_d(B) \geq \sigma_d(B)/2 \right\} \, .$$

By Proposition 4.0.1 each $(A,B) \in \mathcal{G}$ is stabilizable (since A is stable) and B has full column rank. Therefore by Lemma 6.2.19, for any $(A,B) \in \mathcal{G}$ we have that $\mathsf{dare}(A, B, I_n, 0)$ has a unique positive definite solution and its derivative is well defined. By the compactness of $\mathcal{G}$ and the continuity of h and its derivative, we define the finite constants

$$C_K := \sup_{A,B \in \mathcal{G}} \|h(A,B)\| \, , \quad C_{\mathrm{deriv}} := \sup_{A,B \in \mathcal{G}} \|[D_{(A,B)} h(A,B)]\| \, .$$

We can now Taylor expand $J(K)$ around $K_\star$ and obtain:

$$\begin{aligned}
J(\widehat{K}(N)) - J_\star &= \frac{1}{2} \mathrm{Hess} J(\tilde{K})[\widehat{K}(N) - K_\star, \widehat{K}(N) - K_\star] \\
&\leq \frac{1}{2} \left(\sup_{\|\tilde{K}\| \leq C_K + \|K_\star\|} \|\mathrm{Hess} J(\tilde{K})\| \right) \|\widehat{K}(N) - K_\star\|_F^2 \\
&\leq \frac{d}{2} \left(\sup_{\|\tilde{K}\| \leq C_K + \|K_\star\|} \|\mathrm{Hess} J(\tilde{K})\| \right) C_{\mathrm{deriv}}^2 (\|\widehat{A}(N) - A\|^2 + \|\widehat{B}(N) - B\|^2) \, .
\end{aligned}$$

Hence for N sufficiently large, by Lemma 6.2.14 we have

$$N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] \le CN^p(\mathbb{E}[\|\widehat{A}(N) - A\|^{2p}] + \mathbb{E}[\|\widehat{B}(N) - B\|^{2p}]) \le CN^p(\frac{1}{N^p}) = C\,.$$

On the event $\mathcal{E}^c_{\mathsf{Bdd}} \cap \mathcal{E}_{\mathsf{Alg}}$. In this case, we use the bounds given by $\mathcal{E}_{\mathsf{Alg}}$ to bound the controller $\widehat{K}(N)$. Lemma 6.2.19 ensures that the solution $\widehat{V} = \mathsf{dare}(\widehat{A}(N), \widehat{B}(N), I_n, 0)$ exists and satisfies $\widehat{V} \succeq I_n$. Let the finite constant C_P be

$$C_P := \sup_{\rho(A)\le\varrho, \|A\|\le\zeta, \|B\|\le\psi, \sigma_d(B)\ge\gamma} \|\mathsf{dare}(A, B, I_n, 0)\|\,.$$

We can then bound $\|\widehat{K}(N)\|$ as follows. Dropping the indexing of N,

$$\|\widehat{K}\| = \|(\widehat{B}^\mathsf{T}\widehat{V}\widehat{B})^{-1}\widehat{B}^\mathsf{T}\widehat{V}\widehat{A}\| \le \frac{1}{\sigma_{\min}(\widehat{B}^\mathsf{T}\widehat{V}\widehat{B})}\|\widehat{B}^\mathsf{T}\widehat{V}\widehat{A}\| \le \frac{C_P\psi\zeta}{\gamma^2}\,.$$

Therefore:

$$N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}^c_{\mathsf{Bdd}}\cap\mathcal{E}_{\mathsf{Alg}}}] \le N^p \cdot \left(\sup_{\|K\|\le\frac{C_P\psi\zeta}{\gamma^2}} (J(K) - J_\star)^p\right) \mathbb{P}(\mathcal{E}^c_{\mathsf{Bdd}}) \le CN^p\mathbb{P}(\mathcal{E}^c_{\mathsf{Bdd}})\,.$$

By Lemma 6.2.14, we can choose N large enough such that $\mathbb{P}(\mathcal{E}^c_{\mathsf{Bdd}}) \le 1/N^p$ so that $N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}^c_{\mathsf{Bdd}}\cap\mathcal{E}_{\mathsf{Alg}}}] \le C$.

On the event $\mathcal{E}^c_{\mathsf{Bdd}} \cap \mathcal{E}^c_{\mathsf{Alg}}$. This case is simple. We simply invoke Lemma 6.2.14 to choose an N large enough such that $\mathbb{P}(\mathcal{E}^c_{\mathsf{Bdd}}) \le 1/(N(J(0) - J_\star))^p$.

Putting it together. If we take N as the maximum over the three cases described above, we have hence shown that for all N greater than this constant:

$$\mathbb{E}[Z_N^p] \le C\,.$$

This shows the desired uniform integrability condition for Z_N. The asymptotic bound now follows from Lemma 6.2.12.

6.2.7.3 Proof of Theorem 6.2.5

The proof works by applying Lemma 6.2.15 with the function $F(\theta) = J_W(K)$ with $W = \sigma_\eta^2 BB^\mathsf{T} + \sigma_w^2 I_n$ and $G(\theta) = J(K)$. We first need to verify the hypothesis of the lemma. We

define the convex domain Θ as $\Theta = \{K \in \mathbb{R}^{d\times n} : \|K\| \le \zeta\}$. Note that $K_\star$ is in the interior of Θ, since we assume that $\|K_\star\| < \zeta$. Recall that the policy gradient $g(K;\xi)$ is:

$$g(K;\xi) = \frac{1}{\sigma_\eta^2}\sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} \Psi_t \ , \ \ \xi = (\eta_0, w_0, \eta_1, w_1, ..., \eta_{T-1}, w_{T-1}) \ .$$

It is clear that x_t is a polynomial in (K,ξ). Furthermore, all three of the Ψ_t's we study are also polynomials in (K,ξ). Hence $[D_K g(K;\xi)]$ is a matrix with entries that are polynomial in (K,ξ). Therefore, for every ξ, for all fixed $K_1, K_2 \in \Theta$,

$$\|g(K_1;\xi) - g(K_2;\xi)\|_F \le \sup_{K\in\Theta} \|[D_K g(K;\xi)]\|_F \|K_1 - K_2\|_F \ .$$

Hence squaring and taking expectations,

$$\mathbb{E}_\xi[\|g(K_1;\xi) - g(K_2;\xi)\|_F^2] \le \mathbb{E}_\xi\left[\sup_{K\in\Theta} \|[D_K g(K;\xi)]\|_F^2\right] \|K_1 - K_2\|_F^2 \ .$$

We can now define the constant $L := \mathbb{E}_\xi\left[\sup_{K\in\Theta}\|[D_K g(K;\xi)]\|_F^2\right]$. To see that this quantity L is finite, observe that $\|[D_K g(K;\xi)]\|_F^2$ is a polynomial of ξ with coefficients given by K (and A, B). Since K lives in a compact set Θ, these coefficients are uniformly bounded and hence the their moments are bounded. In Section 6.2.7.1, we showed that the function $J_\Sigma(K)$ satisfies the $\mathsf{RSC}(m,\Theta)$ condition with $m = 2(T-1)\sigma_w^2\sigma_{\min}(B)^2$. Also it is clear that the high probability bound on $\|g(K;\xi)\|_F$ can be achieved by standard Gaussian concentration results. Hence by Lemma 6.2.15, and in particular Equation 6.2.12,

$$\begin{aligned}\liminf_{N\to\infty} N\cdot\mathbb{E}[J(\widehat{K}) - J_\star] &\ge \frac{\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]}{8(T-1)\sigma_w^2\sigma_{\min}(B)^2\lambda_{\max}((\nabla^2 J(K_\star))^{-1}(\nabla^2 J_\Sigma(K_\star) - \frac{m}{2}I_{nd}))} \\ &= \frac{\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]}{8(T-1)\sigma_{\min}(B)^2(\sigma_w^2 + \sigma_\eta^2\|B\|^2)} \ . \end{aligned} \tag{6.2.18}$$

Above, the inequality holds since we have that,

$$\begin{aligned}\nabla^2 J(K_\star) &= 2(T-1)(\sigma_w^2 I_n \otimes B^\mathsf{T}B) \ , \\ \nabla^2 J_\Sigma(K_\star) &= 2(T-1)((\sigma_w^2 I_n + \sigma_\eta^2 BB^\mathsf{T})\otimes B^\mathsf{T}B) = \nabla^2 J(K_\star) + 2(T-1)\sigma_\eta^2(BB^\mathsf{T}\otimes B^\mathsf{T}B) \ ,\end{aligned}$$

and therefore,

$$\begin{aligned}(\nabla^2 J(K_\star))^{-1}(\nabla^2 J_\Sigma(K_\star) - \frac{m}{2}I_{nd}) &= I_{nd} + \frac{\sigma_\eta^2}{\sigma_w^2}(BB^\mathsf{T}\otimes I_d) - \frac{\sigma_{\min}(B)^2}{2}(I_n\otimes(B^\mathsf{T}B)^{-1}) \\ &\preceq I_{nd} + \frac{\sigma_\eta^2}{\sigma_w^2}(BB^\mathsf{T}\otimes I_d) \ .\end{aligned}$$

The remainder of the proof is to estimate the quantity $\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]$. Note that at $K = K_\star$, $x_t = B\eta_{t-1} + w_{t-1}$ since the dynamics are cancelled out. Define $c_{t\to T} := \sum_{\ell=t}^T \|x_\ell\|^2$.

At $K = K_\star$, we have $c_{t\to T} = \sum_{\ell=t-1}^{T-1} \|B\eta_\ell + w_\ell\|^2$. Observe that we have for $t_2 > t_1$, for any h that depends on only $(\eta_{t_1}, w_{t_1}, \eta_{t_1+1}, w_{t_1+1}, ...)$:

$$\begin{aligned}\mathbb{E}[\langle \eta_{t_1}, \eta_{t_2}\rangle \langle x_{t_1}, x_{t_2}\rangle h] &= \mathbb{E}[\langle \eta_{t_1}, \eta_{t_2}\rangle (\langle B\eta_{t_1-1}, B\eta_{t_2-1}\rangle + \langle w_{t_1-1}, w_{t_2-1}\rangle \\ &\qquad + \langle B\eta_{t_1-1}, w_{t_2-1}\rangle + \langle B\eta_{t_2-1}, w_{t_1-1}\rangle) h] \\ &= 0\,.\end{aligned}$$

As a consequence, we have that as long as Ψ_t only depends on $(\eta_t, w_t, \eta_{t+1}, w_{t+1}, ...)$:

$$\begin{aligned}\mathbb{E}[\|g(K;\xi)\|_F^2] &= \frac{1}{\sigma_\eta^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|^2 \|x_t\|^2 \Psi_t^2] + \frac{2}{\sigma_\eta^4} \sum_{t_2 > t_1 = 1}^{T-1} \mathbb{E}[\langle \eta_{t_1}, \eta_{t_2}\rangle \langle x_{t_1}, x_{t_2}\rangle \Psi_{t_1} \Psi_{t_2}] \\ &= \frac{1}{\sigma_\eta^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|^2 \|x_t\|^2 \Psi_t^2]\,.\end{aligned}$$

Simple baseline. Recall that the simple baseline is to set $b_t(x_t; K) = \|x_t\|^2$. Hence, the policy gradient estimate simplifies to $g(K;\xi) = \frac{1}{\sigma_\eta^2} \sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} c_{t+1\to T}$. Since we have that $c_{t+1\to T}$ at optimality only depends only on $(\eta_t, w_t, \eta_{t+1}, w_{t+1}, ...)$, we compute $\mathbb{E}[\|g(K_\star;\xi)\|_F^2]$

as follows:

$$
\begin{aligned}
&\mathbb{E}[\|g(K_\star;\xi)\|_F^2] \\
&= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}[\|\eta_t\|^2\|x_t\|^2 c_{t+1\to T}^2] \\
&= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}\Bigg[\|\eta_t\|^2\|B\eta_{t-1}+w_{t-1}\|^2 \\
&\qquad \times\left(\sum_{\ell=t}^{T-1}\|B\eta_\ell+w_\ell\|^4 + 2\sum_{\ell_2>\ell_1=t}^{T-1}\|B\eta_{\ell_1}+w_{\ell_1}\|^2\|B\eta_{\ell_2}+w_{\ell_2}\|^2\right)\Bigg] \\
&= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}[\|B\eta_{t-1}+w_{t-1}\|^2\|\eta_t\|^2\|B\eta_t+w_t\|^4] \\
&\qquad + \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell=t+1}^{T-1}\mathbb{E}[\|B\eta_{t-1}+w_{t-1}\|^2\|\eta_t\|^2\|B\eta_\ell+w_\ell\|^4] \\
&\qquad + \frac{2}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell_2>t}\mathbb{E}[\|B\eta_{t-1}+w_{t-1}\|^2\|\eta_t\|^2\|B\eta_t+w_t\|^2\|B\eta_{\ell_2}+w_{\ell_2}\|^2] \\
&\qquad + \frac{2}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell_2>\ell_1=t+1}^{T-1}\mathbb{E}[\|B\eta_{t-1}+w_{t-1}\|^2\|\eta_t\|^2\|B\eta_{\ell_1}+w_{\ell_1}\|^2\|B\eta_{\ell_2}+w_{\ell_2}\|^2] \\
&= \frac{2}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell_2>\ell_1=t+1}^{T-1}\mathbb{E}[\|B\eta_{t-1}+w_{t-1}\|^2\|\eta_t\|^2\|B\eta_{\ell_1}+w_{\ell_1}\|^2\|B\eta_{\ell_2}+w_{\ell_2}\|^2] + o(T^3) \\
&= \frac{2}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell_2>\ell_1=t+1}^{T-1}\sigma_\eta^2 d(\mathbb{E}[\|B\eta_0+w_0\|^2])^3 + o(T^3) \\
&\asymp T^3\frac{1}{\sigma_\eta^2}d(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)^3 + o(T^3)\,.
\end{aligned}
$$

Value function baseline. Recall that the value function at time t for a particular policy K is defined as:

$$
V_t^K(x) = \mathbb{E}\left[\sum_{\ell=t}^{T}\|x_\ell\|^2\,\middle|\,x_t = x\right]\,.
$$

We now consider policy gradient with the value function baseline $b_t(x_t;K) = V_t^K(x_t)$:

$$
g(K;\xi) = \frac{1}{\sigma_\eta^2}\sum_{t=1}^{T-1}\eta_t x_t^\mathsf{T}(c_{t\to T} - V_t^K(x_t))\,.
$$

Recalling that under $K_\star$ the dynamics are cancelled out, we readily compute:

$$V_t^{K_\star}(x) = \|x\|^2 + (T-t)(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)\,.$$

Therefore:

$$g(K_\star;\xi) = \frac{1}{\sigma_\eta^2}\sum_{t=1}^{T-1}\eta_t x_t^\mathsf{T}(c_{t+1\to T} - (T-t)(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n))\,.$$

Define $\beta := \sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n$. We compute the variance as:

$$\begin{aligned}
\mathbb{E}[\|g(K_\star;\xi)\|_F^2] &= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}\Bigg[\|\eta_t\|^2\|B\eta_{t-1}+w_{t-1}\|^2 \\
&\qquad \times\left(\sum_{\ell=t}^{T-1}(\|B\eta_\ell + w_\ell\|^2-\beta)^2 + 2\sum_{\ell_2>\ell_1=t}^{T-1}(\|B\eta_{\ell_1}+w_{\ell_1}\|^2-\beta)(\|B\eta_{\ell_2}+w_{\ell_2}\|^2-\beta)\right)\Bigg] \\
&= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}[\|\eta_t\|^2\|B\eta_{t-1}+w_{t-1}\|^2(\|B\eta_t+w_t\|^2-\beta)^2] \\
&\qquad + \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell=t+1}^{T-1}\mathbb{E}[\|\eta_t\|^2\|B\eta_{t-1}+w_{t-1}\|^2(\|B\eta_\ell+w_\ell\|^2-\beta)^2] \\
&= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\sum_{\ell=t+1}^{T-1}\mathbb{E}[\|\eta_t\|^2\|B\eta_{t-1}+w_{t-1}\|^2(\|B\eta_\ell+w_\ell\|^2-\beta)^2] + o(T^2) \\
&\asymp T^2\frac{d}{\sigma_\eta^2}(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)(\mathbb{E}[\|B\eta_\ell+w_\ell\|^4]-\beta^2) + o(T^2) \\
&\overset{(a)}{\asymp} T^2\frac{d}{\sigma_\eta^2}(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)(\sigma_\eta^4\|B^\mathsf{T}B\|_F^2+\sigma_w^4 n+\sigma_w^2\sigma_\eta^2\|B\|_F^2) + o(T^2)\,,
\end{aligned}$$

Above, (a) follows because:

$$\mathbb{E}[\|B\eta_\ell+w_\ell\|^4] = 2(\sigma_\eta^4\|B^\mathsf{T}B\|_F^2+\sigma_w^4 n+2\sigma_w^2\sigma_\eta^2\|B\|_F^2) + (\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)^2\,.$$

Ideal advantage baseline. Let us first compute $Q_t^{K_\star}(x_t,u_t)$. Under $K_\star$, $x_{\ell+1} = B\eta_\ell + w_\ell$. So we have:

$$\begin{aligned}
Q_t^{K_\star}(x_t,u_t) &= \|x_t\|^2 + \mathbb{E}_{w_t}[\|Ax_t+Bu_t+w_t\|^2] + (T-t-1)(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n) \\
&= \|x_t\|^2 + \|Ax_t+Bu_t\|^2 + \sigma_w^2 n + (T-t-1)(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)\,.
\end{aligned}$$

Recalling that $V_t^{K_\star}(x) = \|x\|^2 + (T-t)(\sigma_\eta^2\|B\|_F^2+\sigma_w^2 n)$,

$$A_t^{K_\star}(x_t,u_t) = Q_t^{K_\star}(x_t,u_t) - V_t^{K_\star}(x_t) = \|Ax_t+Bu_t\|^2 - \sigma_\eta^2\|B\|_F^2\,.$$

Therefore, if $u_t = K_\star x_t + \eta_t$, we have $A_t^{K_\star}(x_t, u_t) = \|B\eta_t\|^2 - \sigma_\eta^2\|B\|_F^2$. Since $A_t^{K_\star}(x_t, u_t)$ depends only on η_t,

$$\begin{aligned}\mathbb{E}[\|g(K_\star;\xi)\|_F^2] &= \frac{1}{\sigma_\eta^4}\sum_{t=1}^{T-1}\mathbb{E}[\|\eta_t\|^2\|x_t\|^2(\|B\eta_t\|^2 - \sigma_\eta^2\|B\|_F^2)^2] \\ &= \frac{1}{\sigma_\eta^4}(T-1)(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)\mathbb{E}[\|\eta_1\|^2(\|B\eta_1\|^2 - \sigma_\eta^2\|B\|_F^2)^2]\,.\end{aligned}$$

We have that $\mathbb{E}[\|\eta_1\|^2] = \sigma_\eta^2 d$, $\mathbb{E}[\|B\eta_1\|^2\|\eta_1\|^2] = \sigma_\eta^4(d+2)\|B\|_F^2$, and $\mathbb{E}[\|B\eta_1\|^4\|\eta_1\|^2] = \sigma_\eta^6((d+4)\|B\|_F^4 + (2d+8)\|B^\mathsf{T}B\|_F^2)$ (this can be computed using Lemma 6.2.16). Hence,

$$\begin{aligned}&\mathbb{E}[\|\eta_1\|^2(\|B\eta_1\|^2 - \sigma_\eta^2\|B\|_F^2)^2] \\ &= \mathbb{E}[\|B\eta_1\|^4\|\eta_1\|^2 + \sigma_\eta^4\|B\|_F^4\|\eta_1\|^2 - 2\sigma_\eta^2\|B\|_F^2\|B\eta_1\|^2\|\eta_1\|^2] \\ &= \sigma_\eta^6((d+4)\|B\|_F^4 + (2d+8)\|B^\mathsf{T}B\|_F^2) + \sigma_\eta^6\|B\|_F^4 d - 2\sigma_\eta^6\|B\|_F^4(d+2) \\ &= \sigma_\eta^6(2d+8)\|B^\mathsf{T}B\|_F^2\,.\end{aligned}$$

Therefore,

$$\mathbb{E}[\|g(K_\star;\xi)\|_F^2] \asymp T(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)\sigma_\eta^2 d\|B^\mathsf{T}B\|_F^2\,.$$

Putting it together Combining Equation (6.2.18) with the calculations for the variance $\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]$, we obtain:

$$\liminf_{N\to\infty} N\cdot\mathbb{E}[J(\widehat{K}_{\mathrm{pg}}(N)) - J_\star] \gtrsim \frac{1}{\sigma_d(B)^2(\sigma_w^2 + \sigma_\eta^2\|B\|^2)}\times$$

$$\begin{cases} T^2\frac{d}{\sigma_\eta^2}(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)^3 + o(T^2) & \text{(Simple baseline)} \\ T\frac{d}{\sigma_\eta^2}(\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)(\sigma_\eta^4\|B^\mathsf{T}B\|_F^2 + \sigma_w^4 n + \sigma_w^2\sigma_\eta^2\|B\|_F^2) + o(T) & \text{(Value function baseline)} \\ (\sigma_\eta^2\|B\|_F^2 + \sigma_w^2 n)\sigma_\eta^2 d\|B^\mathsf{T}B\|_F^2 & \text{(Advantage baseline)} \end{cases}$$

from which Theorem 6.2.5 follows.

6.2.7.4 Proof of Theorem 6.2.6

Our proof is inspired by lower bounds for the query complexity of derivative-free optimization of stochastic optimization (see e.g. Jamieson et al. [52]).

Recall from (6.2.16) that the function $J(K)$ satisfies the quadratic growth condition $J(K) - J_\star \geq (T-1)\rho^2\sigma_w^2\|K - K_\star\|_F^2$. Therefore for any $\vartheta > 0$,

$$\begin{aligned}
&\inf_{\widehat{K}} \sup_{(A,B)\in\mathscr{G}(\rho,d)} \mathbb{E}[J(\widehat{K}) - J_\star] \\
&\geq \inf_{\widehat{K}} \sup_{(A,B)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}(J(\widehat{K}) - J_\star \geq (T-1)\rho^2\sigma_w^2\vartheta^2) \\
&\geq \inf_{\widehat{K}} \sup_{(A,B)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}((T-1)\rho^2\sigma_w^2\|(-U_\star^\mathsf{T}) - \widehat{K}\|_F^2 \geq (T-1)\rho^2\sigma_w^2\vartheta^2) \\
&= \inf_{\widehat{K}} \sup_{(A,B)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}(\|(-U_\star^\mathsf{T}) - \widehat{K}\|_F \geq \vartheta) \,.
\end{aligned}$$

Above, the first inequality is Markov's inequality and the second is the quadratic growth condition.

We first state a result regarding the packing number of $O(n,d)$, which we define as:

$$O(n,d) := \{U \in \mathbb{R}^{n\times d} : U^\mathsf{T}U = I_d\} \,.$$

Lemma 6.2.21. *Let $\delta > 0$, and suppose that $d \leq n/2$. We have that the packing number M of $O(n,d)$ in the Frobenius norm $\|\cdot\|_F$ satisfies*

$$M(O(n,d), \|\cdot\|_F, \delta d^{1/2}) \geq \left(\frac{c}{\delta}\right)^{d(n-d)} \,,$$

where $c > 0$ is a universal constant.

Proof. Let $G_{n,d}$ denote the Grassman manifold of d-dimensional subspaces of $\mathbb{R}^n$. For two subspaces $E, F \in G_{n,d}$, equip $G_{n,d}$ with the metric $\rho(E,F) = \|P_E - P_F\|_F$, where P_E, P_F are the projection matrices onto E, F respectively. Proposition 8 of Pajor [90] tells us that the covering number $N(G_{n,d}, \rho, \delta d^{1/2}) \geq \left(\frac{c}{\delta}\right)^{d(n-d)}$. But since $M(G_{n,d}, \rho, \delta d^{1/2}) \geq N(G_{n,d}, \rho, \delta d^{1/2})$, this gives us a lower bound on the packing number of $G_{n,d}$. Now for every $E \in G_{n,d}$ we can associate a matrix $E_1 \in O(n,d)$ such that $\mathrm{span}(E_1) = E$. The projector P_E is simply $P_E = E_1E_1^\mathsf{T}$. Now let $E, F \in G_{n,d}$ and observe the inequality,

$$\|P_E - P_F\|_F = \|E_1E_1^\mathsf{T} - F_1F_1^\mathsf{T}\|_F \leq 2\|E_1 - F_1\|_F \,.$$

Hence a packing of $G_{n,d}$ also yields a packing of $O(n,d)$ up to constant factors. □

Now letting $U_1, ..., U_M$ be a 2ϑ-separated set we have by the standard reduction to multiple hypothesis testing that that the risk is lower bounded by:

$$(T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \inf_{\widehat{V}} \mathbb{P}(\widehat{V} \neq V) \geq (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \left(1 - \frac{I(V;Z) + \log 2}{\log M}\right) \,. \tag{6.2.19}$$

where V is a uniform index over $\{1, ..., M\}$ and the inequality is Fano's inequality.

Now we can proceed as follows. First, we let $U_1, ..., U_M$ be elements of $O(n, d)$ that form a $2\vartheta \asymp \sqrt{d}$ packing in the $\|\cdot\|_F$ norm. We know we can let $M \geq e^{d(n-d)}$ by Lemma 6.2.21. Each U_i induces a covariance $\Sigma_i = \sigma_w^2 I_n + \rho^2 \sigma_\eta^2 U_i U_i^\mathsf{T} \preceq (\sigma_w^2 + \rho^2\sigma_\eta^2) I_n$. Furthermore, the closed-loop L_i given by playing a feedback matrix K that satisfies $\|K\| \leq 1$ is:

$$L_i = \rho U_i (U_i + K^\mathsf{T})^\mathsf{T} .$$

It is clear that $\|L_i\| \leq 2\rho$ and hence if $\rho < 1/2$ then this system is stable. Furthermore, we have that $\mathrm{rank}(L_i) \leq d$. With this, we can control:

$$\begin{aligned}
\mathrm{tr}(\mathbb{E}[x_t x_t^\mathsf{T}]) = \mathrm{tr}\left(\sum_{\ell=0}^{t-1} L_i^\ell \Sigma_i (L_i^\ell)^\mathsf{T}\right) &\leq (\sigma_w^2 + \rho^2\sigma_\eta^2) \sum_{\ell=0}^{t-1} \|L_i^\ell\|_F^2 \\
&\leq d(\sigma_w^2 + \rho^2\sigma_\eta^2) \sum_{\ell=0}^{t-1} \|L_i^\ell\|^2 \leq \frac{d(\sigma_w^2 + \rho^2\sigma_\eta^2)}{1 - (2\rho)^2} .
\end{aligned}$$

Hence for one trajectory $Z = (x_0, u_0, x_1, u_1, ..., x_{T-1}, u_{T-1}, x_T)$, conditioned on a particular K,

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{i|K}, \mathbb{P}_{j|K}) &\leq \sum_{t=0}^{T-1} \frac{1}{2\sigma_w^2} \mathbb{E}_{x_t \sim \mathbb{P}_{i|K}}[\|(L_i - L_j) x_t\|^2] \\
&\leq \frac{8\rho^2}{\sigma_w^2} \sum_{t=0}^{T-1} \mathrm{tr}(\mathbb{E}[x_t x_t^\mathsf{T}]) \\
&\leq \frac{8(\sigma_w^2 + \rho^2\sigma_\eta^2)\rho^2 T d}{\sigma_w^2 (1 - (2\rho)^2)} .
\end{aligned}$$

This allows us to bound the KL between the distributions involving all the iterations as:

$$\mathrm{KL}(\mathbb{P}_i, \mathbb{P}_j) = \sum_{\ell=1}^{N} \mathbb{E}_{K_\ell \sim \mathbb{P}_i}[\mathrm{KL}(\mathbb{P}_{i|K_\ell}, \mathbb{P}_{j|K_\ell})] \leq \frac{8(\sigma_w^2 + \rho^2\sigma_\eta^2)\rho^2 N T d}{\sigma_w^2 (1 - (2\rho)^2)} .$$

Assuming $d(n-d)$ is greater than an absolute constant, we can set ρ to be (recall we have N different rollouts):

$$\rho^2 \asymp \frac{\sigma_w^2}{\sigma_w^2 + \sigma_\eta^2} \frac{n-d}{TN} ,$$

and bound $\frac{I(V;Z) + \log 2}{\log M} \leq 1/2$. The result now follows from plugging in our choice of ρ into (6.2.19).

Chapter 7

Experiments

7.1 Model-based Methods

Here, we show the empirical performance of the methods described in Chapter 5. We focus our experiments on a particular example system. Consider the LQR problem instance specified by

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B = I_3, \quad S = 10^{-3} I_3, \quad R = I_3 \,. \tag{7.1.1}$$

The dynamics correspond to a marginally unstable graph Laplacian system where adjacent nodes are weakly connected, each node receives direct input, and input size is penalized relatively more than state. Dynamics described by graph Laplacians arise naturally in consensus and distributed averaging problems. For this system, we estimate the dynamics with the estimator (3.0.3), using inputs with variance $\sigma_u^2 = 1$ and noise with variance $\sigma_w^2 = 1$. The error bounds of $\|\widehat{A} - A\|$ and $\|\widehat{B} - B\|$ are estimated via a simple bootstrap procedure described in Dean et al. [31].

Using the estimates of the system in (7.1.1), we synthesize controllers using two robust control schemes: the convex problem in (5.2.27) with filters of length $L = 32$ and V set to 0, and the common Lyapunov (CL) relaxation of the static synthesis problem (5.2.28). Once the FIR responses $\{\Phi_x(k)\}_{k=1}^F$ and $\{\Phi_u(k)\}_{k=1}^F$ are found, we need a way to implement the system responses as a controller. We represent the dynamic controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ by finding an equivalent state-space realization (A_K, B_K, C_K, D_K) via Theorem 2 of Anderson and Matni [9]. In what follows, we compare the performance of these controllers with the nominal LQR controller (described in Section 5.1), and explore the trade-off between robustness, complexity, and performance.

The relative performance of the nominal controller is compared with robustly synthesized controllers in Figure 7.1. For both robust synthesis procedures, two controllers are compared: one using the true errors on A and B, and the other using the bootstrap estimates of the

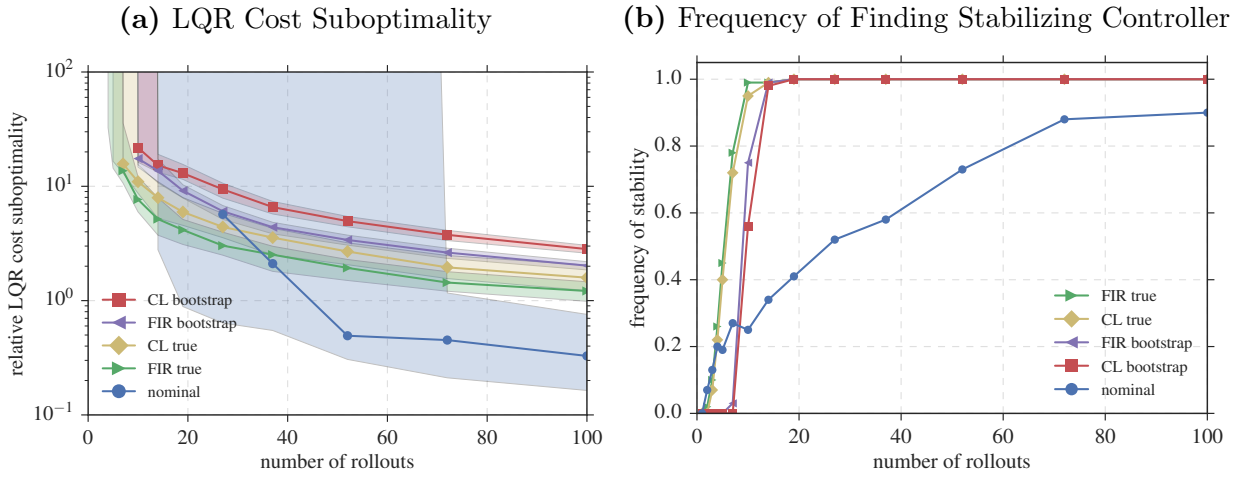

Figure 7.1: The performance of controllers synthesized on the results of the 100 identification experiments is plotted against the number of rollouts. Controllers are synthesis nominally, using FIR truncation, and using the common Lyapunov (CL) relaxation. In (a), the median suboptimality of nominal and robustly synthesized controllers are compared, with shaded regions displaying quartiles, which go off to infinity in the case that a stabilizing controller was not found. In (b), the frequency that the synthesis methods found stabilizing controllers.

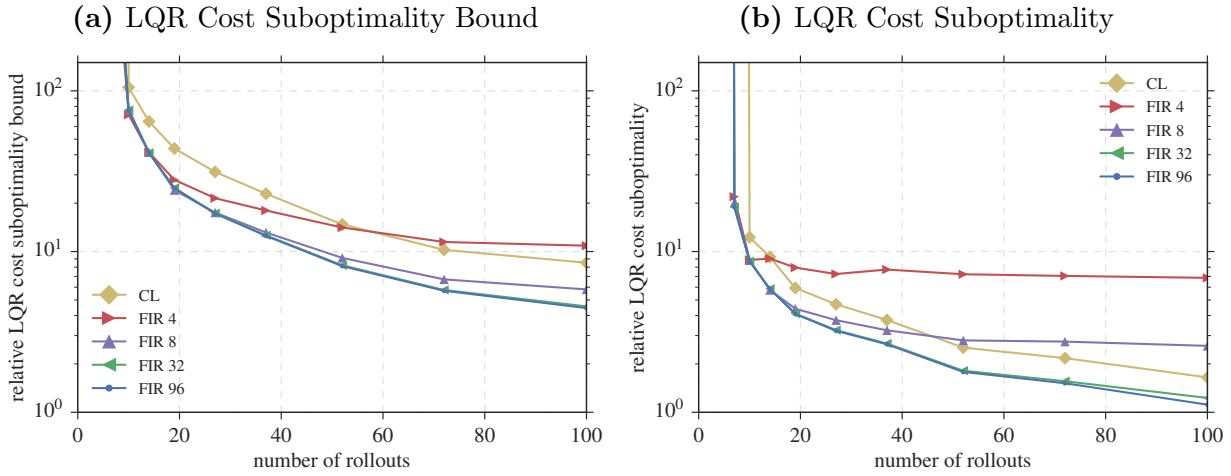

Figure 7.2: The performance of controllers synthesized with varying FIR filter lengths on the results of 10 of the identification experiments using true errors. The median suboptimality of robustly synthesized controllers does not appear to change for FIR lengths greater than 32, and the common Lyapunov (CL) synthesis tracks the performance in both upper bound and actual cost.

errors. The robust static controller generated via the common Lyapunov approximation performs slightly worse than the more complex FIR controller, but it still achieves reasonable control performance. Moreover, the conservative bootstrap estimates also result in worse control performance, but the degradation of performance is again modest.

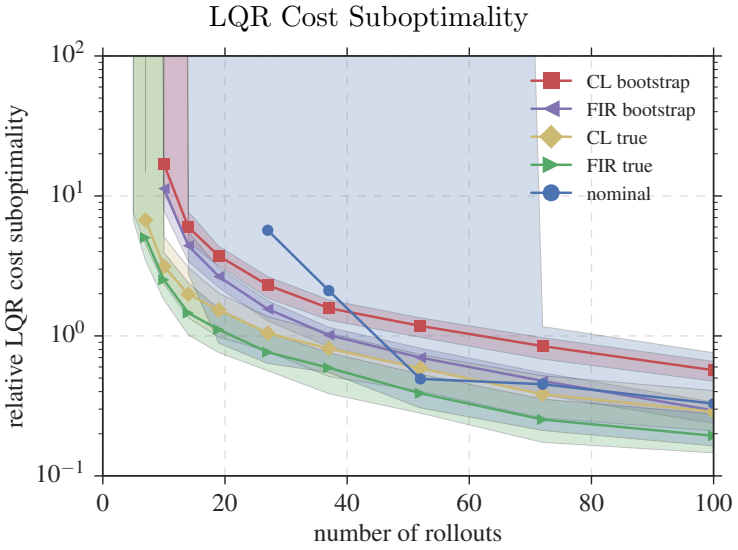

Figure 7.3: The performance of controllers synthesized on the results of 100 identification experiments is plotted against the number of rollouts. The plot compares the median suboptimality of nominal controllers with fixed-γ robustly synthesized controllers ($\gamma = 0.999$).

Furthermore, the experiments show that the nominal controller often outperforms the robust controllers *when it is stabilizing.* On the other hand, the nominal controller is not guaranteed to stabilize the true system, and as shown in Figure 7.1, it only does so in roughly 80 of the 100 instances after $N = 60$ rollouts. It is also important to note a distinction between stabilization for nominal and robust controllers. When the nominal controller is not stabilizing, there is no indication to the user (though sufficient conditions for stability can be checked using our result in Corollary 5.2.4 or structured singular value methods [93]). On the other hand, the robust synthesis procedure will return as infeasible, alerting the user by default that the uncertainties are too high. We observe similar results when we fix the number of trials but vary the rollout length.

Figure 7.2 explores the trade-off between performance and complexity for the computational approximations, both for FIR truncation and the common Lyapunov relaxation. We examine the tradeoff both in terms of the bound on the LQR cost (given by the value of the objective) as well as the actual achieved value. It is interesting that for smaller numbers of rollouts (and therefore larger uncertainties), the benefit of using more complex FIR models is negligible, both in terms of the actual costs and the upper bound. This trend makes sense: as uncertainties decrease to zero, the best robust controller should approach the nominal controller, which is associated with infinite impulse response (IIR) transfer functions. Furthermore, for the experiments presented here, FIR length of $L = 32$ seems to be sufficient to characterize the performance of the robust synthesis procedure in (5.2.18). Additionally, we note that static controllers are able to achieve costs of a similar magnitude.

The SLS framework guarantees a stabilizing controller for the true system provided that

the computational approximations are feasible for *any* value of γ between 0 and 1, as long as the system errors $(\varepsilon_A, \varepsilon_B)$ are upper bounds on the true errors. Figure 7.3 displays the controller performance for robust synthesis when γ is set to 0.999. Simply ensuring a stable model and neglecting to optimize the nominal cost yields controllers that perform nearly an order of magnitude better than those where we search for the optimal value of γ. This observation aligns with common practice in robust control: constraints ensuring stability are only active when the cost tries to drive the system up against a safety limit. We cannot provide end-to-end sample complexity guarantees for this method and leave such bounds as an enticing challenge for future work.

7.2 Model-free Methods

In this section we look at the performance of the model-free methods described in Section 1.3 and Chapter 6: policy gradients, derivative-free optimization, and policy iteration. We will compare these model-free methods to the model-based nominal control (Section 5.1) as a baseline.

We consider the stable system:

$$A = \begin{bmatrix} 0.95 & 0.01 & 0 \\ 0.01 & 0.95 & 0.01 \\ 0 & 0.01 & 0.95 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0.1 \\ 0 & 0.1 \\ 0 & 0.1 \end{bmatrix}, \quad S = I_3, \quad R = I_2. \tag{7.2.1}$$

We choose an LQR problem where the A matrix is stable, since the model-free methods we consider need to be seeded with an initial stabilizing controller; using a stable A allows us to start at $K_0 = 0_{2\times 3}$. We fix the process noise $\sigma_w = 1$. As before, the model-based nominal method learns (A, B) using (3.0.3), exciting the system with Gaussian noise that has variance $\sigma_u = 1$.

For policy gradients and derivative-free optimization, we use the projected stochastic gradient descent (SGD) method with a constant step size μ as the optimization procedure. After every iteration, we project the iterate K_t onto the set $\{K : \|K\|_F \leq 5\|K_\star\|_F\}$, where $K_\star$ is the optimal LQR controller (we assume the value $\|K_\star\|_F$ is known for simplicity). We tune the parameters of each algorithm as follows. We consider a grid of step sizes μ given by $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ and a grid of σ_η's given by $[1, 10^{-1}, 10^{-2}, 10^{-3}]$. We fix the rollout horizon length $T = 100$ and choose the pair of (σ_η, μ) in the grid which yields the lowest cost after 10^6 timesteps. This resulted in the pair $(\sigma_\eta, \mu) = (1, 10^{-5})$ for policy gradients and $(\sigma_\eta, \mu) = (10^{-3}, 10^{-4})$ for DFO. We use the two point evaluation for derivative-free optimization (1.3.13), so each iteration requires $2T$ timesteps. For policy gradient, we evaluate two different baselines. One baseline, which we call the *simple* baseline, uses the empirical average cost $\frac{1}{T}\sum_{t=1}^{T} c_t$ from the previous iteration as a constant baseline. The second baseline, which we call the *value function* baseline, uses the value function $V(x) = x^\mathsf{T} V(K) x$ with $V(K) = \mathsf{dlyap}(A + BK, S + K^\mathsf{T} R K)$ as the baseline. We note that using this baseline requires exact knowledge of the dynamics (A, B); it can however be

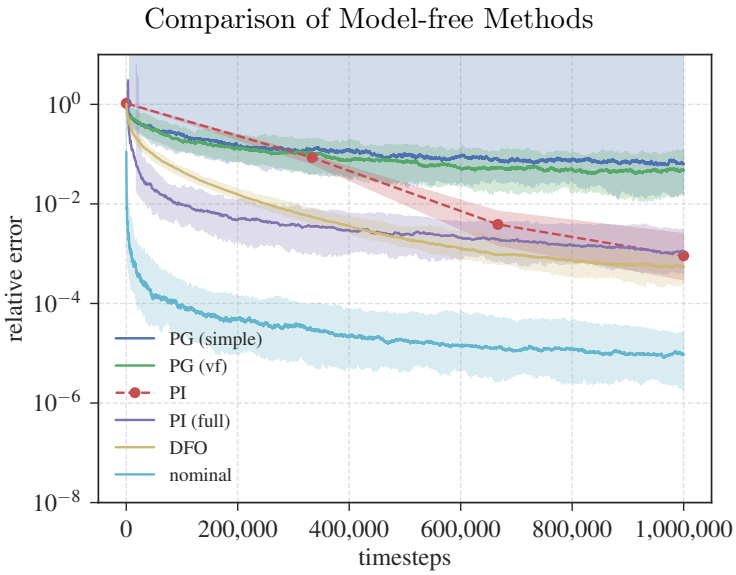

Figure 7.4: The performance of various model-free methods compared with the nominal (Section 5.1) controller. The shaded regions represent the lower 10th and upper 90th percentile over 100 trials, and the solid line represents the median performance. Here, PG (simple) is policy gradients with the simple baseline, PG (vf) is policy gradients with the value function baseline, PI is the modified policy iteration (Algorithm 2), PI (full) is the standard policy iteration (Algorithm 1), and DFO is derivative-free optimization.

estimated from data at the expense of additional sample complexity (c.f. Section 6.1). For the purposes of this experiment, we simply assume the baseline is available to us.

For policy iteration, we use the least-squares policy iteration (LSPI) algorithm described in Section 6.1. We evaluate both variants presented in Algorithm 1 and Algorithm 2. For every iteration of LSTD-Q, we project the resulting Q-function parameter matrix onto the set $\{Q : Q \succeq \gamma I\}$ with $\gamma = \min\{\lambda_{\min}(S), \lambda_{\min}(R)\}$. For Algorithm 1, we choose $N = 15$ by picking the $N \in [5, 10, 15]$ which results in the best performance after $T = 10^6$ timesteps. For Algorithm 2, we set $(N, T) = (3, 333333)$ which yields the lowest cost over the grid $N \in [1, 2, 3, 4, 5, 6, 7]$ and T such that $NT = 10^6$.

Figure 7.4 shows the results of these experiments, plotting the relative error $(J(\widehat{K}) - J_\star)/J_\star$ versus the number of timesteps. We see that the nominal method is more sample efficient than the other model-free methods considered. We also see that the value function baseline is able to dramatically reduce the variance of the policy gradient estimator compared to the simple baseline. The DFO method performs the best out of all the model-free methods considered on this example after 10^6 timesteps, although the performance of policy iteration is comparable.

Chapter 8

Conclusion

In this thesis, we studied the sample complexity of learning to control the Linear Quadratic Regulator. We looked at both model-based methods, which use trajectory data to build an estimate of the dynamics and design a controller from the estimated model, and also model-free methods which learn intermediate representations or directly search for the parameters of the optimal controller. We saw both theoretically and experimentally that for our particular setup, model-based methods substantially outperformed model-free methods.

8.1 Future Work

This thesis raises many questions for future study. Here, we outline several possible directions for future research.

Partial observability. In this thesis, we studied settings with full state observation. In practice it is often not possible to fully observe the state. The canonical model for partially observed linear systems is given by the following model:

$$x_{t+1} = Ax_t + Bu_t + w_t\ ,\ \ w_t \sim \mathcal{N}(0, W)\ , \tag{8.1.1}$$
$$y_t = Cx_t + v_t\ ,\ \ v_t \sim \mathcal{N}(0, V)\ . \tag{8.1.2}$$

This is a specific instance of a *partially observed Markov Decision Process* (POMDP). Consider the following optimal control problem:

$$\min_{u_t(\cdot)} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t\right] \ \text{ s.t. } (8.1.1), (8.1.2)\ , \tag{8.1.3}$$

where $u_t(\cdot)$ is allowed to only depend on $y_1, ..., y_{t-1}$. This problem, known as the Linear Quadratic Gaussian (LQG) problem, can also be solved exactly. Furthermore, the optimal solution has an elegant structure known as the *separation principle*. In particular, the optimal

feedback u_t is given by $u_t = K\hat{x}_t$, where K is the optimal LQR solution for (A, B, S, R), and $\hat{x}_t = \mathbb{E}[x_t|y_1, ..., y_{t-1}]$. The estimate $\hat{x}_t$ can be solved for using *Kalman filtering*.

This setup naturally raises the question of how to learn the optimal LQG controller from input/output observations of (8.1.1)-(8.1.2). This question is much more delicate than the analogous question studied in this thesis for LQR because there are an infinite number of realizations that generate the same input/output map represented by (8.1.1)-(8.1.2). Nevertheless, partial progress has been made on this problem, namely that of constructing estimates $(\widehat{A}, \widehat{B}, \widehat{C})$ of the true (A, B, C) up to a similarity transform [88, 99, 106, 114]. However, it remains open how to best use these estimate $(\widehat{A}, \widehat{B}, \widehat{C})$ to design an LQG controller that has a corresponding sub-optimality guarantee. A partial solution to this was given by Boczar et al. [19] using the System Level Synthesis machinery from Section 5.2, under the assumption that the dynamics are represented as a single input, single output (SISO) finite impulse response (FIR) filter of known order. The general case, however, is still open.

Adversarial disturbances. Throughout this thesis, we assumed that the process noise w_t was stochastic zero-mean and independent across time. While this assumption helped to simplify the analysis, it is not always realistic in practice. For instance, if we consider our linear dynamics arising from local linearization of non-linear dynamics, then the process noise would represent the linearization error which we certainly do not expect to be uncorrelated across time. A better model would be where the process noise sequence is chosen in a non-stochastic (possibly adversarial) manner. In this situation, we would still like to be able to learn how to control an unknown system, although we will need to change the type of guarantee we are after. A particular style of guarantee that has received a lot of attention in the machine learning community is that of *regret minimization*, where a proposed adaptive algorithm is compared to the best non-adaptive algorithm which has perfect knowledge of the adversary's behavior. One particular problem formulation is as follows. The goal is to design an adaptive algorithm $\mathcal{A} = \{u_t(\cdot)\}$ to minimize the following regret:

$$\mathsf{Regret}(T; \mathcal{A}) := \sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t - \min_{K: u_t = K x_t} \sum_{t=1}^{T} x_t^\mathsf{T} S x_t + u_t^\mathsf{T} R u_t \,, \tag{8.1.4}$$

where the comparator on the RHS searches over fixed static feedback policies with the knowledge of the process noise $\{w_t\}$ sequence chosen by the adversary. Partial progress in this formulation has been made recently. Abbasi-Yadkori et al. [3] study the closely related problem of tracking an unknown target position that moves in an adversarial manner, under the assumption that the algorithm has perfect knowledge of the dynamics (A, B). Cohen et al. [29] study the problem where the cost matrices (S, R) are allowed to vary over time in an adversarial manner. Here, it is also assumed that the (A, B) are known. Agarwal et al. [7] study the formulation posed in (8.1.4), where the sequence $\{w_t\}$ is only assumed to be uniformly bounded. They give an algorithm based on online convex optimization that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. Once again, it is assumed in Agarwal et al. [7] that the dynamics

matrices (A, B) are known. Extending this work to where the (A, B) are not known, and even allowed to vary in some way over time, is interesting future work.

Non-linear dynamics. In general, many systems which appear in the physical world are not linear. What kinds of guarantees can be made in a non-linear setting? There are many possible approaches to making progress on this problem. Perhaps the most natural way to get started is to look at techniques such as iterative LQR, differential dynamic programming [112], and model predictive control [21], and see if it possible to analyze how the model mismatch affects control performance. There is also a long line of research in RL that uses Gaussian Process (GP) regression to estimate either the dynamics or the value function (see e.g. [34, 37, 45, 96] and the references within). GP regression is appealing because there are known powerful concentration inequalities (see e.g. Srinivas et al. [108]) which can be used to construct data-driven confidence intervals. There are also a few other approaches which are less well-known in the machine learning communities that are also worth exploring. One approach is based on the Koopman operator from dynamical systems theory [27]. Brunton et al. [26] provide an overview of how the Koopman operator can be used in the context of non-linear optimal control. The use of the Koopman operator for data-driven control is an active area of current research. Another possibility is to model the input/output behavior of non-linear systems using Volterra series [77], which is a non-linear extension of the finite impulse response model.

It is worth mentioning however that regardless of approach taken, to obtain non-trivial PAC style bounds one most likely needs to restrict both the function class of the non-linear system and the function class of the controller considered. The question then becomes, what is an interesting model class that is rich enough to capture practical applications, but structured enough to be amenable to theoretical analysis? Given this, it is most likely that true progress on the non-linear front will only be achieved by new insights that combine both theory and practice together.

Bibliography

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret Bounds for the Adaptive Control of Linear Quadratic Systems. In *Conference on Learning Theory*, 2011.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems. In *Conference on Learning Theory*, 2011.

[3] Yasin Abbasi-Yadkori, Peter Bartlett, and Varun Kanade. Tracking Adversarial Targets. In *International Conference on Machine Learning*, 2014.

[4] Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Model-Free Linear Quadratic Control via Reduction to Expert Prediction. In *AISTATS*, 2019.

[5] Marc Abeille and Alessandro Lazaric. Thompson Sampling for Linear-Quadratic Control Problems. In *AISTATS*, 2017.

[6] Marc Abeille and Alessandro Lazaric. Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems. In *International Conference on Machine Learning*, 2018.

[7] Naman Agarwal, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. Online Control with Adversarial Disturbances. In *International Conference on Machine Learning*, 2019.

[8] Brian D. O. Anderson and John B. Moore. *Optimal Control: Linear Quadratic Methods*. 2007.

[9] James Anderson and Nikolai Matni. Structured State Space Realizations for SLS Distributed Controllers. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[10] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

[11] Karl J. Åström and Björn Wittenmark. *Adaptive Control*. 2013.

[12] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

[13] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

[14] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. 2007.

[15] Dimitri P. Bertsekas. Value and Policy Iterations in Optimal Control and Adaptive Dynamic Programming. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):500–509, 2017.

[16] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. In *Conference on Learning Theory*, 2018.

[17] Patrick Billingsley. *Probability and Measure*. 1995.

[18] Sergio Bittanti and Marco C. Campi. Adaptive Control of Linear Time Invariant Systems: The "Bet on the Best" Principle. *Communications in Information and Systems*, 6(4):299–320, 2006.

[19] Ross Boczar, Nikolai Matni, and Benjamin Recht. Finite-Data Performance Guarantees for the Output-Feedback Control of an Unknown System. In *57th IEEE Conference on Decision and Control*, 2018.

[20] Vladimir I. Bogachev. *Gaussian Measures*. 2015.

[21] Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive Control for Linear and Hybrid Systems*. 2017.

[22] Justin Boyan. Least-Squares Temporal Difference Learning. In *International Conference on Machine Learning*, 1999.

[23] Richard P. Braatz, Peter M. Young, John C. Doyle, and Manfred Morari. Computational Complexity of μ Calculation. *IEEE Transactions on Automatic Control*, 39(5):1000–1002, 1994.

[24] Steven J. Bradtke. *Incremental Dynamic Programming for On-Line Adaptive Optimal Control*. PhD thesis, University of Massachusetts Amherst, 1994.

[25] Steven J. Bradtke and Andrew G. Barto. Linear Least-Squares Algorithms for Temporal Difference Learning. *Machine Learning*, 22(1–3):33–57, 1996.

[26] Steven L. Brunton, Bingni W. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control. *PLOS ONE*, 11(2):1–19, 2016.

[27] Marko Budišić, Ryan M. Mohr, and Igor Mezić. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

[28] Marco C. Campi and Erik Weyer. Finite Sample Properties of System Identification Methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

[29] Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazić, Yishay Mansour, and Kunal Talwar. Online Linear Quadratic Control. In *International Conference on Machine Learning*, 2018.

[30] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret. *arXiv:1902.06223*, 2019.

[31] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator. *arXiv:1710.01688*, 2017.

[32] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator. In *Neural Information Processing Systems*, 2018.

[33] Deepmind. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`, 2019.

[34] Marc P. Deisenroth and Carl E. Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *International Conference on Machine Learning*, 2011.

[35] John C. Doyle, Nikolai Matni, Yuh-Shyang Wang, James Anderson, and Steven Low. System Level Synthesis: A Tutorial. In *IEEE 56th Annual Conference on Decision and Control*, 2017.

[36] Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*. 2007.

[37] Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with Gaussian processes. In *International Conference on Machine Learning*, 2005.

[38] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-Based Adaptive Regulation of Linear-Quadratic Systems. *arXiv:1711.07230*, 2017.

[39] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite Time Identification in Unstable Linear Systems. *Automatica*, 96:342–353, 2018.

[40] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized Policy Iteration with Nonparametric Function Spaces. *Journal of Machine Learning Research*, 17(139):1–66, 2016.

[41] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

[42] Claude-Nicolas Fiechter. PAC Adaptive Control of Linear Systems. In *Conference on Learning Theory*, 1997.

[43] Pascal M. Gahinet, Alan J. Laub, Charles S. Kenney, and Gary A. Hewer. Sensitivity of the Stable Discrete-Time Lyapunov Equation. *IEEE Transactions on Automatic Control*, 35(11): 1209–1217, 1990.

[44] Alexander Goldenshluger. Nonparametric Estimation of Transfer Functions: Rates of Convergence and Adaptation. *IEEE Transactions on Information Theory*, 44(2):644–658, 1998.

[45] Robert C. Grande, Thomas J. Walsh, and Jonathan P. How. Sample Efficient Reinforcement Learning with Gaussian Processes. In *International Conference on Machine Learning*, 2014.

[46] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient Descent Learns Linear Dynamical Systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[47] Elad Hazan and Cyril Zhang. Learning Linear Dynamical Systems via Spectral Filtering. In *Neural Information Processing Systems*, 2017.

[48] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral Filtering for General Linear Dynamical Systems. In *Neural Information Processing Systems*, 2018.

[49] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random Design Analysis of Ridge Regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[50] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems. In *Neural Information Processing Systems*, 2012.

[51] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[52] Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query Complexity of Derivative-Free Optimization. In *Neural Information Processing Systems*, 2012.

[53] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning Provably Efficient? In *Neural Information Processing Systems*, 2018.

[54] Galin L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[55] Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. 2000.

[56] Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[57] Mihail M. Konstantinov, Petko Hr. Petkov, and Nicolai D. Christov. Perturbation analysis of the discrete riccati equation. *Kybernetika*, 29(1):18–29, 1993.

[58] Mihail M. Konstantinov, Da-Wei Gu, Volker Mehrmann, and Petko Hr. Petkov. *Perturbation Theory for Matrix Equations*, volume 9. 2003.

[59] Sanjay Krishnan, Roy Fox, Ion Stoica, and Ken Goldberg. DDCO: Discovery of Deep Continuous Options for Robot Learning from Demonstrations. In *Conference on Robot Learning*, 2017.

[60] Harold Kushner and George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. 2003.

[61] Vitaly Kuznetsov and Mehryar Mohri. Learning Theory and Algorithms for Forecasting Non-Stationary Time Series. In *Neural Information Processing Systems*, 2015.

[62] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[63] Michail G. Lagoudakis and Ronald Parr. Least-Squares Policy Iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

[64] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-Sample Analysis of Least-Squares Policy Iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

[65] Hosoo Lee and Yongdo Lim. Invariant metrics, contractions and nonlinear matrix equations. *Nonlinearity*, 21(4):857–878, 2008.

[66] Sergey Levine and Vladlen Koltun. Learning Complex Neural Network Policies with Trajectory Optimization. In *International Conference on Machine Learning*, 2014.

[67] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[68] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4–5):421–436, 2018.

[69] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

[70] Bo Lincoln and Anders Rantzer. Relaxed Dynamic Programming. *IEEE Transactions on Automatic Control*, 51(8):1249–1260, 2006.

[71] Lennart Ljung. *System Identification: Theory for the User*. 1999.

[72] Jan R. Magnus. The expectation of products of quadratic forms in normal variables: the practice. *Statistica Neerlandica*, 33(3):131–136, 1979.

[73] Dhruv Malik, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, , and Martin J. Wainwright. Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems. In *AISTATS*, 2019.

[74] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. In *Neural Information Processing Systems*, 2018.

[75] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty Equivalent Control of LQR is Efficient. *arXiv:1902.07826*, 2019.

[76] Henry B. Mann and Abraham Wald. On the Statistical Treatment of Linear Stochastic Difference Equations. *Econometrica*, 11(3–4):173–220, 1943.

[77] V. John Mathews and Giovanni L. Sicuranza. *Polynomial Signal Processing*. 2000.

[78] Nikolai Matni, Yuh-Shyang Wang, and James Anderson. Scalable system level synthesis for virtually localizable systems. In *IEEE 56th Annual Conference on Decision and Control*, 2017.

[79] Alexandre Megretski and Anders Rantzer. System Analysis via Integral Quadratic Constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

[80] Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An Analysis of Reinforcement Learning with Function Approximation. In *International Conference on Machine Learning*, 2008.

[81] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[82] Abdelkader Mokkadem. Mixing properties of ARMA processes. *Stochastic Processes and their Applications*, 29(2):309–315, 1988.

[83] B. Molinari. The Stabilizing Solution of the Discrete Algebraic Riccati Equation. *IEEE Transactions on Automatic Control*, 20(3):396–399, 1975.

[84] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[85] OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

[86] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning Dexterous In-Hand Manipulation. *arXiv:1808.00177*, 2018.

[87] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[88] Samet Oymak and Necmiye Ozay. Non-asymptotic Identification of LTI Systems from a Single Trajectory. *arXiv:1806.05722*, 2018.

[89] Andrew Packard and John C. Doyle. The Complex Structured Singular Value. *Automatica*, 29(1):71–109, 1993.

[90] Alain Pajor. Metric Entropy of the Grassmann Manifold. *Convex Geometric Analysis*, 34: 181–188, 1998.

[91] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile Off-Road Autonomous Driving Using End-to-End Deep Imitation Learning. In *Robotics: Science and Systems*, 2018.

[92] Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008.

[93] Li Qiu, Bo Bernhardsson, Anders Rantzer, Edward J. Davison, Peter M. Young, and John C. Doyle. A Formula for Computation of the Real Stability Radius. *Automatica*, 31(6):879–890, 1995.

[94] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *International Conference on Machine Learning*, 2012.

[95] Anders Rantzer. Concentration Bounds for Single Parameter Adaptive Control. In *American Control Conference*, 2018.

[96] Carl E. Rasmussen and Malte Kuss. Gaussian Processes in Reinforcement Learning. In *Neural Information Processing Systems*, 2004.

[97] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[98] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, 2019.

[99] Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. Finite-Time System Identification for Partially Observed LTI Systems of Unknown Order. *arXiv:1902.01848*, 2019.

[100] Kathrin Schäcke. On the Kronecker Product. Master's thesis, University of Waterloo, 2004.

[101] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016.

[102] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Neural Information Processes Systems*, 2018.

[103] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[104] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

[105] Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. In *Conference on Learning Theory*, 2018.

[106] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning Linear Dynamical Systems with Semi-Parametric Least Squares. In *Conference on Learning Theory*, 2019.

[107] James C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. 2003.

[108] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *International Conference on Machine Learning*, 2010.

[109] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC Model-Free Reinforcement Learning. In *International Conference on Machine Learning*, 2006.

[110] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-Based Reinforcement Learning in Contextual Decision Processes. In *Conference on Learning Theory*, 2019.

[111] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-Real: Learning Agile Locomotion For Quadruped Robots. In *Robotics: Science and Systems*, 2018.

[112] Yuval Tassa, Nicolas Mansard, and Emo Todorov. Control-Limited Differential Dynamic Programming. In *International Conference on Robotics and Automation*, 2014.

[113] Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[114] Anastasios Tsiamis and George J. Pappas. Finite Sample Analysis of Stochastic System Identification. *arXiv:1903.09122*, 2019.

[115] John N. Tsitsiklis and Benjamin Van Roy. Feature-Based Methods for Large Scale Dynamic Programming. *Machine Learning*, 22(1):59–94, 1996.

[116] John N. Tsitsiklis and Benjamin Van Roy. On Average Versus Discounted Reward Temporal-Difference Learning. *Machine Learning*, 49(2):179–191, 2002.

[117] Stephen Tu and Benjamin Recht. Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

[118] Stephen Tu and Benjamin Recht. The Gap Between Model-Based and Model-Free Methods on the Linear Quadratic Regulator: An Asymptotic Viewpoint. In *Conference on Learning Theory*, 2019.

[119] Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-Asymptotic Analysis of Robust Control from Coarse-Grained Identification. *arXiv:1707.04791*, 2017.

[120] Anirudh Vemula, Wen Sun, and J. Andrew Bagnell. Contrasting Exploration in Parameter and Action Space: A Zeroth-Order Optimization Perspective. In *AISTATS*, 2019.

[121] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

[122] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. 2018.

[123] Mathukumalli Vidyasagar and Rajeeva L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3–4):421–430, 2008.

[124] Yuh-Shyang Wang, Nikolai Matni, and John C. Doyle. A System Level Approach to Controller Synthesis. *arXiv:1610.04815*, 2016.

[125] Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M. Rehg, Byron Boots, and Evangelos A. Theodorou. Information Theoretic MPC for Model-Based Reinforcement Learning. In *International Conference on Robotics and Automation*, 2017.

[126] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–246, 1992.

[127] Bin Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[128] Fuzhen Zhang. *The Schur Complement and its Applications*, volume 4 of *Numerical Methods and Algorithms*. 2005.

[129] Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. 1995.

[130] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-Sample Analysis for SARSA and Q-Learning with Linear Function Approximation. *arXiv:1902.02234*, 2019.