Statistical Complexity and Regret in Linear Control

by

Max Simchowitz

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Co-chair
Professor Benjamin Recht, Co-chair
Professor Joshua Blumenstock

Spring 2021

Abstract

Statistical Complexity and Regret in Linear Control

by

Max Simchowitz

Doctor of Philosophy in Engineering- Electrical Engineering and Computer Science

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor Benjamin Recht, Co-chair

The field of linear control has seen broad application in fields as diverse as robotics, aviation, and power-grid maintenance. Accordingly, decades of research have been dedicated to learning to control systems when the dynamics are not known a priori, but must be inferred from data. However, much of the classical work considered asymptotic regimes, which shed little light on precisely how much data are required to attain desired control performance in the presence of statistical noise.

In this thesis, we apply modern advances in statistical and online learning theory to enrich our understanding of the statistical complexity of data-driven control. In the first half of the thesis, we study estimation of system parameters, known as system identification. We show that the stability of the system - long thought to characterize the sample complexity of estimation - is often inversely proportional to the number of trajectories required to recovery the system parameters to specified tolerance.

We then turn to adaptive control: the problem in which a learning agent interacts with an unknown dynamical system in hopes of attaining control performance comparable to if she had foreknowledge of the system dynamics. Adopting the conventions of modern online learning theory, we study the problem from the perspective of regret, or cumulative performance relative to an idealized benchmark. Building on the system identification results, we present upper and lower bounds that demonstrate that a remarkably simple strategy enjoys optimal regret for the adaptive control of the linear quadratic regulator. We then propose a unified framework - based on online convex optimization - which enjoys similar regret guarantees in the considerably more general setting of nonstochastic control, accommodating partial observation, time-varying costs, and adversarially-chosen disturbances. Surprisingly, we find that the best-attainable regret (with respect to a suitable benchmark of stabilizing

LTI control policies) exhibits the same scaling in the problem horizon as the optimal regret for the online LQR problem.

To Mom, Dad, and Natasha

Contents

Acknowledgments

There are so many without whom this thesis would have never materialized, and I am deeply grateful for their mentorship, advice, guidance, patience and friendship.

The subject matter of this thesis, and of much of the prior work which informed it, is due to my advisor, Benjamin Recht, and his vision for uniting learning and control. Noting my tendency to get lost in the weeds, he pushed me towards research which said something new, which (at least attemped to) solved a "real problem", which did not let technicality serve as a substitute for taste. He was exacting. But I could not have any of this without him.

I am equally grateful to my co-advisor, Michael I. Jordan, who bought me a decaf espresso on admit weekend, sat me down, and invited me to join his group, SAIL, should I accept Berkeley's offer. My SAIL colleagues have done more than I could have imagined to broaden my horizons as a researcher, rivalled only by Mike's capacious wisdom. I also have to thank my undergraduate advisors, David Blei (Mike's former student at Berkeley) and Sanjeev Arora (another Berkeley alumn), for their incredible mentorship, and for their encouragement to come to Berkeley.

Kevin Jamieson, now faculty at University of Washington, was like a third advisor to me. During his post-doc in Ben's group, he helped me formalize my first research problems, work through the math, and see projects to their end. Working with him was my first introduction to active learning - the paradigm that data is not simply given but sought after. This perspective has informed every single one of my research projects since.

Towards the end of my program, I had the great fortune of spending a summer with Elad Hazan and his group at Google Princeton. Somewhere between a caring PI and old-world uncle, he instilled in me optimism and excitement for research at a time I needed the upliftment most. It was in his lab that I began to work on *non-stochastic control* - a paradigm that he, along with Sham Kakade, Naman Agrawal, Brian Bullins, and Karan Singh, had developed - which consitutes one of the major problem settings considered in this thesis.

After my time with Elad, I was fortunate to intern at Microsoft Research in New York under Akshay Krishnamurthy, then visit Sasha Rakhlin's group at MIT, and finally, intern again at MSR-NY, this time under the (remote) supervision of Alex Slivkins. Though the research undertaken during these visits is not included in this thesis, my writing, thinking, and perspective have grown immeasurably from the time spent with my hosts. Last but not least, a postdoc with my soon-to-be supervisor Russ Tedrake has been a light and the end of the tunnel throughout the writing of this dissertation. Russ also brought to my attention that the word "optimal" has quite a different meaning in the control community than it does in the learning community; clarying this distinction was invaluable in both the writing of this thesis and in presentations of its constituent work.

I have had so many incredible collaborators over the years and I am so happy to have worked with Horia Mania, Stephen Tu, Ross Bozcar, Dylan Foster, and Karan Singh on the material on this thesis. To Horia I owe a special gratitude. When I was freshman bumbling through Princeton's math curriculum, after having taken a year off and with very little

background in proof-writing, Horia patiently helped me through my the analysis lectures in our dining hall. And, years later, it was Horia's generosity in letting me on board what would ultimately become "Learning Without Mixing" that brought me into Ben's learning and control program.

Each paper included in thesis omits an uncredited collaborator - Ben's Modest Yachts Slack channel. In addition to my colleagues listed above, Eric Sparks, Jason Lee, Sarah Dean, Laurent Lessard, Mahdi Soltanolkotabi, and Vaishaal Shankar were always an immense help. I'd especially like to thank Eric Jonas for always being there when I needed to think through my future, Nikolai Matni for helping me navigate decades and decades of control literature, John 'Coach' Miller for commiserating with me about the vicissitudes of academic life, and Esther Rolf for being a brilliant collaborator, a great friend, and for teaching me to (at least begin to) situate my research in the real world.

Outside the program, I've had so many dear friends who made my time at Berkeley unforgettable. Surviving a PhD would have been impossible without the long calls with Jasper Ryckman about "eyy guys", 2am Peggy Gu tracks with Allan Jabri at Arch Street, deep-cut memes from Gavin Cook, nonsensical internet-speak texts from Nikhilesh Sigatapu, weekend debriefs with Lekha Khanchinadam and Divya Farias, dim-sum with Joe Girton, nights out at Starline and lazy Sundays at 2727 with Laila Riazi, Thomas Krasnowiecki, E.J. Rosen, Pawel Koscielny, Julian Self, and Lauren Lubell. And the friends not in the Bay: I wouldn't have visited Google Princeton and MSR-NY had I not dearly missed Conor McGrory, Margaret Spencer, Godfrey Furchgott, David Sahar, and Byrd Pinkerton. And I would not have survived February's conference deadline season (for which the writing begins sometime around the winter holidays) were it not for kickbacks at Mia's with Eli Petzold and Gaby Leslie, and the highlight of every year - New Years Eve - with Minnie Schedeen, Graham Parkes, and Chiara Towne.

Finally, my family. I am forever grateful for my parents' unwavering encouragement, to my cousin Maya for always being by phone when I needed to talk, to Auntie Chantie who always had a place for me to stay in the city, "Aunt" Claudia, my second mom, and to Natasha — the coolest sister and best role model a big brother could ask for.

Chapter 1

Introduction

In a control task, one manipulates a dynamical system so as to achieve a certain end. This specification, though broad, involves three components. First, a *dynamical* system, that is, a process in which a system state evolves over time. Second, a mechanism of manipulating the system, specified by a sequence of *control inputs*. And finally, the goal of the task - the desired behavior that the inputs are selected to elicit. Control problems are ubiquitous in science and engineering. Their scope encompasses aeronautics, heating and ventillation, robotics, power distribution, medical device engineering, and popular contemporary culinary methods.

When mathematical laws describing the evolution of system dynamics are known a priori, and when task performance can be summarized by a scalar objective cost, it is purely a computational matter to derive the optimal control *law*. This law describes how a current input should be selected given all available system observations from the past, so as to minimize cost.

In many, if not, most cases of interest, there is an aspect of the dynamics which is not known in advance. For example, the system may be driven by *disturbances*, or exogenous perturbations not determined by the control input. This uncertainty can be rather benign: when a probabilistic model for these disturbances exists, one can still derive (probabilistically) optimal control laws.

Many more challenging sources of uncertainty may exists. For example, a model for the disturbances may not be known, or worse, the disturbances may be selected by a malicious adversary whose sole aim is to compromise performance. In other cases, the scalar objective function which summarizes control performance may not be known in advance: for example, if an aerial drone attempts to track the position of a moving target, it is only after observing the target's position that we know where the drone should go (or should have gone). Finally, the dynamical laws governing the time-evolution of the system may be known only coarsely, or not at all.

Learning, or modeling from past-collected data, is a powerful tool to address unknown and uncertain aspects of control problems. Learning has been most widely applied to estimation of the dynamical laws of the system, though in principle it can also be used to model

disturbances and costs.

The field of *system identification* studies the problem of learning these laws. The field of *adaptive control* studies how to bootstrap system identification into control policies which use learning to make their choice of control inputs. At the risk of upsetting one or both communities, the field called *adaptive control* by control theorists is designated *reinforcement learning* by statisticians and computer scientists. Here, the emphasis is placed on the learning, more so than the control.

This thesis: the complexity of learning and control

The purpose of this thesis is to understand the *complexity* of learning for control problems. Complexity is a broad term, but refers roughly to the question "how much data must be collected in order to achieve a certain level of control performance." Formal definitions are given later in the chapter.

This thesis takes incremental steps towards understanding the sample complexity of control in a range of linear control settings, including both system identification and adaptive control. For these tasks, we present algorithms which achieve desired performance using as little data as possible, and in some cases, can mathematically verify that no procedure can use substantially less. These contributions demonstrate the power of remarkably *simple* algorithms, built upon commonplace learning and optimization primitives like ordinary least squares, gradient descent, and Newton's method.

As part of a broader research program, characterizing the sample complexity of control tasks promises far reaching implications, since accurate error bounds are indispensable for designing robust and high-performing control systems. It would also illuminate which properties of control environments enable efficient learning, which make learning tractable but challenging, and which thwart the prospect of learning altogether.

More broadly, a deeper understanding of the sample complexity of control would inform other statistical problems where one wishes to learn from non-i.i.d. or time-correlated data. Indeed, dynamics makes learning for control more challenging than traditional statistical estimation. This is because observations and inputs at a given time may be correlated with, or exert direct effects upon, observations in the future. These correlations preclude the conventional statistical and learning-theoretic arsenals developed for identically and independently distributed data.

Thesis roadap

This thesis consists of five of the author's publications on linear system identification and adaptive contron.

We begin in Section 1.1 below, by formally defining linear time invariant systems, the collection of control systems to which our results apply.

Section 1.2 of this introduction summarizes the our findings for linear system identification, which are presented in Part I, Chapters 3 and 4 of this thesis. Chapter 3 studies

system identification under full observation. It demonstrates that the "mixing" property, long thought to be critical to system identification, is not only unnecessary, but at times inversely related to the sample complexity. Chapter 4 extends the analysis to partially observed settings. Chapter 3 and Chapter 4 are based on the publications Simchowitz et al. [2018] and Simchowitz et al. [2019], respectively.

Section 1.3 outlines our contributions to online control, a learning theoretic variant of adaptive control where performance is measured in terms of "regret". The contributions in this section are presented in full in Part II, Chapters 5 to 7. Chapter 5 characterizes the optimal regret rates for the broadly studied online LQR problem [Abbasi-Yadkori and Szepesvári, 2011]. Chapter 6 presents low-regret algorithms for the considerably more general "nonstochastic control" formulation, first studied in Agarwal et al. [2019a]. Finally, Chapter 7 shows that, under additional assumptions and with a carefully designed algorithm, regret guarantees in the more challenging nonstochastic control setting can be just as strong those attainable in the simpler online LQR game. These findings reveal that the difficulty of adaptive control, when measured in regret, is determined primarily by *knowledge of system dynamics*. Chapters 5 to 7 present Simchowitz and Foster [2020], Simchowitz et al. [2020], and Simchowitz [2020], respectively.

To emphasize connections between the different settings under study, we centralize our discussion of related work in Chapter 2. The chapter does its best contextualizes the contributions of this thesis within the vast and still-rapidly growing body of scholar work at the intersection of learning theory and control. Chapter 2 further exposes how the technical ideas in this thesis both draw upon and contribute to the broader learning theory literature.

Finally, Chapter 8 provides concluding remarks, with an eye towards nonlinear control.

1.1 LTI Systems and Mathematical Notation

This thesis studies discrete time, linear time invariant, or *LTI*, dynamical systems. These systems admit efficient controller design, are statistically tractable, are frequently used as first-order approximations to more general, non-linear dynamics.

An LTI system is governed by a state $\mathbf{x}_t \in \mathbb{R}^{d_x}$, which evolves linearly in a *control input* $\mathbf{u}_t \in \mathbb{R}^{d_u}$ and *process noise* variable $\mathbf{w}_t \in \mathbb{R}^{d_x}$, via

$$\mathbf{x}_{t+1} = A_\star \mathbf{x}_t + B_\star \mathbf{u}_t + \mathbf{w}_t, \tag{1.1}$$

where $A_\star \in \mathbb{R}^{d_x \times d_x}$ and $B_\star \in \mathbb{R}^{d_x \times d_u}$ are fixed matricies. For simplicity, we will assume throughout that the system is initialized at the origin, that is $\mathbf{x}_1 \equiv 0$. When the learning agent may observe the state $\mathbf{x}_t$ directly, we call the system *fully observed*.

In many scenarios, the learning agent does not have access to the state $\mathbf{x}_t$, but instead observes a noisy, linear observation $\mathbf{y}_t \in \mathbb{R}^{d_y}$

$$\mathbf{y}_t = C_\star \mathbf{x}_t + \mathbf{e}_t, \tag{1.2}$$

where $\mathbf{e}_t \in \mathbb{R}^{d_y}$ is called the *sensor noise*, and $C_\star \in \mathbb{R}^{d_y \times d_x}$ is another fixed matrix. We call this setting *partially observed* although, it encompasses the fully observed setting when $d_x = d_y$, $C_\star = I$, and $\mathbf{e}_t \equiv 0$.

We posit a learning agent or *learner* who can select the $\mathbf{u}_t$ at will, as an arbitrary function of internal randomness, past inputs, and either past states (full observation) or past outputs (partial observation). We frequently refer to the noise terms $\mathbf{e}_t$ and $\mathbf{w}_t$ as *disturbances*, which are not under the learner's control.

Notational Conventions

We let $\mathbb{R}, \mathbb{C}, \mathbb{Z}, \mathbb{N}$ denote the reals, complex numbers, integers and positive integers, respectively. $\mathbb{R}^d$ is the space of d-dimensional real vectors, $\mathbb{R}^{d_1 \times d_2}$ the space of $d_1 \times d_2$ dimensional real matrices, $\mathbb{S}^d$ the space of real positive symmetric matrices $X \in \mathbb{R}^{d \times d}$, $\mathbb{S}^d_+$ the space of real positive semidefinite matrices, and $\mathbb{S}^d_{++}$ the space of real strictly positive definite matrices. We let $X \succeq Y$ if $X \in \mathbb{S}^d$ dominates $Y \in \mathbb{S}^d$ in the standard (Lowner) PSD ordering. Given a vector $v \in \mathbb{R}^d$, $\|v\|$ denotes the Euclidean norm unless otherwise noted. For matrices, $\|X\|_{\mathrm{op}}$ and $\|X\|_{\mathrm{F}}$ denote operator and Frobenius norms, respectively.

$\mathcal{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\| = 1\}$ denotes the sphere, and for a compact and convex set $\mathcal{C} \subset \mathbb{R}^d$, $\mathrm{Proj}_{\mathcal{C}}(v) = \arg\min_{v' \in \mathcal{C}} \|v - v'\|$ denotes Euclidean projection. Given a sequence $v_1, v_2, \ldots$, we let (v_t) denote the (possibly infinite) sequence, and for indices $s \le t$, we let $v_{s:t} = (v_s, v_{s+1}, \ldots, v_t)$ and $v_{t:s} = (v_t, v_{t-s}, \ldots, v_s)$ denote subsequences.

We write $f(n) \lesssim g(n)$ if $f \le Cg$ for a universal constant C independent of n; $\gtrsim$ is used similarly. We use Big-Oh assymptotic notation throughout only for informal statements: $f(n) = \mathcal{O}(n)$ if $f(n) \le Cg(n)$, where C potentially supresses dependence on certain problem parameters. We use Oh-Tilde notation $\tilde{\mathcal{O}}(n)$ to indicate supression of logarithmic factors. Informally, we shall also write $X \gtrsim Y$ if there is a constant $c > 0$ such that PSD matrices X, Y satisfy $X \succeq cY$.

The notation $a := b$ means that quantity a is taken to be equal to quantity b by defintion. Given a probability distribution $\mathcal{D}$, we write $X_1, X_2, \ldots \overset{\text{i.i.d}}{\sim} \mathcal{D}$ if $X_1, X_2, \ldots$ for random variables that are drawn independently and indentically (i.i.d.) from a distribution $\mathcal{D}$. Given a set $\mathcal{X}$, we write $X_1, X_2, \ldots \overset{\text{unif}}{\sim} \mathcal{X}$ to denote that the draws are drawn i.i.d. from a cannonical uniform distribution on $\mathcal{X}$. For example, $X \overset{\text{unif}}{\sim} \{-1, 1\}^n$ denotes uniform sampling from the hypercube.

1.2 System identification

Part I of this thesis considers the problem of linear *system identification*: estimation of a model of system dynamics from observations of trajectories generated from those dynamics. Our aim is to understand the question of *sample complexity*:

How much data is required to construct an accurate model of the system dynamics?

We focus on estimation of a system from a single trajectory of inputs via ordinary least squares. We show that, for both fully and partially observed systems, the *stability* property long thought to characterize the sample complexity of learning may in fact have no bearing on it. And, in the case of fully observed systems, stability and sample complexity may exhibit an inverse relationship.

Stability We ground our analysis in a discussion of system *stability*. There are many definitions of stability, most of which are equivalent for linear dynamical systems. The spectral radius of a square matrix $A \in \mathbb{R}^{d\times d}$ is notated as $\rho(A)$, denotes the largest magnitude (possibly complex) eigenvector of A. $\rho(A)$ also admits a definition in terms of the limiting powers of A, via

$$\rho(A) := \lim_{n\to\infty} \|A^n\|_{\text{op}}^{1/n}.$$

We say that a dynamical system with state transition matrix $A_\star$ is *stable* if $\rho(A_\star) < 1$, marginally stable if $\rho(A_\star) = 1$ (that is, $A_\star$ has admits an eigenvalue of unit magnitude), and *unstable* if $\rho(A_\star) < 1$. Stated colloquially, stable systems decay, unstable systems explode, and marginally stable systems remain bounded above and below. For stable systems, the following informal approximation is a good rule of thumb:

$$\|A^n\|_{\text{op}} \approx \rho(A)^n$$

Stability, Mixing, and Noise Accumulation The most well-established technique in the statistics literature for dealing with non-independent, time-series data is the use of *mixing-time* arguments. Informally, the mixing time of a discrete-time dynamical process $\mathbf{z}_1, \mathbf{z}_2, \ldots$ refers to the smallest natural number $h \in \mathbb{N}$ such that $\mathbf{z}_t$ and $\mathbf{z}_{t+h}$ are "almost statistically independent" for all times t. A formal definition can be found in Yu [1994].

Estimation in systems with small mixing times h has long been regarded to be easy, because each subsequence $(\mathbf{z}_k, \mathbf{z}_{k+h}, \mathbf{z}_{k+2h})$ contains (nearly) statistical independent data, and comprise an $1/h$-fraction of the entire sequence. Thus, traditional estimation rates in mixing systems coincide with those for estimation rates in i.i.d. systems, up to a factor proporitional to the mixing time [Mohri and Rostamizadeh, 2007a,b, Kuznetsov and Mohri, 2017, McDonald et al., 2017].

For stable systems ($\rho(A_\star) < 1$), mixing time h is roughly related to the spectral radius of the dynamical system $\rho(A_\star)$ via the relationship $h \approx \frac{1}{1-\rho(A_\star)}$. To see this, we unfold the linear dynamics:

$$\mathbf{x}_{t+1} = \sum_{n=0}^{t-1} A_\star^n (B_\star \mathbf{u}_t + \mathbf{w}_t).$$

The definition of spectral definition ensures that $\|A_\star^n\| \approx \rho(A_\star)^n$, hence all $n \gg \frac{1}{1-\rho(A_\star)}$, we can neglect the effect of inputs $\mathbf{u}_{t-n}$ and disturbances $\mathbf{w}_{t-n}$ on the current state $\mathbf{x}_{t+1}$. In particular, as the spectral radius $\rho(A_\star)$ approaches 1 (from below), the mixing time diverges. Therefore, a fundamental limitation of mixing-time arguments is that the bounds all degrade as the mixing-time increases. This has two implications for linear system identification: (a) none of prior work correctly captured the qualitative behavior as the $A_\star$ matrix reaches instability, and (b) these techniques cannot be applied to the regime where $A_\star$ is unstable, for which estimation is not only well-posed, but should be quite easy.

The same computation reveals that nearly-unstable systems have yet another undesirable property. Because they have long-run dependencies on past noise terms, they *accumulate noise*, driven the states to be larger and larger in magnitude. To see this, consider the scalar system driven by Gaussian noise with zero input: $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. Then, $\mathbf{x}_t \sim \mathcal{N}(0, t-1)$, so the variance of the system state grows linear in t.

Learning without stability in Fully Observed Systems

In Chapter 3, we study the estimation of the transition matrix $A_\star$ via the ordinary least squares estimator

$$\widehat{A}_{\text{ls}} = \inf_{A \in \mathbb{R}^{d_x \times d_x}} \sum_{t=1}^{N} \|\mathbf{x}_{t+1} - A\mathbf{x}_t\|^2$$

trained to a single trajectory of a fully observed linear dynamical system, with zero input and i.i.d. Gaussian noise $\mathbf{w}_t$. The sample complexity is determined by the decay of the operator-norm error $\|\widehat{A}_{\text{ls}} - A_\star\|_{\text{op}}$ as a function of the sample size N. The work in this chapter is based on Simchowitz et al. [2018].

Intuition based on mixing-time arguments would suggest that the sample complexity of this problem worsens as $\rho(A_\star)$ approaches unity, because the mixing time of the resulting dynamical problem $(\mathbf{x}_1, \mathbf{x}_2, \dots)$ diverges. Chapter 3 shows that this is demonstrable not the case. We summarize our findings in the following informal theorem.

Informal Theorem 1. *For all systems with $\rho(A_\star) \le 1$, the error $\|\widehat{A}_{\text{ls}} - A_\star\|_{\text{op}}$ decays as $1/\sqrt{N}$ with high probability, suppressing logarithmic factors. Notably, this is true even if $\rho(A_\star) = 1$, and thus the system* does not mix.

Moreover, for many systems of interest, the mixing time improves *as $\rho(A_\star)$ approaches* 1. *For example, if $A_\star = \rho I$, $\rho \in (0, 1)$, the error degrades as* $\max\{1/N, \sqrt{(1-\rho)/N}\}$, *which is* decreasing *in ρ.*

In other words, it is possible that *less data* is needed to achieve a given error tolerance ϵ as $\rho(A_\star)$ approaches unity from below.

The key insight is that sample complexity is driven not by mixing, but by the relative magnitude of the states $\mathbf{x}_t$, which serve as the regression covariates, to the disturbances $\mathbf{w}_t$,

which serve as the regression noise. The ratio of these two magnitudes informally corresponds to the *signal-to-noise ratio* (SNR), with larger SNRs corrsponding to small estimation errors.

As the spectral radius approaches 1, the accumulation of noise – often a hindrance in control tasks – plays to our advantage by increasing the SNR, thereby aiding learning. Making this argument precise requires two technical ingredients – a martingale concentration inequality due to Abbasi-Yadkori et al. [2011], popular in the learning community but perhaps less commonplace amongst control theorists, and an original martingale *anti-concentration* inequality, inspired by the small-ball method of Mendelson [2014].

Learning without Stability in Partially Observed Systems

Our previous result pertained to a rather simple setting: full state observation, no control input, and i.i.d. Gaussian noise. In Chapter 4, we unpack the role of stability in a potentially more challenging setting. The work in this chapter is based on Simchowitz et al. [2019].

We consider a partially observered LTI system

$$\begin{aligned} \mathbf{x}_{t+1} &= A_\star \mathbf{x}_t + B_\star \mathbf{u}_t + \mathbf{w}_t, \\ \mathbf{y}_t &= C_\star \mathbf{x}_t + D_\star \mathbf{u}_t + \mathbf{e}_t, \end{aligned}$$

where the disturbances $\mathbf{w}_t$ and $\mathbf{e}_t$ may be chosen to be any arbitrary fixed sequence, unknown to the learner, but independent of the learners choice of inputs. For a given parameter length $p \in \mathbb{N}$, our goal is to recovery the system *Markov operator*

$$G_{\star;p} = [D_\star \mid C_\star B_\star \mid C_\star A_\star B_\star \mid \cdots \mid C_\star A_\star^{p-2} B_\star];$$

The Markov operator represents the linear response form inputs to outputs in the absence of disturbance: $\mathbf{y}_p = G_{\star;p}\mathbf{u}_{p;1}$, where $\mathbf{u}_{p:1}$ denotes the vector obtained by concatenating $(\mathbf{u}_p, \mathbf{u}_{p-1}, \ldots, \mathbf{u}_1)$. Recovery of the Markov parameter is also sufficient for recovery of the dynamical matrices $(A_\star, B_\star, C_\star, D_\star)$, up to an (unidentifiable) similarity transformation [Oymak and Ozay, 2019].

We identify an *semi-parametric* relationship

$$\mathbf{y}_t = G_{\star;p}\mathbf{u}_{t:t-p+1} + \boldsymbol{\delta}_t,$$

where $\boldsymbol{\delta}_t$ depends on inputs $\mathbf{u}_s$ for $s \le t-p$, as well as disturbances ($\mathbf{w}_s$ and $\mathbf{e}_s$ for $1 \le s \le t$. Importantly, in $\mathbf{u}_t$ are drawn i.i.d. from any mean-zero distribution, then $\mathbb{E}[\mathbf{u}_{t:t-p+1}\boldsymbol{\delta}_t^\top]$ is identically zero. Importantly, this relationship holds *even if* the noise process is biased.

We leverage this observation to analyzes the least squares estimator

$$\widehat{G} \in \arg\min_G \sum_{t=p+1}^{N} \|\mathbf{y}_t - G_\star \mathbf{u}_{t;t-p+1}\|^2,$$

where $\mathbf{u}_t$ are i.i.d. Rademacher vector (uniform $\{-1,1\}^{d_u}$) random variables. Our analysis shows that

Informal Theorem 2. *Due to the semi-parametric relationship,* $\|\widehat{G} - G_{\star;p}\|_{\mathrm{op}}$ *scales as* $1/\sqrt{N}$ *with high probability, provided the error terms* $\|\boldsymbol{\delta}_t\|$ *are uniformly bounded.*

Unfortunately, any uniform bound on $\|\boldsymbol{\delta}_t\|$ requires system stability, degrading as $\rho(A_\star) \to 1$ and failing to hold for marginally stable systems with $\rho(A_\star) = 1$. We remedy this my proposing *prefiltered least squares*, a novel two-scale least squares algorithm which ameliorates the effect of large noise.

Informal Theorem 3. *Via the two-scale least squares algorithm, we obtain an estimator* $\widehat{G}$ *such that* $\|\widehat{G} - G_{\star;p}\|_{\mathrm{op}}$ *scales as* $1/\sqrt{N}$ *with high probability, for all systems with* $\rho(A_\star) \leq 1$.

The analysis relies on a term we dub the "oracle error", which measures how accurate the errors $\boldsymbol{\delta}_t$ can be predicted from past observations $\mathbf{y}_s, s \leq t - p$. We expose how the oracle error can be bounded both in terms of the minimal polynomial of the matrix $A_\star$, and in terms of solution to a Kalman filtering problem, both of which guarantee $1/\sqrt{N}$ error for marginally stable systems. This oracle error therefore provides a novel measure of problem difficulty, orthogonal to stability, for system identification in partially observed systems with potentially adversarial disturbances.

1.3 Online Control

In Part II, we transition from system identification to study the problem of online control. Online control is a formultion of the adaptive control problem in which the agent must control attain near-optimal performance single dynamical trajectory relative to a benchmark of desirable control policies, chosen with full system knowledge. The difference between the algorithm performance and optimal benchmark controller cost is termed *regret*. Regret therefore measures the gap in performance inccured as a consequence of having to rely on learning, rather than a prior knowledge of the system dynamics (or potentially, costs and disturbances).

The regret is realized over a time horizon of length T, and the aim is to achieve regret which grows as slowly as possible in T; *optimal regret* refers to regret bound which which exhibits the slowest possible rate of growth. Any regret which grows sublinearly in T is considered non-vacuous.

Our main contributions are as follows.

- We characterize the optimal regret rate in the broadly-studied online LQR problem (Chapter 5). This problem studies adaptive control of an unknown, fully observed dynamical system, with fixed quadratic costs and independent, Gaussian disturbances.
- We demonstrate that sublinear (in T) regret is attainable in a vastly more general "non-stochastic setting." (Chapter 6) This setting accomodates partially observed systems, changing convex costs, and possibly adversarially selected disturbances. Regret upper

bounds are establish both when the learner has foreknowledge of system dynamics, and when she does not.

- We show that that, with appropriate regularity condition including strong convexity of the control costs (Eq. (1.3)), the optimal regret guarantees in the nonstochastic setting general setting exhibit the same scaling in T as in the far more restrictive LQR setting. (Chapter 7)

Our findings are summarized in Table 1.1, with $\tilde{\mathcal{O}}(\cdot)$ as informal asymptotic notation suppressing logarithmic factors. Taken together, they imply that the optimal regret in online control is primilarily driven by whether or not the learner has foreknowledge of system dynamics.

Regret Rate		
Setting	Known	Unknown
Online LQR (Chapter 5)	~ 0 (definition)	$\Omega(\sqrt{T})$ (Theorem 5.1)
Nonstochastic Control *Strongly Convex Cost* (Chapter 7)	$\text{poly}(\log T)$ (Theorem 7.1)	$\tilde{\mathcal{O}}(\sqrt{T})$ (Theorem 7.2)
Nonstochastic Control *General Convex Loss* (Chapter 6)	$\tilde{\mathcal{O}}(\sqrt{T})$ (Theorem 6.2)	$\tilde{\mathcal{O}}(T^{2/3})$ (Theorem 6.3)

Table 1.1: A sumary of regret rates in various settings of interest. Note that the fast rates under the "*Strongly Convex Cost*" row require certain caveats detailed in Remarks 1.1 and 1.2.

Online LQR

Chapter 5 studies the adaptive control of the linear quadratic regulator, or Online LQR problem. Its findings are based on Simchowitz and Foster [2020].

LQR takes as its point of departure the clasical LQR problem characterized by Kalman [1960], which we sketch as follows. The LQR dynamics evolves according to Eq. (1.1), with i.i.d. Gaussian noise $\mathbf{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$. The learner is tasked with finding a control policy so

as to minimize a running quadratic cost in states and inputs,

$$J_T := \sum_{t=1}^{T} \mathbf{x}_t^\top Q \mathbf{x}_t + \mathbf{u}_t^\top R \mathbf{u}_t,$$

where Q and R are fixed, positive definite matrices. Up to probabilistic fluctuations and transient factors, the optimal LQR policy is to select $\mathbf{x}_t = K_\star \mathbf{u}_t$, where $K_\star$ is a fixed matrix depending on $A_\star, B_\star, Q, R$.[1] This matrix can be computed efficiently by solving the discrete algebraic Riccati equation or DARE (see e.g. [Zhou et al., 1996]). T

In online LQR, the goal is to do nearly as well as $K_\star$, but without knowledge of the dynamical matrices $A_\star$ and $B_\star$. The performance of an adaptive algorithm alg measured in regret. Informally, the

$$\text{LQR-Reg}_T := J_T(\mathsf{alg}) - T \cdot J_\infty(K_\star),$$

where $J_T(\mathsf{alg})$ is the total cost incurred by running an adaptive algorithm alg, and $J_\infty(K_\star) = \lim_{T\to\infty} \frac{1}{T} J_T(K_\star)$ is the infinite horizon limit of the cost that *would have* been incurred had the learner selected the optimal control policy $K_\star$.

Regret measures the relative degradation in performance that results from lack of system knowledge. The goal is to have the regret scale sublinearly in T, so that the regret is second order to the running total cost. The smaller the regret, the less the learner is penalized for not having foreknowledge of the dynamics.

Numerous prior had studied this regret problem, aiming to characterize what the optimal scaling in T. The state of the art had demonstrate that $\sqrt{T}$ regret was attainable with computationally efficient algorithms [Cohen et al., 2019], and under an additional condition called controllability [Zhou et al., 1996, Chapter 3], Mania et al. [2019] demonstrated that $\sqrt{T}$ regret could be attained with only a modest dependence on problem dependence, and with a remarkably simple strategy. This strategy is called *certainty equivalence*, and consists of

- $N \ll T$ steps of injecting exploratory random noises
- Constructing estimates of $(\hat{A}, \hat{B})$ of $(A_\star, B_\star)$ via ordinary least squares
- Selecting inputs $\mathbf{u}_t = \hat{K}\mathbf{x}_t$ for times $t > N + 1$, where $\hat{K}$ is the *certainty equivalent controller*; that is, the controller synthesized by selecting the optimal policy as if $(\hat{A}, \hat{B})$ were the ground truth.

Given that this remarkable simple strategy had enjoyed near state-of-the-art performance (with the caveat of requiring controllability), it is natural to ask if one could do better? Can a more sophisticated titration of learning and control attain lower regret?

[1] This requires the pair $(A_\star, B_\star)$ to be stabilizable; a necessary technical assumption that we make throughout our discussion of the online control problem.

The main finding of Chapter 5 is that this is not the case. First, we present a lower bound which demonstrates that no algorithm can beat $\sqrt{T}$ regret, unless it faces a highly degenerate instance:

Informal Theorem 4. *For every sufficiently non-degerate problem instance $(A_\star, B_\star)$ (and sufficently non-degenerate costs Q, R), the regret of any adaptive control algorithm must be as large as $\sqrt{T d_x d_u^2}$ in expectation, where again T is the problem horizon, d_x the state dimension, and d_u^2 the input dimension.*[2]

This lower bound precludes the possibility of any significant improvement on the certainty equivalence strategy of Mania et al. [2019]. In addition, it suggests a certain optimal dependence on *problem dimension*, painting an even finer picture of the sample complexity of online control. It is natural to understand if this dependence on problem dimension can be attained, and if so, whether a design principle as simple as certainty equivalence can attain it it. Our second result answers in the affirmative:

Informal Theorem 5. *There exists an adaptive control algorithm whose regret is at most $\sqrt{T \log(T) d_x d_u^2}$ with high probability against any (stabilizable) control system. The algorithm strongly resembles the simple certainty-equivalence recipe, but with continual exploration so as to refine dependence on problem dimension. Furthermore, the analysis removes the controllability condition necessitated by prior work.*

Together, our results demonstrates that relatively simple adaptive algorithms are sufficient for near-optimal regret. Crucially, our analysis exploits a fundamental tension between control and indentification in linear systems, first described by Polderman [1986], and summarized in Polderman [1989].

Removing the controllbility assumption from the analysis of certainty equialence is a contribution of independent technical interest. Chapter 5 exposes a new strategy for bounding perturbations to the DARE we call the self-bounding ODE method, which yields the desired controllability-free bounds.

Nonstochastic Control

Online LQR constitutes a relatively benign adaptive control setting, because all learner uncertainty arises from lack of knowledge of system dynamics. As in our study of system identification above, we attempt to extend our findings to a considerably more general domain. To this end, Chapter 6 takes up study of the nonstochastic control problem, first introduced by Agarwal et al. [2019a]. The chapter summarizes the findings in the first half of Simchowitz et al. [2020].

In nonstochastic control, the disturbance sequence is not stochastic, nor are the costs fixed in advance. Instead, they a new cost function $\ell_t(\cdot)$ is revealed to the learner at each

[2]This lower bound, and the upper bound to follow, suppress a dependence on $\|P_\star\|_{\mathrm{op}}$, the opertor norm of the solution to the DARE, as well as relatively minor dependence on one or two other problem parameters.

time t, and distrubances $\mathbf{w}_t$ and $\mathbf{e}_t$may be chosen adversarially. Extending Agarwal et al. [2019a], we also consider partially observed dynamics determined by Eqs. (1.1) and (1.2).

The performance of a given policy π is measured as

$$J_T(\pi) = \sum_{t=1}^{T} \ell_t(\mathbf{y}_t^\pi, \mathbf{u}_t^\pi),$$

where $\ell_t(y, u)$ is a convex loss taking in an output and control input argument, and where $(\mathbf{y}_t^\pi, \mathbf{u}_t^\pi)$ describe the output and inputs which would be visited (a) by policy π, (b) under the dynamics Eqs. (1.1) and (1.2), and (c) holding the sequence of realized disturbances $\mathbf{w}_1, \ldots, \mathbf{w}_T$ and $\mathbf{e}_1, \ldots, \mathbf{e}_T$ as given (that is, fixed across π).

Nonstochastic Regret We consider regret with respect to a restricted class $\Pi_\star$ of dynamic, *LTI* control policies; informally, these are policies which evolve according to their own set of fixed linear dynamic equations. The regret of our chosen, adaptive strategy alg with respect to $\Pi_\star$ is then

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_\star) = J_T(\mathsf{alg}) - \inf_{\pi \in \Pi_\star} J_T(\pi).$$

Note that this notion of regret evaluates alg against the optimal choice of policy $\pi \in \Pi_\star$, given full knowledge of system dynamics Eqs. (1.1) and (1.2), disturbances $\mathbf{w}_t$ and $\mathbf{e}_t$, and cost functions ℓ_t. Thus, low regret can be interpreted in the spirit of a more adaptive notion of $\mathcal{H}_\infty$ control [Zhou et al., 1996], where one adapts to the realized sequences of disturbances, rather than the game-theoretic worst case. Due to known lower bounds [Li et al., 2019], it is necessary to consider comparisons with restricted policy classes, or else suffer regret linear in T.

Nonstochastic Regret Bounds It was shown in Agarwal et al. [2019a] that one can obtain $\sqrt{T}$ regret with respect to a restricted class of static feedback policies $\mathbf{u}_t = K\mathbf{x}_t$, provided the system state is fully observed. This is achieved by using SLS [Wang et al., 2019] to construct a convex relaxation of the set of such policies, and applying an online learning framework - online convex optimization with memory framework [Anava et al., 2015] - to this relaxation.

The results in this chapter extend the guarantee to partially observed systems, and to richer policy benchmarks. In doing so, we develop disturbance response control (Drc), which provides a more expressive language for convex control relaxations, encompassing the classical input-output Zames [1981] and Youla-Kucěra parametrizations [Youla et al., 1976, Kučera, 1975]. We state our guarantee as a reduction

Informal Theorem 6. *The* Drc *parametrizations yields reduction from nonstochastic control with partial observed to online convex optimization with memory. When instantiated with online gradient descent as the learning subroutine, the reduction yields* $\tilde{\mathcal{O}}(\sqrt{T})$ *regret, provided that the dynamics are* known *to the learner at the outset of the game.*

Note that nonstochastic problem is still non-trivial, because the learner must contend with unknown disturbances and cost-functions, revealed online. Indeed, Simchowitz et al. [2020] demonstrates that no algorithm can best $\sqrt{T}$ regret, even in the most benign special cases.

We extend our reduction to unknown systems as follows:

Informal Theorem 7. *When preceeded by an initial state of uniform exploration with random-sign inputs, the* Drc *parametrization yields a reduction from nonstochastic control with partial observed to online convex optimization with memory applicable to* unknown *system dynamics. When instantiated with online gradient descent as the learning subroutine, and $N \approx T^{2/3}$ steps of initial uniform exploration, the reduction yields $\tilde{\mathcal{O}}(T^{2/3})$ regret.*

$T^{2/3}$ regret had been obtained by Hazan et al. [2019] assuming full state observation, under a strong controllability assumption. In contrast, our reduction handles partial observation and does not require controllability. These generalizations are possible because we learn the Markov operator directly, rather than recovering system matrices. Conveniently, analysis of the estimation phase is a direct application of Chapter 4.

Making Nonstochastic Control as Easy as Stochastic

There is a significant gap between the $\sqrt{T}$ optimal online LQR regret rate, and the $T^{2/3}$ regret upper bound derived in Chapter 6 for nonstochastic control of unknown dynamical systems. The gap of known systems is even more stark: in nonstochastic control, the reduction in Chapter 6 claim $\sqrt{T}$ regret; online LQR with a known system, on the other hand, is just regular old LQR, so we can synthesize the optimal control policy and pay essentially *zero* regret.[3] This poses a natural question:

> *When viewed from the perspective of regret, is nonstochastic control fundamentally more challenging than adpative control with stochastic noise, namely online LQR?*

Presenting material from Simchowitz [2020] and the latter half of Simchowitz et al. [2020], Chapter 7 demonstrates that the short answer is "no."

To level the playing field, between online LQR and nonstochastic control, we assume that the losses ℓ_t are not just convex, but *strongly convex*, that is the minimimal eigenvalue of the Hessians are at most some $\alpha > 0$:

$$\forall (y, u), \quad \lambda_{\min}(\nabla^2 \ell_t(y, u)) \geq \alpha. \tag{1.3}$$

It is clear that such assumption holds in LQR because the costs are positive definite quadratic functions. Moreover, it has been widely observed in online learning that, when loss functions

[3] Depending on the way regret is defined, there may be some transient effects making the regret constant, or perhaps logarithmic, in the time horizon.

such exhibit curvature, improved regret bounds are often possible [Hazan et al., 2007, Vovk, 2001]. It remains unknown whether one can improve upon $T^{2/3}$ regret for unknown system dyanimcs without this assumption, and Simchowitz et al. [2020] prove that, in its absence, $\sqrt{T}$ regret is unimprovable for known system dyanmics.

While promising, the strong convexity of the costs $\ell_t(y, u)$ does not imply strong convexity of the DRC-parametrized control problem. This is in part due to the nonstochastic noise, which may excite some system modes, but not others. We overcome this challenge with a second-order online learning algorithm we call Online Semi-Newton Step (Semi-ONS), inspired by the Online Newton Step algorithm of Hazan et al. [2007]. Semi-ONS carefuly accounts for how the loss curvature propogates through the DRC parametrization. Plugging this algorithm into the reduction of Chapter 6, we find:

Informal Theorem 8. *For known systems, instantiating the* DRC*-reduction with* Semi-ONS *as the online learning procedure yields* $\mathrm{poly}(\log T)$ *regret on a time horizon* T*, provided the losses enjoy the curvature property* (1.3)*.*

Logarithmic regret is typically the very best one can hope for in any non-vacuous online learning setting, and indeed one can show that no algorithm can improve upon $\log T$ regret for our setting (even with, say, stochastic noise but adversarial costs).

For unknown systems, the analogous reduction match the rates of online LQR:

Informal Theorem 9. *For unknown systems, instantiating the* DRC*-reduction with an initial estimation phase of length* $N \approx \sqrt{T}$*, and with* Semi-ONS *as the online learning procedure, yields* $\tilde{\mathcal{O}}(\sqrt{T})$ *regret, provided the losses enjoy the curvature property* (1.3)*, and the complementary property that* $\lambda_{\max}(\nabla^2 \ell)$ *is uniformly* upper bounded.

Our findings, together with those in Chapters 5 and 6 are summarized in Table 1.1. In sum, we conclude that, from the perspective of regret, nonstochastic control is (almost) as easy as stochastic adaptive control. There are two important caveats to our findings, adressed in the remarks below.

In addition there are settings where stochastic control does seem easier than nonstochastic: online linear quadratic Gaussian control (LQG) [Lale et al., 2020a], online LQR with knowlege of the $B_\star$ matrix [Cassel et al., 2020], and nonstochastic control of an unknown system with a fixed general (possible non-strongly convex) convex cost [Plevrakis and Hazan, 2020]. All three are discussed further in the related work chapter, Chapter 2. Finally, understanding whether or not $\sqrt{T}$ regret is attainable in the full non-stochastic model with general convex costs remains an open problem.

Remark 1.1. The claimed fast rates for nonstochastic control require one of two additional condition: either the system can be stabilized by static feedback, or the nonstochastic noise includes a stochastic component with non-degenerate covariance; we call this latter noise model *semi-adversarial noise*. The first condition always holds for any stabilizable fully observed system, but may fail under partial observation. Thus, the optimal regret

in non-stochastic control of a fully observed system matches that of online LQR. A more general condition than static-feedback stabilizability suffices for partially observed systems (Lemma 7.1), which can be verified in restrictive cases (Lemma 7.16). The second option - semi-adversarial noise - applies to fully and partially observed systems alike.

Remark 1.2. From the perspective of *optimal control*, nonstochastic control is considerably more challenging; indeed, prohibitvely so. This is because in online LQR, the exact optimal policy was a static feedback controller $\mathbf{u}_t = K\mathbf{x}_t$, and we could obtain low regret relative (or "compete with") such policies. In nonstochastic control, we consider a benchmark of linear dynamic control policies. And while this benchmark does include the static policies optimal for LQR instances, it may not contain the global, unconstrained optimal control policy *for a given nonstochastic control problem* [Li et al., 2019]. Thus, in online LQR, we compete with the optimal control law, whereas in nonstochastic control, we do not.

Chapter 2

Related Work

Mirroring the structure of this thesis, we divide our discussion of prior work into system identification and adaptive control. We then conclude with a broader discussion of how the algorithmic and analytic ideas in this thesis relate to the broader learning theory community.

2.1 System Identification

Part I of this thesis addresses the problem of system identification of linear time-invariant (LTI) dynamical systems; that is, learning an approximate model of the dynamics from observed trajectories.

Identifying LTI systems from data has a decades-old history in both the time-series and system identification communities (see Ljung [1999], Verhaegen [1993], Galrinho [2016] and references therein) with least squares estimation being a central tool for dozens of algorithms, many of them similar in spirit to the two stage least square. A complementary viewpoint comes from a family of techniques techniques known broadly as *subspace identification* (e.g. Qin [2006]), which take a singular value decomposition (SVD) of the raw data.

Early Results on Nonasymptotic System Identification Results

Prior to the work in this thesis, the results on non-asymptotic system identification had been somewhat sparse. Some of the earlier non-asymptotic literature in system identification include Campi and Weyer [2002] and Vidyasagar and Karandikar [2008]. The results provided in this line of work are often quite conservative, featuring quantities which are exponential in the degree of the system. Furthermore, the rates given are often difficult to interpret.

Prior Work on System Identification with Full State Observation

Our findings in Chapter 3 for fully observed systems qualitatively match the behavior of the rate given for least squares in Dean et al. [2017], in that the key spectral quantity governing

the rate of convergence is the minimum eigenvalue of the finite-time controllability Gramian. The major difference is that our analysis analyzes a single trajectory, whereas the the analysis in Dean et al. [2017] uses multiple independent trajectories, and discards all but the last state-transition in each trajectory. This decouples the covariates, and reduces the analysis to that of random design linear regression with independent covariates. We note, however, that the analysis in Dean et al. [2017] applies even when $A_\star$ is unstable, provided each individual trajectory length is not too large.

Another closely related work is the analysis by Rantzer [2018] of scalar systems $\mathbf{x}_t = a_\star \mathbf{x}_t + \mathbf{w}_t$, $a_\star \in \mathbb{R}$. This work demonstrates similar improvements in estimation performance as the magnitude of $a_\star$ grows.

Most directly related to our work on fully observed systems are works by Faradonbeh et al. [2018a, 2017], who study the linear system identification problem by proving a non-asymptotic rate on the convergence of the OLS estimator to the true system matrices. In the regime where $A_\star$ is stable, these papers recover a similar rate as our result. The major difference is that the dependence of their analysis on the spectral properties of $A_\star$ are qualitatively suboptimal, and difficult to interpret precisely.

The material presented in Chapter 3 is drawn from Simchowitz et al. [2018]. Subsequent to that publication, Sarkar and Rakhlin [2019] extended the analysis to potentially unstable systems. The uncovered an interesting regularity condition: that unstable linear systems with repeat unstable eigenvalues (e.g., $A_\star = 2 \cdot I$) pose significant challenges for the ordinary least squares estimator. Other works have shown a promising role for active learning and experiment design, both to improve estimation rates [Wagenmaker and Jamieson, 2020], optimize downstream controller synthesis [Wagenmaker et al., 2021], and accomodate certain classes of nonlinear systems [Mania et al., 2020].

Partial State Observation

Shah et al. [2012] pose the problem of recovering a single-input, single-output (SISO) LTI system from linear measurements in the frequency domain as a sparse recovery problem, proving polynomial sample complexity for recovery in the $\mathcal{H}_2$-norm. Hardt et al. [2016] show that under fairly restrictive assumptions on the $A_\star$ matrix, projected gradient descent recovers the state-space representation of an LTI system with only a polynomial number of samples. The analysis from both Shah et al. [2012] and Hardt et al. [2016] degrade polynomially in $\frac{1}{1-\rho(A_\star)}$, where $\rho(A_\star)$ is the spectral radius of underlying $A_\star$. Hazan et al. [2017, 2018] provide online prediction bounds for prediction in LTI systems, though these bounds degrade for marginally stable systems.Note that Hardt et al. [2016], Hazan et al. [2017, 2018] consider prediction. As noted in our discussion of Dean et al. [2017], strict stability can be removed in many of these settings as well, at the expense of requiring a number of independent trajectories which grows with desired accuracy [Oymak, 2018].

Our findings in Chapter 4 build on on the analysis due to Oymak and Ozay [2019], both analyzing recovery of the Markov parameter of an LTI system from input-output data via least squares. The main differences are that our work 1. introduces a two-stage least

squares procedure to remove the dependence on system stability, 2. extends to possibly non-stochastic, and 3. focuses solely on recovery of the Markov operator, and not system matrices. Notably, Oymak and Ozay [2019] analyze the classical Ho-Kalman algorithm [Ho and Kalman, 1966] for parameter recovery. An analysis of this algorithm was refined by Sarkar et al. [2019].

Our presentation of Chapter 4 is based on the publication [Simchowitz et al., 2019]. Subsequent to its publication, Tsiamis and Pappas [2019] provide an elegant analysis of system identification under purely Gaussian noise in the presence of a Kalman filter, and [Rashidinejad et al., 2020] draws interesting connections between learning with long-memory systems and the notion of "Kolmogorov complexity".

2.2 Adaptive Control

Adaptive control studies how a controller system ought to improve its own performance in a given control task by incorporating past data. Most conventionally, adaptive control is posed in the setting where certain aspects of the system, typically system dynamics, are unknown, and must be inferred during the control process. Adaptive control has been studied at length in the linear [Stengel, 1994], nonlinear [Krstic et al., 1995], and robust control [Ioannou and Sun, 2012] settings.

Part II of this thesis considers more the problem of online control: a modern, learning-theoretic formulation of adaptive control in which adaptivity is measured by *regret* compared to an idealized control benchmark given a priori knowledge of various control system properties and task desideratum. The following discussion describes work in this area. We remark that data-driven control design has been studied more broadly in batch, PAC settings [Fiechter, 1997, Dean et al., 2017], in the context of model-free methods [Fazel et al., 2018], and from the perspective of experiment design [Wagenmaker et al., 2021].

Online LQR The online LQR setting we study in Chapter 5 was introduced by Abbasi-Yadkori and Szepesvári [2011]. This work considers the problem of controlling an unknown linear system under stationary stochastic noise. They showed that an algorithm based on the optimism in the face of uncertainty (OFU) principle enjoys $\sqrt{T}$, but their algorithm is computationally inefficient and their regret bound depends exponentially on dimension. The problem was revisited by Dean et al. [2018], who showed that an explicit explore-exploit scheme based on ε-greedy exploration and certainty equivalence achieves $T^{2/3}$ regret efficiently, and left the question of obtaining $\sqrt{T}$ regret efficiently as an open problem. This issue was subsequently addressed by Faradonbeh et al. [2018b] and Mania et al. [2019], who showed that certainty equivalence obtains $\sqrt{T}$ regret, and Cohen et al. [2019], who achieve $\sqrt{T}$ regret using a semidefinite programming relaxation for the OFU scheme. The regret bounds in Faradonbeh et al. [2018b] do not specify dimension dependence, and (for

$d_x \geq d_u$), the dimension scaling of Cohen et al. [2019] can be as large as $\sqrt{d_x^{16}T}$;[1] Mania et al. [2019] incurs an almost-optimal dimension dependence of $\sqrt{d_x^3 T}$ (suboptimal when $d_u \ll d_x$), but at the expense of imposing a strong controllability assumption.

The question of whether regret for online LQR could be improved further (for example, to $\log T$) remained open, and was left as a conjecture by Faradonbeh et al. [2018c]. Our lower bounds resolve this conjecture by showing that $\sqrt{T}$-regret is optimal. Moreover, by refining the upper bounds of Mania et al. [2019], our results show that the asymptotically optimal regret is $\widetilde{\Theta}(\sqrt{d_u^2 d_x T})$, and that this achieved by certainty equivalence. Beyond attaining the optimal dimension dependence, our upper bounds also enjoy refined dependence on problem parameters, and do not require a-priori knowledge of these parameters.

The work presented in Chapter 5 is drawn from the author's publication Simchowitz and Foster [2020]. Concurrent work by Cassel et al. [2020] also demonstrate that $\sqrt{T}$ regret is unavoidable in online LQR, but does not characterize the optimal dimension dependence. It also observes that logarithmic regret *is* possible the matrix $B_\star$ is known in advance; this insight can also be derived from the intuitions in Chapter 5; this observation is extended further in Ziemann and Sandberg [2020]. Another concurrent work, Abeille and Lazaric [2020] provides a computationally efficient implements of the optimisim-algorithm first introdced by Abbasi-Yadkori and Szepesvári [2011] a decade earlier. While this algorithm appears to suffer worse dimension-dependence than the algorithm analyzed in Chapter 5, it may have a more benign dependence on other problem parameters.

Finally, a parallel line of research provides Bayesian and frequentist regret bounds for online LQR based on Thompson sampling [Ouyang et al., 2017, Abeille and Lazaric, 2017], with Abeille and Lazaric [2018] demonstrating $\sqrt{T}$-regret for the scalar setting. Unfortunately, Thompson sampling is not computationally efficient for the LQR.

Nonstochastic Control

Chapters 6 and 7 consider the non-stochastic control problem, which departs from LQR by considering adversarially-chosen disturbances and cost functions. These chapters are based on the author's publications Simchowitz et al. [2020] and Simchowitz [2020].

Recent work first departed from online LQR by considered adversarially chosen costs under known stochastic or noiseless dynamics [Abbasi-Yadkori et al., 2014, Cohen et al., 2018]. The setting we consider in this paper was established in Agarwal et al. [2019a], who obtain $\sqrt{T}$-regret in the more general and challenging setting where the Lipschitz loss function *and the perturbations* are adversarially chosen. The key insight behind this result is combining an improper controller parametrization know as disturbance-based control with recent advances in online convex optimization with memory due to Anava et al. [2015]. Follow up work Gradu et al. [2020] extend to linear, time-varying systems and adaptive regret. Analogous problems have also been studied in the tabular MDP setting [Even-Dar

[1] The regret bound of Cohen et al. [2019] scales as $d_x^3\sqrt{T} \cdot (J_\infty(K_\star))^5$; typically, $J_\infty(K_\star)$ scales linearly in d_x

et al., 2009, Zimin and Neu, 2013, Dekel and Hazan, 2013, Rosenberg and Mansour, 2019, Jin et al., 2019].

Under the considerably stronger condition of controllability, the recent work by Hazan et al. [2019] attains $T^{2/3}$ regret for adversarial noise/losses when the system is *unknown*. The material presented in Chapter 6 extends this prior work to partially-observed systems. This requires accomodating both a richer control benchmark of dynamic LTI controllers, and accomodating a much richer class of controller parametrizations. Unlike Hazan et al. [2019], our algorithms do not recover explicit state-space representations of the dynamics, and therefore circumvent controllability assumptions. Chapter 7 further improves the $T^{2/3}$ regret bound of Hazan et al. [2019] to $\sqrt{T}$ when the control costs are strongly convex.

Importantly, all aforementioned work consider regret with respect to a restricted class of control policies: static feedback in Agarwal et al. [2019b], Hazan et al. [2019], and dynamic LTI controllers in Chapters 6 and 7. This turns out to be necessary: Li et al. [2019] demonstrate that it is impossible to attain sublinear (in T) regret with respect to the unconstrained optimal policy, even in benign problems with no process noise and full state observation, but adversarially-chosen state and control costs. They show that this negative result can be circumvented if the algorithm has access to a finite lookahead of costs H steps into the future, obtaining $T \cdot \exp(-\Omega(H))$ regret when such lookahead is available. Concurrent work by Plevrakis and Hazan [2020] also demonstrates $\sqrt{T}$ regret in a(n incomparable) setting with a *fixed* convex cost which is *not necessarily strongly convex*, and with i.i.d. stochastic noise.

Logarithmic Regret for Online Control of Known Systems Abbasi-Yadkori et al. [2014] were the first to attain logarithmic regret in a restricted class of adversarial online tracking problems. Work by Agarwal et al. [2019b] achieves logarithmic pseudo-regret for strongly convex, adversarially selected losses and well-conditioned stochastic noise. This is entirely superseded by this author's publication Simchowitz et al. [2020], which improves the bound from pseudo-regret to actual regret (under a slightly stronger smoothness assumption), and extends to partial observation and semi-adversarial noise. This was further extended to fully adversarial noise in Simchowitz [2020], provided the system can be stabilized with a restricted family of control policies. These findings are both outlined in Chapter 7.

Simchowitz [2020] supersedes another of the author's publications, Foster and Simchowitz [2020], which proposes a creative alternative approach for logarithmic regret in online control. Derived from the performance-difference lemma [Kakade, 2003], that work derives "online learning with advantages" - which yields logarithmic regret with truly adversarial noise, but fixed quadratic cost functions and with full observation. Notably, Foster and Simchowitz [2020] characterizes the unconstrained optimal policies for any *fixed* disturbance sequence, which was received later study [Goel and Hassibi, 2020, Yu et al., 2020].

The Curious Case of Online LQG

LQG, or linear quadratic Gaussian control, is the natural extension of LQR to partially observed systems. Mania et al. [2019] were the first to study this problem in the online

setting, presenting perturbation bounds which suggest $T^{2/3}$ regret. There were later improved $\sqrt{T}$ by Lale et al. [2020b], matching the optimal rate for LQR. The fast rates of Chapter 7 — more precisely, bounds for restricted systems and adversarial noise due to Simchowitz [2020] and for unrestricted system and adversarial noise due to [Simchowitz et al., 2020] — also imply such a $\sqrt{T}$ rate. As $\sqrt{T}$ regret is the optimal regret rate obtainable in online LQR (Chapter 5), one may be led to conclude that there is no more to the story.

Suprisingly, for LQG with both non-denegerate process and observation noise, Lale et al. [2020a] attain $\mathrm{poly}(\log T)$ regret, demonstrating that in the presence of significant observation noise, LQG is in fact *easier* than LQR (with no observation noise) in terms of regret. This is because the process and observation noise provide continual exploration, allowing the learner to greedily exploit all current knowledge, without succumbing to the explore-exploit tradeoffs established for LQR in Chapter 5.

2.3 Provenance of Techniques

The work in this thesis would not have been possible without the significant advances made by the learning community in the past two decades. Here, we explain how our algorithmic and analytic ideas emerge from and contribute to that body of literature.

Statistical Learning Theory

Chapters 3 to 5 all rely on modern advances in statistical learning techniques. To analyzes the performance of the ordinary least squares estimator, all three chapters apply the self-normalized martingale bound of Abbasi-Yadkori et al. [2011], originally due to de la Pena et al. [2009]. Chapter 3 also proposes the "block-martingale small-ball" technique, a tool for lower bounding the eigenvalues of covariance matrices, which draws its inspiration for Mendelson [2014]'s small-ball method; this technique proves useful again in Chapter 5. Finally, the lower bounds in Chapter 5 build on well-known lower bound technique for adaptive sensing based on Assouad's lemma [Arias-Castro et al., 2012] (see also Assouad [1983], Yu [1997]) in order to obtain optimal dimension dependence.

Prefiltering and Variance Reduction

The key algorithmic idea in Chapter 4 is the two-staged least squares procedure we term "prefiltered least squares". One can regard prefiltered least squares as a specific instance of a prefiltered autoregressive model (such as ARX or ARMAX); much work has been done on explicit filtering and debiasing schemes for these types of models Spinelli et al. [2005], Ding [2013], Zheng [2004], Guo and Huang [1989], Zhang [2011], Wang [2011], Galrinho et al. [2014]. However, analyses of these schemes are often (i) asymptotic, (ii) for strictly stable systems only, or (iii) use a limited noise model.

Beyond linear systems, our prefiltering step bears similarities to the *instrumental variables* technique in used in controls [Viberg et al., 1997], econometrics [Hansen and Singleton, 1982] and causal statistics [Angrist et al., 1996], which is used more for debiasing than for denoising. More broadly, variance reduction has become an indispensable component of reinforcement learning [Weaver and Tao, 2001, Greensmith et al., 2004, Tucker et al., 2017, Sutton and Barto, 1998], including the theoretical study of tabular Markov Decision Processes [Kakade et al., 2018, Sidford et al., 2018].

Convex Parameterization of Linear Controllers

The algorithm and analysis presented in Chapters 6 and 7 rely on disturbance response control (DRC), a template of convex parametrizations for online control.

Convex or *lifted* parameterizations have a rich history in the control literature. Our DRC parametrization encompasses input-ouput parametrizations Zames [1981], Rotkowitz and Lall [2005], Furieri et al. [2019] as well as classical Youla or Youla-Kučera parametrization [Youla et al., 1976, Kučera, 1975], and approximations to the Youla parametrization which require only *approximate* knowledge of the system. More recently, Goulart et al. [2006] propose a parametrization over state-feedback policies, and Wang et al. [2019] introduce a generalization of Youla called system level synthesis (SLS); SLS is equivalent to the parametrizations adopted by Agarwal et al. [2019a] et seq., and underpins the $T^{2/3}$-regret algorithm of Dean et al. [2018] for online LQR with an unknown system; one consequence of our work is that convex parametrizations can achieve the optimal $\sqrt{T}$ in this setting. However, it is unclear if SLS (as opposed to input-output or Youla) can be used to attain sublinear regret under partial observation and adversarial noise.

Online learning and online convex optimization.

Chapters 6 and 7 also make extensive use of techniques from the field of online learning and regret minimization in games [Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz et al., 2012, Hazan, 2019]. Of particular interest are techniques for coping with policy regret and online convex optimization for loss functions with memory called "online convex optimization with memory" [Anava et al., 2015]. [Arora et al., 2012] has proposed alternative notions of policy regret pertinent to other online learning settings.

Fast Rates in Online Learning Logarithmic regret bounds are ubiquitous in online learning and optimization problems with strongly convex loss functions [Vovk, 2001, Hazan et al., 2007, Rakhlin and Sridharan, 2014]. Agarwal et al. [2019b] demonstrate that for the problem of controlling an *known* linear dynamic system with adversarially chosen, strongly convex costs, logarithmic regret is also attainable. Our $\sqrt{T}$ lower bound in Chapter 5 shows that the situation for the online LQR with an *unknown* system parallels that of bandit convex optimization, where Shamir [2013] showed that $\sqrt{T}$ is optimal even for strongly convex

quadratics. That is, in spite of strong convexity of the losses, issues of partial observability prevent fast rates in both settings.

Fast-rates (e.g. logarithmic regret) have also been observed in more general families of loss functions, notably exp-concave losses [Vovk, 2001, Hazan et al., 2007]. However, no such fast rates for these general families are known when the loss functions have memory, as in Anava et al. [2015]'s online-convex-optimization-with-memory setting described above. In Chapter 7, our work intervenes in this problem by formulating a refinement setting we term "online convex optimization with affine memory", under which fast rates are possible. In adition, it describes a Euclidean-movement lower bounds, rooted in the broader learning-with-switching-costs literature [Altschuler and Talwar, 2018, Chen et al., 2019, Dekel et al., 2014], which suggest that fast rates for losses with memory may be challenging in general.

Robustness in Convex Optimization When deriving rates for systems with *unknown dynamics* in Chapters 6 and 7, we rely on the robustness of online learning procedures to misspecification of their loss functions. Notably, for the fast $\sqrt{T}$-rate for unknown systems described in Chapter 7, we show that online optimization methods can exhibit quadratic sensitivity – $T\epsilon^2$ – to ϵ-bounded errors in the online loss functions, provided those losses exhibit sufficient curvature. This generalizes a known robustness result for batch stochastic convex optimization [Devolder et al., 2014].

Part I

System Identification

Chapter 3

Learning Without Mixing

This chapter focused on identifying a discrete-time *linear dynamical system* from an observed trajectory. Such systems are described by two parameter matrices $A_\star$ and $B_\star$, and the dynamics evolve according to the law $\mathbf{x}_{t+1} = A_\star \mathbf{x}_t + B_\star \mathbf{u}_t + \mathbf{w}_t$, where $\mathbf{x}_t \in \mathbb{R}^{d_x}$ is the state of the system, $\mathbf{u}_t$ is the input of the system, and $\mathbf{w}_t \in \mathbb{R}^{d_x}$ denotes unobserved process noise.

Prior to the work presented in this chapter, the relationship between the matrix $A_\star$ and the statistical rate for estimating this matrix was poorly understood. We note that the larger the state vectors $\mathbf{x}_t$ are in comparison to the process noise, the larger the *signal-to-noise ratio* for estimating $A_\star$ is. As a result, larger matrices $A_\star$ (larger in an appropriate sense, discussed below) lead to states $\mathbf{x}_t$ of larger norm, which in turn should make the estimation of $A_\star$ easier. However, it is difficult to theoretically formalize this intuition because the sequence of measurements $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ used for estimation is not i.i.d. and it is dependent on the noise $\mathbf{w}_1, \ldots, \mathbf{w}_{N-1}$. Even the computationally straightforward ordinary least-squares (OLS) estimator is difficult to analyze. Standard analyses for OLSon random design linear regression [Hsu et al., 2012] cannot be used due to the dependency between the covariates $\mathbf{x}_t$ and the process noise $\mathbf{w}_t$.

In the statistics and machine learning literature, correlated data is usually dealt with using mixing-time arguments [Yu, 1994], which relies on fast convergence to a stationary distribution that allows correlated samples to be treated roughly as if they were independent. While this approach has been successfully used to develop generalization bounds for time-series data [Mohri and Rostamizadeh, 2008], a fundamental limitation of mixing-time arguments is that the bounds deteriorate when the underlying process is slower to mix. In the case of linear systems, this behavior is qualitatively incorrect. For linear systems, the rate of mixing is intimately tied to the eigenvalues of the matrix $A_\star$, specifically the *spectral radius* $\rho(A_\star)$. When $\rho(A_\star) < 1$ (i.e. when the system is *stable*), the process mixes to a stationary distribution at a rate that deteriorates as $\rho(A_\star)$ approaches the boundary of one. However, as discussed above, as $\rho(A_\star)$ increases we expect estimation to become easier due to better signal-to-noise ratio, and not harder as mixing-time arguments suggest.

We address these difficulties and offer a new statistical analysis of the ordinary least-squares (OLS) estimator of the dynamics $\mathbf{x}_{t+1} = A_\star \mathbf{x}_t + \mathbf{w}_t$ with no inputs, when the spectral

radius of $A_\star$ is at most one ($\rho(A_\star) \leq 1$, a regime known as *marginal stability*). Our results show that the statistical performance of OLSis determined by the minimum eigenvalue of the (finite-time) controllability Gramian $\Gamma_T = \sum_{s=0}^{T-1} A_\star^s (A_\star^\top)^s$. The controllability Gramian is a fundamental quantity in the theory of linear systems; the eigenvalues of the Gramian quantify how much white process noise $\mathbf{w}_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2 I)$ can excite the system. We show that a larger $\lambda_{\min}(\Gamma_T)$ leads to faster estimation of $A_\star$ in operator norm, and we also prove that up to log factors the OLSestimator is minimax optimal.

Organization

Section 3.1 formalizes the problem of system identification with i.i.d. Gaussian noise and full state observation, and describes the ordinary least squares estimator. It also introduces the key concepts of stability and mixing, which ground the discussion throughout the chapter.

Section 3.2 then asks to what extent mixing and stability have a bearing on the sample complexity of system identification. We explain hypotheses from past work that mixing is *essential* to system identification. We then introduce the *Gramian* matrix, which leads to a heuristic calculation that mixing may in fact be anti-correlated with estimator performance. This is expounded upon for an illustrative example of systems whose dynamical matrix $A_\star$ is equal to the identity scaled by a factor $\rho \in (0, 1]$: $A_\star = \rho I$.

Section 3.3 states our main findings. Theorem 3.1 presents an upper bound on the error of the least squares estimator. It shows that indeed the Gramian matrix, *and not the mixing property*, are the key to characterizing the sample complexity. We also develop corollaries for special cases of interest, including an in-depth discussion of learning rates for systems of the form $A_\star = \rho I$. A lower bound corroborates the optimality of our findings in this special case.

Finally, Section 3.4 provides proofs of our findings. We emphasize a novel technique - the *martingale small-ball method* - which facilliates our analysis.

3.1 Learning the State Transition Matrix

In this chapter, we consider the estimation of the state-transition matrix $A_\star \in \mathbb{R}^{d_x \times d_x}$ given access to a single, length-N trajectory of samples for times $t = 1, 2, \ldots, N$:

$$\mathbf{x}_{t+1} = A_\star \mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_w^2), \quad \mathbf{x}_0 \equiv 0 \tag{3.1}$$

Generalization to non-indentity covariance matrices is also possible, but omitted for simplicity.

The above setting is perhaps the simplest variant of the *system identification* problem, studied at length by .... We study the performance of the *least squares* estimator

$$\widehat{A}_{\text{ls}} \in \arg\min_{A \in \mathbb{R}^{d_x^2}} \sum_{t=1}^{N} \|\mathbf{x}_t - A\mathbf{x}_t\|_2^2, \tag{3.2}$$

and analyze recovery in the operator norm.

Problem 3.1. What is the performance of the least-squares estimator, $\widehat{A}_{\text{ls}}$. Specifically, for what constant $\mathcal{C}_{\text{ls}}$ depending on $A_\star$ can we ensure that, with high probability,

$$\|A_\star - \widehat{A}_{\text{ls}}\|_{\text{op}}^2 \lesssim \frac{\mathcal{C}_{\text{ls}}}{N} \quad ? \tag{3.3}$$

Problem Problem 3.1 can be alternatively framed in terms of *sample complexity*. That is, for a given tolerance ϵ, how long a trajectory N is required to ensure $\|A_\star - \widehat{A}_{\text{ls}}\|_{\text{op}}^2 \leq \epsilon$; this amounts to inverting the bound in (3.3). The results are visually more appealing when presented as in (3.3), but we shall slightly abuse language and characterize such upper bounds as assessing sample complexity.

Stability and Marginal Stability

This chapter investigates the extent to which *stability* determines the samples complexity of estimation. Stability is determined by the *spectral radius* of a matrix A.

Definition 3.1. Given a square matrix $A \in \mathbb{R}^{d_x \times d_x}$, its *spectral radius* is defined as the limit

$$\rho(A) := \lim_{n \to \infty} \|A^n\|_{\text{op}}^{1/n}. \tag{3.4}$$

A square matrix A is said to be Schur stable, or simply *stable*, if its spectral radius is strictly less than one, $\rho(A) < 1$. A is said to be *unstable* if $\rho(A) > 1$. We say that A is *marginally stable* if $\rho(A) = 1$.

By representing A in Jordan normal form, one can show that $\rho(A)$ corresponds to its largest magnitude eigenvalue, keeping in mind that eigenvalues of A may be complex. For example, given any $\rho \in \mathbb{C}$, the matrix $A = \rho I$ has $\rho(A) = |\rho|$. Thus, the norms of powers A^n of a stable A shrink geometrically, and grow geometrically if A is unstable.

Powers of marginally stable A do not shrink. Indeed, if A is marginally stable, then it has at least one eigenvalue λ of magnitude one, and for the corresponding eigenvector v, $\|A^n v\| = \|\lambda^n v\| = |\lambda|^n \|v\| = \|v\|$. In fact, the norms of powers of A may grow *polynomially*, as may be verifed when A is a Jordan block with eigenvalue 1. One can establish the following estimate via the Jordan cannonical form. Recall that every square matrix A can be expressed as

$$A = SJS^{-1},$$

where J is a block-diagonal *Jordan* matrix, with *Jordan blocks* of the form

$$B_{k,\lambda} := \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 & \\ 0 & \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \dots & 0 & \lambda \end{bmatrix} \in \mathbb{R}^{k,k}, \tag{3.5}$$

where $\lambda \in \mathbb{C}$ are elements of the spectrum of A. Using an explicit formula for powers of $B_{k,\lambda}$ (see [Simchowitz et al., 2018, Appendix A3] for further details), one can compute the upper and lower estimates on the norms of powers of A in terms of spectral radius.

Lemma 3.1. *For any square matrix $A \in \mathbb{R}^{d_x \times d_x}$, there exists constants $c_1, c_2 > 0$ and $c_3 \in [0, d_x]$ depending on A such that, for any $n \in \mathbb{N}$,*

$$c_1 \rho(A)^n \le \|A^n\|_{\mathrm{op}} \le c_2 n^{c_3} \rho(A)^n \tag{3.6}$$

In particular, c_3 can be chosen so that $c_3 + 1$ is the size of the largest Jordan block $B_{k,\lambda}$ in the Jordan-decomposition of A.

In this chapter, we assume that $A_\star$ is either stable or marginally stable. Formally,

Assumption 3.1. We assume that $\rho(A_\star) \le 1$.

An excellent follow-up due to Sarkar and Rakhlin [2019] studies the performance of the ordinary least-squares estimator even in potentially unstable systems.

3.2 What property determines sample complexity?

Learning with Mixing

Prior work has focused on the setting where the matrix $A_\star$ is *stable*, $\rho(A_\star) < 1$, so that $\|A_\star^n\|_{\mathrm{op}}$ shrinks as $\rho(A_\star)^{\Omega(n)}$, in view of Lemma 3.1. In this regime, the system exhibits a behavior known as *mixing*. Informally, for a suitably long delay parameter H and any time t, the iterates $\mathbf{x}_t$ is approximately independent of $\mathbf{x}_1, \dots, \mathbf{x}_{t+H}$. No see this, observe that we can write

$$\mathbf{x}_t = \sum_{i=0}^{t-1} A_\star^i \mathbf{w}_{t-i} = \left(\sum_{i=0}^{H-1} A_\star^i \mathbf{w}_{t-i} \right) + \left(\sum_{t=H}^{t-1} A_\star^i \mathbf{w}_{t-i} \right). \tag{3.7}$$

For large enough H, the second term in the last line is vanishingly small, on the order of $\rho(A_\star)^{\Omega(n)}$. Hence, $\mathbf{x}_t$ is roughly independent of $(\mathbf{w}_1, \dots, \mathbf{w}_{t-H})$, over which the past iterates $\mathbf{x}_1, \dots, \mathbf{x}_{t-H}$ are deterministic functions.

Via mixing arguments, past work has shown that one can essentially "block up" the learning problem into independent blocks j from iterates $(\mathbf{x}_j, \mathbf{x}_{j+H}, \mathbf{x}_{j+2H}, \dots)$, and use the fact that the iterates within blocks are essentially independent. Thus, mixing provides a means of reducing learning with dynamic and dependent data to learning with i.i.d. data. For brevity, we have kept this discussion rather informal, and defer to Yu [1994], Mohri and Rostamizadeh [2008] for precise arguments. In light of these arguments, one might conjecture that mixing is essential to establishing statistically efficient learning:

Hypothesis 3.1. *Because mixing measures how correlated the data in the rollout are, the mixing time $\rho(A_\star)$ characterizes the performance of the least squares estimator. That is, the scaling $\mathcal{C}_{\mathrm{ls}}$ in Eq. (3.3)* grows *with $\rho(A_\star)$.*

The Gramian Matrix

Directly from the form of the least squares estimator, we one can

$$\begin{aligned}
\widehat{A}_{\text{ls}} &= \left(\sum_{t=1}^{t} \mathbf{x}_{t+1}\mathbf{x}_t^\top\right)\left(\sum_{t=1}^{N} \mathbf{x}_t\mathbf{x}_t^\top\right)^{-1} && \text{(from Eq. (3.2))} \\
&= \left(\sum_{t=1}^{N} \mathbf{w}_t\mathbf{x}_t^\top\right)\left(\sum_{t=1}^{N} \mathbf{x}_t\mathbf{x}_t^\top\right)^{-1} + \left(\sum_{t=1}^{N} A_\star\mathbf{x}_t\mathbf{x}_t^\top\right)\left(\sum_{t=1}^{N} \mathbf{x}_t\mathbf{x}_t^\top\right)^{-1} && \text{(from Eq. (3.1))} \\
&= \left(\sum_{t=1}^{N} \mathbf{w}_t\mathbf{x}_t^\top\right)\mathbf{\Sigma}_N^{-1} + A_\star, \quad \text{where } \mathbf{\Sigma}_N := \mathbf{x}_t\mathbf{x}_t^\top. && (3.8)
\end{aligned}$$

The first term on the last line, $\sum_{t=1}^{t} \mathbf{w}_t\mathbf{x}_t^\top$ is mean zero, since $\mathbb{E}[\mathbf{w}_t \mid \mathbf{x}_t] = 0$. Hence, the error $\widehat{A}_{\text{ls}} - \widehat{A}_{\text{ls}}$ scales with the inverse magnitude of the covariance matrix $\mathbf{\Sigma}_N := \sum_{t=1}^{N} \mathbf{x}_t\mathbf{x}_t^\top$.

Hence, a natural conjecture is the expected state covariance of the dyanmics in Eq. (3.1), which we call the *Gramian*, determines the sample complexity.

Definition 3.2 (Gramian)**.** For $t \in \mathbb{N}$, we define time-t *Gramian* Γ_t under $A_\star$ as

$$\Gamma_t := \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] = \sum_{s=0}^{t-1} (A_\star^s)(A_\star^s)^\top$$

Intuitively, the Gramian Γ_N measure the magnitude, or strength of the "signal" $\mathbf{x}_t$; and the "noise" variance scales as σ_w^2, regardless of the dynamics. Hence, the larger Γ_N, the larger the signal-to-noise ratio. Moreover, we observe that

$$\mathbb{E}[\mathbf{\Sigma}_N] = \sigma_w^2 \sum_{t=1}^{N} \Gamma_t, \tag{3.9}$$

so that the magnitudes of Γ_t determine the magnitude of the (expected) covariance. This motivates the following hypothesis.

Hypothesis 3.2. *The sample complexity constant $\mathcal{C}_{\text{ls}}$ is* inversely *proportional to the size of the Gramian; for example, $\mathcal{C}_{\text{ls}} \propto 1/\sigma_{\min}(\Gamma_N)$. This is consistent with the observation that, the larger Γ_N, the larger the signal-to-noise ratio.*

As we shall make frequent reference to the Gramian, the following fact is useful to bear in mind, and can be checked via Lemma 3.1.

Fact 3.2. *For any $1 \le s \le t$, we have $I \preceq \Gamma_s \preceq \Gamma_t$. Moreover, $\lim_{t\to\infty} \Gamma_t$ exists (is finite) if and only $\rho(A_\star) < 1$.*

A Canonical Example

The hypothesis that the sample complexity is dictated by the Gramian is at odds with the hypotheses that sample complexity degrades with longer mixing time. We illustrate this with a simple example. Consider

$$A_\star = \rho I, \quad \rho \in [0,1] \tag{3.10}$$

Then $\rho(A_\star) = \rho$, and thus the mixing time roughly scales proportionally to the parameter ρ. On the other hand, we can compute that

$$\Gamma_t = I \cdot \sum_{s=0}^{t-1} \rho^{2s} \eqsim \begin{cases} \frac{1}{1-\rho} \cdot I_{d_x} & \rho \leq 1 - \frac{1}{N} \\ N \cdot I & \rho > 1 - \frac{1}{N}, \end{cases} \tag{3.11}$$

where $\eqsim$ denotes Lowner inequality up to universal constants.[1] In particular, we see that the Gramian scales linearly in ρ up to a limit $1 - \frac{1}{N}$, at which which saturates point the Grammian saturates at $N \cdot I$. Stated in terms of the expected covariance matrix, Eq. (3.9) entails

$$\frac{1}{\sigma_w^2}\mathbb{E}[\mathbf{\Sigma}_N] \eqsim \begin{cases} \frac{N}{1-\rho} \cdot I_{d_x} & \rho \leq 1 - \frac{1}{N} \\ N^2 \cdot I & \rho > 1 - \frac{1}{N}, \end{cases} \tag{3.12}$$

As a consequence of the above computations, we find

> *There is a fundamental tension between the mixing time in the system, and the magnitude of the Gramian. In particular, Hypotheses 3.1 and 3.2 are incompatible.*

A third hypothesis might interpolate between Hypotheses 3.1 and 3.2. That is, the magnitude of the Gramian determines the scaling of $\mathcal{C}_{\text{ls}}$ for N sufficiently large; however, this occurs after a burn-in time which dependends on the mixing properties of $A_\star$.

Hypothesis 3.3. *There may be a long* burn-in *time N before one can establish meaningful guarantees of the form Eq. (3.3). However, $\mathcal{C}_{\text{ls}}$ has no meaningful dependence on $\rho(A_\star)$ for large N, and is in fact determined by Γ_N in the limit.*

One possible argument for Hypothesis 3.3 is that slow mixing systems induces states $\mathbf{x}_t$ which are highly correlated across time. This correlation may force the *empirical* covariance matrix $\mathbf{\Sigma}_N := \sum_{t=1}^N \mathbf{x}_t \mathbf{x}_t$ to be poorly conditioned, even if its *expectation* may favor faster statistical estimation (in view of Eq. (3.12)).

If Hypothesis 3.3 is true, however, then for marginally stable systems - i.e. $\rho(A_\star) = 1$ - the burn-in time may be infinite, and thus the least squares estimator may be inconsistent; that is $\widehat{A}_{\text{ls}} \not\to A_\star$ ast $N \to \infty$. This would be despite the fact that the covariance matrix would grow superlinearly in expectation (Eq. (3.12)).

[1] That is, $X \eqsim Y$ if $c_1 X \preceq Y \preceq c_2 X$ for universal constants c_1, c_2.

3.3 Main Results

We state our bounds in terms of the time-averaged Gramian,

$$\tilde{\Gamma}_t := \frac{1}{t}\sum_{i=1}^{t} \Gamma_t \in \mathbb{S}_{++}^{d_x} \tag{3.13}$$

The matrix $\tilde{\Gamma}_t$ is roughly on the order of $t\Gamma_t$; more precisely, it lies between the PSD bound $\lfloor \frac{1}{2} \rfloor \tilde{\Gamma}_{\lfloor t/2 \rfloor} \preceq \tilde{\Gamma}_t \preceq \Gamma_t$. Our main theorem says that the sample complexity is essentially governed this term:

Theorem 3.1. *There exists universal constants $c_0, c_1 > 0$ such that the following holds. For each $k, N \in N$ and $\delta \in (0, 1/e)$, define the effective dimension*

$$d_{\delta,k,N} := d_x \log \frac{d_x}{\delta} + \log\det(I + \tilde{\Gamma}_k^{-1}\tilde{\Gamma}_N). \tag{3.14}$$

Then, for any k such that (k, N, δ) satisfy $N \geq c_0 \cdot k \cdot d_{\delta,k,N}$, it holds that

$$\mathbb{P}\left[\|A_\star - \widehat{A}_{\mathrm{ls}}\|_{\mathrm{op}}^2 \leq c_1 \cdot \frac{d_{\delta,k,N}}{N\lambda_{\min}(\tilde{\Gamma}_k)}\right] \leq \delta.$$

Note that the above bound is independent of σ_w^2, which follows from the invariance of the dynamics Eq. (3.1) to scaling of the states. Moreover, note that $\lambda_{\min}(\tilde{\Gamma}_k) \geq 1$, so the bound is at most $c_1 \cdot \frac{d_{\delta,k,N}}{N}$.

The effective dimension $d_{\delta,k,N}$ governs both the magnitude of the error bound on $A_\star - \widehat{A}_{\mathrm{ls}}$, and the burn-in, or minimal time N for which the bound applies. The upper bound is optimized by selecting k to be the largest integer for which $N \geq c_0 \cdot k \cdot d_{\delta,k,N}$. This is because, as k grows, $\tilde{\Gamma}_k$ incorporates more terms in the sum, and thus has a larger minimal eigenvalue. We now specialize our bound to three important settings; further discussion is provided in Simchowitz et al. [2018]. The last setting considers our paradigmatic example with $A_\star = \rho I$, and reveals our most suprising finding: that learning performance can *improve* as mixing degrades.

Stable Systems Let us begin with an analysis which appeals to system stability. For stable systems (i.e. $\rho(A_\star) < 1$), Γ_k, and thus $\tilde{\Gamma}_k$, converge to a limit

$$\Gamma_\infty = \lim_{k\to\infty} \Gamma_k = \lim_{k\to\infty} \tilde{\Gamma}_k.$$

Hence, there is some k_0 such that $\tilde{\Gamma}_{k_0} \succeq \frac{1}{2}\Gamma_\infty$, and by monotonicity, $\tilde{\Gamma}_N \preceq \Gamma_N \preceq \Gamma_\infty$. In particular, $\tilde{\Gamma}_{k_0}^{-1}\tilde{\Gamma}_N \preceq 3I$. Hence, for this k_0,

$$d_{\delta,k_0,N} \leq d_x \log \frac{3d_x}{\delta}.$$

This yields the following corollary.

Corollary 3.1. *Consider a stable system, $\rho(A_\star) < 1$. Then the limit $\lim_{t\to\infty} \Gamma_t = \Gamma_\infty$ is defined, let k_0 be as in the above discussion. Then, for $N \geq c_0 k_0 \cdot d_x \log \frac{3d_x}{\delta}$,*

$$\|A_\star - \widehat{A}_{\mathrm{ls}}\|_{\mathrm{op}}^2 \lesssim \frac{d_x \log(d_x/\delta)}{N\lambda_{\min}(\Gamma_\infty)} \ \ w.p.\ 1-\delta.$$

Note that this bound depends on this number k_0, which may be related the stablity/spectral radus of the system. Indeed, for the paradigmatic example $A_\star = \rho I$, we chan check that k_0 is on the order of $\frac{1}{1-\rho}$, and thus the burn-in time for Corollary 3.1 depends on the mixing time.

Marginally Stable Systems We now evaluate our system allowing for $\rho(A_\star) = 1$. This analysis does not depend on the parameter k_0 in the above analysis, which may be related to the system's mixing time.

Because $\tilde{\Gamma}_k$ is non-decreasing in k, $d_{\delta,k,N}$ is non-increasing in k. In particular, since $\tilde{\Gamma}_k \succeq I$, we can use the estimate in Lemma 3.1 to establish the following upper bound. Indeed, for any matrix with $\rho(A_\star) \leq 1$, Lemma 3.1 implies that, for a constant c_2, and for $c_3 + 1$ denoting the size of the largest Jordan block of $A_\star$,

$$\|\tilde{\Gamma}_N\|_{\mathrm{op}} \leq \sum_{s=0}^{N-1} \|A_\star^s\|^2 \leq c_2^2 N^{1+2c_3}.$$

Hence, the effective dimension may be bounded by

$$d_{\delta,k,N} \leq d_x \log \frac{\|\tilde{\Gamma}_N\| d_x}{\delta} \leq (1+2c_3) d_x \log N + d_x \log \frac{c_2^2 d_x}{\delta},$$

and thus grows at most logarithmically in N. Therefore our bound applies for $N = \tilde{\mathcal{O}}((1+c_3)d_x)$, even for marginally stable matrices.

Scaled Identity Matrices Lastly, we specialize our bound for the paradigmatic setting of scalar matrices, $A_\star = \rho I$, for $|\rho| \leq 1$. Extending the computation in Eq. (3.11) to $\rho \in [-1, 1]$, we see that we can select a block length of size $k = \Theta(\frac{N}{d\log(d/\delta)})$. This gives an upper bound of

$$\|A_\star - \widehat{A}_{\mathrm{ls}}\|_{\mathrm{op}}^2 \lesssim \begin{cases} \frac{(1-|\rho|)d\log(d/\delta)}{N} & |\rho| \leq 1 - \frac{cd\log(d/\delta)}{N} \\ \left(\frac{d\log(d/\delta)}{N}\right)^2 & |\rho| > 1 - \frac{cd\log(d/\delta)}{N} \end{cases}.$$

In particular,

- For $|\rho|$ bounded away from 1, the learning rate exhibits a linear speed, with the error decaying as $1 - |\rho|$.

- At the extreme, $|\rho| = 1$, we achieve an *fast rate* of $\tilde{\mathcal{O}}(\frac{d^2}{N^2})$, rather than the traditional $1/N$ scaling for standard least squares regression.

Therefore we find, that paradoxically, the learning rate *improves* as $|\rho| \to 1$, that is, as the mixing property degrades, and the learning rate is best for $|\rho| = 1$, when the system *does not mix*. One can verify that the same Gramian computations pertain to scaled-orthonormal systems $A_\star = \rho O$, where $O^\top O = I$, and thus the same learning rates apply as well.

Lower Bounds for Linear System Identification

We have seen in Theorem 3.1 and in the subsequent examples that the estimation of linear dynamical systems is easier for systems which are easily excitable. It is natural to ask what is the best possible estimation rate one can hope to achieve. To make explicit the dependence of the lower bounds on the spectrum of Γ_t, we consider the minimax rate of estimation over the set $\rho \cdot \mathscr{O}(d)$, where $\rho \in \mathbb{R}$ and $\mathscr{O}(d)$ denotes the orthogonal group. In this case, we can define an *scalar* Gramian $\gamma_t(\rho) := \sum_{s=0}^{t-1} |\rho|^{2s}$, so that $\Gamma_t := \gamma_t(\rho) \cdot I$. We now show that the estimation rate of the least squares estimator given above is optimal up to log factors for $|\rho| \le 1 - \mathcal{O}(d/T)$:

Theorem 3.2. *Fix $d \ge 2$, $\rho \in \mathbb{R}$, $\delta \in (0, 1/4)$, and $\epsilon \le \frac{\rho}{2048}$. Then, there exists a universal constant c_0 such for any estimator $\widehat{A}$,*

$$\sup_{O \in \mathscr{O}(d)} \mathbb{P}_{\rho O}\left[\left\|\widehat{A}(T) - \rho O\right\|_{\mathrm{op}} \ge \epsilon\right] \ge \delta \;\; \textit{for any } T \textit{ such that } \; T\gamma_T(\rho) \;\le\; \frac{c_0\left(d + \log\left(1/\delta\right)\right)}{\epsilon^2},$$

where $\mathscr{O}(d)$ is the orthogonal group of $d \times d$ real matrices.

This is proven in Simchowitz et al. [2018] as Theorem 2.3. We can interpret it by considering the following regimes:

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \ge \begin{cases} \Omega\left(\sqrt{\frac{(d+\log(1/\delta))\cdot(1-|\rho|)}{N}}\right) & \text{if } \; |\rho| \;\le\; 1 - \frac{1}{N}, \\ \Omega\left(\frac{\sqrt{d+\log(1/\delta)}}{T}\right) & \text{if } \; 1 - \frac{1}{N} < |\rho| < 1 + \frac{1}{N} \\ \Omega\left(\frac{\sqrt{d+\log(1/\delta)}}{N|\rho|^N}\right) & \text{if } \; 1 + \frac{1}{N} \le |\rho|. \end{cases}$$

Comparing to our corresponding upper bounds for scaled identity matrices, we see that for $|\rho| \le 1 - \mathcal{O}(d/N)$, our upper and lower bounds coincide up to logarithmic factors. In the regime $\rho \in [1 - \mathcal{O}(d/N), 1]$, our upper and lower bounds differ by a factor of $\mathcal{O}(\sqrt{d + \log(1/\delta)})$.

3.4 Proof of Results

We establish our results in in a more general sequential regression setting, where a response variable $\mathbf{y}_t$ is linear in a covariate $\mathbf{z}_t$, and perturbed by martingale subgaussian noise $\boldsymbol{\epsilon}_t$. We proceed to sketch the general setting, analysis, and culminate in a general bound,Theorem 3.3. Ommitted proofs are deferred to Section 3.5. Our setting is as follows.

Setting 3.1 (Martingale Least Squares)**.** Let $(\mathcal{F}_t)_{t\geq 1}$ denote a filtration, and let $(\mathbf{z}_t)_{t\geq 1}$ be a sequence in $\mathbb{R}^d$ which is adapted to the filtration. Moreover, let

$$\mathbf{y}_t = \theta_\star^\top \mathbf{z}_t + \boldsymbol{\epsilon}_t, \tag{3.15}$$

where $\theta_\star \in \mathbb{R}^{d\times m}$ is a fixed parameter, where $\mathbf{y}_t \in \mathbb{R}^m$ is called the *response*, $\boldsymbol{\epsilon}_t \in \mathbb{R}^m$ the *noise*, and where $\boldsymbol{\epsilon}_t \mid \mathcal{F}_{t-1}$ is σ^2-subgaussian, that is

$$\mathbb{E}[\exp(\lambda v^\top \boldsymbol{\epsilon}_t) \mid \mathcal{F}_t] \leq \exp\left(-\frac{\sigma^2\lambda^2\|v\|^2}{2}\right), \quad \forall v \in \mathbb{R}^d \tag{3.16}$$

The dynamics in Eq. (3.1) are a special case, with $\mathbf{z}_t \leftarrow \mathbf{x}_t$, $\mathbf{y}_t \leftarrow \mathbf{x}_{t+1}$, $\boldsymbol{\epsilon}_t \leftarrow \mathbf{w}_t$, and $\theta_\star = A_\star^\top$. Generalizing our system identification problem, we analyze the ordinary least squares estimator,

$$\widehat{\theta}_{\mathrm{ls}} = \min_{\theta\in\mathbb{R}^d} \sum_{t=1}^{N} \|\mathbf{y}_t - \theta^\top \mathbf{z}_t\|_2^2, \tag{3.17}$$

and aim to bound the operator norm error $\|\theta_\star - \widehat{\theta}_{\mathrm{ls}}\|_{\mathrm{op}}$. Our analysis of $\widehat{\theta}_{\mathrm{ls}}$ proceeds in two steps:

- First, we use the self-normalized martingale inequality Abbasi-Yadkori et al. [2011, Theorem 2] to establish that

$$\|\theta_\star - \widehat{\theta}_{\mathrm{ls}}\|_{\mathrm{op}}^2 \leq \widetilde{\mathcal{O}}\left(\frac{m+d}{\lambda_{\min}(\boldsymbol{\Lambda}_N}\right), \tag{3.18}$$

  where $\boldsymbol{\Lambda}_N$ is the empirical covariance of the covariates $(\mathbf{z}_t)$.

- Second, we impose a novel condition on the covariates called the *Block-Martingale Small-Ball* condition, or BMSB. Under the BMSB, we show that $\boldsymbol{\Lambda}_N$ can be lower bounded with high probability. Notably, this condition does not rely on mixing arguments.

Combing steps 1 and 2 yields a high-probability error bound on $\|\theta_\star - \widehat{\theta}_{\mathrm{ls}}\|_{\mathrm{op}}^2$, culminating Theorem 3.3. Finally, we demonstrate that the covariates $(\mathbf{x}_t)$ in the dynamical system (3.1) do indeed satisfy the BMSB, and derive Theorem 3.1 as a direct corollary of our more general result.

The Self-Normalized Bound

The first step in our proof is stating an error bound on $\widehat{\theta}_{\text{ls}}$ in terms of the empirical covariance matrix. Recall that the error of least squares estimate can be expressed as

$$\widehat{\theta}_{\text{ls}} - \theta_\star = \mathbf{\Lambda}_N^{-1} \left(\sum_{t=1}^{N} \mathbf{z}_t \boldsymbol{\epsilon}_t \right), \qquad \text{where } \mathbf{\Lambda}_N = \sum_{t=1}^{N} \mathbf{z}_t \mathbf{z}_t^\top \tag{3.19}$$

The right-hand side of Eq. (3.19) has been studied at length in the online learning community, and there is a standard result which controls its magnitude:

Lemma 3.3 (Self-Normalized Martingale Inequality, Theorem 2 in Abbasi-Yadkori et al. [2011]). *Fix any PSD matrix $V \succ 0$, and a confidence paramater $\delta > 0$, and suppose the response dimension is $m = 1$. Then, with probability $1 - \delta$,*

$$\left\| (\mathbf{\Lambda}_N + V)^{-1/2} \left(\textstyle\sum_{t=1}^{N} \mathbf{z}_t \boldsymbol{\epsilon}_t \right) \right\|_2^2 \le 2\sigma^2 \log \left(\tfrac{1}{\delta} \cdot \det(V^{-1/2}(V + \mathbf{\Lambda}_N)V^{-1/2}) \right). \tag{3.20}$$

In view of Eq. (3.19), the above lemma directly implies an error bound on $\widehat{\theta}_{\text{ls}}$ in the cause where $\mathbf{y}_t$ and $\boldsymbol{\epsilon}_t$ are scalar, i.e. $m = 1$. When $m \ge 1$, we can bootstrap the $m = 1$ case via a covering argument. This yields the following, intermediate error bound on $\widehat{\theta}_{\text{ls}}$ in terms of $\lambda_{\min}(\mathbf{\Lambda}_t)$:

Lemma 3.4. *Consider the Setting 3.1. For a PSD matrix $V \succeq 0$, and define the event $\mathcal{E}_{\succeq V} := \{\mathbf{\Lambda}_N \succeq V\}$. Then for all $\delta > 0$, the following bound holds with probability $1 - \delta$ on $\mathcal{E}_{\succeq V}$:*

- *If the response dimension m is equal to 1, (i.e. $\mathbf{y}_t \in \mathbb{R}$ for all t, and $\theta_\star \in \mathbb{R}^d$)*

$$\|\widehat{\theta}_{\text{ls}} - \theta_\star\|_2^2 \le \frac{4\sigma^2}{\lambda_{\min}(\mathbf{\Lambda}_N)} \left(\log \frac{1}{\delta} + \log\det(I + \mathbf{\Lambda}_N V^{-1}) \right) \tag{3.21}$$

- *For response dimensions $m \ge 1$,*

$$\|\widehat{\theta}_{\text{ls}} - \theta_\star\|_{\text{op}}^2 \le \frac{16\sigma^2}{\lambda_{\min}(\mathbf{\Lambda}_N)} \left(m \log 5 + \log \frac{1}{\delta} + \log\det(I + \mathbf{\Lambda}_N V^{-1}) \right) \tag{3.22}$$

For a sense of scaling, observe that if we take $V = \lambda I$, then we can bound $\log\det(I + \mathbf{\Lambda}_N V^{-1}) \le d \log \frac{\|\mathbf{\Lambda}_N\|_{\text{op}}}{\lambda}$, or $\tilde{\mathcal{O}}(d)$.

The Block-Martingale Small Ball Condition

In view of Lemma 3.4, it remains to lower bound $\lambda_{\min}(\mathbf{\Lambda}_N)$ with high probability. Of course, this requires some additional condition on the covariates $(\mathbf{z}_t)_{t \ge 1}$. Indeed, the degeneracy

$\mathbf{z}_1 = \mathbf{z}_2 = \cdots = \mathbf{z}_N$ would still satisfy the conditions of the martingale least squares setting, Setting 3.1, but clearly could make learning $\theta_\star$ impossible.

Suppose we aim to show that $\mathbf{\Lambda}_N \gtrsim N\Gamma$, where Γ is a target covariance matrix. The condition we formulate roughly says that, on any interval of steps $t \in \{j, j+1, \ldots, j+k-1\}$, there is a constant probability that the average magnitude of $\langle v, \mathbf{z}_t\rangle^2$ will exceed $v^\top \Gamma v$, conditioned on the past with at least constant probability. We term this property the *Block-Martingale Small-Ball Condition.*

Definition 3.3 (Block-Martingale Small-Ball Condition (BMSB))**.** Let $(m_t)_{t\geq 1}$ be a non-negative scalar sequence which is adapted to the filtration $(\mathcal{F}_t)$. We say (m_t) satisfies the (simplied) Block-Martingale Small-Ball condition (BMSB) with parameters (k, ν, p) if, almost surely, the following holds for all $j \in \mathbb{N}$,

$$\mathbb{P}\left[\frac{1}{k}\sum_{i=0}^{k-1} m_{j+i} \geq \nu \mid \mathcal{F}_{j-1}\right] \geq p.$$

We say that the vector valued sequence $(\mathbf{z}_t)_{t\geq 1}$ satisfies the BMSB ith parameter (k, Γ, p) if, for any fixed $v \in \mathbb{R}^d$ of norm $\|v\| = 1$, the non-negative-scalar sequence $m_{t;v} := \langle v, \mathbf{z}_t\rangle^2$ satisfies the BMSB condition with parameters $(k, v^\top \Gamma v, p)$.

The "small-ball" nomenclature is inspired by the seminal work of [Mendelson, 2014], who use a similar lower bound on covariates to derive excess risk bounds on the empirical risk minimizer in settings with possibly heavy-tailed covariates. The formulation in Definition 3.3 intentionally (but only slightly) differs from the original formulation of the BMSB in Simchowitz et al. [2018] in that it is somewhat more direct.

The key insight behind the condition is that, for any direction $v \in \mathbb{R}^d$, we do not need $v^\top \mathbf{\Lambda}_N v$ to concentrate around its expectation. Instead, it suffices to obtain a lower tail inequality, i.e. a lower bound on $v^\top \mathbf{\Lambda}_N v$. For any given k, we can write

$$v^\top \mathbf{\Lambda}_N v = \sum_{t=1}^{N} \langle v, \mathbf{z}_t\rangle^2 \geq \sum_{q=1}^{\lfloor N/k\rfloor - 1} \left(\sum_{t=k(q-1)+1}^{qk} \langle v, \mathbf{z}_t\rangle^2 \right).$$

Since each term $\langle v, \mathbf{z}_t\rangle^2$ is *nonnegative*, we can lower bound $v^\top \mathbf{\Lambda}_N v \gtrsim N \cdot v^\top \Gamma N v$ as long as we can argue that a constant number of length k-chunks in the paranthetical above are at least $k v^\top \Gamma v$. This is precisely what the BMSB condition affords, via a Chernoff bound. For simplicity, we state this consequence for nonnegative scalar sequences first:

Lemma 3.5. *Let $(m_t)_{t\geq 1}$ satisfy the scalar BMSM with parameters (k, ν, p). Then, for any $N \geq 4k$,*

$$\mathbb{P}\left[\sum_{t=1}^{N} m_t \geq \frac{pN}{4} \cdot \nu\right] \geq 1 - e^{\frac{-pN}{16k}}.$$

No pass from scalar lower bounds to PSD lower bounds, we require a covering argument. The following lemma contains the two essential ingredients, cited from Simchowitz et al. [2018] without proof.

Lemma 3.6. *Let $Q \in \mathbb{S}^d_{++}$ be positive definite, and let $\Lambda_0 \preceq \Lambda_1 \in \mathbb{S}^d_{++}$ be such that $Q \preceq \Lambda_1$. Define the ellipsoid $\mathscr{E}_{\Lambda_0} := \{w \in \mathbb{R}^d : w^\top \Lambda_0 w \leq 1\}$, and let $\mathcal{N}_{\Lambda_0 \to \Lambda_1}$ denote a 1/4 net of $\mathscr{E}_{\Lambda_0}$ in the metric induced by the norm $\|w\|_{\Lambda_1} = \sqrt{w^\top \Lambda_1 w}$, that is,*

$$\forall w \in \mathscr{E}_{\Lambda_0}, \quad \exists w' \in \mathcal{N}_{\Lambda_0 \to \Lambda_1} \ s.t. \ \|w - w'\|_{\Lambda_1} \leq 1/4. \tag{3.23}$$

Then,

(a) (Simchowitz et al. [2018, Lemma 4.1]) If $w^\top Q w \geq \Lambda_0$ for all $w \in \mathcal{N}_{\Lambda_0 \to \Lambda_1}$, then $Q \succeq \frac{1}{2}\Lambda_0$.

(b) (Simchowitz et al. [2018, Lemma D.1]) There exists such a net $\mathcal{N}_{\Lambda_0 \to \Lambda_1}$ of cardinality at most

$$\log |\mathcal{N}_{\Lambda_0 \to \Lambda_1}| \leq d \log 9 + \log \det(\Lambda_0^{-1} \Lambda_1).$$

In words, item (a) of the above lemma stimulates that it suffices to establish a PSD lower bound over the net $\mathcal{N}_{\Lambda_0 \to \Lambda_1}$ in order to establish such a bound over $\mathbb{R}^d$. Item (b) bounds the cardinality of this net roughly in terms of the ratio of the volumes of the ellipsoids induced by matrices Λ_0 and Λ_1. Leveraging the above covering bounds, we complete the BMSM argument.

Lemma 3.7. *Suppose that $(\mathbf{z}_t)_{t \geq 1}$ satisfy the BMSB condition with parameters (k, Γ, p). Given $\Gamma_+ \geq \Gamma$, and define the event $\mathcal{E}_+ := \{\mathbf{\Lambda}_N \preceq N\Gamma_+\}$. Then, for $N \geq \frac{32k}{p}(d \log \frac{36}{p} + \log \det(\Gamma_+^{-1}\Gamma))$, the following inequality holds:*

$$\mathbb{P}\left[\{\mathbf{\Lambda}_N \not\succeq \frac{pN}{8} \cdot \Gamma\} \cap \mathcal{E}_{\max}\right] \leq e^{-pN/32k}. \tag{3.24}$$

Proof. Define the matrix $\Lambda_0 := \frac{pN}{4}\Gamma$, and let $\Lambda_1 = \Gamma_+$. Let $\mathcal{N}$ be the net stipulated by Lemma 3.6, taken to have log-cardinality at most $d \log 9 + \log \det(\Lambda_1^{-1}\Lambda_0) = d \log(36/p) + \log \det(\Gamma_+^{-1}\Gamma)$. By Lemma 3.5 and a union bound bound, it holds that with probability at least

$$1 - \exp\left\{-\frac{pN}{16k} + d \log(36/p) + \log \det(\Gamma^{-1}\Gamma_+)\right\}. \tag{3.25}$$

that, for all $w \in \mathcal{N}$, $w^\top \mathbf{\Lambda}_N w \geq \frac{pN}{4} w^\top \Gamma w$. Lemma 3.6 then ensures that, on the intersection of this event, and the event $\{\mathbf{\Lambda}_N \preceq N\Gamma_+\}$, we have that $\mathbf{\Lambda}_N \succeq \frac{pN}{8}\Gamma$. Moreover, by our condition that $N \geq \frac{32k}{p}(d \log(36/p) + \log \det(\Gamma^{-1}\Gamma_+))$ ensures that the quantity in Eq. (3.25) is at least $1 - \exp(-\frac{pN}{32k})$. □

Note that Lemma 3.7 requires very weak concentration of $\mathbf{\Lambda}_N$. Indeed, consider scalar upper and lower bounds $\Gamma = \lambda_0 I$ and $\Gamma_+ = \lambda_1 I$. Then, the above condition on N reduces to

$$N \geq \frac{32kd}{p} \cdot \log \frac{36\lambda_1}{p\lambda_0},$$

which is at most *logarithmic* in the ration of λ_1 to λ_0.

A General Bounds for Martingale Regression

Putting together the consequence of the self-normalized tail bound (Lemma 3.4), and our covariance lower bound derived from the BMSB condition (Lemma 3.7), we derive a general bound for the martingale least squares in Setting 3.1.

Theorem 3.3. *Consider the setting Setting 3.1. Fix a PSD matrices $0 \preceq \Gamma \preceq \Gamma_+$. Suppose that $(\mathbf{z}_t)_{t\geq 1}$ satisfy the BMSB condition with parameters (k, Γ, p), and define the event $\mathcal{E}_+ := \{\frac{1}{N}\mathbf{\Lambda}_N \preceq \Gamma_+ I\}$, Then, for $N \geq N \geq \frac{32k}{p}(d \log \frac{36}{p} + \log\det(\Gamma^{-1}\Gamma_+))$,*

$$\mathbb{P}\left[\left\{\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_{\mathrm{op}}^2 \leq \frac{\sigma^2}{N} \cdot \frac{\mathcal{C}_\delta}{\lambda_{\min}(\Gamma)}\right\} \cap \mathcal{E}_+\right] \leq \delta + e^{-pN/32k},$$

where we define

$$\mathcal{C}_\delta = \frac{128}{p}\left(m \log 5 + \log\frac{1}{\delta} + \log\det(I + \tfrac{8}{p}\Gamma^{-1}\Gamma_+)\right).$$

Proof. Set $V = \frac{pN}{8}\Gamma$. Then, Lemma 3.7 ensures that the event $\mathcal{E}_{\succeq V} := \{\mathbf{\Lambda}_N \succeq V\}$ holds with probability at least $1 - e^{-pN/32k}$. The bound follows directly from Lemma 3.4, upper bounding $\mathbf{\Lambda}_N \preceq N\Gamma_+$ on $\mathcal{E}_+$. □

At the expense of only slight looseness, we can appeal to Markov's inequality to further simplify the above bound.

Lemma 3.8. *Let $\mathbf{Q} \in \mathbb{S}^d_{++}$ be a random positive definitve matrix. Then, with probability at least $1 - \delta$, $\mathbf{Q} \preceq \frac{d}{\delta}\mathbb{E}[\mathbf{Q}]$.*

Proof. If $\mathbf{Q} \preceq \frac{d}{\delta}\mathbb{E}[\mathbf{Q}]$ fails, then the operator norm, and hence the trace, of $\mathbb{E}[\mathbf{Q}]^{-1/2}\mathbf{Q}\mathbb{E}[\mathbf{Q}]^{-1/2}$ must be at least d/δ. By Markov's inequality, the probability this trace exceeds d/δ is at most $\frac{\delta}{d} \cdot \mathbb{E}[\mathrm{tr}(\mathbb{E}[\mathbf{Q}]^{-1/2}\mathbf{Q}\mathbb{E}[\mathbf{Q}]^{-1/2})]$. Swapping traces and expectations and simplify, the resultant quantity is exactly δ. □

In most cases, the top eigenvalue of $\mathbf{\Lambda}_N$ will enjoy far sharper concentration than that ensured by Markov's inequality. however, Corollary 3.2 is sharp enough for our purposes. Indeed, Lemma 3.8 directly yields the following corollary:

Corollary 3.2. *Consider the setting Setting 3.1. Suppose that $(\mathbf{z}_t)_{t\geq 1}$ satisfy the BMSB condition with parameters (k, Γ, p). Further, define $\bar{\Gamma} := \mathbb{E}[\mathbf{\Lambda}_N]/N$, and the effective dimension*

$$d_\delta := d \log \frac{36d}{p\delta} + \log\det(I + \Gamma^{-1}\bar{\Gamma}). \tag{3.26}$$

Then, as soon as $N \geq \frac{32k}{p} \cdot d_\delta$, we have

$$\mathbb{P}\left[\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_{\mathrm{op}}^2 \lesssim \frac{\sigma^2(m + d_\delta)}{pN\lambda_{\min}(\Gamma)}\right] \leq 3\delta$$

Note that when $\Gamma \succeq \lambda_0 I$ and $\bar{\Gamma} \preceq \lambda_1 I$, we can upper bound of $\log\det(I + \Gamma^{-1}\bar{\Gamma}) \leq d\log(1 + \frac{\lambda_1}{\lambda_0})$, so that d_δ depends at most *logarithmic* in the relative scaling of the matrices Γ and $\bar{\Gamma}$.

Specializing to System Identification

Let us conclude by specializing to system identification for the dynamics in Eq. (3.1). The reduction holds with the following substitutions:

$$\theta_\star \leftarrow A_\star, \text{ and } \forall t \geq 1, \quad \mathbf{z}_t \leftarrow \mathbf{x}_t, \quad \mathbf{y}_t \leftarrow \mathbf{x}_{t+1}, \quad \boldsymbol{\epsilon}_t \leftarrow \mathbf{w}_t. \tag{3.27}$$

No do so, we need to show that the sequence $(\mathbf{x}_t)$ satisfies a BMSM condition. We do so by verifying a general condition under which the BMSB condition holds:

Definition 3.4. We say that the sequence $(\mathbf{z}_t)_{t\geq 1}$ is α-Paley-Zygmud if, for all $j \geq 1$ and $i \geq 0$, and all $v \in \mathbb{R}^d$, $\mathbb{E}[\langle v, \mathbf{z}_{j+i}\rangle^4 \mid \mathcal{F}_{j-1}] \leq \alpha\mathbb{E}[\langle v, \mathbf{z}_{j+i}\rangle^2 \mid \mathcal{F}_{j-1}]^2$.

Intuitively, the Paley-Zygmud stays that the tails of the random variables $v^\top \mathbf{z}_t$ aren't too fat; thus, if their expectation is large, then they must be large with at least constant probably. Notatble, for the Gaussian noise model in Eq. (3.1) yields a Paley-Zygmud constant of 3:

Lemma 3.9. *The sequence $(\mathbf{x}_t)_{t\geq 1}$ in Eq. (3.1) is 3-Paley-Zygmud.*

Proof. Fix j, i and let $Y = \langle v, \mathbf{x}_{j+i}\rangle \mid \mathcal{F}_{j-1}$ denote the conditional distribution of $\langle v, \mathbf{x}_{j+i}\rangle$. The linear dynamics and Gaussian noise in Eq. (3.1) mean that Y is scalar Gaussian random variable, say with some mean and variance μ_Y and σ_Y^2. From the moment formula for Gaussian variables, $\mathbb{E}[Y^2] = \mu_Y^2 + \sigma_Y^2$, and $\mathbb{E}[Y^4] = \mu_Y^4 + 6\mu_Y^2\sigma_Y^2 + 3\sigma_Y^2 \leq 3(\mu_Y^2 + \sigma_Y^2)^2$. Hence, $\mathbb{E}[Y^4] \leq 3\mathbb{E}[Y^2]^2$, as needed. □

The following lemma now shows that for Paley-Zygmud sequences, the BMSB holds with parameter Γ

Lemma 3.10. *Suppose that, $(\mathbf{z}_t)_{t\geq 1}$ is α-Paley-Zygumd, and for a given $k \in \mathbb{N}$ and all $j \in \mathbb{N}$,*

$$\frac{1}{k}\sum_{i=0}^{k-1} \mathbb{E}\left[\mathbf{z}_{j+i}\mathbf{z}_{j+i}^{\top} \mid \mathcal{F}_{j-1}\right] \succeq \Gamma.$$

Then, $(\mathbf{z}_t)_{t\geq 1}$ satisfies the $(k, \frac{1}{2}\Gamma, \frac{1}{4\alpha})$ BMSB condition.

By combining Lemmas 3.9 and 3.10 and Corollary 3.2, we can prove Theorem 3.1:

Proof of Theorem 3.1. Recall that $(\mathcal{F}_n)_{n\geq 0}$ denotes the filtration generated by states $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. For any indices $i \geq 0$, $j \in \mathbb{N}$, the conditional distribution of the state $\mathbf{x}_{j+i} \mid \mathcal{F}_{j-1}$ is equal to the distribution of the sum of two random variables $\sigma_w(Y'_{i,j} + Y''_{i,j})$, and $Y''_i = \mathbb{E}[\mathbf{x}_{j+i} \mid \mathcal{F}_{j-1}]$, and where one can show that $Y'_i \sim \mathcal{N}(0, \Gamma_{i+1})$. In particular, $\mathbb{E}[Y'_{i,j}(Y''_{i,j})^{\top}] = 0$, and thus

$$\mathbb{E}\left[\mathbf{x}_{j+i}\mathbf{x}_{j+i}^{\top} \mid \mathcal{F}_{j-1}\right] = \sigma_w^2\mathbb{E}[Y'_i(Y'_i)^{\top}] + \sigma_w^2\mathbb{E}[Y''_i(Y''_i)^{\top}] \succeq \sigma_w^2\mathbb{E}[Y'_i(Y'_i)^{\top}] = \sigma_w^2\Gamma_{i+1}. \tag{3.28}$$

Therefore, we find the lower bounds

$$\frac{1}{k}\sum_{i=0}^{k-1} \mathbb{E}\left[\mathbf{x}_{j+i}\mathbf{x}_{j+i}^{\top} \mid \mathcal{F}_{j-1}\right] \succeq \sigma_w^2 \frac{1}{k}\sum_{i=0}^{k-1} \Gamma_{i+1} := \sigma_w^2 \cdot \tilde{\Gamma}_k.$$

From Lemma 3.9, the sequence $(\mathbf{x}_t)$ is 3-Paley Zygmud. Hence, applying Lemma 3.10 with $\Gamma \leftarrow \tilde{\Gamma}_k$, we mind that the sequence $(\mathbf{x}_t)$ satisfies the $(k, \frac{1}{4}\tilde{\Gamma}_k, 1/12)$-BMSM condition. We can also compute that

$$\mathbb{E}[\mathbf{\Sigma}_N] = \sum_{t=1}^{N} \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^{\top}] = \sum_{t=1}^{N} \Gamma_t = N \cdot \sigma_w^2\tilde{\Gamma}_N.$$

In particular, taking

$$d_{\delta,k,N} := d_x \log\frac{d_x}{\delta} + \log\det(I + \tilde{\Gamma}_k^{-1}\tilde{\Gamma}_N). \tag{3.29}$$

Applying Corollary 3.2 with $\Gamma \leftarrow \tilde{\Gamma}_k$ and $\Gamma_+ \leftarrow \tilde{\Gamma}_N$ implies that there are universal constants c_0, c_1 such that

$$\mathbb{P}\left[\|A_\star - \widehat{A}_{\mathrm{ls}}\|_{\mathrm{op}}^2 \leq c_1 \cdot \frac{d_{\delta,k,N}}{N\lambda_{\min}(\tilde{\Gamma}_k)}\right] \leq 3\delta,$$

provided that $N \geq c_0 k d_{\delta,k,N}$. By re-scaling constants, we can replace the 3δ with δ. This concludes the proof. □

3.5 Omitted Proofs

Proof of Lemma 3.4. On $\mathcal{E}_{\succ V}$, we have

$$\begin{aligned}\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2 &\le \tfrac{1}{\lambda_{\min}(\mathbf{\Lambda}_N)^{1/2}} \cdot \left\|\mathbf{\Lambda}_N^{-1/2}\left(\textstyle\sum_{t=1}^N \mathbf{z}_t\boldsymbol{\epsilon}_t\right)\right\|_2. \\ &\overset{(i)}{\le} \tfrac{\sqrt{2}}{\lambda_{\min}(\mathbf{\Lambda}_N)^{1/2}} \cdot \left\|(\mathbf{\Lambda}_N + V)^{-1/2}\left(\textstyle\sum_{t=1}^N \mathbf{z}_t\boldsymbol{\epsilon}_t\right)\right\|_2 \\ &\overset{(i)}{\le} \tfrac{2\sigma}{\lambda_{\min}(\mathbf{\Lambda}_N)^{1/2}} \cdot \sqrt{\log\left(\tfrac{1}{\delta}\cdot\det(V^{-1/2}(V+\mathbf{\Lambda}_N)V^{-1/2})\right)}.\end{aligned}$$

where (i) uses that on $\mathcal{E}_{\min}$, $V + \mathbf{\Lambda}_N \preceq 2\mathbf{\Lambda}_N$, and (ii) uses Lemma 3.3. We can then rewrite

$$\log\left(\tfrac{1}{\delta}\cdot\det(V^{-1/2}(V+\mathbf{\Lambda}_N)V^{-1/2})\right) = \log\tfrac{1}{\delta} + \log\det(I + V^{-1}\mathbf{\Lambda}_N).$$

For $m \ge 1$, we use a covering argument. Let $\mathcal{N}$ denote a $1/2$ net of $\mathbb{R}^m$, which can be taken to have cardinality $\log|\mathcal{N}| \le m\log 5$ by .... By a standard covering argument (cite),

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_{\mathrm{op}} = \max_{v\in\mathcal{S}^{m-1}} \|(\widehat{\theta}_{\mathrm{ls}} - \theta_\star)v\|_2 \le 2\max_{v\in\mathcal{N}} \|(\widehat{\theta}_{\mathrm{ls}} - \theta_\star)v\|_2 \tag{3.30}$$

We observe that, by the $m = 1$ case, we have that with probability $1-\delta$,

$$\|(\widehat{\theta}_{\mathrm{ls}} - \theta_\star)v\|_2^2 \le \frac{4\sigma^2}{\lambda_{\min}(\mathbf{\Lambda}_N)}\left(\log\frac{1}{\delta} + \log\det(I + V^{-1}\mathbf{\Lambda}_N)\right)$$

Naking a union bound over $v \in \mathcal{N}$, and applying Eq. (3.30) concludes. □

Proof Lemma 3.5. Let $(m_t)_{t\ge1}$ satisfy the BMSM with parameters (k, ν, p). For $q \in \{1, 2, \dots, q_{\max}\}$, where $q_{\max} := \lfloor \frac{N}{k} \rfloor - 1$, define the random indicators and filtration $\tilde{\mathcal{F}}_q$

$$B_q := \mathbb{I}\left\{\sum_{i=1}^k m_{j+i-1} \ge k\nu\right\}, \tilde{\mathcal{F}}_q = \mathcal{F}_{(q+1)k}$$

Then, (B_q) is $\tilde{\mathcal{F}}_q$-adapted, and $\mathbb{E}[B_q \mid \tilde{\mathcal{F}}_{q-1}] \ge p$. By a Martingale Chernoff bound,

$$\mathbb{P}\left[\sum_{q=1}^{q_{\max}} B_q \ge \frac{pq_{\max}}{2}\right] \ge 1 - e^{-\frac{pq_{\max}}{8}} = 1 - e^{-\frac{p\lfloor N/k-1\rfloor}{8}} \ge 1 - e^{\frac{N/k-2}{8}}.$$ [2]

On the other hand, if $\sum_{q=1}^{q_{\max}} B_q \ge \frac{\nu q_{\max}}{2}$, then by nonnegativity of the (m_t) sequence,

$$\sum_{t=1}^N m_t \ge \sum_{q=1}^{q_{\max}}\sum_{i=1}^k m_{qk+i-1} \ge \frac{q_{\max}\nu}{2}\cdot k = \frac{\nu}{2}\cdot k(\lfloor N/k\rfloor - 1) \ge \frac{\nu}{2}(N-2k).$$

Hence,

$$\mathbb{P}\left[\sum_{t=1}^N m_t \ge \frac{\nu p}{2}(N-2k)\right] \ge 1 - e^{p\frac{N/k-2}{8}}.$$

Naking $N \ge 4k$ concludes. □

Proof of Lemma 3.10. Fix an index $j \in \mathbb{N}$, and a vector $v \in \mathbb{R}^d$. For $i = 0, 1, \ldots, k-1$, introduce the random variables $Y_i = \langle v, \mathbf{z}_{j+i} \rangle^2$. Also, introduce the shorthand $\mathbb{E}_j[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{j-1}]$, and $\mathbb{P}_j[\cdot]$ analogously.

By the (unconditional) Paley-Zygmud inequality, the following holds for any $\nu \geq 0$:

$$\mathbb{P}_j \left[\sum_{i=0}^{k-1} Y_i \geq \frac{1}{2} \mathbb{E}_j \left[\sum_{i=0}^{k-1} Y_i \right] \right] \geq \frac{1}{4} \frac{\mathbb{E}_j[\sum_{i=0}^{k-1} Y_i]^2}{\mathbb{E}_j[(\sum_{i=0}^{k-1} Y_i)^2]}.$$

Moreover, since $(\mathbf{z}_t)_{t \geq 1}$ are α-Paley Zygmud, $\mathbb{E}_j[Y_i^2] \leq \alpha \mathbb{E}_j[Y_i]^2$, and thus, by Cauchy-Schwartz,

$$\mathbb{E}_j[(\sum_{i=0}^{k-1} Y_i)^2] = \sum_{i,i'=0}^{k-1} \mathbb{E}_j[Y_i Y_{i'}] \leq \sum_{i,i'=0}^{k-1} \sqrt{\mathbb{E}_j[Y_i^2] \mathbb{E}_j[Y_{i'}^2]} \leq \alpha \sum_{i,i'=0}^{k-1} \mathbb{E}_j[Y_i] \mathbb{E}_j[Y_{i'}].$$

The right most term is precisely $\alpha \mathbb{E}_j[\sum_{i=0}^{k-1} Y_i]^2$. Hence, combining the above two displays,

$$\mathbb{P}_j \left[\sum_{i=0}^{k-1} Y_i \geq \frac{1}{2} \mathbb{E}_j \left[\sum_{i=0}^{k-1} Y_i \right] \right] \geq \frac{1}{4\alpha}.$$

Moreover, by the assumption of the lemma,

$$\mathbb{E}_j \left[\sum_{i=0}^{k-1} Y_i \right] = v^\top \mathbb{E}_j \left[\sum_{i=0}^{k-1} \mathbf{z}_{j+i} \mathbf{z}_{j+i}^\top \right] v \geq v^\top (k\Gamma) v,$$

where the last line invokes the assumption of the lemma. Thus, for all vectors $v \in \mathbb{R}^d$ and indices $j \in \mathbb{N}$,

$$\mathbb{P}_j \left[\frac{1}{k} \sum_{i=0}^{k-1} Y_i \geq \frac{1}{2} v^\top \Gamma v \right] \geq \frac{1}{4\alpha}.$$

This is precisely the definition of the $(k, \frac{1}{2}\Gamma, \frac{1}{4\alpha})$ BMSB. □

Chapter 4

SysID under Partial Observation

The previous chapter studied a relatively benign setting of system identification under i.i.d. Gaussian noise and with fully observed system states. In this chapter, we aim to understand if the same results are attainable more generally: with partially observed states and arbitrary noise models. These generalizations are necessary to capture many control systems of interest, where states are not fully observed, and noise may be worst case [Zhou et al., 1996].

We demonstrate that the dynamics of such systems can still be estimated, provided the spectral radius of the true dynamical matrix $A_\star$ is less than or *equal to* 1: $\rho(A_\star) \leq 1$. Thus, we find that even in this more expansive learning setting, lack of stability does not preclude system identification. Unlike the previous chapter, however, we do not identify cases when the sample complexity *improves* as the system becomes less stable.

Organization

Section 4.1 describes our formal learning setting, defining partially observed system dynamics and introducing an adversarial noise model for learning. Section 4.2 presents an analysis of the ordinary least-squares estimator in this setting with i.i.d. Rademacher inputs, and demonstrates consistency for strictly stable ($\rho(A_\star) < 1$) systems. The bound is given formally in Theorem 4.1 and subsequent corollaries.

We find that our analysis of least squares breaks down for marginally stable systems ($\rho(A_\star) = 1$) due to accumulation of disturbances and past inputs. Section 4.3 presents a two-stage least squares algorithm we call *pre-filtered least squares*, which implements an initial preprocessing phase to mitigate these effects. Theorem 4.2 analyzes this estimator, and establish learning rates in terms of a certain "oracle error", which describes how well noise accumulation can be predicted from prior observations. Section 4.4 discusses many choices for bounding the oracle error, including by relating the bound to the performance of an optimal Kalman filter. The discussion is left informal, but demonstrations that the oracle error is well behaved even for marginally stable systems.

The remaining two sections of the chapter prove Theorem 4.1 and Theorem 4.2, respectively.

4.1 Formal problem setting

In this chapter, we consider the problem of system identification of an unknown linear dynamical system under partial observation. Starting from initial state $\mathbf{x}_1 = 0$, we consider the following dynamics for time steps $1 \le t \le N$:

$$\begin{aligned} \mathbf{x}_{t+1} &= A_\star \mathbf{x}_t + B_\star \mathbf{u}_t + \mathbf{w}_t \\ \mathbf{y}_t &= C_\star \mathbf{x}_t + D_\star \mathbf{u}_t + \mathbf{e}_t, \end{aligned} \tag{4.1}$$

Under partial observation, the system state $\mathbf{x}_t$ remains hidden, and the learner observes only the inputs $\mathbf{u}_t$ and outputs $\mathbf{y}_t$. Here, we call $\mathbf{w}_t$ the *process noise*, which affects the evolution of the hidden state, and $\mathbf{e}_t$ the *observation noise*, which perturbs the observed outputs.

Markov Parameters

Our goal is to recover the first p Markov paramters in the operator norm:

$$G_{\star;p} := [D_\star \mid B_\star C_\star \mid C_\star A_\star B_\star \mid \dots C_\star A_\star^{p-2} B_\star]. \tag{4.2}$$

The recovery of the Markov parameters is sufficient to recover the system matrices $(A_\star, B_\star, C_\star, D_\star)$ via the Ho-Kalman [Ho and Kalman, 1965] algorithm. Quantitative sensitivity analysis was provided by Oymak and Ozay [2019], and refined to order-optimal rates by Sarkar et al. [2019]. For simplicity, we assume that the total sample length N is divisible by p.

Noise and Input Model

We assume that the disturbances $(\mathbf{w}_t)$ and $(\mathbf{e}_t)$ are selected by an *oblivious* adversary, which means they are selected to be an arbitrary sequence, but are selected without knowledge of the inputs $\mathbf{u}_t$ selected by the adversary. The adversary may randomize in its selection of the disturbance, and we let $\mathcal{F}_0$ denote the sigma-algebra generated by their possibly random selection. We assume that the noise terms satisfy a uniform bound in ℓ_2:

Assumption 4.1. We assume that, for all $1 \le t \le N$, $\|\mathbf{e}_t\|_2 \le B$ and $\|\mathbf{w}_t\| \le B$.

Our upper bounds hold independently of this assumption, but Assumption 4.1 is useful to obtain intuition for their scaling. Note that for stochastic noise with sub-Gaussian tails, B can be chosen to grow as $\sqrt{\log N}$ with the rollout length.

The inputs are up to the learner's discretion to select. For simplicity, we select inputs to be independent, Rademacher random variables[1]

$$\mathbf{u}_1, \dots, \mathbf{u}_N \overset{\text{i.i.d}}{\sim} \mathrm{Unif}(\{-1, 1\}^{d_u}). \tag{4.3}$$

[1]The analysis extends to when $\mathbf{u}_t$ are scaled Rademacher random variables; scaling may be desirable to ensure inputs and disturbance magnitudes are on the same order.

Crucially, the random input are uncorrelated with the adversarial disturbaces; as explained below, this facillitates consistent estimation of the Markov parameters even with non-mean-zero noise. We remark that the inputs in Simchowitz et al. [2019] were selected to be *Gaussian*; the random signed inputs chosen here slightly simplify the analysis because they are bounded uniformly.

Together, the noise an input model yield a natural filtration structure. Recalling that $\mathcal{F}_0$ is the sigma-algebra generated by the noise, we let $\mathcal{F}_t$ denote the sigma-algebra geneted by $\mathcal{F}_0$ and the inputs $\mathbf{u}_1, \dots, \mathbf{u}_t$. Note that the sequence $(\mathcal{F}_t)_{t \geq 1}$ form a filtration.

Learning without mixing (again)

As in Chapter 3, we aim to find learning rates for $G_{\star;p}$ which *do not* require that $A_\star$ is stable matrix (see Section 3.1 for a review of the spectral radius and stability). In particular, we cannot appear to the decay powers of $A_\star$.

For the fully observed setting considered in Chapter 3, we found that the absence of stability was little hinderance, because the learning rates were determined by the magnitude of the covariance matrix of states in the trajectory. In fact, for the special case of $A_\star = \rho I$, we found that performance of the least squares estimator improved as $\rho \to 1$.

With only partial settings, we do not directly observe the covariance matrix of states, and hence it is not clear if it is possible to achieve faster learning rates with less stable systems. Moreover, the presence of potentially adversarially noise precludes method-of-moment based approaches (e.g. Yule-Walker, Shumway et al. [2000]), which might be able to take advantage of greater levels of excitation. Therefore, we content ourselves with the following goal.

> *Is it possible to learn the Markov operator of a possibly marginally stable system, even under potentially adversarially disturbances?*

4.2 Learning via Least Squares

In this section, we explain how to use least squares estimation to recover the Markov operator. This leverages an important *semi-parametric* relationship detailed below, where the possibly adversarial noise is uncorrelated with the Rademacher inputs. Leveraging this observation, we establish a general guarantee Theorem 4.1, which enables $\widetilde{\mathcal{O}}\left(1/\sqrt{N}\right)$ estimation rates whenever the noise terms are uniformly bounded, and the system is strictly stable ($\rho(A_\star) < 1$). Extension to unstable systems requires a new algorithmic tool, and is discussed in the following section.

Semi-Parametric Relationship

For any time t, Markov parameters describe the linear relation between the output $\mathbf{y}_t$, and the past inputs $t, t-1, \dots, t-p+1$. Specifically, define $\mathbf{u}_{t:t-p+1} = (\mathbf{u}_t, \mathbf{u}_{t-1}, \dots, \mathbf{u}_{t-p+1})$ as

the concatenation of the past p inputs. Further, define

$$\mathbf{y}_t^{\text{nat}} := \mathbf{e}_t + \sum_{i=1}^{t-1} C_\star A_\star^{t-i-1} \mathbf{w}_t, \quad \mathbf{y}_t^{\mathbf{u};p} = \sum_{i=1}^{t-p} C_\star A_\star^{t-i-1} B_\star \mathbf{u}_t, \quad \boldsymbol{\delta}_t = \mathbf{y}_t^{\text{nat}} + \mathbf{y}_t^{\mathbf{u};p}$$

The first term, $\mathbf{y}_t^{\text{nat}}$, we call *Nature's Y* because it represents the response of the linear dynamical system (4.1) to noise, in the absense of input. The term $\mathbf{y}_t^{\mathbf{u};p}$ captures the contributions of past inputs before times $t-p$. Defining $\boldsymbol{\delta}_t$ as the sum of these two sources of error, we have

$$\mathbf{y}_t = G_{\star;p} \mathbf{u}_{t:t-p+1} + \boldsymbol{\delta}_t, \quad \forall t \geq p. \tag{4.4}$$

Observe that $\boldsymbol{\delta}_t$ depends only on past disturbances, and on $\mathbf{u}_1, \ldots, \mathbf{u}_{t-p}$. On the other hand, $\mathbf{u}_{t:t-p+1}$ depends on inputs $\mathbf{u}_{t-p+1}, \ldots, \mathbf{u}_t$. Thus, for the choice of Rademacher inputs in Eq. (4.3), the error terms $\boldsymbol{\delta}_t$ are independent of the inputs being regressed upon:

$$\mathbb{E}[\boldsymbol{\delta}_t \mathbf{u}_{t:t-p+1}^\top] = 0. \tag{4.5}$$

This is called an *semi-parametric* relationship (see e.g. Chernozhukov et al. [2016] or Krishnamurthy et al. [2018]), because the disturbances (which partly constitute $\boldsymbol{\delta}_t$) need not be mean zero, but are nevertheless uncorrelated with the regressor inputs due to Eq. (4.3). The semi-parameteric structure is ubiquitous in the economics community, where the random inputs to the system are an "instrumental variable".

Semi-Parametric Least Squares

As has been observed in prior work, the semi-parametric relationship faciliates efficient estimation via least squares, provided that $\boldsymbol{\delta}_t$ satisfy some uniform bound. Specifically, consider the least squares estimator

$$\widehat{G}_{\text{LS}} := \min_{G \in \mathbb{R}^{d_y \times p d_u}} \sum_{t=p+1}^{N} \|\mathbf{y}_t - G \mathbf{u}_{t:t-p+1}\|_2^2. \tag{4.6}$$

Further, let $\boldsymbol{\Delta}$ denote the matrix whose rows are $\boldsymbol{\delta}_{p+1}, \boldsymbol{\delta}_{p+2}, \ldots, \boldsymbol{\delta}_N$. We show that this estimate enjoys the following guarantee:

Theorem 4.1. *Given failure probability $\delta \in (0,1)$ and parameter $\lambda > 0$, suppose $N \geq N_0(\delta) := 8 d_u p^2 \log(p^2 d_u/\delta)$. Then, the following holds probability at least $1-\delta$,*

$$\|\widehat{G}_{\text{LS}} - G_{\star;p}\|_{\text{op}} \lesssim \frac{(\lambda + \|\boldsymbol{\Delta}\|_{\text{op}}) \sqrt{p(\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}} + p d_u + d_y)}}{N}, \tag{4.7}$$

where $\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}} := \log(1 + \frac{\|\boldsymbol{\Delta}\|_{\text{op}}^2}{\lambda}) + \log(1/\delta)$ is a logarithmic factor.

Proof Sketch. The above theorem is proved Section 4.5; it follows by invoking a standard decomposition for the least squares error, where the covariance matrix is given in terms on the rademacher inputs $\mathbf{u}_{t:t-p+1}$ and the error terms $\boldsymbol{\delta}_t$. We show that the minimum eigenvalue of the covariance is $\Omega(N)$ with high probability, taking care to account for the fact that the covariates are correlated (due to the fact that, e.g. $\mathbf{u}_{t:t-p+1}$ and $\mathbf{u}_{t+1:t-p}$ overlap). It remains to handle the correlation between covariates $\mathbf{u}_{t:t-p+1}$ and $\boldsymbol{\delta}_t$. Here, invoke use the semi-parametric relationship Eq. (4.5) to show that it is a mean-zero random process. More precisely, we can show that it scales as $\mathcal{O}\left(\|\boldsymbol{\Delta}\|_{\mathrm{op}}\right)$ using the martingale subgaussian bounded encoured in Chapter 3. □

For a sense of scaling, consider what happens when $\|\boldsymbol{\delta}_t\| \leq R$ for all t. In that case, we achieve the following corollary by setting λ to be the upper bound $\|\boldsymbol{\Delta}\|_{\mathrm{op}} \leq R\sqrt{N}$:

Corollary 4.1. *Let $\delta \in (0,1)$, suppose that $N \geq N_0(\delta)$, and assume further that, for all $t \in [N]$, $\|\boldsymbol{\delta}_t\| \leq R$. Then, with probability $1-\delta$,*

$$\|\widehat{G}_{\mathrm{LS}} - G_{\star;p}\|_{\mathrm{op}} \lesssim R\sqrt{\frac{p(\log(1/\delta) + pd_u + d_y)}{N}}. \tag{4.8}$$

Thus, when $\boldsymbol{\delta}_t$ are uniformly bounded, we recover the standard $1/\sqrt{N}$ rate, up to dimension factors. This is true for stable systems, provided the noise terms are bounded. The following proposition follows from direct computation.

Proposition 4.1. *Suppose that, for all $k \geq 0$, $\|A^k\| \leq \mathcal{C}_A \rho^k$ for a given $\rho \in (0,1)$. Further define, the parameter $\mathcal{C}_1 := \max\{\|B_\star\|_{\mathrm{op}}, \|C_\star\|_{\mathrm{op}}, \|B_\star\|_{\mathrm{op}} \cdot \|C_\star\|_{\mathrm{op}}, 1\}$. Then,*

$$\max_t \|\boldsymbol{\delta}\|_t \leq \mathcal{C}_1(\max_t \|\mathbf{e}_t\|_2 + \frac{\mathcal{C}_A}{1-\rho}(1 + \max_t \|\mathbf{w}_t\|))$$

On the other hand, in marginally stable systems $\|A^k\|$ does not decay to zero, and consequently $\|\boldsymbol{\delta}_t\|$ may grow with time as noise accumlates. Indeed, if $A = 1$ is scalar and $\mathbf{w}_t = 1$ for all 1, then $\boldsymbol{\delta}_t$ can grow as $\Omega(t)$, making the above bound vacuous. This challenged is adressed in Section 4.3.

Remark 4.1. Theorem 4.1 is less sharp that the corresponding result in corresponding guarantee in Simchowitz et al. [2019], both in terms of the p dependence in the error bound, and the minimal sample size N_0. Both improvements rely on careful chaining arguments - the latter due to a lower bound on the spectrum of the outer product of circulant matrices (due to Oymak and Ozay [2019], established for Gaussian inputs $\mathbf{u}_t$), and the former due to a similar chaining bound for handling martingale error terms Simchowitz et al. [2019, Appendix E]. In contrast, the bound presented in this thesis admits a rather brief and self-contained proof, which we give now.

4.3 Learning without mixing or full observation

Inspecting Corollary 4.1, the bound becomes vacuous as soon as the upper bound R on $\max_t \|\boldsymbol{\delta}_t\|$ scales as $\sqrt{N}$. While this is not a worry when $\rho(A) < 1$ (in view of Proposition 4.1), the problem arises readily when $\rho(A) = 1$.

Example 4.1. Consider the most benign scalar case, where $A = B = C = 1$, the process noise $\mathbf{e}_t \equiv 0$ is identically zero, and $\mathbf{w}_t \overset{\text{i.i.d}}{\sim} \text{Unif}\{(-1, 1)\}$. Then, for $t \geq p$, $\boldsymbol{\delta}_t$ is the sum of $2t - p$ Rademacher random variables, and thus its variance scales as $\Omega(t)$. Hence, with constant probability, each $\boldsymbol{\delta}_t$ is $\Omega(\sqrt{t})$, which is on the order of $\sqrt{N}$ for, say, all $t \geq N/2$.

In what follows, we present a two-stage estimator called *pre-filtered* least squares, which we show is consistent even when the spectral radius of $A_\star$ is one.

Pre-Filtered Least Squares

To adress the challenge of noise accumulation, we propose a simple prefiltering step. To begin, let us designate a vector of features $\mathbf{k}_t$. For concreteness, we select features as the concatenated sequence of the past inputs

$$\mathbf{k}_t = (\mathbf{y}_{t-p}, \mathbf{y}_{t-p-1}, \ldots, \mathbf{y}_{t-p-(L-1)}) \in \mathbb{R}^{d_y \times L}, \quad t \geq p + L., \tag{4.9}$$

though in principle, the following analysis applies to any choice of features for which $\mathbf{k}_t$ is $\mathcal{F}_{t-p}$ measurable for all t.

The first stage of the prefiltering algorithm predicts outputs $\mathbf{y}_t$ from the features $\mathbf{k}_t$. For simplicity, we select a linear predictor:

$$\hat{\phi} = \min_{\phi \in \mathbb{R}^{d_y \times L}} \sum_{t=p+L}^{N} \|\mathbf{y}_t - \phi \cdot \mathbf{k}_t\|_2^2 + \mu \|\phi\|_{\text{F}}^2. \tag{4.10}$$

The goal of $\hat{\phi}$ is to coarsely predict $\mathbf{y}_t$ from the past features $\mathbf{k}_t$. Since $\mathbf{k}_t$ are $\mathcal{F}_{t-p}$ measurable, $\hat{\phi}$ picks out the $\mathcal{F}_{t-p}$-measurable part of $\mathbf{y}_t$, captures most of the magnitude of $\boldsymbol{\delta}_t$ when the system is marginally stable.

The second stage of the prefiltering algorithm is to train a least squares classifer to the residual differnces between $\mathbf{y}_t$ and the course prefiltering estimate, $\hat{\phi} \cdot \mathbf{k}_t$:

$$\widehat{G}_{\text{pf}} \in \operatorname*{arg\,min}_{G \in \mathbb{R}^{d_y \times p d_u}} \sum_{t=p+L}^{N} \|(\mathbf{y}_t - \hat{\phi} \cdot \mathbf{k}_t) - G \cdot \mathbf{u}_{t:t-p+1}\|_2^2. \tag{4.11}$$

Estimation Rates for Prefiltered Least Squares

Going forward, we overload notation slightly from Section 4.2, and let $\boldsymbol{\Delta}$ denote the matrix whose rows are the errors $\boldsymbol{\delta}_t$, and let $\mathbf{K}$ denote the analogous matrix with rows $\mathbf{k}_t$. We let $\boldsymbol{\Delta}_\phi := \boldsymbol{\Delta} - \phi^\top \cdot \mathbf{K}$ denote the matrix whose rows are the residuals $\boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t$. To state the bound, we require the following terms.

Oracle Error First, we define Opt_μ as the best-possible (regularized) operator error for predicting the errors $\boldsymbol{\delta}_t$ from features $\mathbf{k}_t$:

$$\mathsf{Opt}_\mu := \inf_{\phi \in \mathbb{R}^{d_y \times L}} \|\boldsymbol{\Delta} - \phi^\top \cdot \mathbf{K}\|_{\mathrm{op}}^2 + \mu \|\phi\|_{\mathrm{op}}^2. \tag{4.12}$$

We call Opt_μ the *oracle error*, because it represents the performance of a predictor ϕ selected by an oracle with access to the error terms $\boldsymbol{\delta}_t$. In general, $\mathsf{Opt}_\mu = \Omega(\sqrt{N})$. We will show that the term is in fact always scales as $\widetilde{\mathcal{O}}\left(\sqrt{N}\right)$. Note that Opt_μ is highly-data dependent; it represents the best prediction of the errors $\boldsymbol{\delta}_t$ for the *realized* noise sequence.

Overfitting Error The second term in our analysis is the *overfitting error* . It arises from the fact that the predictor $\hat{\phi}$ in the prefiltering step is trained not to $\boldsymbol{\delta}_t$ but to $\mathbf{y}_t$:

$$\mathsf{Ovfit}_{\mu,\delta} := \|G_{\star;p}\|_{\mathrm{op}} \cdot p^2 \left(\log\det(1 + \tfrac{\mathbf{K}^\top \mathbf{K}}{\mu}) + d_u + \log \tfrac{1}{\delta}\right) \tag{4.13}$$

Importantly, the overfitting term is typically much smaller than the oracle error Opt_μ. Indeed, for non-unstable steps, $\|G_{\star;p}\|_{\mathrm{op}} = \mathrm{poly}(p)$ and $\|\mathbf{K}\|_{\mathrm{op}} = \mathcal{O}\left(\mathrm{poly}(N)\right)$, provided the disturbances $\mathbf{w}_t, \mathbf{e}_t$ are uniformly bounded as in Assumption 4.1.[2], and thus $\mathsf{Ovfit}_{\mu,\delta}$ grows at most logarithmicaly in the time horizon.

Statement of the bound Finally, we state the dimension-quantity that arises in the analysis:

$$\dim_{\mathrm{eff}} := p^2 d_u + p \log \tfrac{1}{\delta} + p(1 + L d_y) \log_+ \frac{\|\boldsymbol{\Delta}_{\hat{\phi}}\|_{\mathrm{op}}^2 + \sqrt{N}\|\mathbf{K}\|_{\mathrm{op}} + \mu^{-1/2}\|\mathbf{Y}\|_{\mathrm{F}}}{\lambda} \tag{4.14}$$

Again, note that the logarithmic factors grow at most logarithmically in N for non-unstable systems. Our bound is as follows:

Theorem 4.2. *For the matrix $\boldsymbol{\Delta}_{\hat{\phi}}$ whose rows are the residuals $\mathbf{y}_t - \boldsymbol{\delta}_t \cdot \hat{\phi}$, the following bounds hold together with probability $1 - \delta$:*

(a) $\boldsymbol{\Delta}_{\hat{\phi}} \le \mathsf{Ovfit}_{\mu,\delta} + \mathsf{Opt}_\mu$

(b) The error of the prefiltered least squares estimator is bounded by

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\mathrm{op}} \lesssim \left(\lambda + \mathsf{Opt}_\mu + \mathsf{Ovfit}_{\mu,\delta,\mathbf{K}}\right) \frac{\sqrt{\dim_{\mathrm{eff}}}}{N}$$

In particular, if Assumption 4.1, then letting $\widetilde{\mathcal{O}}(\cdot)$ suppress all but polynomial dependence in the problem horizon and selecting $\lambda = 1$, we have

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\mathrm{op}} = \widetilde{\mathcal{O}}\left(\frac{\mathsf{Opt}_\mu}{N}\right). \tag{4.15}$$

[2]This can be verified by invoking the fact that $\|A_\star^n\|_{\mathrm{op}} \le \mathrm{poly}(n)$ for $\rho(A_\star) \le 1$, established in Lemma 3.1. Moreover, this only requires Assumption 4.1 to hold for $B = \mathrm{poly}(N)$

Proof Sketch. The proof of Theorem 4.2 is deferred to Section 4.6. It involves three main steps. First, it leverages the non-prefiltered bound to establish an error estimate on the least squares regressor trained to residuals $\mathbf{y}_t - \phi \cdot \mathbf{k}_t$ for ϕ *fixed*. Then, to handle the adaptively selected prefilter $\hat{\phi}$, it establishes a uniform bound over all predictors ϕ in terms of the norm of their errors $\|\boldsymbol{\Delta}_\phi\|_{\text{op}} = \|\boldsymbol{\Delta} - \phi^\top \cdot \mathbf{K}\|_{\text{op}}$. Finally, we show that $\|\boldsymbol{\Delta}_{\hat{\phi}}\|_{\text{op}}$ can be bounded by the sum of Opt_μ and the overfit term. □

4.4 Bounding the oracle error

As per the above discussion, the overfitting error $\mathsf{Ovfit}_{\mu,\delta}$ grows at most logarithmically in N under reasonable assumptions on the noise. Hence, to establish consistent estimation rates, it remains to understand the oracle error term Opt_μ. Ideally, want to ensure that Opt_μ scales as

$$\mathsf{Opt}_\mu \sim \sqrt{N},$$

because this translates into $\widetilde{\mathcal{O}}\left(1/\sqrt{N}\right)$ estimation rates via Theorem 4.2. For this, it suffices that there exists a predictor ϕ of reasonable norm which predicts the error terms with constant accuracy:

$$\exists \phi : \forall t, \quad \|\boldsymbol{\delta}_t - \phi \cdot \mathbf{y}_t\|_{\text{op}} = \mathcal{O}(1). \tag{4.16}$$

In this chapter, we provide (somewhat informal) constructions of predictors ϕ which witness the above guarantee.

To facilitate the analysis, define the terms for $1 \le s \le t$

$$\mathbf{x}_{t|s} = A_\star^{t-s}\mathbf{x}_s, \quad \mathbf{y}_{t|s} = C_\star \mathbf{x}_{t|s} \tag{4.17}$$

In works, $\mathbf{x}_{t|s}$ and $\mathbf{y}_{t|s}$ correspond to the state and output that would arise if, for all times $s' \ge s$, the disturbances $\mathbf{w}_{s'}$ and $\mathbf{e}_{s'}$ and the inputs $\mathbf{u}_{s'}$ are indentically zero. Set $q = p + L$, and define the feature vector

$$\mathbf{k}_{t|t-q} := (\mathbf{y}_{t-p|t-q}, \mathbf{y}_{t-p-1|t-q}, \mathbf{y}_{t-q|t-q}) \in \mathbb{R}^{Ld_y}. \tag{4.18}$$

We can represent the error of a predictor ϕ as

$$\boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t = (\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q}) + \phi \cdot (\mathbf{k}_{t|t-q} - \mathbf{k}_t) + (\boldsymbol{\delta}_t - \mathbf{y}_{t|t-q}). \tag{4.19}$$

Using Lemma 3.1, one can verify the following estimate:

Lemma 4.2. *If $\rho(A_\star) \le 1$, and the error terms $\max_t \|\mathbf{w}_t\|, \|\mathbf{e}_t\| \le B$ (Assumption 4.1), then there exists constants $c_1, c_2 > 0$ depending on the system such that $\|\mathbf{k}_{t|t-q} - \mathbf{k}_t\|$ and $\|\boldsymbol{\delta}_t - \mathbf{y}_{t|t-q}\|$ are bounded by at most $(1+B)c_1(p+L)^{c_2}$. Thus,*

$$\boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t \le \|\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q}\| + (1 + \|\phi\|_{\text{op}})(1+B)c_1(p+L)^{c_2}.$$

Moreover, c_2 is at most a constant times the size of the largest Jordan block of $A_\star$.

We omit the proof of the lemma, but it follows from the fact that the differences in $\mathbf{k}_{t|t-q} - \mathbf{k}_t$ and $(\boldsymbol{\delta}_t - \mathbf{y}_{t|t-q}$ only contain the contributions of inputs and disturbances of a window of length at most $q = p + L$. Hence, it remains to understand whether there are predictors ϕ of reasonable magnitude for which $(\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q})$ is small.

Polynomial Interpolation

One way to construct a predictor ϕ for which $\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q}$ is small is by considering polynomially interpolation. Indeed, suppose that $L \geq a \cdot p$ iand consider an a-order monomial with terms for only for powers which are integer multiples of p, written in terms of coeffiects $f_1, f_2, \ldots, f_a$ via the following expansion:

$$f(z) = z^a + f^{(1)} z^{(a-1)} + f^{(2)} z^{(a-2)} + \cdots + f^{(a)} \tag{4.20}$$

Define ϕ_f to be the corresponding predictor with blocks $(\phi^{(0)}, \phi^{(2)}, \ldots \phi_f^{(L)})$, where

$$\phi^{(i)} = \begin{cases} f^{(1+i/p)} \cdot I_{d_y} & i \mod p = 0, i \leq ap \\ 0 & \text{otherwise} \end{cases}$$

Then, we compute

$$\begin{aligned} \mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q} &= A^{p+L} \mathbf{x}_{t;t-q} - \sum_{i=1}^{l} \phi^{(i)} \mathbf{y}_{t-i|t-q} \\ &= A_\star^{(a+1)p} \mathbf{x}_{t;t-q} - \sum_{i=1}^{a} f^{(i)} A_\star^{ip} \mathbf{x}_{t;t-q} = A_\star^p \cdot f(A_\star^p) \mathbf{x}_{t;t-q}. \end{aligned}$$

In particular, if f can be selected so that $f(A^p)$ is small, then $\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q}$ is small as well. In the original work, Simchowitz et al. [2019], this gives rise to a rather technical quantity termed *phase rank*. For simplicity, we give a rather crude upper bound which captures the same spirit.

The Caley Hamilton theorem guarantees that there is a minimal polynomial $f_{\min}$ of $A_\star^p$ degree at most d_x for which $f_{\min}(A_\star^p)$ is exactly 0. Specifically, the minimal polynomial of $A_\star^p$ is of the form

$$f_{\min}(z) = \prod_{\lambda \in \text{spec}(A_\star)} (z - \lambda^p)^{\alpha_\lambda}, \tag{4.21}$$

where α_λ is the algebraic multiplicity of the eigenvalue $\lambda \in \text{spec}(A_\star)$. In particular, if $L \geq p \cdot \deg(f_{\min})$, then there is a ϕ such that $\mathbf{y}_{t|t-q} - \phi \cdot \mathbf{k}_{t|t-q}$ is identicaly zero. The exists of this polynomial therefore imples that, for $L \geq p \cdot \deg(f_{\min})$, there exists a predictor such, under the condition of Lemma 4.2,

$$\|\mathbf{y}_t - k \cdot \phi\| \leq (1 + \|f_{\min}\|_{\ell_1}) \cdot (1 + B) c_1 (p + L)^{c_2} \quad \text{for all } t, \tag{4.22}$$

where $\|f_{\min}\|_{\ell_1} = 1 + \sum_i |f^{(i)}_{\min}|$. This is enough to ensure that Ovfit_μ scales as $\sqrt{N}$ with N.

Unfortunately, the coefficients minimal polynomial may be quite large in magnitude, leading to poor upper bounds; in the worst case, its coefficients can be exponentially large in the dimension d, even for a stable system. For example, the matrix $A_\star$ with d distinct eigenvalues above $1 - \frac{1}{p}$ have

$$|f_{\min}(-1)| \geq (1 + (1 - \frac{1}{p})^p)^d \geq (1 + e)^d, \tag{4.23}$$

and thus the ℓ_1 norm of its coefficients is exponentially large. This is what gave rise to the refined notion of *phase rank* described by Simchowitz et al. [2019].

Bounds via detectability

A second approach to controlling the error is via *detectability*. Informally, the pair $(A_\star, C_\star)$ is detectable if one can identify the state $\mathbf{x}_t$ along all non-stable eigendirections of $A_\star$. For a full discussion of detectability, we refer the reader to Zhou et al. [1996, Chapter 3].

One useful consequence of detectability is that there exists a matrix $F \in \mathbb{R}^{d_x \times d_y}$ such that $A_\star - FC_\star$ is stable. For detectable systems, one such chose F so as to solve the infinite-horizon *Kalman Filtering*; see ... for more details.

For now, let us suppose there exists an L such that $A_\star - LC_\star$ is stable, and let $c_F \geq 1$ and $\rho_F \in (0, 1)$ be such that

$$\|(A_\star - LC_\star)^n\| \leq c_F \rho_F^n, \quad \forall n \geq 0. \tag{4.24}$$

We use the existence of such an F to construct a good predictor ϕ. To do so, we construct an *state observer sequence* of the states $\mathbf{x}_{s|t-q}$, defined as

$$\tilde{\mathbf{x}}_{t-q|t-q} = 0, \quad \tilde{\mathbf{x}}_{s+1|t-q} = A_\star \tilde{\mathbf{x}}_{s|t-q} + F(\mathbf{y}_{s|t-q} - C_\star \tilde{\mathbf{x}}_{s|t-q}), \quad s \geq t - q. \tag{4.25}$$

Observe then that, defining the error $\delta\tilde{\mathbf{x}}_{t-q|t-q} := \mathbf{x}_{t-q|t-q} - \tilde{\mathbf{x}}_{t-q|t-q}$, we have

$$\delta\tilde{\mathbf{x}}_{t-q|t-q} = \mathbf{x}_{t-q|t-q} := \mathbf{x}_{t-q}, \quad \delta\tilde{\mathbf{x}}_{s+1|t-q} = (A_\star - FC_\star)\delta\tilde{\mathbf{x}}_{s|t-q},$$

so that, by Eq. (4.24),

$$\|\mathbf{x}_{t-p|t-q} - \tilde{\mathbf{x}}_{t-q|t-q}\|_2 \leq c_F \rho_F^L \|\mathbf{x}_{t-q}\|_{t-q}, \tag{4.26}$$

In particular, let us take $\phi = [\phi^{(1)} \mid \phi^{(2)} \mid \dots \phi^{(L)}]$ with

$$\phi^{(i)} = C_\star A_\star^p (A_\star - FC_\star)^{i-1} F. \tag{4.27}$$

Then, by unfolding the recursion in Eq. (4.25),

$$\begin{aligned}
\mathbf{y}_{t|t-q} - \phi^\top \cdot \mathbf{k}_{t|t-q} &= C_\star A_\star^p (\mathbf{x}_{t-p|t-q} - \sum_i (A_\star - FC_\star)^{i-1} F \mathbf{y}_{t-p-i|t-q}) \\
&= C_\star A_\star^p (\mathbf{x}_{t-p|t-q} - \tilde{\mathbf{x}}_{t-p|t-q}) \\
&= C_\star A_\star^p (A_\star - FC_\star)^L \mathbf{x}_{t-q|t-q}
\end{aligned}$$

Hence, from Eq. (4.24),

$$\|\phi\|_{\mathrm{op}} \le \|C_\star\|_{\mathrm{op}}\|A_\star^p\|_{\mathrm{op}}\|F\|\frac{c_F}{1-\rho_L}$$
$$\|\mathbf{y}_{t|t-q} - \phi^\top \cdot \mathbf{k}_{t|t-q}\| \le \|C_\star\|_{\mathrm{op}}\|A_\star^p\|_{\mathrm{op}} \cdot c_F\rho_F^L \cdot \|\mathbf{x}_{t-q}\|.$$

In particular, for when $\rho(A_\star) \le 1$, $\|A_\star^p\| = \mathrm{poly}(p)$ we have $\|\mathbf{x}_t\| = \mathrm{poly}(t)$ and thus, for $L = \mathcal{O}\left(\log_{\rho_F}(N)\right)$, the following (informal) bounds hold:

$$\|\phi\|_{\mathrm{op}} = \mathcal{O}\left(\mathrm{poly}(p)\right)$$
$$\|\mathbf{y}_{t|t-q} - \phi^\top \cdot \mathbf{k}_{t|t-q}\| \le N^{-\Omega(1)}.$$

Hence, Lemma 4.2 ensures that, for such a choice of L,

$$\boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t = \mathcal{O}\left((1+B)(p+L)^{\mathcal{O}(1)}\right)\forall t, \tag{4.28}$$

witnessing the desired constant magnitude of error in Eq. (4.16).

4.5 Proof of Theorem 4.1

We begin with the standard least squares error decomposition. Let $\mathbf{U}$ denote the matrix whose rows are $\mathbf{u}_{t:t-p+1}$ for $t = p+1, \ldots, N$, and recall that $\boldsymbol{\Delta}$ denotes the matrix marix whose rows are $\boldsymbol{\delta}_t$ for the same indices. Then, when $\mathbf{U}$ is full row rank, the standard least squares error decomposition yields

$$(\widehat{G}_{\mathrm{LS}} - G_{\star;p})^\top = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\boldsymbol{\Delta}. \tag{4.29}$$

Hence,

$$\|(\widehat{G}_{\mathrm{LS}} - G_{\star;p})^\top\|_{\mathrm{op}} \le \lambda_{\min}(\mathbf{U}^\top\mathbf{U})^{-1}\|\mathbf{U}^\top\boldsymbol{\Delta}\|_{\mathrm{op}},$$

which is of course vacuous when $\mathbf{U}$ is row-rank deficient.

The bound the terms $\mathbf{U}^\top\mathbf{U}$ and $\mathbf{U}^\top\boldsymbol{\Delta}$, let us break the time steps $t \in \{p+1, p+2, \ldots, N\}$ into subsequences. Define $K = (N/p) - 1$, and let k range from $1, \ldots, K$, and i range from 1 to p, and set

$$\tilde{\mathbf{u}}_{k,i} = \mathbf{u}_{t:t-p+1} \text{ for } t = pk + i, \quad \tilde{\boldsymbol{\delta}}_{k,i} = \boldsymbol{\delta}_{pk+i}.$$

Then, we can express both terms of interest as double sums:

$$\mathbf{U}^\top\mathbf{U} = \sum_{i=1}^{p}\sum_{k=1}^{K}\tilde{\mathbf{u}}_{k,i}\tilde{\mathbf{u}}_{k,i}^\top, \quad \text{and} \quad \mathbf{U}^\top\boldsymbol{\Delta} = \sum_{i=1}^{p}\sum_{k=1}^{K}\tilde{\mathbf{u}}_{k,i}\tilde{\boldsymbol{\delta}}_{k,i}. \tag{4.30}$$

Importantly, since the noise is selected i.i..d, $(\tilde{\mathbf{u}}_{k,i})$ are a sequence of independent random variables as k ranges and i is fixed. Thus, we can use standard techniques to bound the inner summation, and conclude by handling the outer summation appropriately.

Lower bounding $\lambda_{\min}(\mathbf{U}^\top\mathbf{U})$. For a fixed i, the inner summation $\sum_{k=1}^{K}\tilde{\mathbf{u}}_{k,i}\tilde{\mathbf{u}}_{k,i}^\top$ is a sum over outer products independent isotropic random vectors. Set $\mathbf{Z}_{k,i}\tilde{\mathbf{u}}_{k,i}\tilde{\mathbf{u}}_{k,i}^\top$. Since $\|\tilde{\mathbf{u}}_{k,i}\|_2^2 = pd_u$ with probability 1, and since $\mathbb{E}[\mathbf{Z}_{k,i}] = I$, we observe that

$$0 \preceq \mathbf{Z}_k \preceq pd_u \ (\forall k), \quad \lambda_{\min}\left(\mathbb{E}\left[\sum_{k=1}^{K} Z_{k,i}\right]\right) = K.$$

Thus, by the Matrix Chernoff inequality ([Tropp, 2012, Theorem 1.1]),

$$\mathbb{P}[\lambda_{\min}(\sum_{k=1}^{K} Z_{k,i}) \geq K/2] \geq (pd_u)\cdot(2/e)^{K/2pd_u} \geq (pd_u)e^{Kpd_u/7} \tag{4.31}$$

In particular, for $K \geq 7d_u p\log(p^2d_u/\delta)$, for which it suffices that $N \geq 8d_u p^2\log(p^2d_u/\delta)$, the above holds with probability at least $1-\delta/p$. By a union bound,

$$\mathbf{U}^\top\mathbf{U} = \sum_{i=1}^{p}\sum_{k=1}^{K} Z_{k,i} \succeq pK/2I = (N-p)/2I \succeq N/3\cdot I \tag{4.32}$$

where in the last line we use that $N \geq 8p$.

Upper bounding $\mathbf{U}^\top\boldsymbol{\Delta}$ To bound $\|\mathbf{U}^\top\boldsymbol{\Delta}\|_{\text{op}}$, we apply a union bound. Fix unit norm vectors $v \in \mathcal{S}^{t-p+1-1}$ and $w \in \mathcal{S}^{d_y-1}$, and consider

$$v^\top\mathbf{U}^\top\boldsymbol{\Delta}w = \sum_{i=1}^{p}\sum_{k=1}^{K}\langle v, \tilde{\mathbf{u}}_{k,i}\rangle\cdot\langle w, \boldsymbol{\delta}_{k,i}\rangle.$$

For any given $i \in [p]$, we apply the (scalar) self-normalized martingale inequality (Lemma 3.3), letting $\langle v, \tilde{\mathbf{u}}_{k,i}$ play the role of the *noise* $\boldsymbol{\epsilon}_t$, and $\langle w, \boldsymbol{\delta}_{k,i}\rangle$. play the role of $\mathbf{z}_t$. By the following lemma, $\boldsymbol{\epsilon}_t$ are 1-subGaussian random vectors:

Lemma 4.3. *Let $\mathbf{z} \sim \text{Unif}(\{-1,1\}^n)$ be a random vector with i.i.d. Rademacher entries. Then, for all $v \in \mathbb{R}^n$, $\mathbb{E}[\exp(\langle v, \mathbf{z}\rangle)] \leq \exp(\|v\|^2/2)$; that is, $\mathbf{z}$ is a 1-subGaussian random vector.*

Proof. First, consider the scalar case. By Hoeffdings inequality, it follows that $\mathbb{E}[\exp(v\mathbf{z})] \leq e^{v^2/2}$. For the vector case, let $v[i]$ and $\mathbf{z}[i]$ denote the coordinates of v and $\mathbf{z}$, respectively. Since the entries of $\mathbf{z}$ are independent, $\mathbb{E}[\exp(\langle v, \mathbf{z}\rangle)] = \prod_{i=1}^{n}\mathbb{E}[\exp(v[i]\mathbf{z}[i])] \leq \prod_{i=1}^{n}\exp(v[i]^2/2) = \exp(\|v\|^2/2)$, as needed. □

Thus for any parameter $\lambda > 0$, and defining the scalar quantity $V_{K,i;w} = \sum_{k=1}^{K}\langle w, \boldsymbol{\delta}_{k,i}\rangle^2 \leq \|\boldsymbol{\Delta}\|_{\text{op}}^2$, Lemma 3.3 entails that

$$(\sum_{k=1}^{K}\langle v, \tilde{\mathbf{u}}_{k,i}\rangle\cdot\langle w, \boldsymbol{\delta}_{k,i}\rangle)^2 \leq 2(\lambda + V_{K,i,w})(\log(1 + \frac{\|\boldsymbol{\Delta}\|_{\text{op}}^2}{\lambda}) + \log(1/\delta)). \tag{4.33}$$

Thus, by Cauchy-Schwartz, and the above concentration inequality, a union bound yields the following with probability $1-\delta$:

$$\begin{aligned} v^\top \mathbf{U}^\top \mathbf{\Delta} w &\le p \sum_{i=1}^{p} (\sum_{k=1}^{K} \langle v, \tilde{\mathbf{u}}_{k,i} \rangle \cdot \langle w, \boldsymbol{\delta}_{k,i} \rangle)^2 \\ &\le 2p \sum_{i=1}^{p} (\lambda + V_{K,i,w})(\log(1 + \frac{\|\mathbf{\Delta}\|_{\mathrm{op}}^2}{\lambda}) + \log(p/\delta)). \end{aligned}$$

Reparameterizing λ to λ/p, we have

$$\sum_{i=1}^{p} (\lambda/p + V_{K,i,w}) = \lambda + \sum_{i=1}^{p} V_{K,i,w} = \lambda + \sum_{i=1}^{p} \sum_{k=1}^{K} \langle w, \boldsymbol{\delta}_{k,i} \rangle^2 = \lambda + \|\mathbf{\Delta} w\|_2^2 = \lambda + \|\mathbf{\Delta}\|_{\mathrm{op}}^2.$$

Thus, with probability $1-\delta$,

$$v^\top \mathbf{U}^\top \mathbf{\Delta} w \le 2p(\lambda + \|\mathbf{\Delta}\|_{\mathrm{op}}^2)(\log(1 + \frac{\|\mathbf{\Delta}\|_{\mathrm{op}}^2}{\lambda}) + \log(p^2/\delta)).$$

By a standard covering argument, (see e.g. [Vershynin, 2018, Corollary 4.2.13]), this entails that with probability $1-\delta/2$

$$\begin{aligned} \|\mathbf{U}^\top \mathbf{\Delta}\|_{\mathrm{op}} &= \sup_{v,w:\|v\|=\|w\|=1} v^\top \mathbf{U}^\top \mathbf{\Delta} w \\ &\lesssim \sqrt{p(\lambda + \|\mathbf{\Delta}\|_{\mathrm{op}}^2) \left(\log(1 + \frac{\|\mathbf{\Delta}\|_{\mathrm{op}}^2}{\lambda}) + \log(p^2/2\delta) + p d_u + d_y \right)} \\ &\lesssim \sqrt{p(\lambda + \|\mathbf{\Delta}\|_{\mathrm{op}}^2)(\mathcal{L}_{\delta,\lambda,\mathbf{\Delta}} + p d_u + d_y)}, \end{aligned}$$

where $\mathcal{L}_{\delta,\lambda,\mathbf{\Delta}} := \log(1 + \frac{\|\mathbf{\Delta}\|_{\mathrm{op}}^2}{\lambda}) + \log(1/\delta)$.

Concluding Combining the two bounds, for any fixed $\lambda > 0$, the following holds with probability at least $1 - \delta/2 - \delta/2 = 1 - \delta$,

$$\begin{aligned} \|(\widehat{G}_{\mathrm{LS}} - G_{\star;p})^\top\|_{\mathrm{op}} &\le \lambda_{\min}(\mathbf{U}^\top \mathbf{U})^{-1} \|\mathbf{U}^\top \mathbf{\Delta}\|_{\mathrm{op}} \\ &\lesssim \frac{\sqrt{p(\lambda + \|\mathbf{\Delta}\|_{\mathrm{op}}^2)(\mathcal{L}_{\delta,\lambda,\mathbf{\Delta}} + p d_u + d_y)}}{N}. \end{aligned}$$

□

4.6 Proof of Theorem 4.2

For any given $\phi \in \mathbb{R}^{d_y \times Ld_y}$, define $\boldsymbol{\delta}_{\phi;t} := \boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t$. Let $\boldsymbol{\Delta}_\phi$ denote the matrix whose rows are $\boldsymbol{\delta}_{\phi;t}$, for $t \in \{p+L, p+L+1, \ldots, N\}$. Recall that $\mathbf{K}$ denotes the matrix whose rows are $\mathbf{k}_t$ for the same indices.

Since both $\boldsymbol{\delta}_t$ and $\phi \cdot \mathbf{k}_t$ are both $\mathcal{F}_t$ measurable, the analysis of Theorem 4.1 applies to the least squares estimator which predicts the residuals $\mathbf{y}_t - \phi \cdot \mathbf{k}_t$ for any *fixed* predictor ϕ:

$$\widehat{G}_\phi \in \operatorname*{arg\,min}_{G \in \mathbb{R}^{d_y \times pd_u}} \sum_{t=p+L}^{N} \|(\mathbf{y}_t - \phi \cdot \mathbf{k}_t) - G \cdot \mathbf{u}_{t:t-p+1}\|_2^2, \tag{4.34}$$

which is defined using our putative fixed predictor ϕ, rather than the solution $\hat{\phi}$. Unpacking the proof of Theorem 4.1, consider the event

$$\mathcal{E}_{\mathrm{cond}} = \{\mathbf{U}^\top \mathbf{U} \succeq cNI\}, \qquad \text{for some universal constant } c > 0, \tag{4.35}$$

We see that for $N \geq N_0(\delta)$ (where N_0 is defined in Theorem 4.1), $\mathcal{E}_{\mathrm{cond}}$ holds with failure probability at most $1 - \delta/2$. Morevover, from the proof of Theorem 4.1, the following guarantee holds with an additional failure probability of $1 - \delta/2$ on $\mathcal{E}_{\mathrm{cond}}$ (for all $\lambda > 0$):

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\mathrm{op}} \lesssim \frac{(\lambda + \|\boldsymbol{\Delta}_\phi\|_{\mathrm{op}})\sqrt{p(\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}_\phi}(\phi) + pd_u + d_y)}}{N}, \tag{4.36}$$

where we recall $\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}}(\phi) := \log(1 + \frac{\|\boldsymbol{\Delta}_\phi\|_{\mathrm{op}}}{\lambda}) + \log\frac{1}{\delta}$.

The above provides a bound for fixed ϕ; it remains to analyze the performance when $\phi = \hat{\phi}$ is selected during the prefiltering step.

A Uniform Bound

Because $\hat{\phi}$ is chosen in a data-dependent fashion, we analyze its performance by establish an *uniform* bound on the performance of all predictors simultaneously. We show that the performance of each predictor ϕ is simultaneously controlled by its individual error $\boldsymbol{\Delta}_\phi$:

Lemma 4.4. *Fix parameters $\delta, \lambda > 0$, and recalling $\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}}$ from Theorem 4.1, define the dimension quantity.*

$$\dim_k(\phi) := p(\mathcal{L}_{\delta,\lambda,\boldsymbol{\Delta}_\phi} + kLd_y + pd_u). \tag{4.37}$$

Finally, define the random variable $k_0 := \lceil \log \frac{\sqrt{N}\|\mathbf{K}\|_{\mathrm{op}}}{\lambda} \rceil$. Then, the bounds with probability at least $1 - 3\delta/4$ for all $k \geq k_0$ and all predictors ϕ of norm $\|\phi\|_{\mathrm{F}} \leq \lambda e^k$ simultaneously:

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\mathrm{op}} \lesssim \frac{(\lambda + \|\boldsymbol{\Delta}_\phi\|_{\mathrm{op}})\sqrt{\dim_k(\phi)}}{N}. \tag{4.38}$$

Proof. For the chosen regularizer $\lambda > 0$, and defining the error terms $\epsilon_k = e^{-k}$, let $\mathcal{T}_k$ be an $\lambda\epsilon_k$-net of the λ/ϵ_k Frobenius norm ball around 0 in $\mathbb{R}^{d_y \times L}$. By a standard covering argument (see e.g. [Vershynin, 2018, Corollary 4.2.13]), we can choose the net $\mathcal{T}_k$ to be small enough to satisfy

$$\log |\mathcal{T}_k| \lesssim kLd_y. \tag{4.39}$$

By union bounding over all $k \geq 1$, with failure probability $\delta_k = \delta\epsilon_k/4$ assigned to each $k \in \mathbb{N}$, the following holds uniformly over all k:

$$\forall k \in \mathbb{N}, \tilde{\phi} \in \mathcal{T}_k, \quad \|\widehat{G}_{\tilde{\phi}} - G_{\star;p}\|_{\text{op}} \lesssim \frac{(\lambda + \|\mathbf{\Delta}_{\tilde{\phi}}\|_{\text{op}})\sqrt{\dim_k(\tilde{\phi})}}{N}$$

Denote the event $\mathcal{E}_\star$. Let $k_0 := \lceil \log(\sqrt{N}\|\mathbf{K}\|_{\text{op}}/\lambda) \rceil$ (note that k_0 is a random variable), $k \geq k_0$, and consider any ϕ with $\|\phi\|_{\text{F}} \leq \lambda e^k$. Then, there exists a $\tilde{\phi} \in \mathcal{T}_k$ such that

$$\|\mathbf{\Delta}_\phi - \mathbf{\Delta}_{\hat{\phi}}\|_{\text{op}} = \|\mathbf{K}(\phi - \hat{\phi})^\top\|_{\text{op}} \leq \|\mathbf{K}\|_{\text{op}}\|\phi - \hat{\phi}\|_{\text{op}} \leq \|\mathbf{K}\|_{\text{op}} \cdot \lambda\epsilon^{-k} \leq \lambda,$$

where the final equalities use the definition of the covering and the fact that $k \geq k_0 \geq \lceil \log(\|\mathbf{K}\|_{\text{op}}/\lambda) \rceil$. In particular, for this pair of $\phi, \tilde{\phi}$, we verify that

$$\|\mathbf{\Delta}_{\tilde{\phi}}\|_{\text{op}} \leq \lambda + \|\mathbf{\Delta}_\phi\|_{\text{op}}, \quad \dim_k(\tilde{\phi}) \lesssim \dim_k(\phi)$$

Thus, on $\mathcal{E}_\star$,

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\text{op}} \lesssim \|\widehat{G}_\phi - \widehat{G}_{\tilde{\phi}}\|_{\text{op}} + \frac{(\lambda + \|\mathbf{\Delta}_\phi\|_{\text{op}})\sqrt{\dim_k(\phi)}}{N}.$$

Again, using the least squares decomposition in Eq. (4.29), we find that, on $\mathcal{E}_{\text{cond}}$,

$$\|\widehat{G}_\phi - \widehat{G}_{\tilde{\phi}}\|_{\text{op}} = \|\mathbf{U}^\dagger(\Delta_\phi - \Delta_{\tilde{\phi}})\| \lesssim \frac{\|\mathbf{K}\|_{\text{op}}}{\sqrt{N}}\|\phi - \hat{\phi}\|_{\text{op}} \leq \frac{\lambda}{N}.$$

Hence, we find that for all $k \geq k_0$, and all ϕ with $\|\phi\|_{\text{F}} \leq \|\phi\|_{\text{F}} \leq \lambda e^k$, on the event $\mathcal{E}_\star$ it holds that

$$\|\widehat{G}_\phi - G_{\star;p}\|_{\text{op}} \lesssim \frac{(\lambda + \|\mathbf{\Delta}_\phi\|_{\text{op}})\sqrt{\dim_k(\phi)}}{N}.$$

□

Bounding the prefiltered residuals

The next step in the analysis is to bound the operator norm of the matrix consisting of prefiltered residuals, $\|\mathbf{\Delta}_{\hat{\phi}}\|_{\text{op}}$. Note that $\hat{\phi}$ is trained by regressing the features $\mathbf{k}_t$ to $\mathbf{y}_t$,

whereas the terms we need to bound are the rows $\boldsymbol{\delta}_{t;\hat{\phi}} := \boldsymbol{\delta}_t - \hat{\phi} \cdot \mathbf{k}_t$ of $\boldsymbol{\Delta}_{\hat{\phi}}$. We execute this decoupling by using the explicit form of the ridge estimator. Recall that $\hat{\phi}$ is the ridge estimator of $\mathbf{y}_t$ from features $\mathbf{k}_t$. Define $\tilde{\phi}$ to be the same but regressing to the (unknown) perturbations $\boldsymbol{\delta}_t$, i.e.

$$\tilde{\phi} = \min_{\phi} \sum_{t=L+p}^{N} \|\boldsymbol{\delta}_t - \phi \cdot \mathbf{k}_t\|_2^2 + \mu \|\phi\|_{\mathrm{F}}^2. \tag{4.40}$$

Using the explicit form of the least squares estimators,

$$\begin{aligned}
\boldsymbol{\Delta}_{\hat{\phi}} &= \boldsymbol{\Delta} - \hat{\phi} \cdot \mathbf{K} = \boldsymbol{\Delta} - \mathbf{Y}^\top \mathbf{K}(\mathbf{K}^\top \mathbf{K} + \mu I)^{-1} \mathbf{K} \\
&= \boldsymbol{\Delta} - \boldsymbol{\Delta}^\top \mathbf{K}(\mathbf{K}^\top \mathbf{K} + \mu I)^{-1} \mathbf{K} + G_\star \mathbf{U}^\top \mathbf{K}(\mathbf{K}^\top \mathbf{K} + \mu I)\mathbf{K} \\
&= (\boldsymbol{\Delta} - \tilde{\phi}\mathbf{K}) + G_\star \mathbf{U}^\top \mathbf{K}(\mathbf{K}^\top \mathbf{K} + \mu I)^{-1} \mathbf{K},
\end{aligned}$$

Thus, $\|\boldsymbol{\Delta}_{\hat{\phi}}\|_{\mathrm{op}}$ can be bounded by the sum of two terms:

$$\|\boldsymbol{\Delta}_{\hat{\phi}}\|_{\mathrm{op}} \le \underbrace{\|\boldsymbol{\Delta} - \tilde{\phi}\mathbf{K}\|_{\mathrm{op}}}_{\text{(a)}} + \|G_\star\|_{\mathrm{op}} \underbrace{\|\mathbf{U}^\top \mathbf{K}(\mathbf{K}^\top \mathbf{K} + \mu I)^{-1} \mathbf{K}\|_{\mathrm{op}}}_{\text{(b)}}. \tag{4.41}$$

Term (a) is just the error of the ridge regression *trained to* $\boldsymbol{\Delta}$, and term (b) can be regarded as how much the regression *overfits* to $\mathbf{K}$. To adress (a) we invoke the following lemma

Lemma 4.5 (Term (a))**.** *Let $\tilde{\phi}$ denote the idealized ridge estimator, trained to the residuals $\boldsymbol{\delta}_t$ as in (4.40). The following inequality holds deterministically:*

$$\|\boldsymbol{\Delta} - \tilde{\phi} \cdot \mathbf{K}\|_{\mathrm{op}}^2 \le \inf_{\phi \in \mathbb{R}^{d_y \times L}} \|\boldsymbol{\Delta} - \phi \cdot \mathbf{K}\|_{\mathrm{op}}^2 + \mu \|\phi\|_{\mathrm{op}}^2 := \mathsf{Opt}_\mu \tag{4.42}$$

Proof. For any fixed $v \in \mathcal{S}^{d_y - 1}$, we have

$$\begin{aligned}
\|v^\top(\boldsymbol{\Delta} - \tilde{\phi}\mathbf{K})\|_2^2 &\le \left\| \begin{bmatrix} v^\top(\boldsymbol{\Delta} - \tilde{\phi}\mathbf{K}) \\ \sqrt{\mu} v^\top \tilde{\phi} \end{bmatrix} \right\|^2 \\
&= \sum_{t=L+p}^{N} \langle v, \boldsymbol{\delta}_t - \tilde{\phi}\mathbf{k}_t \rangle^2 + \mu \langle v, \tilde{\phi} \rangle^2.
\end{aligned}$$

Using the direct form of $\tilde{\phi}$, we can verify that $\tilde{\phi}$ is in fact the unconstrained minimizer of the second line in the above display over all linear predictors $\phi \in \mathbb{R}^{d_y \times L}$. Thus,

$$\|v^\top(\boldsymbol{\Delta} - \tilde{\phi}\mathbf{K})\|_2^2 \le \inf_{\phi \in \mathbb{R}^{d_y \times L}} \sum_{t=L+p}^{N} \langle v, \boldsymbol{\delta}_t - \tilde{\phi}\mathbf{k}_t \rangle^2 + \mu \langle v, \tilde{\phi} \rangle^2.$$

Since the operator norm is the supremum of the RHS of the above over all $v \in \mathcal{S}^{d_y-1}$,

$$\|\mathbf{\Delta} - \tilde{\phi}\mathbf{K}\|_2^2 \le \sup_{v \in \mathcal{S}^{d_y-1}} \inf_{\phi \in \mathbb{R}^{d_y \times L}} \sum_{t=L+p}^{N} \langle v, \boldsymbol{\delta}_t - \tilde{\phi}\mathbf{k}_t\rangle^2 + \mu\langle v, \tilde{\phi}\rangle^2.$$

Swapping the infinum and supremum establishes the inequality (again, using the variational form of the operator norm) concludes. □

Lemma 4.6 (Term (b))**.** *With probability* $1 - \delta/4$*, we have*

$$\|\mathbf{U}^\top \mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}\|_{\mathrm{op}}^2 \lesssim p^2 \left(\log\det(1 + \tfrac{\mathbf{K}^\top\mathbf{K}}{\mu}) + d_u + \log\tfrac{1}{\delta}\right) =: \mathsf{Ovfit}_{\mu,\delta,\mathbf{K}}/\|G_{\star;p}\|_{\mathrm{op}}$$

Proof. We bound Term (b) by analogy to our bound on the cross term $\mathbf{U}^\top\mathbf{\Delta}$ in the proof of the non-prefiltered least squares bound, Theorem 4.1. First, some simplifications. We have

$$\begin{aligned}
\|\mathbf{U}^\top \mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}\|_{\mathrm{op}}^2 &= \max_{v \in \mathcal{S}^{d_u-1}} v^\top\mathbf{U}^\top\mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}^\top\mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}^\top\mathbf{U}v \\
&\le \max_{v \in \mathcal{S}^{d_u-1}} v^\top\mathbf{U}^\top\mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}(\mathbf{K}^\top\mathbf{K} + \mu I)(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}^\top\mathbf{U}v \\
&= \max_{v \in \mathcal{S}^{d_u-1}} v^\top\mathbf{U}^\top\mathbf{K}(\mathbf{K}^\top\mathbf{K} + \mu I)^{-1}\mathbf{K}^\top\mathbf{U}v \\
&= \max_{v \in \mathcal{S}^{d_u-1}} \|v^\top\mathbf{U}^\top\mathbf{K}\|_{(\mathbf{K}^\top\mathbf{K}+\mu I)^{-1}}^2.
\end{aligned}$$

Next, we recap the *blocking* argument from Theorem 4.1. Define $K = \frac{N-L_0-p}{p}$, and let k range from $1, \dots, K$, and i range from 1 to p, and set

$$\tilde{\mathbf{u}}_{k,i} = \mathbf{u}_{t:t-p+1} \text{ for } t = L + pk + i, \quad \tilde{\mathbf{k}}_{k,i} = \mathbf{k}_{pk+i+l}.$$

For any fixed vector $v \in \mathcal{S}^{d_u-1}$, define

$$v^\top\mathbf{U}^\top\mathbf{K} = \sum_{i=1}^{p}\sum_{k=1}^{K}\langle v, \tilde{\mathbf{u}}_{k,i}\rangle\mathbf{k}_{k,i}$$

Hence, setting $\mathbf{K}_i$ to denote the submatrix of $\mathbf{K}$ whose rows are are $\mathbf{k}_i$, and noting $\mathbf{K}_i^\top\mathbf{K}_i \preceq \mathbf{K}^\top\mathbf{K}$,

$$\begin{aligned}
\|v^\top\mathbf{U}^\top\mathbf{K}\|_{(\mathbf{K}^\top\mathbf{K}+\mu I)^{-1}}^2 &\le p\sum_{i=1}^{p}\left\|\sum_{k=1}^{K}\langle v, \tilde{\mathbf{u}}_{k,i}\rangle\mathbf{k}_{k,i}\right\|_{(\mathbf{K}^\top\mathbf{K}+\mu I)^{-1}}^2 \\
&\le p\sum_{i=1}^{p}\left\|\sum_{k=1}^{K}\langle v, \tilde{\mathbf{u}}_{k,i}\rangle\mathbf{k}_{k,i}\right\|_{(\mathbf{K}_i^\top\mathbf{K}_i+\mu I)^{-1}}^2.
\end{aligned}$$

Using the self-normalized martingale bound (Lemma 3.3) to each summard in i, we have that with probability $1-\delta$, the above is at most

$$2p\sum_{i=1}^{p}\left(\log\det(1+\frac{\mathbf{K}_i^\top\mathbf{K}_i}{\mu})+\log\frac{1}{\delta}\right)\le 2p^2\left(\log\det(1+\frac{\mathbf{K}^\top\mathbf{K}}{\mu})+\log\frac{1}{\delta}\right).$$

Via another union bound to establish uniform convergence over $v\in\mathcal{S}^{d_u-1}$, we conclude that with probability, say, $1-\delta/4$ (after appropriate δ rescaling),

$$\|v^\top\mathbf{U}^\top\mathbf{K}\|^2_{(\mathbf{K}^\top\mathbf{K}+\mu I)^{-1}}\lesssim p^2\left(\log\det(1+\frac{\mathbf{K}^\top\mathbf{K}}{\mu})+d_u+\log\frac{1}{\delta}\right),$$

as needed. □

Thus, from (4.41),

$$\|\mathbf{\Delta}_{\hat\phi}\|_{\mathrm{op}}\le \mathsf{Opt}_\mu+\mathsf{Ovfit}_{\mu,\delta,\mathbf{K}}\quad\text{w.p. } 1-\delta/4. \tag{4.43}$$

There is one last step required in the proof: bounding the Frobenius norm of ϕ. This is straightforward:

Lemma 4.7. $\|\hat\phi\|_{\mathrm{F}}^2\le\|\mathbf{Y}\|_{\mathrm{F}}^2/\mu$.

Proof. The risk of the zero-predictor in Eq. (4.10) is $\|\mathbf{Y}\|_{\mathrm{F}}^2$. Hence, if $\|\hat\phi\|_{\mathrm{F}}>\|\mathbf{Y}\|_{\mathrm{F}}^2/\mu$, its risk is strictly greater. □

Concluding the proof

We now collect the requisite ingredients for the follow. Let

$$k_\star=\lceil\log(\tfrac{\sqrt{N}\|\mathbf{K}\|_{\mathrm{op}}\vee\mu^{-1/2}\|\mathbf{Y}\|_{\mathrm{F}}}{\lambda}\rceil$$

By definition, $k_\star\ge k_0$, where k_0 was defined in Lemma 4.4, and moreover, $\lambda e^{k_\star}\ge\|\hat\phi\|_{\mathrm{F}}^2$ by Lemma 4.7. Therefore, Lemma 4.4 ensures that, with probability $1-\frac{3\delta}{4}$,

$$\|\widehat{G}_{\hat\phi}-G_{\star;p}\|_{\mathrm{op}}\lesssim\frac{(\lambda+\|\mathbf{\Delta}_{\hat\phi}\|_{\mathrm{op}})\sqrt{\dim_{k_\star}(\hat\phi)}}{N}.$$

We can explicitly express

$$\begin{aligned}\dim_{k_\star}(\hat\phi)&=p(\mathcal{L}_{\delta,\lambda,\mathbf{\Delta}_{\hat\phi}}+k_\star Ld_y+pd_u)\\&=p(\log(1+\frac{\|\mathbf{\Delta}\|_{\mathrm{op}}}{\lambda})+\log(1/\delta)+\lceil\log(\tfrac{\sqrt{N}\|\mathbf{K}\|_{\mathrm{op}}\vee\mu^{-1/2}\|\mathbf{Y}\|_{\mathrm{F}}}{\lambda}\rceil Ld_y+pd_u)\\&\lesssim p^2d_u+p\log\tfrac{1}{\delta}+p(1+Ld_y)\log\left(e+\tfrac{\|\mathbf{\Delta}_{\hat\phi}\|_{\mathrm{op}}^2+\sqrt{N}\|\mathbf{K}\|_{\mathrm{op}}+\mu^{-1/2}\|\mathbf{Y}\|_{\mathrm{F}}}{\lambda}\right):=\dim_{\mathrm{eff}}.\end{aligned}$$

Hence,

$$\|\widehat{G}_{\hat\phi}-G_{\star;p}\|_{\mathrm{op}}\lesssim\frac{(\lambda+\|\mathbf{\Delta}_{\hat\phi}\|_{\mathrm{op}})\sqrt{\dim_{\mathrm{eff}}}}{N}.$$

Finally, we invoke Eq. (4.43) to bound $\|\mathbf{\Delta}_{\hat\phi}\|_{\mathrm{op}}$. □

Part II

Online Control

Chapter 5

Online LQR

In the previous two chapters, we consider the problem of *system identification*; estimating the dynamical parameters in an appropriate matrix norm. In this chapter and the following, we turn our attention to *adaptive control*, where the parameters of the system are identifed with the aim of synthesizing control policies which attain high performance.

This chapter focuses on the problem of online LQR, a simple continuous control problem where a learning agent interacts with an unknown linear dynamical system, yet attempts to attain control performance comparable to if the system parameters were known in hindsight. The metric of performance is *regret*, or comparison with the best system performance given full knowledge of system parameters.

In many reinforcement learning settings, including tabular MDPs, MDPs with linear function approximation, and rich contextual decision processes, careful exploration is essential for sample efficiency [Lykouris et al., 2019, Jiang et al., 2017, Jin et al., 2018, Azar et al., 2017]. In this chapter we ask if the same is true of the widely studied online LQR setting [Abbasi-Yadkori and Szepesvári, 2011, Dean et al., 2018, Cohen et al., 2019, Mania et al., 2019, Faradonbeh et al., 2018c]. Recently, however, it was shown that for the online variant of the LQR problem, relatively simple exploration strategies suffice to obtain the best-known performance guarantees [Mania et al., 2019]. In this paper, we address a curious question raised by these results: Is sophisticated exploration helpful for LQR, or is linear control in fact substantially easier than the general reinforcement learning setting? More broadly, we aim to shed light on the question:

To what extent to do sophisticated exploration strategies improve learning in online linear-quadratic control?

In this chapter, we show that a surprisingly simple exploration modifying Mania et al. [2019] yields optimal regret for the online LQR problem, both in problem horizon and system dimension. The policy alternates between injecting Gaussian noise, and estimating the system parameters via least squares. Interestingly, the algorithm does not apply any exploration bonuses typical of reinforcement algorithms for MDPs and their variants. That such

a simple policy is optimal that linear control problems are qualitatively different from other reinforcement learning problems studied in the literature.

Organization

Unlike the previous two chapters, this chapter will not aim to provide complete proofs of all the constituent results; space constraints do not permit it. Instead, it focuses on isolating the key techniques, and establishing the intution behind the main results.

The chapter is organized as follows. In Section 5.1, we present the *Linear Quadratic Regulator* or *LQR*, a classical problem in linear control, and describe its solution in terms of the Discrete Algebraic Ricatti Equation, or DARE. The section is not rushed, so as to allow the reader to develop adequate intuition for the classical problem.

In Section 5.2, we introduce *online LQR*, an adaptive control problem where the learner attempts to identify the best control policy for a dynamical system with unknown parameters, and performance by *regret*, or suboptimality compared to the best control policy given full system knowledge. Here we motivated our contributes via contrast to prior work.

Section 5.3 informally our main results: that the optimal regret in online LQR scales as $\sqrt{Td_xd_u^2}$, where d_x is the state dimension, and d_u the input dimension. It also describes the intuition beind both upper and lower bounds. Section 5.4 formally exposes our main results: a lower bound, an algorithm and upper bound, a novel perturbation bound for certainty-equivalent control based on a technique we term the "self-bounding ODE method". The subsequents provide details for the perturbation, upper bound, and lower bound, respectively, with additional proofs relegated to Section 5.8.

5.1 The Linear Quadratic Regulator

This chapter concerns itself with the particular problem of the adaptive control of *linear quadratic regulator*, or LQR. In LQR, we consider a fully observed linear dynamical systems, whose stated evolves according to Gaussian perturbations, much like in . . . :

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{w}_t, \quad \mathbf{x}_1 \equiv 0, \quad \mathbf{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I) \tag{5.1}$$

The choice of identity noise covariance and 0 initial state is non-essential, and chosen for simplicity. The analysis may be extended, for example, to i.i.d. subgaussian noises with non-degenerate and well-conditioned conditioned covariance.[1] We let $\theta = (A, B) \in \mathbb{R}^{d_x \times (d_x + d_u)}$ denote the system parameter.

Control Policies The choice of inputs $\mathbf{u}_t$ is specified by a control policy. Formally, control policy is mapping π mapping past states and inputs to inputs. This mapping is allowed to

[1] The problem may become qualitatively difference if the noise covariance is rank-deficient or very ill-conditioned.

be *randomized*, by taking some random seed argument ξ, drawn before the game, to encode randomness. We express the choice of input at time t as

$$\mathbf{u}_t = \pi(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}, t, \xi) \tag{5.2}$$

Given such a policy π, we let $\mathbb{E}_{\theta,\pi}$ denote expectations with respect to the dynamics induced by $\theta = (A, B)$ and inputs selected by the policy π, under the random Gaussian noise in Eq. (5.1).

One popular selection of policies are state feedback controllers, which return an input that is linear in the current state. Formally, given $K \in \mathbb{R}^{d_x \times d_u}$, the associated state feedback policy is $\pi^K(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}, t, \xi) = K\mathbf{x}_t$. Such a policy induces the following *closed-loop* dynamics

$$\mathbf{x}_{t+1} = (A + BK)\mathbf{x}_t + \mathbf{w}_t, \tag{5.3}$$

Of particular importance are state feedback controllers which stabilize the system (A, B).

Definition 5.1. We say that $K \in \mathbb{R}^{d_u \times d_x}$ *stabilizes* $\theta = (A, B)$ if $A + BK$ is a stable matrix, that is, $\rho(A + BK) < 1$, where again $\rho(\cdot)$ denotes the spectral radius (Definition 3.1). We say that θ is *stabilizable* if the exists some $K \in \mathbb{R}^{d_u \times d_x}$ which stabilizes θ.

Note that, if K stabilizes θ, the covariance matrix of $\mathbf{x}_t$ under dynamics (6.6) reaches a steady-state; that is, $\lim_{t\to\infty} \mathbb{E}_{\pi^K,\theta}[\mathbf{x}_t\mathbf{x}_t^\top]$ is some bounded operator. Conversely, if K does not stabilizes θ the variance of the iterates becomes infinite in the limit. For further details, see the discussion surround the Gramian operator in Chapter 3.

Costs and Cost Functionals At each time t, the learner suffers a quadratic cost $\ell(\mathbf{x}_t, \mathbf{u}_t)$ given by

$$\ell(x, u) = x^\top Q x + u^\top R u, \tag{5.4}$$

where $Q \in \mathbb{S}_{++}^{d_x}$ and $R \in \mathbb{S}_{++}^{d_u}$. The learners goal is to minimize $\sum_{t=1}^T \ell(\mathbf{x}_t, \mathbf{u}_t)$. More precisely, for the costs ℓ in Eq. (5.4) above, the running cost of a policy $\mathbf{J}_T(\pi; \theta$ is

$$\mathbf{J}_T(\pi; \theta) = \sum_{t=1}^{T} \ell(\mathbf{x}_t, \mathbf{u}_t), \quad \text{subject to}$$
$$\text{the dynamics in Eq. (5.1) and inputs } \mathbf{u}_t \text{ in Eq. (5.2).}$$

Note that $\mathbf{J}_T$ is a random variable, due to the random noise $\mathbf{w}$, and possible randomness in the policy. When considering state feedback policies π^K, we shall also consider their infinite horizon cost

$$J_\infty(K; \theta) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}[\mathbf{J}_T(\pi^K; \theta)] \tag{5.5}$$

This cost captures the long-run average expected cost of the state feedback policy K.

Since the costs satisfy $\ell(x, u) > c\|x\|^2$ for some $c > 0$, and since the noise $\mathbf{w}_t$ has full covariance, one can directly verify that $J_\infty(K)$ is finite if and only if K stabilizes θ, and in this case, the limit in its definition converges.

The Optimal LQR Control Law

Due to the mean-zero Gaussian noise and quadratic costs, the optimal infinte horizon control laws for LQR are state feedback policies. Formally, when $\theta = (A, B)$ is stabilizable, there is a unique policy $K_\infty(\theta)$ such that

$$\pi^{K_\infty(\theta)} \in \arg\inf_{\pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\mathbf{J}_T(\pi;\theta)]. \tag{5.6}$$

In particular, $K_\infty(\theta)$ is the unique minimizer of the function $K \mapsto J_\infty(K;\theta)$. The optimal state feedback controller can be computed explicitly in two steps. First, one solves for the unique matrix $P_\infty(\theta) \in \mathbb{S}^{d_x}_{++}$ satisfying the discrete algebraic Riccati equation, or DARE:

$$\begin{aligned} P_\infty(\theta) \text{ solves } P &= A^\top PA - (A^\top PB)(R + B^\top PB)^{-1}(B^\top PA) + Q = 0 \\ \text{subject to} \quad & P \succ 0. \end{aligned} \tag{5.7}$$

The DARE arises as the fixed point of the dynamic program which computes the optimal finite time controlly policy, and its solution can be computed by either a fixed point iteration [Pappas et al., 1980], or an SDP [Cohen et al., 2018]. In particular, the matrix $P_\infty(\theta)$ corresponds to an infinite horizon-cost to go, or *value function*.

The optimal control law is closed form in terms of $P_\infty(\theta)$:

$$K_\infty(\theta) = (R + B^\top P_\infty(\theta)B)^{-1}(B^\top P_\infty(\theta)A). \tag{5.8}$$

One can show that this optimal law is invariant to change of basis of the state, and hence is the same optimal control law for *any* choice of noise covariance.

5.2 Online LQR

In the above discussion, we described the LQR problem, and how to synthesize the optimal (finite horizon) control policy. Importantly, the optimal control policy depends on the costs matrix Q and R, as well as the system dynamics $\theta = (A, B)$.

In online LQR, the learner does not know the dynamics in advance, but must learn them on the fly. Throughout, we let $\theta_\star = (A_\star, B_\star)$ denote a true instance to be learned, and use $K_\star = K_\infty(\theta_\star)$ and $P_\star = P_\infty(\theta_\star)$ to denote the associated optimal control policy and value function. The learner adopts a so-called *adaptive* policy, denoted by alg, which tries to compete with the optimal policy given knowledge of the true system dynamics. Formally, the learner aims for low regret:

$$\textsc{LqrReg}_T(\mathsf{alg};\theta_\star) := \mathbf{J}_T(\mathsf{alg};\theta_\star) - T \cdot J_\infty(K_\star;\theta_\star). \tag{5.9}$$

We require that alg is a general, randomized control policy of the form Eq. (5.2), which *may not* have explicit dependence on the true parameter $\theta_\star$, but may incorporate knowledge of the matrices R and Q. To ensure that the regret is non-vacuous, we assume that the true instance $\theta_\star$ is stabilizable:

Assumption 5.1. $\theta_\star = (A_\star, B_\star)$ is stabilizable, and hence $K_\star = K_\infty(\theta_\star)$ and $P_\star = P_\infty(\theta_\star)$ are well-defined.

In words, $\textsc{LqrReg}_T(\mathsf{alg})$ measures the relative suboptimality of the adaptive control algorithm alg relative to the optimal infinite horizon average cost, $J_\infty(K_\star)$, scaled by the horizon T. This second term is often referred to as the *comparator*, because it represents the benchmark against which we aim to compete.[2] We note that $\textsc{LqrReg}_T(\mathsf{alg})$ is a random variable, because $\mathbf{J}_T(\mathsf{alg})$ accounts for random fluctuations of the noise. The policy alg is called adaptive because its goal is to balance control actions which minimize cost with exploration to ascertain the true system parameters $(A_\star, B_\star)$. Our aim is to understand:

What is the optimal regret attainable, and what sorts of algorithm design principles match it?

Throughout, make a standard assumption in past literature that we are given access to an initial stabilizing controller K_0, that is, $\rho(A_\star + B_\star K_0) < 1$. Hence, $J_\infty(K_0; \theta_\star) < \infty$.

Quantities of Interest

For simplicity, we impose the following normalization on our cost matrices:

Assumption 5.2. We assume that $Q \succeq I$ and $R \succeq I$.

Our bounds are be stated in terms of 4 primary quantities. First, the problem horizon length T, which is the typically the most salient parameter in an online parameter. Second, the state and input dimension d_x and d_u; we regard these parameters separately, and establish bounds that are optimal even in a typical regime where $d_u \ll d_x$.

The last two key parameters in our bounds are $\|B_\star\|_{\mathrm{op}}$ (the norm of the B-matrix), and $\|P_\star\|_{\mathrm{op}}$. Because these terms consider operator norms (instead of say, Frobenius or nuclear norms), we regard them as dimension free, and will suppress them from informal Big-Oh notation. The norm $\|P_\star\|_{\mathrm{op}}$ is gives an upper bound on the decay of the system under the optimal controller. For example, a classical Lyapunov argument reveals that $\|(A_\star + B_\star K_\star)^i\|_{\mathrm{op}} \le \kappa_\star \gamma_\star^i$ for $\kappa_\star = \|P_\star\|_{\mathrm{op}}$ and $\gamma_\star = (1 - \|P_\star\|_{\mathrm{op}}^{-1})^{-1}$; this decay property was termed $(\kappa_\star, \gamma_\star)$ by Cohen et al. [2018].[3]

Prior work and Certainty Equivalence

Prior to the work in this thesis, it was known that one could attain regret that scaled as $\sim \sqrt{T}$ with the time horizon. This regret was first attained by Abbasi-Yadkori and Szepesvári [2011], albeit with a computationally inefficient algorithm. Moreover, their regret quarantine

[2] The comparator term can be replaced by, say, $\inf_\pi \mathbb{E}[\mathbf{J}_T(\pi)]$, up to constants independent of the horizon T.

[3] From the normalization conditions, one can show that $\|P_\star\| \ge 1$.

had an inexplicit, and possibly exponential dependence on system dimension. Dean et al. [2018] presented an approach based on robust control which attained a larger $\sim T^{2/3}$ regret, but via a computationally efficient algorithm, and with polynomial dependence on relevant problem parameters. Subsequently, three concurrent works provided algorithms which were simultaneously computationally efficient, attain $\sim \sqrt{T}$ regret, and enjoyed polynomial dependence on problem parameters [Mania et al., 2019, Cohen et al., 2019, Faradonbeh et al., 2018c].

Of the algorithms proposed above, the simplest - both computationally and conceptually - is due to Mania et al. [2019]. They adopted the *certainty equivalence* strategy, which comprises of two steps:

- For the first N steps $t = 1, 2, \ldots, N$, execute a random input $\mathbf{u}_t \sim \mathcal{N}(0, I)$, and produce a least squares estimates $\widehat{\theta} = (\hat{A}, \hat{B})$ of $\theta_\star$ using the current trajectory data.
- From these estimates $\widehat{\theta}$, synthesize the *certainty equivalent* control policy $\hat{K} = K_\infty(\widehat{\theta})$ and execute it for all remaining time steps.

Mania et al. [2019]'s analysis of certainty equivalence also produced the sharpest dimension dependence amongst previous work: $\hat{\mathcal{O}}(\sqrt{d^3 T})$, where $d = \max\{d_u, d_x\}$. In contrast, the bounds due to Cohen et al. [2019] have a large polynomial scaling in d in the worst case. However, Cohen et al. [2019]'s algorithm enjoys guaranteed performance for *any* stabilizable system, whereas the analysis due to Mania et al. [2019] requires the pair $\theta_\star$ to be controllable.

Is intelligent exploration needed?

Given that some of the strongest guarantees for online LQR are obtained by a remarkably simple algorithm, one might hope that more intelligent exploration could yield improved performance.

For one, the cost structure in online LQR is strongly convex, and it is well know that in many online decision making settings with strongly convex costs, one can obtained regret that is at most *logarithmic* in the time horizon. Intuitively, logarithmic regret arises because the curvature of the costs allow the learner to rule out suboptimal decisions rapidly.

A second reason to hope that a more intelligent algorithm may enjoy superior performance is that the version of certainty equivalence proposed by Mania et al. [2019] is an *explore-exploit* scheme: after a predetermined exploration window N, it ceases to collect new information. Explore-exploit algorithms are known to be suboptimal in many settings of interests, and perhaps the same is true here.

Finally, it is possible that the certainty equivalence is a poor way to synthesize control policies. For example, the analysis provided by Mania et al. [2019] relies on system controllability, rather than *stabilizability*. Controllability requires that any target state $x \in \mathbb{R}^{d_x}$ can

be reached for the 0 state in finitely many steps. For example, the system

$$A_\star = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{bmatrix} B_\star = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

is not controllable, because no input can affect the second system coordinate. However, the system is *stabilizable*, because, for the 0 feedback law $K = 0$, $\rho(A_\star + BK) = \rho(A_\star) = \frac{1}{2}$. Perhaps other techniques - like optimism-under-uncertainty [Abbasi-Yadkori and Szepesvári, 2011], system level synthesis [Dean et al., 2018], or SDP relaxations [Cohen et al., 2019] are necessary to relax this requirement.

5.3 Informal Results and Techniques

The Optimal Regret Scaling

In this chapter, we show that the optimal regret scaling for LQR attainable by any adaptive LQR algorithm does in fact scale with the square root of the time horizon:

$$\inf_{\mathsf{alg}} \text{LQR-Reg}_T(\mathsf{alg}; \theta_\star) = \widetilde{\Theta}(\sqrt{d_x d_u^2 T} \cdot \text{poly}(\|P_\star\|_{\text{op}}, \|B_\star\|_{\text{op}})), \tag{5.10}$$

provided that the horizon T is sufficiently large. Not only does this regret bound characterize T dependence, it also characterizes the optimal dependence on dimension. When viewed in the context of the prior literature, this result shows that:

- No algorithm can obtain regret which grows slower than the square root of the time horizon T. Our lower bound holds in a strong, local sense about any (sufficiently non-degerate) problem instance. That is, $\sqrt{T}$ regret is the *typical* regret scaling for LQR, rather than a consequence of a pathological problem instance.
- The basic certainty equivalence approach analyzed by Mania et al. [2019] is nearly optimal. Moreover, the algorithm that we analyze and which yields the optimal dimension dependence is a slight variation thereof. Instead of N steps of exploration upfront, the algorithm proceeds in doubling epochs k of length $\tau_k = 2^k$, injecting noise with variance proportional to $1/\sqrt{\tau_k}$ during said epoch. At the end of each epoch, we re-estimate the system parameters $(A_\star, B_\star)$ and execute the resulting certainty equivalent controller is synthesized for the remaining interval $t \in \{2^k, 2^k + 1, \ldots, 2^{k+1} - 1\}$.

Together, these bounds demonstrate that the optimal regret for LQR is attained by a relatively simple algorithm - one whose exploration is simply isotropic noise.

Both our upper and lower bounds are motivated by the following question: Suppose that the learner is selecting near optimal control inputs $\mathbf{u}_t \approx K_\star \mathbf{x}_t$, where $K_\star = K_\infty(A_\star, B_\star)$ is the optimal controller for the system $(A_\star, B_\star)$. What information can she glean about the system?

Intuition: Lower Bound

To understand the regret scaling in Eq. (5.10), consider the following thought experiment. Suppose that the learner (who does not known $(A_\star, B_\star)$) is given the optimal LQR controller K from an untrusted source, and asked to verify that K is the optimal, or say a near-optimal, control policy: that is, $J_\infty(K) \geq J_\infty(K_\star) - \epsilon$.

Let us consider what happens when the K given to our learner is indeed $K_\star$. A natural starting point may be to begin to execute the associated control policy $\mathbf{x}_t = K_\star \mathbf{u}_t$. However, if the learner does so, she cannot estimate the system $(A_\star, B_\star)$ entirely. To see why, observe that if she exectures the optimal control policy, the range of the covariance matrix

$$\mathbf{\Lambda} := \sum_{t=1}^{T} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{bmatrix}$$

lies in the range of the subspace $\mathcal{V}_{K_\star} := \{(x, u) := u = Kx\} \subset \mathbb{R}^{d_x + d_u}$. Thus, the learner cannot estimate $\theta_\star = (A_\star, B_\star)$ along row-directions perpendicular to $\mathcal{V}_{K_\star}$.

Said another way, if the learner executes the policy $\mathbf{x}_t = K_\star \mathbf{u}_t$, she is observing the trajectory whose dynamics are $\mathbf{x}_{t+1} = (A_\star + B_\star K_\star)\mathbf{x}_t + \mathbf{w}_t$. Hence, she cannot disambiguate $(A_\star, B_\star)$ from any other instance

$$\hat{\theta} = (\hat{A}, \hat{B}), \quad \text{satisfying} \quad \hat{A} + \hat{B}K_\star = A_\star + B_\star K_\star. \tag{5.11}$$

Recall that in this thought experiment, the learner wants to verify that $K_\star$ is a near optimal controller, but she does not know that $\theta_\star$ is the ground truth instance. Hence, she may need to rule out other instances $\widehat{\theta}$ consistent with the same closed loop trajectory Eq. (5.11). More precisely, she may need to rule out the instances $\widehat{\theta}$ for which $K_\star$ is not ϵ-suboptimal.

Importantly, it turns out that disambiguating *all* instances of the form Eq. (5.11) is necessary to verify that $K_\star$ is near optimal. To make this precise, we can always express such an instance in terms of some perturbation matrix $\Delta \in \mathbb{R}^{d_x \times d_u}$, via

$$\hat{B} = B_\star + \Delta, \quad \hat{A} = A_\star - \Delta K_\star.$$

We show that, for the optimal controller $\hat{K} = K_\infty(\hat{\theta})$ for the perturbed instance is quantitatively far from $K_\star$:

Lemma 5.1 (Derivative Computation (Proposition 2 in Abeille and Lazaric [2018]))**.** *Let $(A_\star, B_\star)$ be stabilizable Then*

$$\frac{d}{dt} K_\infty(A_\star - t\Delta K_\star, B_\star + t\Delta)\big|_{t=0} = -(R + B_\star^\top P_\star B_\star)^{-1} \cdot \Delta^\top P_\star A_{\mathrm{cl},\star},$$

where we recall $A_{\mathrm{cl},\star} := A_\star + B_\star K_\star$.

In particular, as long as $\sigma_{\min}(A_{\mathrm{cl},\star}) > 1$, and Δ is small enough for a Taylor expansion to hold,

$$\|\hat{K} - K_\star\|_{\mathrm{F}}^2 = \|K_\infty(\hat{\theta}) - K_\infty(\theta_\star)\|_{\mathrm{F}}^2 \geq \Omega(\|\Delta\|_{\mathrm{F}}^2).$$

From a standard suboptimality decomposition for the infinite horizon LQR cost (see e.g. Fazel et al. [2018, Lemma 6]), this implies that

$$J_\infty(\hat{K}; \theta_\star) - J_\infty(K_\star; \theta_\star) = \Omega(\|\Delta\|_{\mathrm{F}}^2).$$

Hence, the learner really does need to estimate $\theta_\star$ in all directions perpendicular to the subspace $\{[\hat{A} \mid \hat{B}] : \hat{A} + B\hat{K} = A_\star + B_\star K_\star]\}$; and an ϵ-Frobenius estimation error roughly translates to a $T\epsilon^2$ control cost.

On the otherhand, the above matrix subspace is of dimension $d_u \times d_x$, and to estimate all directions, one needs to provide excitation from all d_u directions of inputs. One can then show that ϵ-estimation error in the Frobenius requires a regret (i.e. deviation from $K_\star$) of $d_x \cdot d_u^2/\epsilon^2$. Together, these two yield a trade off

$$\min_\epsilon \left\{ T\epsilon^2 + \frac{d_x \cdot d_u^2}{\epsilon^2} \right\} = \sqrt{d_x d_u^2 T},$$

which is precisely the promised regret scaling. This technique is based on Assouad's lemma [Assouad, 1983], and more specifically, an adaptation thereof to adaptive estimation problems [Arias-Castro et al., 2012].

Intuition: Upper Bound

The intuition behind the upper bound is similar. Consider playing some control policy $\mathbf{u}_t = K\mathbf{x}_t + \mathbf{z}_t$, where $\mathbf{z}_t$ is exploratory Gaussian noise with variance σ^2. First suppose that $\mathbf{z}_t \equiv 0$, and consider N steps of exploratory. Then from Chapter 3, one can estimate $A_\star + B_\star K$ with accuracy that decays like $1/N$. There are d_x^2 parameters, so the dimension dependence of the Frobenius norm error is d_x^2/N. This leaves the remaining $d_x d_u$ parameters unaccounted for, and this where we require $\sigma^2 > 0$. For such a σ^2, the remaining parameters are estimated at a rate of $\frac{d_x d_u}{\sigma^2}$. Hence, the least square error scales like

$$\|\hat{\theta} - \theta_\star\|_{\mathrm{F}}^2 \leq \frac{d_x d_u}{N\sigma^2} + \frac{d_x^2}{N}.$$

A crucial part of the argument is showing that the certainty equivalent control policy $\hat{K} = K_\infty(\hat{\theta})$ satisifes

$$J_\infty(K_\infty(\hat{\theta}); \theta_\star) - J_\infty(K_\star; \theta) \leq \|\hat{\theta} - \theta_\star\|_{\mathrm{F}}^2.$$

Hence, the above exploration strategy leads to a control cost of

$$J_\infty(K_\infty(\hat{\theta}); \theta_\star) - J_\infty(K_\star; \theta) \leq \frac{d_x d_u}{N\sigma^2} + \frac{d_x^2}{N}.$$

On the other hand, the noise injection costs adds $\sigma^2 N$ to the regret. Balancing $\sigma^2 N$ and $\frac{d_x d_u}{N\sigma^2}$ yields the same $\sqrt{d_x d_u^2 N}$ scaling.

Novel Perturbation Bounds

Both our upper and lower bounds make use of novel perturbation bounds to control the change in P_∞ and K_∞ when we move from a nominal instance $\theta_\star$ to a nearby instance $\hat{\theta}$. For our upper bound, these are used to show that a good estimator for the nominal instance leads to a good controller, while for our lower bounds, they show that the converse is true. The self-bounding ODE method allows us to prove perturbation guarantees that depend only on the norm of the value function $\|P_\infty(\theta_\star)\|_{\mathrm{op}}$ for the nominal instance, which is a weaker assumption that subsumes previous conditions. The key observation underpinning the method is that the norm of the directional derivative of $\frac{\mathrm{d}}{\mathrm{d}s} P_\infty(\theta(s))\big|_{s=u}$ at a point $ts = u$ along a line $\theta(s)$ is bounded in terms of the magnitude of $\|P_\infty(\theta(u))\|$; we call this the *self-bounding* property. From this relation, we show that bounding the norm of the derivatives reduces to solving a scalar ordinary differential equation, whose derivative saturates the scalar analogue of this self-bounding property. Notably, this technique does not require that the system be controllable, and in particular does not yield guarantees which depend on the smallest singular value of the controllability matrix as in Mania et al. [2019]. Moreover, given estimates $\hat{\theta}$ and an upper-bound on their deviation from the true system $\theta_\star$, our bound allows the learner to check whether the certainty-equivalent controller synthesized from $\hat{\theta}$ stabilizes the true system and satisfies the preconditions for our perturbation bounds.

5.4 Formal Results

This section formally states our main results: the lower bound, a matching upper bound, and a novel perturbation bound for certainty equivalent control which enables both.

Lower Bound

We provide a *local minimax* lower bound, which captures the difficulty of ensuring low regret on both a *nominal instance* $\theta_\star = [A_\star \mid B_\star]$ and on the hardest nearby alternative. For a distance parameter $\varepsilon > 0$, we define the local minimax complexity at scale ε as

$$\mathcal{R}_T(\varepsilon; \theta_\star) := \min_{\mathsf{alg}} \max_{\theta=(A,B)} \left\{ \mathbb{E}[\textsc{Lqr}\text{-}\mathrm{Reg}_T(\mathsf{alg}; \theta)] : \|A - A_\star\|_{\mathrm{F}}^2 \vee \|B - B_\star\|_{\mathrm{F}}^2 \le \varepsilon \right\}.$$

Local minimax complexity captures the idea certain instances $(A_\star, B_\star)$ are more difficult than others, and allows us to provide lower bounds that scale only with control-theoretic parameters of the nominal instance. Of course, the local minimax lower bound immediately implies a lower bound on the global minimax complexity as well.[4] Note further that when ε is sufficiently small, one can find a single controller K_0 which stabilizes all local instances

[4]Some care must be taken in defining the global complexity, or it may well be infinite. One sufficient definition, which captures prior work, is to consider minimax regret over all instances subject to a global bound on $\|P_\star\|$, $\|B_\star\|$, and so on.

under consideration, so the assumption under which our upper bounds hold is satisfied.The following theorem is established in Section 5.7.

Theorem 5.1. *Let $c_1, p > 0$ denote universal constants. For $m \in [d_x]$, define $\nu_m := \sigma_m(A_\star + B_\star K_\star)/\|R + B_\star^\top P_\star B_\star\|_{\mathrm{op}}$. Then if $\nu_m > 0$, we have*

$$\mathcal{R}_T(\varepsilon_T; \theta_\star) \gtrsim \sqrt{d_u^2 m T} \cdot \frac{1 \wedge \nu_m^2}{\|P_\star\|_{\mathrm{op}}^2}, \quad \text{where } \varepsilon_T = \sqrt{d_u^2 m / T},$$

provided that $T \geq c_1(\|P_\star\|_{\mathrm{op}}^p(d_u^2 m \vee \frac{d_x^2 \Psi_{B_\star}^4 (1 \vee \nu_m^{-4})}{m d_u^2} \vee d_x \log(1 + d_x \|P_\star\|_{\mathrm{op}}))$.

Let us briefly discuss some key features of Theorem 5.1.

- The only system-dependent parameters appearing in the lower bound are the operator norm bounds $\Psi_{B_\star}$ and $\|P_\star\|_{\mathrm{op}}$, which only depend on the nominal instance. The latter parameter is finite whenever the system is stabilizable, and does not explicitly depend on the spectral radius or strong stability parameters.

- The lower bound takes $\varepsilon_T \propto T^{-1/2}$, so the alternative instances under consideration converge to the nominal instance $(A_\star, B_\star)$ as $T \to \infty$.

- The theorem can be optimized for each instance by tuning the dimension parameter $m \in [d_x]$: The leading $\sqrt{d_u^2 m T}$ term is increasing in m, while the parameter ν_m scales with $\sigma_m(A_{\mathrm{cl},\star})$ and thus is decreasing in m. The simplest case is when $\sigma_m(A_{\mathrm{cl},\star})$ is bounded away from 0 for $m \gtrsim d_x$; here we obtain the optimal $\sqrt{d_u^2 d_x T}$ lower bound. In particular, if $d_u \leq d_x/2$, we can choose $m = \frac{1}{2} d_x$ to get $\sigma_m(A_{\mathrm{cl},\star}) \geq \sigma_{\min}(A_\star)$.

Upper Bound

To attains regret that scales lie $\sqrt{d_x d_u^2 T}$, we propose certainty equivalence with continual exploration. The pseudocode is given in Algorithm 5.1.

The algorithm proceeds in phases $k = 0, 1, 2, \ldots$, consisting of time steps $t = \tau_k, \tau_k + 1, \ldots, \tau_{k+1} - 1$, where $\tau_k = 2^k$ is the phase length. The algorithms initial phases are dictated by the routine $\mathsf{safeSet}(K_0, \delta)$ (Line 1), whose description is given in ... in .... The goal of this phase is to construct a confidence ball $\mathcal{B}_{\mathrm{safe}}$ of the system parameters, within which certainty equivalence is guaranteed to find stabilizing controllers. This phase uses the initial controller K_0 during the estimation phase, and confidence parameter $\delta > 0$ to control the probability that the confidence ball is invalid. We let k_{safe} denote the phase immediately following the completion of the safeSet routine.

Subsequently, each phase constructs a certainty equivalent controller $\hat{K}_k$ using a current system estimate $\tilde{\theta}_k$ (Line 5), and executes inputs according to that controller perturbed by Gaussian noise (Line 7). The variance of the noise scales as the square-root of the phase length $\sigma_{\mathrm{in}}^2 \tau_k^{1/2}$, where σ_{in} being a parameter selected by the the safeSet routine to optimize

Algorithm 5.1 Certainty Equivalent Control with Continual Exploration

Input: Stabilizing controller K_0, confidence parameter δ.
// Denote $\tau_k = 2^k$
// k_{safe} is adaptively chosen by $\mathsf{safeSet}(K_0, \delta)$.
// $\mathcal{B}_{\mathrm{safe}} \subset \mathbb{R}^{d_x(d_x+d_u)}$ is an operator-norm confidence ball for $\theta_\star$.
Execute routine $\mathsf{safeSet}(K_0, \delta)$ for $t = 1, 2, \ldots, \tau_{k_{\mathrm{safe}}} - 1$, and obtain $(k_{\mathrm{safe}}, \mathcal{B}_{\mathrm{safe}}, \sigma^2_{\mathrm{in}})$.
Set $\tilde{\theta}_{k_{\mathrm{safe}}}$ to be any element of $\mathcal{B}_{\mathrm{safe}}$.
for phases $k = k_{\mathrm{safe}}, k_{\mathrm{safe}} + 1, \ldots,$ **do**
    Synthesize $\hat{K}_k = K_\infty(\tilde{\theta}_k)$. // Certainty equivalence step
    for steps $t = \tau_k, \tau_k + 1, \ldots, \tau_{k+1} - 1$ **do**
        Select $\mathbf{u}_t = \hat{K}_k + \sigma_{\mathrm{in}} \tau_k^{1/4} \mathbf{g}_t$, where $\mathbf{g}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$.
    Let $\widehat{\theta}_{k+1} = (\hat{A}, \hat{B})$ denote the least squares esimate in Eq. (5.12).
    Let $\tilde{\theta}_{k+1}$ denote the Euclidean (Frobenius) projection of $\widehat{\theta}_{k+1}$ on $\mathcal{B}_{\mathrm{safe}}$.

performance. Taking $\sigma^2_{\mathrm{in}} = 1/\sqrt{d_x}$ suffices if one is not concerned about dependence on other system parameters.

Finally, we update our estimate of the system parameters, setting

$$\widehat{\theta}_{k+1} = \operatorname*{arg\,min}_{\theta=(A,B)} \sum_{t=\tau_k}^{\tau_{k+1}-1} \|\mathbf{x}_{t+1} - A\mathbf{x}_t + B\mathbf{u}_t\|_2^2, \tag{5.12}$$

to be the ordinary least-squares estimate of the system parameters. [5] Finally, we project $\widehat{\theta}_{k+1}$ onto the safe set $\mathcal{B}_{\mathrm{safe}}$ to obtain $\tilde{\theta}_{k+1}$, which ensures that the synthesized controller in the following phase is indeed stabilizing.

Theorem 5.2. *When Algorithm 5.1 is invoked with stabilizing controller K_0 and confidence parameter $\delta \in (0, 1/T)$, it guarantees that with probability at least $1 - \delta$,*

$$\begin{aligned}\textsc{Lqr}\text{-}\mathrm{Reg}_T[\mathsf{alg}; \theta_\star] \lesssim \sqrt{d_u^2 d_x T \cdot \Psi_{B_\star}^2 \|P_\star\|_{\mathrm{op}}^{11} \log \frac{1}{\delta}} \\ + r d^2 \cdot \mathcal{P}_0 \Psi_{B_\star}^6 \|P_\star\|_{\mathrm{op}}^{11} (1 + \|K_0\|_{\mathrm{op}}^2) \log \frac{d \Psi_{B_\star} \mathcal{P}_0}{\delta} \log^2 \frac{1}{\delta},\end{aligned}$$

where $\mathcal{P}_0 := J_\infty(K_0; \theta_\star)/d_x$ is the normalized cost of K_0, $d = d_x + d_u$, and $r = \max\{1, \frac{d_u}{d_x}\}$, which is 1 is the typical setting $d_u \le d_x$.

Ignoring dependence on problem parameters, the upper bound of Theorem 5.2 scales asymptotically as $\sqrt{d_u^2 d_x T}$, matching our lower bound. Like the lower bound, the theorem depends on the instance $(A_\star, B_\star)$ only through the operator norm bounds $\Psi_{B_\star}$ and

[5] Due to the Gaussian noise, the minimizer $\widehat{\theta}_{k+1}$ is unique with probability 1 provided that $\tau_k \ge d_x + d_u$

$\|B_\star\|_{\mathrm{op}}$. Similar to previous work [Dean et al., 2018, Mania et al., 2019], the regret bound has additional dependence on the stabilizing controller K_0 through $\|K_0\|_{\mathrm{op}}$ and $\mathcal{P}_0$, but these parameters only affect the lower-order terms.

In the interest of brevity, we do not provide a full proof of Theorem 5.2. Instead, Section 5.6 provides an informal sketch of the proof, highlighting two aspects of our proof which differ from prior work: a novel regret decomposition based on the Hanson-Wright inequality [Rudelson and Vershynin, 2013], and a two-scale least squares estimate which is essential to obtaining the correct dimnesion dependence. This section also provides pseudocode for the safeSet procedure in Algorithm 5.2. The proof relies on a novel perturbation bound, which we describe presently.

A Novel Perturbation Bound

Our upper bound is facilliated by a novel perturbation bound, which differs from past bounds [Mania et al., 2019] in that it does not require controllability, and depends on a single control-theoretic parameter: $\|P_\star\|_{\mathrm{op}}$. [6]

The setup of the perturbation bound is as follows. Consider a ground truth instance $\theta_\star = (A_\star, B_\star)$ with optimal controller $K_\star = K_\infty(\theta_\star)$ and value function $P_\star = P_\infty(\theta_\star)$. Given an alternative instance $\widehat{\theta} = (\hat{A}, \hat{B})$ in the vicinity of $\theta_\star$, we control the suboptimality of the certainty-equivalent controller $K_\infty(\widehat{\theta})$ under $\theta_\star$.

Our perturbation bounds hold in both operator and Frobenius norm, and we let $\|\cdot\|_\circ$ denote either the operator or Frobenius norm, with $\circ \in \{\mathrm{op}, \mathrm{F}\}$. We define the operator and Frobenius errors of $\widehat{\theta}$ via

$$\varepsilon_\circ := \max\{\|A_\star - \hat{A}\|_\circ, \|B_\star - \hat{B}\|_\circ\}, \quad \circ \in \{\mathrm{op}, \mathrm{F}\}. \tag{5.13}$$

Finally, for any stabilizable θ, we define the parameters

$$\mathcal{C}_{\mathrm{safe}}(\theta) = 54\|P_\infty(\theta)\|_{\mathrm{op}}^5, \quad \text{and} \quad \mathcal{C}_{\mathrm{est}}(\theta) = 142\|P_\infty(\theta)\|_{\mathrm{op}}^8. \tag{5.14}$$

Our main theorem is as follows

Theorem 5.3. *Assume the normalizations $Q \succeq I$ and $R \succeq I$. Let $\theta_\star$ be a nominal instance, and $\widehat{\theta}$ an estimate, and let $\varepsilon_{\mathrm{op}}$ and ε_{F} the error bounds in Eq. (5.13). Then if $\varepsilon_{\mathrm{op}} \le 1/\mathcal{C}_{\mathrm{safe}}(\theta_\star)$, the following inequalities hold:*

1. $\|P_\infty(\widehat{\theta})\|_{\mathrm{op}} \le c\|P_\infty(\theta_\star)\|_{\mathrm{op}}$ *and* $\|K_\infty(\theta_\star) - K_\infty(\widehat{\theta})\|_{\mathrm{op}} \le c\|P_\infty(\theta_\star)\|_{\mathrm{op}}^{-3/2}$ *for a universal constant $c > 0$.*
2. $J_\infty(K_\infty(\widehat{\theta}); \theta_\star) - J_\infty(K_\infty(\theta_\star); \theta_\star) \le C_{\mathrm{est}}(\theta_\star) \cdot \varepsilon_{\mathrm{F}}^2$.

[6] This technique also arises in our lower bound by allowing us to reason about the minimal radius ε within which a first-order Taylor of the optimal control policy is accurate.

In words, when the operator norm error between the instances is smaller than $1/\mathrm{poly}(\|P_\star)\|)_{\mathrm{op}}$, the value functions are on the same order, the controllers are close, and the optimal policy under $\theta_\star$ has performance under $\theta_\star$ which scales with the *square* of the Frobenius norm of the difference. As in Mania et al. [2019], the square-norm scaling is essential for $\sqrt{T}$-regret. Establishing an upper bound purely in Frobenius norm is essential to obtain the optimal dimension scaling in the upper bound. This is described at length in Section 5.6.

The second item of Theorem 5.3 follows from the first, and a suboptimality decomposition due to Fazel et al. [2018, Lemma 5]. In Section 5.5, we focus on the proof of an essential component of the bound, and expose a novel technique — the self-bounding ODE method – to that end. A strengthening of Theorem 5.3 can be bound in Simchowitz and Foster [2020, Theorem 5], which, among other things, permits a closeness condition is stated in terms of $\widehat{\theta}$: $\varepsilon_{\mathrm{op}} \le \mathcal{O}(1) \cdot \mathcal{C}_{\mathrm{safe}}(\widehat{\theta})$. Such a bound is appealing because its condition can be verified given only an estimate $\widehat{\theta}$ and confidence radius $\varepsilon_{\mathrm{op}}$, without knowledge of the ground-truth instance.

5.5 The Self-Bounding ODE Method

This section provides an in depth description of the self-bounding ODE method, the key technical ingredient in the proof of Theorem 5.3. For concreteness, we focus our efforts on the proof of the following proposition:

Proposition 5.2. *Define the parameter* $\gamma := 8\|P_\infty(\theta_\star)\|_{\mathrm{op}}^2 \varepsilon_{\mathrm{op}}$. *Then, for any* $\gamma < 1$, *we have*

1. $\|P_\infty(\hat{\theta})\|_{\mathrm{op}} \le (1-\gamma)^{-1/2}\|P_\infty(\theta_\star)\|_{\mathrm{op}}$
2. *For the norms* $\circ \in \{\mathrm{op}, \mathrm{F}\}$,

$$\|K_\infty(\hat{\theta}) - K_\infty(\theta_\star)\|_\circ \le 7(1-u)^{-7/4}\|P_\infty(\theta_\star)\|_{\mathrm{op}}^{7/2} \cdot \varepsilon_\circ.$$

The self-bounding method begins with the following simple observation. Construct the curves for $s \in [0,1]$

$$\begin{aligned}(A(s), B(s)) = \theta(s) &:= s \cdot \hat{\theta} + (1-s) \cdot \theta_\star \\ K(s) &:= K_\infty(\theta(s)), \\ P(s) &:= P_\infty(\theta(s)),\end{aligned} \tag{5.15}$$

Now,suppose that $K(s)$ is well-defined and continuously differentiable with derivative $K'(s)$ for each $s \in [0,1]$. Then, for the norms $\|\cdot\|_\circ$ of interest, we can bound

$$\|K_\star - \hat{K}\|_\circ \le \max_{s\in[0,1]} \|K'(s)\|_\circ.$$

Our aim is to argue that $\|K'(s)\|_\circ$ is on the order of $\varepsilon_\circ$, but a couple questions remain. First of all, how does one bound $K'(s)$? Second, why should $K'(s)$ exist? And lastly, how does one guarantee that $K(s)$ is even *defined*, that is, $\theta(s)$ is stabilizable, for all $s \in [0,1]$?

We now observe from Eq. (5.8) that $K(s)$, where defined, is an analytic function of the corresponding $P(s) := P_\infty(\theta(s))$. Moreover, if $P(s)$ is continuously differentiable at s, one bound $K'(s)$ in terms of $P(s)$, $P'(s)$, $\theta'(s)$, and $\theta(s)$. In Section 5.8, we establish the following bound:

Lemma 5.3. *For all s such that $P(s)$ is continuously differentiable,*

$$\|K'(s)\|_\circ \le 3\|P(s)\|_{\mathrm{op}}\varepsilon_\circ + \|P(s)\|_{\mathrm{op}}^{1/2}\|P'(s)\|_\circ.$$

Thus, the technical challenge amounts to bounding $\|P(s)\|_{\mathrm{op}}$, and arguing that $P(s)$ is smooth, and that $\|P'(s)\|_\circ$ is well defined, and on the order of $\varepsilon_\circ$.

Stability via the implicit function theorem

Define s_{stab} as the largest $s_{\mathsf{stab}} \in [0,1)$ such that $\theta(s)$ is stabilizable for all $s \in [0, s_{\mathsf{stab}})$. We will represent the solution $P(s)$ as the solution of an implicit function to show that $P'(s)$ exists for all $s \in [0, s'_{\mathsf{stab}})$, and if $\|P'(s)\|_{\mathrm{op}}$ is uniformly bounded on $[0, s_{\mathsf{stab}})$, then $s_{\mathsf{stab}} = 1$; i.e. $\theta(s)$ is stabilizable for all $s \in [0,1]$. This argument allows us only to reason about the growth of $P(s)$ where defined. To

From Eq. (5.7), we see that $P_\infty(\theta)$, where defined, is the unique solution of an analytic implicit equation. That is, we can express $P_\infty(\theta)$ as the solution of $\mathcal{F}_{\mathsf{dare}}(P,\theta) = 0$, where $\mathcal{F}_{\mathsf{dare}}$ is jointly analytic in both arguments. For directions $\Delta_P \in \mathbb{S}^{d_x}$, we consider the directional derivatives

$$\mathcal{F}'_{\mathsf{dare}}(s)[\Delta_P] := \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{F}_{\mathsf{dare}}(P + s\cdot\Delta_P, \theta(s))\big|_{P=P(s)}. \tag{5.16}$$

That is, $\mathcal{F}'_{\mathsf{dare}}(s)[\cdot] : \mathbb{S}^{d_x} \to \mathbb{S}^{d_x}$ is the linear operator corresponding to the directional derivatives of the function $\mathcal{F}_{\mathsf{dare}}$ function along direction Δ_P.

We argue that this operator is full rank for all stabilizable $\theta(s)$. To apply the implicit function theorem, we need a little more background. Given $X \in \mathbb{R}^{d_x \times d_x}$ and $Y \in \mathbb{S}^{d_x}$ Discrete Algebraic Ricatti Operator is

$$\mathsf{dlyap}(A,Y) \text{ solves } A^\top X A + Y = 0 \text{ over } X \in \mathbb{S}^{d_x}.$$

When A is stable, i.e. $\rho(A) < 1$, then $\mathsf{dlyap}(A,Y)$ is given by the explicit and unique solution as

$$\mathsf{dlyap}(A,Y) = \sum_{i\ge 0}(A^\top)^i Y A^\top. \tag{5.17}$$

The dlyap operator also characterizes the P-matrices. Specifically, one can show that, for any stabilizable $\theta = (A, B)$,

$$P_\infty(\theta) = \mathsf{dlyap}(A + BK_\infty(\theta), Q + K_\infty(\theta)^\top R K_\infty(\theta)). \tag{5.18}$$

Less obviously, the derivative of $P_\infty(\cdot)$ can also be expressed in terms of the dlyap operator.

Lemma 5.4 (Lemma 3.1 in Simchowitz and Foster [2020])**.** *Let $\theta(s)$ be the curve in Eq. (5.15). For $s \in [0, 1]$ such $\theta(s)$ is stabilizable, define $K(s) = K_\infty(\theta(s))$ and $P = P_\infty(s)$, and set*

$$\begin{aligned} A_{\mathrm{cl}}(s) &:= A(s) + B(s)K(s), \quad \Delta_{A_{\mathrm{cl}}}(s) = A'(s) + B'(s)K(s) \\ \mathcal{Q}_1(s) &:= A_{\mathrm{cl}}(s)^\top P(s)\Delta_{A_{\mathrm{cl}}}(s) + \Delta_{A_{\mathrm{cl}}}(s)^\top P(s) A_{\mathrm{cl}}(s) \end{aligned}$$

Then, for all s for which $\theta(s)$ is stabilizable, we observe $A_{\mathrm{cl}}(s)$ is stable, and

$$\mathcal{F}'_{\mathsf{dare}}(s)[\Delta_P] = A_{\mathrm{cl}}(s)^\top \Delta_P A_{\mathrm{cl}}(s) + \mathcal{Q}_1(s). \tag{5.19}$$

Therefore, $\mathcal{F}'_{\mathsf{dare}}(s)[\cdot]$ is a full rank for all s. Hence, by the implicit function theorem, $P'(s)$ exists and is given by $P'(s) = \mathsf{dlyap}(A_{\mathrm{cl}}(s), \mathcal{Q}_(s))$.

The above lemma has two immediate consequences:

- First, since $\mathcal{F}_{\mathsf{dare}}(\cdot, \theta)$ is continuous, if $P'(s)$ is defined and uniformly bounded (in any norm, say, operator) for all $s \in [0, s_{\mathsf{stab}})$, then a $\lim_{s \to s_{\mathsf{stab}}} P(s)$ exists, and solves $\mathcal{F}_{\mathsf{dare}}(\cdot, \theta(s)) = 0$. Thus, $\theta(s_{\mathsf{stab}})$ is also stabilizable.
- Second, the implicit function theorem guarantees that, if $\mathcal{F}_{\mathsf{dare}}(P, \theta(s)) = 0$ has a solution at $s = s_{\mathsf{stab}}$, it must have a solution at $s \in (s_{\mathsf{stab}} - \epsilon, s_{\mathsf{stab}} + \epsilon)$ for some small enough ϵ.

Taken together, the two points imply that:

Lemma 5.5. *Let $s_{\mathsf{stab}} \in (0, 1]$ denote the largest s such that $\theta(u)$ is stabilizable for all $u \in [0, s)$. Then, $P'(s)$ exists on $[0, s_{\mathsf{stab}})$, and moreover, if $\|P'(s)\|_{\mathrm{op}}$ is uniformly bounded, then $s_{\mathsf{stab}} = 1$, and $P'(s)$ exists on all $[0, 1]$.*

The Self Bounding Property and Its Consequences

The implicit function gives a strategy for establishing existence of $P'(s)$: namely, proving a uniform bound on $P'(s)$ wherever it is defined. To do so, we leverage the following key self-bounding property.

Proposition 5.6. *For any $s \in [0, s_{\mathsf{stab}})$ and norm $\circ \in \{\mathrm{op}, \mathrm{F}\}$, $\|P'(s)\|_\circ \le 4\|P(s)\|_{\mathrm{op}}^3 \varepsilon_\circ$.*

The proof of Proposition 5.6 relies on various properties of the dlyap operator, and is deferred to Section 5.8. To leverage the proposition, consider a scalar ODE which saturates the upper bound on $P'(s)$. For concreteness, let us take the operator norm:

$$z'(s) = 4z(s)^3\varepsilon_{\mathrm{op}}, \quad z(0) = \|P(0)\|_{\mathrm{op}} = \|P_\star\|_{\mathrm{op}}. \tag{5.20}$$

We can explicitly solve for the solution to Eq. (5.20):

$$\frac{\mathrm{d}z}{z^3} = 4\varepsilon_{\mathrm{op}}\mathrm{d}s, \quad \frac{1}{2z(0)^2} - \frac{1}{2z(s)^2} = 4\varepsilon_{\mathrm{op}}\mathrm{d}s$$

Hence,

$$z(s) = \frac{1}{\sqrt{z(0)^{-2} - 8s\varepsilon_{\mathrm{op}}}} = \frac{1}{\sqrt{\|P_\star\|_{\mathrm{op}}^{-2} - 8s\varepsilon_{\mathrm{op}}}}, \quad \forall s < \frac{1}{8\|P_\star\|_{\mathrm{op}}^2\varepsilon_{\mathrm{op}}}$$

We take a moment to appreciate the fact that, while the computation of $P(s)$ along $s \in [0, 1]$ is quite complicated, the solution to the scalar ODE Eq. (5.20) is simple and closed-form. The crux of the self-bounding ODE method is the following observation:

Proposition 5.7 (Self-Bounding)**.** *Let $z(s)$ denote the solution to the ODE in Eq. (5.20). Then, for all $s \in [0, s_{\mathsf{stab}})$, $\|P(s)\|_{\mathrm{op}} \le z(s)$. Hence, $\|P'(s)\|_{\mathrm{op}} \le 4z(s)^3\varepsilon_{\mathrm{op}}$.*

The above proposition is a specialization of a more general technique, [Simchowitz and Foster, 2020, Theorem 13], which applies to a class of "valid implicit functions" generalizing $\mathcal{F}_{\mathsf{dare}}(P, \theta)$.

In particular, defining the parameter $\gamma = 8\|P_\infty(\theta_\star)\|_{\mathrm{op}}^2\varepsilon_{\mathrm{op}}$, we see that for $\gamma < 1$, then $P(s)$ exists for all $s \in [0, 1]$, and

$$\|P(s)\|_{\mathrm{op}} \le (1-\gamma)^{-1/2}\|P_\star\|_{\mathrm{op}}.$$

This establishes the first part of Proposition 5.2. The second part of the proposition follows from substituting this bound into Lemma 5.3. □

5.6 Upper Bound

In the interest of brevity, we provide only a sketch of the proof of our upper bound; full details are given in Simchowitz and Foster [2020, Section 5]. This section also provides the pseudocode for the safeSet procedure in Algorithm 5.2

The aim of this section is to highlight the three aspects of our analysis which differ from prior work:

1. A regret-decomposition based on the Hanson-Wright inequality.

Algorithm 5.2 safeSet(K_0, δ)

Input: Stabilizing controller K_0, confidence parameter δ.
for $k = 0, 1, \ldots$ **do**
  // Let $\tau_k = 2^k$
  for steps $t = \tau_k, \tau_k + 1, \ldots, \tau_{k+1} - 1$ **do**
    Play input $\mathbf{u}_t = K_0 \mathbf{x}_t + \mathbf{g}_t$, where $\mathbf{g}_t \sim \mathcal{N}(0, I)$ // take $\mathbf{x}_0 = 0$
  Let $\widehat{\theta}_k = (\hat{A}_k, \hat{B}_k)$ be the OLS estimator (Eq. (5.12))
  // Let $\mathbf{\Lambda}_k := \sum_{t=\tau_k}^{\tau_{k+1}-1} (\mathbf{x}_t, \mathbf{u}_t)^{\otimes 2}$ denote phase covariance
  Define $\mathsf{conf}_k := 6\lambda_{\min}(\mathbf{\Lambda}_k)^{-1}(d \log 5 + \log(4k^2 \det(3(\mathbf{\Lambda}_k)/\delta))$ // infinite if $\mathbf{\Lambda}_k \not\succ 0$
  if $1/\mathsf{conf}_k \geq 486 \|P_\infty(\widehat{\theta}_k)\|_{\mathrm{op}}^5$ **then**
    Set $k_{\mathrm{safe}} \leftarrow k + 1$
    Set $\mathcal{B}_{\mathrm{safe}}$ to denote the operator norm ball around $\widehat{\theta}_k$ of radius conf_k
    Set $\sigma_{\mathrm{in}}^2 = \sqrt{d_x} \|P_\infty(\widehat{\theta}_k)\|_{\mathrm{op}}^{9/2} \max\{1, \|\hat{B}_k\|_{\mathrm{op}}\} \sqrt{\log \frac{\|P_\infty(\widehat{\theta}_k)\|_{\mathrm{op}}}{\delta}}$
    Return $(k_{\mathrm{safe}}, \mathcal{B}_{\mathrm{safe}}, \sigma_{\mathrm{in}}^2)$
  else
    Continue

2. The use of our perturbation bound to relate regret to the cumulative Frobenius errors $\|\widehat{\theta}_k - \theta_\star\|_{\mathrm{F}}^2$.

3. A two-scale least-squares error bound which leverages the tools from Chapter 3 to carefully bound $\|\widehat{\theta}_k - \theta_\star\|_{\mathrm{F}}^2$.

This latter step is crucial because as described above, the correct analysis must distinguish between the $d_x \times d_x$directions of $\theta_\star$ along the subspace $(x, u) = (x, \hat{K}_k x)$ which are estimated "quickly", and those $d_x \times d_u$ directions perpendicular to the subspace which are estimated "slowly" due to the noise injection.

Let us begin the analysis. We first argue that the initial safeSet stage terminates with high probability in a small number of phases, so that its total contribution to the regret remains second order. Moreover, with high probability, when the termination occurs, the confidence ball $\mathcal{B}_{\mathrm{safe}}$ only contains parameters θ for which $K_\infty(\theta)$ stabilizes $\theta_\star$, and for which the cost of running $K_\infty(\theta)$ is comparable (but not vanishingly close) to that of $K_\star$. This ensures that the algorithm behaves in a regular fashion for the successive phases $k \geq k_{\mathrm{safe}}$.

The brunt of our regret arises form phases $k \geq k_{\mathrm{safe}}$. Let $\sigma_k = \sigma_k^2/\sqrt{\tau_k}$ denote the variance of the injected input noise at phase k. We show a novel regret decomposition where the regret from those phases is at most

$$\sum_{k=k_{\mathrm{safe}}}^{\log_2 T} \tau_k \left(J_\infty(\hat{K}_k; \theta_\star) - J_\infty(K_\star; \theta_\star) \right) + d_u \tau_k \sigma_k^2 + \widetilde{\mathcal{O}}\left(\sqrt{(d_u + d_x)\tau_k} \right) + (\text{low order terms}).$$

The first term, $J_\infty(\hat{K}_k; \theta_\star) - J_\infty(K_\star; \theta_\star)$, captures the suboptimality incurred by choosing a suboptimal controller $\hat{K}_k$ during thet τ_k step of phase k. The second term captures the

contribution of the Gaussian exploration during phase k. The third term adresses the fluctuation of the random costs around its expectation. Past work has adressed controlled the random fluctuations with an Azuma-Bernstein inequality, which yields fluctuations on the order of $(d_u + d_x)\sqrt{\tau_k}$, contributing $(d_u + d_x)\sqrt{T}$ to the final regret bound. In particular, when $d_x \gg d_u^2$, this yields larger regret than the optimal $\sqrt{d_x d_u^2 T}$ scaling. Using the Hanson-Wright inequality, we obtain a sharper control of the fluctuations on the order of $\widetilde{\mathcal{O}}\left(\sqrt{(d_u + d_x)T}\right)$, which is dominated by the optimal $\sqrt{d_x d_u^2 T}$ scaling. The formal regret decomposition is stated in Simchowitz and Foster [2020, Lemma 5.2]. For further discussion, see Simchowitz and Foster [2020, Appendix G.8].

Next, using our novel perturbation bound (Theorem 5.3), we find that

$$J_\infty(\hat{K}_k; \theta_\star) - J_\infty(K_\star; \theta_\star) \le \mathrm{poly}(\|P_\star\|_{\mathrm{op}}) \cdot \|\widehat{\theta} - \theta_\star\|_{\mathrm{F}}^2.$$

Hence, it remains to control the rate at which the estimates of $\theta_\star$ approach the ground truth. We aim to show that, up to logarithmic factors and polynomial factors in system parameters,

$$\|\widehat{\theta} - \theta_\star\|_{\mathrm{F}}^2 \le d_x \cdot \left(\frac{d_x}{\tau_k} + \frac{d_u}{\tau \sigma_k^2}\right). \tag{5.21}$$

As soon as this holds, it is straightforward to see that we can obtain $\sqrt{\dim d_u^2 T}$ by combining Eq. (5.21) with our perturbation bound and regret decomposition, and tuning σ_k^2 $\sqrt{\tau_k d_x}$ appropriately.

Let us understand Eq. (5.21). We can view parameters θ as matrices with rows $\theta^{(i)}$ for $i = 1, 2, \ldots, d_x$. Since the least squares estimator decouples across rows, the total error is

$$\|\widehat{\theta} - \theta_\star\|_{\mathrm{F}}^2 = \sum_{i=1}^{d_x} \|\widehat{\theta}^{(i)} - \theta_\star^{(i)}\|_{\mathrm{F}}^2,$$

giving us the d_x factor out in front.

Let $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{u}_t)$ denote the covariates used for the least squares estimator. Let $\mathbf{Z}_k$ denote the matrix whose rows are $\mathbf{z}_t$ for $t \in \{\tau_k, \tau_k + 1, \ldots, \tau_{k+1} - 1\}$, and $\mathbf{W}_k$ the matrix whose rows $\mathbf{w}_{t+1}$ for the same indices t. Finally, the appropriate covariance matrix

$$\mathbf{\Lambda}_k := \sum_{t=\tau_k}^{\tau_{k+1}-1} \mathbf{z}_t \mathbf{z}_t^\top = \mathbf{Z}_k^\top \mathbf{Z}_k.$$

Letting $\{e_i\}$ denote the cannonical basis vectors for $\mathbb{R}^{d_x}$, a standard least-squares decomposition gives that, for all $i \in [d_x]$,

$$\widehat{\theta}^{(i)} - \theta_\star^{(i)} = \mathbf{\Lambda}_k^{-1} \mathbf{Z}_k^\top \mathbf{W}_k e_i, \tag{5.22}$$

provided that $\mathbf{\Lambda}_k$ is not rank defficient.

The Two-Scale Property

Let us now elucidate a "two-scale" structure for $\mathbf{\Lambda}_k$. To do so, let $(\mathcal{F}_t)$ denote the filtration generated by all inputs $\mathbf{u}_s$ and disturbances $\mathbf{w}_s$ for times $1 \le s \le t$. Then, for all t in phase k,

$$\begin{aligned}\mathbb{E}\left[\mathbf{z}_t\mathbf{z}_t^\top \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbf{z}_t \mid \mathcal{F}_{t-1}\right]\mathbb{E}\left[\mathbf{z}_t \mid \mathcal{F}_{t-1}\right]^\top + \begin{bmatrix} I \\ \hat{K}_k \end{bmatrix}\begin{bmatrix} I \\ \hat{K}_k \end{bmatrix}^\top + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{\mathrm{in}}^2\sqrt{\tau_k} I_{d_u} \end{bmatrix} \\ &\succeq \begin{bmatrix} I \\ \hat{K}_k \end{bmatrix}\begin{bmatrix} I \\ \hat{K}_k \end{bmatrix}^\top + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{\mathrm{in}}^2\sqrt{\tau_k} I_{d_u} \end{bmatrix},\end{aligned}$$

where the blocks in the last matrix are of dimension d_x and d_u, respectively. In particular, if we define the projection matrix P_k projecting onto the null space of $\begin{bmatrix} I \\ \hat{K}_k \end{bmatrix}$, and $\mathsf{P}_k^\perp := I - \mathsf{P}_k$ denote the projection onto its rowspace, then a careful linear algebraic argument shows that the above covariance is lower bounded in a PSD sense by

$$\mathbb{E}\left[\mathbf{z}_t\mathbf{z}_t^\top \mid \mathcal{F}_{t-1}\right] \succeq \frac{\sigma_k}{2}\mathsf{P}_k + \cdot c_1 \mathsf{P}_k^\perp$$

where c_1 is a constant that is polynomial in problem parameters. The formal statement is given by Simchowitz and Foster [2020, Lemma G.8], and is based on an argument due to Dean et al. [2018]. Using the Gaussianity of the process and input noise, one can show use the block-martingale small-ball argument developed in Chapter 3 to establish that, with high probability

$$\Lambda_k \gtrsim \bar{\Lambda}_{k,\mathrm{low}} := \frac{\tau_k\sigma_k^2}{2}\mathsf{P}_k + c_1\tau_k\mathsf{P}_k^\perp, \qquad \text{(Two-Scale Property)}$$

where c_1 is universal constant. We refer to this as the two scale property because the term, because the first term is considerably smaller than the second term, since $\tau_k\sigma_k \approx \sqrt{d_x\tau_k}$ is considerably smaller than τ_k.

Remark 5.1. We remark on the convenience of using the martingale small-ball method developed in Chapter 3 to lower bound the covariance $\mathbf{\Lambda}_k \succeq \bar{\mathbf{\Lambda}}_{k,\mathrm{low}}$. A standard approach based on concentration would use a matrix Bernstein inequality to argue that $\mathbf{\Lambda}_k = \sum_{t=\tau_k}^{\tau_{k+1}-1}\mathbf{z}_t\mathbf{z}_t^\top$ concentrates around its expectation. Even ignoring dependence issues (these can be adressed with mixing-time arguments), terms $\mathbf{z}_t\mathbf{z}_t^\top$ have variance $\|\mathbb{E}[(\mathbf{z}_t\mathbf{z}_t^\top)^2]\|_{\mathrm{op}}\, d_x$, yielding a leading order deviation on the order of $\sqrt{\tau_k d_x}$ (see e.g. Tropp [2015, Theorem 1.6.2]). This deviation is precisely the same order as the term $\frac{\tau_k\sigma_k^2}{2}\mathsf{P}_k$-term in $\bar{\mathbf{\Lambda}}_{k,\mathrm{low}}$. Hence, an argument based on concentration would need to tinker with σ_k^2 to be large enough to overcome these fluctuations. In contrast, the small-ball technique has no such limitation.

A Two-Scale Self-Normalized Bound

Let us understand how to apply the two-scale property to analyze the least-squares estimate. Suppose for the sake of argument that the noise terms $\mathbf{w}_t$ and independent of the covariates $\mathbf{z}_t$; of course this is false due, but will help illustrate intuition. Then,

$$\|\widehat{\theta}^{(i)} - \theta_\star^{(i)}\|^2 = e_i^\top \mathbf{W}_k^\top \mathbf{Z}_k \mathbf{\Lambda}_k^{-2} \mathbf{Z}_k^\top \mathbf{W}_k e_i = e_i^\top \mathbf{W}_k \mathbf{\Lambda}_k^{-1} \mathbf{W}_k e_i \tag{5.23}$$

By the Hanson-Wright inequality, this errors leading term is on the order of $\mathrm{tr}(\mathbf{\Lambda}_k^{-1})$. Using the two scale property and inverting the matrix $\bar{\Lambda}_{k,\mathrm{low}}$,

$$\|\widehat{\theta}^{(i)} - \theta_\star^{(i)}\|^2 \approx \mathrm{tr}(\bar{\Lambda}_{k,\mathrm{low}}) = \frac{2d_u}{\tau_k \sigma_k^2} + \frac{c_1 d_x}{\tau_k},$$

Summing up over $i \in [d_x]$ yields the error bound in Eq. (5.21). Unfortunately, this simple heuristic argument fails, and we develop a more sophisticated bound in the following section.

We present the following bound, a specialization of Simchowitz and Foster [2020, Lemma E.1]. This bound adapts to the two-scale propety, but respects the martingale structure of the problem.

Lemma 5.8 (Two-Scale OLS Estimate)**.** *Let $d = d_x + d_u$, and let $\mathsf{P} \in \mathbb{R}^{d\times d}$ denote a projection matrix onto a subspace of dimension p, and fix an orthonormal basis of $\mathbb{R}^{d_x}$, $v_1, \ldots, v_{d_x}$, such that $v_1, \ldots, v_p$ form an orthonormal basis for the range of P. Further, fix positive constants $0 < \lambda_1 \le \nu \le \lambda_2$ such that $\nu \le \sqrt{\lambda_1\lambda_2/2}$, and define the event*

$$\mathcal{E} := \{\mathbf{\Lambda}_k \succeq \lambda_1 \mathsf{P} + \lambda_2(I - \mathsf{P})\} \cap \{\|\mathsf{P}\mathbf{\Lambda}_k(I - \mathsf{P})\|_{\mathrm{op}} \le \nu\}$$

Then, with probability $1 - \delta$, if $\mathcal{E}$ holds,

$$\|\widehat{\theta}_k - \theta_\star\|_{\mathrm{F}}^2 \le \frac{12 d_x p \boldsymbol{\kappa}_1}{\lambda_1} \log \frac{3 d_x d \boldsymbol{\kappa}_1}{\delta} + \left(\frac{\nu}{\lambda_1}\right)^2 \cdot \frac{48 d_x (d-p) \boldsymbol{\kappa}_2}{\lambda_2} \log \frac{3 d_x d \boldsymbol{\kappa}_2}{\delta}.$$

where $\boldsymbol{\kappa}_1 := \max_{1\le j\le p} v_j^\top \mathbf{\Lambda}_k v_j/\lambda_1$ and $\boldsymbol{\kappa}_2 := \max_{p+1\le j\le d_x} v_j^\top \mathbf{\Lambda}_k v_j/\lambda_2$.

In our case, we apply the bound with $\mathsf{P} \leftarrow \mathsf{P}_k$ denoting the projection operator described in the two-scale bound; this projects onto a subspace of dimension $p = d_u$. Then, using $\lambda_1 = \frac{\tau_k \sigma_k^2}{2}$ and $\lambda_2 = c_1 \tau_k$ from the two-scale property,

$$\|\widehat{\theta}_k - \theta_\star\|_{\mathrm{F}}^2 \le \widetilde{\mathcal{O}}\left(\boldsymbol{\kappa}_1 \cdot \frac{d_x d_u}{\tau_k \sigma_k^2} + \boldsymbol{\kappa}_2 \cdot (\frac{\nu}{\sigma_k^2 \tau_k})^2 \cdot \frac{d_x^2}{\tau_k}\right),$$

attaining Eq. (5.21), provided that (a) $\boldsymbol{\kappa}_1$, $\boldsymbol{\kappa}_2$ are constants, and (b) that we can choose ν to be on the order of $\tau_k \sigma_k^2 \sim \sqrt{\tau_k}$; this requires some technical effort, but is demonstrated in Simchowitz and Foster [2020, Appendix G].

The key step in the proof of Lemma 5.8 is relative the *square* of the covariance matrice, $\mathbf{\Lambda}_k^2$, to its lower bound $\bar{\mathbf{\Lambda}}_{k,\mathrm{low}}^2$. Even though $\mathbf{\Lambda}_k \succeq \bar{\mathbf{\Lambda}}_{k,\mathrm{low}}$ with high probability, this does not imply any similar relation for the squares; indeed, given positive matrices A and B, $A \preceq B$ does not imply $A^2 \preceq B^2$. To circumvent this, we use the two-scale structure of the covariance lower bound $\bar{\mathbf{\Lambda}}_{k,\mathrm{low}}$ to show that $\mathbf{\Lambda}_k^2 \succeq \bar{\mathbf{\Lambda}}_{k,\mathrm{low}}^2$, provided the cross term $\|\mathsf{P}_k \mathbf{\Lambda}_k (1 - \mathsf{P}_k)\|_{\mathrm{op}}$ is small.

5.7 Lower Bound

We now prove the main lower bound, Theorem 5.1. The proof follows the plan outlined in Section 5.4: We construct a packing of alternative instances, show that low regret on a given instance implies low estimation error, and then deduce from an information-theoretic argument that this implies high regret an alternative instance. Throughout that we assume $\sigma_w^2 = 1$ for simplicity. In the interest of brevity, all constituent lemmas are stated without proof; an unabridged version (with references to full proofs) can be found in [Simchowitz and Foster, 2020, Section 4].

Alternative Instances and Packing Construction

We construct a packing of alternate instances $\theta_e = [A_e \mid B_e]$ which take the form $[A_\star + K_\star \Delta_e \mid B_\star + \Delta_e]$, for appropriately chosen perturbations Δ_e described shortly. As discussed in Section 5.4, this packing is chosen because the learner *cannot* distinguish between alternatives if she commits to playing the optimal policy $\mathbf{u}_t = K_\star \mathbf{x}_t$, and must therefore deviate from this policy in order to distinguish between alternatives. We further recall Lemma 5.1, which describes how the optimal controllers from these instances varying with the perturbation Δ.

Lemma 5.1 (Derivative Computation (Proposition 2 in Abeille and Lazaric [2018]))**.** *Let $(A_\star, B_\star)$ be stabilizable Then*

$$\frac{d}{dt} K_\infty(A_\star - t\Delta K_\star, B_\star + t\Delta)\big|_{t=0} = -(R + B_\star^\top P_\star B_\star)^{-1} \cdot \Delta^\top P_\star A_{\mathrm{cl},\star},$$

where we recall $A_{\mathrm{cl},\star} := A_\star + B_\star K_\star$.

In particular, if A_{cl} is non-degenerate, then to first order, the Frobenius distance between between the optimal controllers for $A_\star, B_\star$ and the alternatives (A_e, B_e) is $\Omega(\|\Delta\|_{\mathrm{F}})$.

To obtain the correct dimension dependence, it is essential that the packing is sufficiently large; a single alternative instance will not suffice. Our goal is to make the packing as large as possible while ensuring that if one can recover the optimal controller for a given instance, they can also recover the perturbation Δ.

Let $n = d_u$, and let $m \le d_x$ be the free parameter from the theorem statement. We construct a collection of instances indexed by sign vectors $e \in \{-1, 1\}^{[n] \times [m]}$. Let $w_1, \dots, w_n$

denote an eigenbasis basis of $(R + B_\star^\top P_\star B_\star)^{-1}$, and $v_1, \dots, v_m$ denote the first m right-singular vectors of $A_{\mathrm{cl},\star} P_\star$. Then for each $e \in \{-1, 1\}^{[n]\times[m]}$, the corresponding instances is $\theta_e = [A_e \mid B_e]$, where

$$[A_e \mid B_e] := [A_\star - \Delta_e K_\star \mid B_\star + \Delta_e], \quad \text{where } \Delta_e = \varepsilon_{\mathrm{pack}} \sum_{i=1}^{n} \sum_{j=1}^{m} e_{i,j} w_i v_j^\top. \tag{5.24}$$

It will be convenient to adopt the shorthand $K_e := K_\infty(\theta_e)$, $P_e = P_\infty(A_e, B_e)$ and $J_e = J_\infty(K_e; \theta_e)$, and

$$\Psi_\star = \max\{1, \|A_\star\|_{\mathrm{op}}, \|B_\star\|_{\mathrm{op}}\} \quad \Psi_e = \max\{1, \|A_e\|_{\mathrm{op}}, \|B_e\|_{\mathrm{op}}\}$$

The following lemma gathers a number of bounds on the error between θ_e and $\theta_\star$ and their corresponding system parameters. Perhaps most importantly, the lemma shows that to first order, K_e can be approximated using the derivative expression in Lemma 5.1.

Lemma 5.9. *There exist universal polynomial functions $\mathfrak{p}_1, \mathfrak{p}_2$ such that, for any $\varepsilon_{\mathrm{pack}} \in (0, 1)$, if $\varepsilon_{\mathrm{pack}}^2 \le \mathfrak{p}_1(\|P_\star\|_{\mathrm{op}})^{-1}/nm$, the following bounds hold:*

1. **Parameter errror:** $\max\{\|A_e - A_\star\|_{\mathrm{F}}, \|B_e - B_\star\|_{\mathrm{F}}\} \le \sqrt{\|P_\star\|_{\mathrm{op}}}\sqrt{mn}\varepsilon_{\mathrm{pack}}$.
2. **Boundedness of value functions:** $\Psi_e \le 2^{1/5}\Psi_\star$ *and* $\|P_e - P_\star\|_{\mathrm{op}} \le 2^{1/5}\|P_\star\|_{\mathrm{op}}$.
3. **Controller error:** $\|K_e - K_\star\|_{\mathrm{F}}^2 \le 2\|P_\star\|_{\mathrm{op}}^3 mn\varepsilon_{\mathrm{pack}}^2$.
4. **First-order error:** $\|K_\star + \frac{d}{dt}K_\infty(\theta_\star + t\theta_e)\big|_{t=0} - K_e\|_{\mathrm{F}}^2 \le \mathfrak{p}_2(\|P_\star\|_{\mathrm{op}})^2(mn)^2\varepsilon_{\mathrm{pack}}^4$.

Notably, item 4 ensures that the first order approximation in Lemma 5.1 is accurate for $\varepsilon_{\mathrm{pack}}$ sufficiently small. Henceforth, we take $\varepsilon_{\mathrm{pack}}$ sufficiently small so as to satisfy the conditions of Lemma 5.9.

Assumption 5.3 (Small $\varepsilon_{\mathrm{pack}}$)**.** $\varepsilon_{\mathrm{pack}}^2 \le \frac{1}{mn}(\mathfrak{p}_1(\|P_\star\|_{\mathrm{op}})^{-1} \wedge \frac{1}{20}\mathfrak{p}_2(\|P_\star\|_{\mathrm{op}})^{-1})$.

Low Regret Implies Estimation for Controller

We now show that if one can achieve low regret on every instance, then one can estimate the infinite-horizon optimal controller K_e. Suppressing dependence on T, we introduce the shorthand

$$\mathrm{PseudoReg}_e[\pi] := \mathbb{E}_{\pi,\theta_\star}[\textsc{Lqr}\text{-}\mathrm{Reg}_T(\pi; \theta_e)].$$

Going forward, we restrict ourselves to algorithms whose regret is sufficiently small on every packing instance; the trivial case where this is not satisfied is handled at the end of the proof.

Assumption 5.4 (Uniform Correctness)**.** For all instances (A_e, B_e), the algorithm π ensures that $\mathrm{PseudoReg}_e[\pi] \le \frac{T}{6d_x\|P_\star\|_{\mathrm{op}}\Psi_\star^2} - \epsilon_{\mathrm{err}}$, where $\epsilon_{\mathrm{err}} := 6\|P_\star\|_{\mathrm{op}}^3\Psi_\star^2$.

We now define an intermediate term which captures which captures the extent to which the control inputs under instance e deviate from those prescribed by the optimal infinite horizon controller K_e on the first $T/2$ rounds:

$$K\text{-Err}_e[\pi] := \mathbb{E}_{\pi,\theta_\star}\left[\sum_{t=1}^{T/2} \|\mathbf{u}_t - K_e\mathbf{x}_t\|^2\right].$$

The following lemma shows that regret is lower bounded by K-$\text{Err}_e[\pi]$, and hence any algorithm with low regret under this instance must play controls close to $K_e\mathbf{x}_t$.

Lemma 5.10. *There is a universal constant $\gamma_{\text{err}} > 0$ such that if Assumptions 5.3 and 5.4 hold and $T \geq \gamma_{\text{err}}\|P_\star\|_{\text{op}}^2\Psi_\star^4$, then*

$$\text{PseudoReg}_e[\pi] \geq \frac{1}{2}K\text{-Err}_e[\pi] - \epsilon_{\text{err}}.$$

In light of Lemma 5.10, the remainder of the proof will focus on lower bounding the deviation K-Err_e. As a first step, the next lemma shows that the optimal controller can be estimated well through least squares whenever K-Err_e is small. More concretely, we consider a least squares estimator which fits a controller using the first half of the algorithm's trajectory. The estimator returns

$$K_{\text{ls}} := \operatorname*{arg\,min}_K \sum_{t=1}^{T/2} \|\mathbf{u}_t - K\mathbf{x}_t\|^2, \tag{5.25}$$

when $\sum_{t=1}^{T/2}\mathbf{x}_t\mathbf{x}_t^\intercal \succeq c_{\text{min}}T\cdot I$, and returns $K_{\text{ls}} = 0$ otherwise.

Lemma 5.11. *If $T \geq c_0 d_x \log(1 + d_x\|P_\star\|_{\text{op}})$ and Assumptions 5.3 and 5.4 hold, and if c_{min} is chosen to be an appropriate numerical constant, then the least squares estimator Eq. (5.25) guarantees*

$$K\text{-Err}_e[\pi] \geq c_{\text{ls}}T\cdot\mathbb{E}_{A_e,B_e,\pi}\left[\|K_{\text{ls}} - K_e\|_{\text{F}}^2\right] - 1,$$

where c_0 and c_{ls} are universal constants.

Henceforth we take T large enough such that Lemma 5.10 and Lemma 5.11 apply.

Assumption 5.5. We have that $T \geq c_0 d_x \log(1 + d_x\|P_\star\|_{\text{op}}) \vee \gamma_{\text{err}}\|P_\star\|_{\text{op}}^2\Psi_\star^4$.

Information-Theoretic Lower Bound for Estimation

We have established that low regret under the instance (A_e, B_e) requires a small deviation from K_e in the sense that K-$\text{Err}_e[\pi]$ is small, and have shown in turn that any algorithm with low regret yields an estimator for the optimal controller K_e (Lemma 5.11). We now

provide necessary condition for estimating the optimal controller, which will lead to the final tradeoff between regret on the nominal instance and the alternative instance. This condition is stated in terms of a quantity related to K-Err_e:

$$K_\star\text{-}\mathrm{Err}_e[\pi] := \mathbb{E}_{A_e,B_e,\pi}\left[\sum_{t=1}^{T/2}\|\mathbf{u}_t - K_\star\mathbf{x}_t\|^2\right].$$

Both $K_\star$-$\mathrm{Err}_e[\pi]$ and K-$\mathrm{Err}_e[\pi]$ concern the behavior of the algorithm under instance (A_e, B_e), but former measures deviation from $K_\star$ ("exploration error") while the latter measures deviation from the optimal controller K_e. Our proof essentially argues the following. Let $(\mathbf{e}, \mathbf{e}')$ be a pair of random indices on the hypercube, where $\mathbf{e}$ is uniform on $\{-1,1\}^{nm}$, and $\mathbf{e}'$ is obtained by flipping a single, uniformly selected entry of $\mathbf{e}$. Moroever, let $\mathbb{P}_{\mathbf{e}}, \mathbb{P}_{\mathbf{e}'}$ denote the respective laws for our algorithm under these two instances. We show that—because our instances take the form $(A_\star - \Delta K_\star, B + \Delta)$—$K_\star$-$\mathrm{Err}_e[\pi]$ captures the KL divergence between these two instances:

$$\mathbb{E}_{\mathbf{e}}K_\star\text{-}\mathrm{Err}_{\mathbf{e}}[\pi] \approx \mathbb{E}_{\mathbf{e},\mathbf{e}'}\mathrm{KL}(\mathbb{P}_{\mathbf{e}}, \mathbb{P}_{\mathbf{e}'}),$$

where the expectations are taken with respect to the distribution over $(\mathbf{e}, \mathbf{e}')$. In other words, the average error $\mathbb{E}_{\mathbf{e}}K_\star$-$\mathrm{Err}_{\mathbf{e}}[\pi]$ corresponds to the average one-flip KL-divergence between instances. This captures the fact that the instances can only be distinguished by playing controls which deviate from $\mathbf{u}_t = K_\star\mathbf{x}_t$.

As a consequence, using a technique based on Assouad's lemma [Assouad, 1983] due to Arias-Castro et al. [2012], we prove an information-theoretic lower bound that shows that any algorithm that can recover the index vector e in Hamming distance on every instance must have $K_\star$-$\mathrm{Err}_e[\pi]$ is large on some instances.

As described above, the following lemma concerns the case where the alternative instance index e is drawn uniformly from the hypercube. Let $\mathbb{E}_{\mathbf{e}}$ denote expectation $\mathbf{e} \overset{\text{unif}}{\sim} \{-1,1\}^{[n]\times[m]}$, and let $\mathsf{d}_{\mathrm{ham}}(e, e')$ denote the Hamming distance.

Lemma 5.12. *Let $\widehat{e}$ be any estimator depending only on $(\mathbf{x}_1, \dots, \mathbf{x}_{T/2})$ and $(\mathbf{u}_1, \dots, \mathbf{u}_{T/2})$. Then*

$$\textit{either} \quad \mathbb{E}_{\mathbf{e}}K_\star\text{-}\mathrm{Err}_{\mathbf{e}}[\pi] \geq \frac{n}{4\varepsilon_{\mathrm{pack}}^2}, \quad \textit{or} \quad \mathbb{E}_{\mathbf{e}}\mathbb{E}_{A_{\mathbf{e}},B_{\mathbf{e}},\mathsf{alg}}\left[\mathsf{d}_{\mathrm{ham}}(\mathbf{e}, \widehat{e})\right] \geq \frac{nm}{4}.$$

To apply this result to the least squares estimator K_{ls}, we apply the following lemma, which shows that any estimator $\widehat{K}$ with low Frobenius error relative to K_e can be used to recover e in Hamming distance.

Lemma 5.13. *Let $\widehat{e}_{i,j}(\hat{K}) := \mathrm{sign}(w_i^\top(\hat{K} - K_\star)v_j)$, and define $\nu_k := \|R + B_\star^\top P_\star B_\star\|_{\mathrm{op}}/\sigma_k(A_{\mathrm{cl},\star})$. Then under Assumption 5.3,*

$$\mathsf{d}_{\mathrm{ham}}(\widehat{e}_{i,j}(\hat{K}), e_{i,j}) \leq \frac{2\left\|\hat{K} - K_e\right\|_{\mathrm{F}}}{\nu_m^2\varepsilon_{\mathrm{pack}}^2} + \frac{1}{20}nm.$$

Combining Lemmas 5.11, 5.12, and 5.13, we arrive at a dichotomy: either the average exploration error $K_\star$-Err$_e[\pi]$ is large, or the regret proxy K-Err$_e[\pi]$ is large.

Corollary 5.1. *Let* $\mathbf{e} \overset{\text{unif}}{\sim} \{-1,1\}^{[n]\times[m]}$. *Then if Assumptions 5.3, 5.4,and 5.5 hold,*

$$\textit{either} \quad \underbrace{\mathbb{E}_{\mathbf{e}} K_\star\text{-Err}_{\mathbf{e}}[\pi] \geq \frac{n}{4\varepsilon_{\text{pack}}^2}}_{\text{(sufficient exploration)}}, \quad \textit{or} \quad \underbrace{\mathbb{E}_{\mathbf{e}} K\text{-Err}_{\mathbf{e}}[\pi] \geq \frac{c_{\text{ls}}}{10} Tnm\nu_m^2\varepsilon_{\text{pack}}^2 - \epsilon_{\text{ls}}}_{\text{(large deviation from optimal)}}. \tag{5.26}$$

Completing the Proof

To conclude the proof, we show that $\mathbb{E}_{\mathbf{e}} K\text{-Err}_{\mathbf{e}} \approx \mathbb{E}_{\mathbf{e}} K_\star\text{-Err}_{\mathbf{e}}$, so that the final bound follows by setting $\varepsilon_{\text{pack}}^2 \approx \sqrt{1/mT}$.

Lemma 5.14. *Under Assumptions 5.3 and 5.4, we have* $\mathbb{E}_{\mathbf{e}} K_\star\text{-Err}_{\mathbf{e}}[\pi] \leq 2\mathbb{E}_{\mathbf{e}} K\text{-Err}_{\mathbf{e}}[\pi] + 4nmT\|P_\star\|_{\text{op}}^4\varepsilon_{\text{pack}}^2$.

Combining Lemma 5.14 with Corollary 5.1, we have

$$\max_e K\text{-Err}_e[\pi] \geq \mathbb{E}_{\mathbf{e}} K\text{-Err}_{\mathbf{e}}[\pi] \geq \left(\frac{n}{8\varepsilon_{\text{pack}}^2} - 2nmT\|P_\star\|_{\text{op}}^4\varepsilon_{\text{pack}}^2\right) \wedge \frac{c_{\text{ls}}}{10} Tnm\nu_m^2\varepsilon_{\text{pack}}^2.$$

The lemma now follows from setting $\varepsilon_{\text{pack}}^2 = \frac{1}{32\|P_\star\|_{\text{op}}^2\sqrt{mT}}$, and relating $\max_e K\text{-Err}_e[\pi]$to $\max_e \text{PseudoReg}_e[\pi]$ via Lemma 5.11, taking care to verify that we can justify Assumption 5.4. See Simchowitz and Foster [2020, Section 4.4] for the complete proof.

5.8 Omitted Proofs

Proof of the Self-Bounding Property (Proposition 5.6)

The previous section tell us that we can ensure stabilizability as soon as we can uniformly bound $P'(s)$. We now elucidate an *self-bounding* property that makes the latter possible. The goal is to take our expression for $P'(s) = \mathsf{dlyap}(A_{\text{cl}}(s), \mathcal{Q}_1(s))$, represent it as $P'(s) \leq c_1\|P(s)\|^{c_2}$ for some constants c_1, c_2. First, we extract two useful facts about the dlyap operator, both of which are direct consequences of Eq. (5.17):

Fact 5.15. *If* $Y \succeq 0$, *then* $\mathsf{dlyap}(A, Y) \succeq Y$. *Moreover, if* $Y \succeq I$, *then* $\mathsf{dlyap}(A, Y) \succeq A^\top A$.

Fact 5.16. $\mathsf{dlyap}(\cdot,\cdot)$ *is linear in the second argument. Moreover, if* $Y \succeq Y'$, *then* $\mathsf{dlyap}(A, Y) \succeq \mathsf{dlyap}(A, Y')$. *Hence,* $\mathsf{dlyap}(A, Y) \preceq \mathsf{dlyap}(A, I)\|Y\|$

First, let us sketch a simple argument for bounding $\|P'(s)\|_{\text{op}}$; a more subtle argument is required for the Frobenius norm.

Lemma 5.17 (Self-Bounding Property: Operator Norm)**.** *Under Assumption 5.2,* $\|P'(s)\|_{\mathrm{op}} \le 4\|P(s)\|_{\mathrm{op}}^3 \varepsilon_{\mathrm{op}}$ *for all* s *such that* $\theta(s)$ *is stabilizable.*

Proof. We shall use the notation $\pm X \preceq Y$ to denote that $X \preceq Y$ and $-X \preceq Y$. Since $\pm\mathcal{Q}_1(s) \preceq \|\mathcal{Q}_1(s)\|_{\mathrm{op}} I$, Fact 5.16 implies that

$$\pm P'(s) = \pm\mathsf{dlyap}(A_{\mathrm{cl}}(s), \mathcal{Q}_1(s)) \preceq \|\mathcal{Q}_1(s)\|_{\mathrm{op}} \cdot \mathsf{dlyap}(A_{\mathrm{cl}}(s), I).$$

Further, under Assumption 5.2, $I \preceq Q + K(s)^\top R K(s)$, so that

$$\mathsf{dlyap}(A_{\mathrm{cl}}(s), I) \preceq \mathsf{dlyap}(A_{\mathrm{cl}}(s), Q + K(s)^\top R K(s)) = P(s, \tag{5.27}$$

where the last line uses Eq. (5.18). Hence, since $\pm X \preceq Y$ implies $\|X\|_{\mathrm{op}} \le \|Y\|_{\mathrm{op}}$,

$$\begin{aligned}\|P'(s)\|_{\mathrm{op}} &\le \|\mathcal{Q}_1(s)\|_{\mathrm{op}}\|P(s)\|_{\mathrm{op}} \\ &\le 2\|A_{\mathrm{cl}}(s)\|_{\mathrm{op}}\|P(s)\|_{\mathrm{op}}^2\|\Delta_{A_{\mathrm{cl}}}(s)\|_{\mathrm{op}} \\ &\le \varepsilon_{\mathrm{op}}(1 + \|K(s)\|_{\mathrm{op}})\|A_{\mathrm{cl}}(s)\|_{\mathrm{op}}\|P(s)\|_{\mathrm{op}}^2.\end{aligned}$$

where in the second inequalities we use the forms of $\mathcal{Q}_1$ and $\Delta_{A_{\mathrm{cl}}}$. Under Assumption 5.2, one can use Fact 5.15 to verify that $\|P(s)\|_{\mathrm{op}} = \|\mathsf{dlyap}(A_{\mathrm{cl}}(s), Q + K(s)^\top R K(s))\|_{\mathrm{op}} \ge \max\{1, \|K(s)\|_{\mathrm{op}}^2, \|A_{\mathrm{cl}}(s)\|_{\mathrm{op}}^2\}$. The bound follows. □

The same bound holds for the Frobenius norm, but the argument is more involved and deferred to the end of the section.

Lemma 5.18 (Self-Bounding Property: Frobenius Norm)**.** *Under Assumption 5.2,* $\|P'(s)\|_{\mathrm{F}} \le 4\|P(s)\|_{\mathrm{op}}^3 \varepsilon_{\mathrm{F}}$*. Hene, for* $\circ \in \{\mathrm{op}, \mathrm{F}\}$*, and all* s *such that* $\theta(s)$ *is stabilizable,*

$$\|P'(s)\|_\circ \le 4\|P(s)\|_{\mathrm{op}}^3 \varepsilon_\circ.$$

Proof. We define the following terms for $\alpha > 0$:

$$\mathcal{Q}_{[\alpha]}(s) := \alpha A_{\mathrm{cl}}(s)^\top P(s)^2 A_{\mathrm{cl}}(s) + \frac{1}{\alpha}\Delta_{A_{\mathrm{cl}}}(s)^\top \Delta_{A_{\mathrm{cl}}}(s). \tag{5.28}$$

From the matrix AM-GM inequality, we have $-\mathcal{Q}_{[\alpha]}(s) \preceq \mathcal{Q}_1(s) \preceq \mathcal{Q}_{1;\alpha}(s)$ for any $\alpha > 0$. Hence, Fact 5.16 implies that

$$\begin{aligned}P'(s) = \mathsf{dlyap}(A_{\mathrm{cl}}(s); \mathcal{Q}_1(s)) \preceq \alpha \underbrace{\mathsf{dlyap}(A_{\mathrm{cl}}(s); A_{\mathrm{cl}}(s)^\top P(s)^2 A_{\mathrm{cl}}(s))}_{:=E_1(s)} \\ + \frac{1}{\alpha}\underbrace{\mathsf{dlyap}(A_{\mathrm{cl}}(s); \Delta_{A_{\mathrm{cl}}}(s)^\top \Delta_{A_{\mathrm{cl}}}(s)}_{:=E_2(s)}.\end{aligned}$$

and similarly for $-P'(s)$. In short, $\pm P'(s) \preceq \alpha \cdot E_1(s) + \frac{1}{\alpha} E_2(s)$ holds $\forall \alpha > 0$. Since $E_1(s), E_2(s) \succeq 0$, [Perdomo et al., 2021, Lemma D.5], this implies that

$$\|P'(s)\|_{\mathrm{F}} \leq 2\sqrt{\|E_1(s)\|_{\mathrm{op}} \mathrm{tr}[E_2(s)]} \tag{5.29}$$

To bound the term $E_1(s)$, we have

$$\begin{aligned}
\|E_1(s)\| &= \|\mathsf{dlyap}(A_{\mathrm{cl}}(s); A_{\mathrm{cl}}(s)^\top P(s)^2 A_{\mathrm{cl}}(s))\|_{\mathrm{op}} \\
&= \|\mathsf{dlyap}(A_{\mathrm{cl}}(s); P(s)^2)\|_{\mathrm{op}} && (\mathsf{dlyap}(A, A^\top Y A) \preceq \mathsf{dlyap}(A, Y)) \\
&= \|P(s)\|_{\mathrm{op}}^2 \|\mathsf{dlyap}(A_{\mathrm{cl}}(s); I)\|_{\mathrm{op}} && \text{(Fact 5.16)} \\
&\leq \|P(s)\|_{\mathrm{op}}^3. && \text{(Eq. (5.27))}
\end{aligned}$$

To bound $\mathrm{tr}[E_2(s)]$, we have

$$\begin{aligned}
\mathrm{tr}[E_2(s)] &= \mathrm{tr}[\mathsf{dlyap}(A_{\mathrm{cl}}(s), \Delta_{A_{\mathrm{cl}}}(s)^\top \Delta_{A_{\mathrm{cl}}}(s))] \\
&= \mathrm{tr}[\sum_{i \geq 0} (A_{\mathrm{cl}}(s)^i)^\top \Delta_{A_{\mathrm{cl}}}(s)^\top \Delta_{A_{\mathrm{cl}}}(s) (A_{\mathrm{cl}}(s)^i)^\top] \\
&\leq \|\Delta_{A_{\mathrm{cl}}}(s)\|_{\mathrm{F}}^2 \sum_{i \geq 0} \|A_{\mathrm{cl}}^i\|_{\mathrm{op}}^2.
\end{aligned}$$

We can then bound $\|\Delta_{A_{\mathrm{cl}}}(s)\|_{\mathrm{F}} \leq \varepsilon_{\mathrm{F}}(1 + \|K(s)\|_{\mathrm{op}}) \leq 2\|P(s)\|_{\mathrm{op}}^{1/2} \varepsilon_{\mathrm{F}}$. By a standard Lyapunov argument [Perdomo et al., 2021, Lemma D.9], one can bound $\sum_{i \geq 0} \|A_{\mathrm{cl}}(s)^i\|_{\mathrm{op}}^2 \leq \|\mathsf{dlyap}(A_{\mathrm{cl}}, I)\|_{\mathrm{op}}^2$, which is at most $\|P(s)\|_{\mathrm{op}}^2$ by Eq. (5.27). Hence,

$$\mathrm{tr}[E_2(s)] \leq 4\|P(s)\|_{\mathrm{op}}^3 \varepsilon_{\mathrm{F}}^2.$$

The bound now follows from our above estimates for E_1 and E_2, and from the inequality Eq. (5.29). □

Derivative Bound on $K'(s)$ (Lemma 5.3)

Introduce $\Delta_A = \hat{A} - A_\star$ and $\Delta_B = \hat{B} - B_\star$, and recall $A_{\mathrm{cl}}(s) = A(s) + B(s)K(s)$ and $\Delta_{A_{\mathrm{cl}}}(s) = \Delta_A + \Delta_B K(S)$. A direct computation [Simchowitz and Foster, 2020, Lemma B.1] reveals that, for $R_0 := (R + B(s)^\top P(s) B(s))$ that

$$K'(s) = -R_0^{-1}(\Delta_B^\top P(s) A_{\mathrm{cl}}(s) + B^\top P(s) \Delta_{A_{\mathrm{cl}}}(s) + B^\top P'(s) A_{\mathrm{cl}}(s)).$$

Hence,

$$\begin{aligned}
\|K'(s)\|_\circ \leq\ & \|R_0^{-1}\|_{\mathrm{op}} \|P(s)\|_{\mathrm{op}} \|\Delta_B\|_\circ \\
& + \|R_0^{-1} B^\top P(s)^{1/2}\|_{\mathrm{op}} \cdot \left(\|P(s)^{1/2}\|_{\mathrm{op}} \|\Delta_{A_{\mathrm{cl}}}(s)\|_\circ + \|P^{-1/2}\|_{\mathrm{op}} \|P'(s)\|_\circ \|A_{\mathrm{cl}}(s)\|_{\mathrm{op}} \right).
\end{aligned}$$

From Assumption 5.2, $R_0 \succeq I$, so $\|R_0^{-1}\|_{\mathrm{op}} \leq 1$. Moreover, since $R_0 \succeq B^\top P(s) B$, we have that $\|R_0^{-1} B^\top P(s)^{1/2}\|_{\mathrm{op}} \leq 1$. Similarly, since $P \succeq Q \succeq I$, $\|P^{-1/2}\|_{\mathrm{op}} \leq 1$. Finally, $\|\Delta_B\|_\circ \leq \varepsilon_\circ$, and using the simplifications in the previous section, $\|\Delta_{A_{\mathrm{cl}}}(s)\|_\circ \leq 2\|P(s)\|_{\mathrm{op}}^{1/2} = 2\|P^{1/2}(s)\|_{\mathrm{op}}$ and $\|A_{\mathrm{cl}}(s)\|_{\mathrm{op}} \leq \|P(s)\|_{\mathrm{op}}^{1/2}$. Applying these simplifications,

$$\|K'(s)\|_\circ \leq 3\|P(s)\|_{\mathrm{op}} \varepsilon_\circ + \|P(s)\|_{\mathrm{op}}^{1/2} \|P'(s)\|_\circ.$$

Proof of Lemma 5.8

We begin with a linear algebraic lemma lower bounding the square of a PSD matrix in the Lowner order.

Lemma 5.19. *Let* $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^\top & X_{22} \end{bmatrix} \succ 0$. *Then, for any parameter* $\alpha > 0$,

$$X^2 \succeq \begin{bmatrix} (1-\alpha_1) X_{11}^2 + (1-\alpha_2^{-1}) X_{12} X_{12}^\top & 0 \\ 0 & (1-\alpha_2) X_{22}^2 + (1-\alpha_1^{-1}) X_{12}^\top X_{12} \end{bmatrix}$$

Proof of Lemma 5.19. We begin by expanding

$$\begin{aligned} \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^\top & X_{22} \end{bmatrix}^2 &= \begin{bmatrix} X_{11}^2 + X_{12} X_{12}^\top & X_{11} X_{12} + X_{12} X_{22} \\ X_{22} X_{12}^\top + X_{12}^\top X_{11} & X_{22}^2 + X_{12}^\top X_{12} \end{bmatrix} \\ &= \begin{bmatrix} X_{11}^2 + X_{12} X_{12}^\top & 0 \\ 0 & X_{22}^2 + X_{12}^\top X_{12} \end{bmatrix} + \begin{bmatrix} 0 & X_{11} X_{12} \\ X_{12}^\top X_{11} & 0 \end{bmatrix} + \begin{bmatrix} 0 & X_{12} X_{22} \\ X_{22} X_{12}^\top & 0 \end{bmatrix}. \end{aligned}$$

Now, for any vector $v = (v_1, v_2)$, and any $\alpha_1 > 0$, we have

$$\begin{aligned} \left\langle v, \begin{bmatrix} 0 & X_{11} X_{12} \\ X_{12}^\top X_{11} & \end{bmatrix} v \right\rangle &= 2 v_1^\top X_{11} X_{12} v_2 \\ &\geq -2 \|v_1^\top X_{11}\| \|X_{12} v_2\| \\ &= -2 \cdot \alpha_1^{1/2} \|v_1^\top X_{11}\| \cdot \alpha_1^{-1/2} \|X_{12} v_2\| \\ &\geq -\alpha_1 \|v_1^\top X_{11}\|^2 - \alpha_1^{-1} \|X_{12} v_2\|^2 \\ &= \left\langle v, \begin{bmatrix} -\alpha_1 X_{11}^2 & 0 \\ 0 & -\alpha_1^{-1} X_{12}^\top X_{12} \end{bmatrix} v \right\rangle. \end{aligned}$$

Similarly, for any v and $\alpha_2 > 0$, we have

$$\left\langle v, \begin{bmatrix} 0 & X_{12} X_{22} \\ X_{22} X_{12}^\top & \end{bmatrix} v \right\rangle \geq \left\langle v, \begin{bmatrix} -\alpha_2^{-1} X_{12} X_{12}^\top & 0 \\ 0 & -\alpha_2 X_{22}^2 \end{bmatrix} v \right\rangle.$$

Thus, for any $\alpha_1, \alpha_2 > 0$,

$$X^2 \succeq \begin{bmatrix} (1-\alpha_1) X_{11}^2 + (1-\alpha_2^{-1}) X_{12} X_{12}^\top & 0 \\ 0 & (1-\alpha_2) X_{22}^2 + (1-\alpha_1^{-1}) X_{12}^\top X_{12} \end{bmatrix}$$

□

The important case of Lemma 5.19 is when the matrix X in question can be lower bounded in terms of the weighted sum of two complementary projection matrices.

Lemma 5.20. *Let P be an orthogonal projection matrix, let $X \succ 0$, and suppose that there exist positive constants $\lambda_1 \le \nu \le \lambda_2$ be such that $X \succeq \lambda_1(I-\mathsf{P})+\lambda_2\mathsf{P}$, and $\|\mathsf{P}X(1-\mathsf{P})\|_{\mathrm{op}} \le \nu$. Then, if $\nu \le \sqrt{\lambda_1\lambda_2}/2$,*

$$X^2 \succeq \frac{1}{4}\lambda_1^2\mathsf{P} + \frac{\lambda_1^2\lambda_2^2}{16\nu^2}(I-\mathsf{P})$$

Proof of Lemma 5.20. Denote the number of rows/columns of X by d. By an orthonormal change of basis, we may assume that P is the projection onto the first $p = \dim(\mathrm{range}(\mathsf{P}))$ cannonical basis vectors. Writing X and P in this basis we have

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^\top & X_{22} \end{bmatrix} \succeq \lambda_1(I-\mathsf{P}) + \lambda_2\mathsf{P} = \begin{bmatrix} \lambda_1 I_{d-p} & 0 \\ 0 & \lambda_2 I_p \end{bmatrix}. \tag{5.30}$$

It suffices to show that, in this basis

$$X^2 \succeq \begin{bmatrix} \frac{\lambda_1^2}{4} I_p & 0 \\ 0 & \frac{\lambda_2^2\lambda_1^2}{16\nu^2} I_{d-p} \end{bmatrix}. \tag{5.31}$$

From Eq. (5.30), $X_{11} \succeq \lambda_1 I_p$, $X_{22} \succeq \lambda_2 I_{p-d}$, and $\|X_{12}\|_{\mathrm{op}} = \|\mathsf{P}X(1-\mathsf{P})\|_{\mathrm{op}}$. Hence, hence the parameters $\nu \ge \lambda_1$ in the lemma satisfies $\nu \ge \|X_{12}\|_{\mathrm{op}}$. Thus,

$$X^2 \succeq \begin{bmatrix} \left\{(1-\alpha_1) + (1-\alpha_2^{-1})(\nu/\lambda_1)^2\right\} \cdot \lambda_1^2 I_p & 0 \\ 0 & \left\{(1-\alpha_2) + (1-\alpha_1^{-1})(\nu/\lambda_2)^2\right\} \cdot \lambda_2^2 I_{d-p} \end{bmatrix}.$$

Set $\alpha_1 = \frac{1}{2}$, and take α_2 to satisfy $(1-\alpha_2^{-1})(\nu/\lambda_1)^2 = -1/4$. Then, we have the following string of implicitations

$$\begin{aligned} 1-\alpha_2^{-1} = -\lambda_1^2/4\nu^2 \quad &\Longrightarrow \quad \alpha_2^{-1} = \frac{4\nu^2+\lambda_1^2}{\nu^2} \quad \Longrightarrow \\ \alpha_2 = \frac{4\nu^2}{4\nu^2+\lambda_1^2} \quad &\Longrightarrow \quad 1-\alpha_2 = \frac{\lambda_1^2}{4\nu^2+\lambda_1^2} \ge \frac{\lambda_1^2}{8\nu^2}, \end{aligned}$$

where in the last line we use $\nu \ge \lambda_1$. For this choice, and using $\nu \le \sqrt{\lambda_1\lambda_2}/2$,

$$X^2 \succeq \begin{bmatrix} \frac{\lambda_1^2}{4} I_p & 0 \\ 0 & \{\frac{\lambda_1^2}{8\nu^2} - \frac{\nu^2}{\lambda_2^2}\} \cdot \lambda_2^2 I_{d-p} \end{bmatrix} \succeq \begin{bmatrix} \frac{\lambda_1^2}{4} I_p & 0 \\ 0 & \frac{\lambda_2^2\lambda_1^2}{16\nu^2} I_{d-p} \end{bmatrix},$$

as needed. □

We are now in a position to prove our desired lemma.

Concluding the proof of Lemma 5.8. For simplicitly, let us drop the dependence on the subscript k for the phase; this bound holds as long as $k \geq k_{\text{safe}}$ is past the safeSet initialization stage.

Let $\mathbf{Z}$ denotes the matrix whose rows are $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{u}_t) \in \mathbb{R}^{d_x+d_u}$ for times $t \in \{\tau_k, \tau_{k-1}, \ldots, \tau_{k+1}-1\}$, and let $\mathbf{W}$ denote the matrix whose rows are the disturbances $\mathbf{w}_t$ for the same times. Let $\widehat{\theta} \leftarrow \widehat{\theta}_k$ denote the least squares estimate in this phase.Finally, we let $\mathbf{W}^{(i)} = \mathbf{W}e_i$ pick out the i-th row of the disturbance matrix. Recall also that $d := d_x + d_u$.

We operate on the event that, for a given projection matrix $\mathsf{P} \in \mathbb{R}^{d\times d}$ to a subspace of dimension p, $\mathbf{\Lambda} \succeq \lambda_1(I-\mathsf{P})+\lambda_2\mathsf{P}$, and that for some $\lambda_1 \leq \nu \leq \sqrt{\lambda_1\lambda_2}/2$, $\nu \geq \lambda_1 \vee \|\mathsf{P}\mathbf{\Lambda}(1-\mathsf{P})\|$. On this event, Lemma 5.20 implies that

$$(\mathbf{Z}^\top\mathbf{Z})^2 = \mathbf{\Lambda}^2 \succeq \frac{1}{4}\lambda_1^2\mathsf{P} + \frac{\lambda_1^2\lambda_2^2}{16\nu^2}(I-\mathsf{P})$$

so after inversion,

$$(\mathbf{Z}^\top\mathbf{Z})^{-2} \preceq 4\lambda_{-1}^2\mathsf{P} + \frac{16\nu^2}{\lambda_1^2\lambda_2^2}(I-\mathsf{P}).$$

Hence, we can render

$$\begin{aligned}
\|\widehat{\theta}-\theta_\star\|_{\mathrm{F}}^2 &= \sum_{i=1}^{d_x} \left\|(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{W}^{(i)}\right\|_2^2 \\
&= \sum_{i=1}^{d_x} \left\langle \mathbf{Z}^\top\mathbf{W}^{(i)}, (\mathbf{Z}^\top\mathbf{Z})^{-2}\mathbf{Z}^\top\mathbf{W}^{(i)} \right\rangle \\
&\leq \sum_{i=1}^{d_x} \left\langle \mathbf{Z}^\top\mathbf{W}^{(i)}, \left(4\lambda_1^{-2}\mathsf{P} + \frac{16\nu^2}{\lambda_1^2\lambda_2^2}(I-\mathsf{P})\right)\mathbf{Z}^\top\mathbf{W}^{(i)} \right\rangle \\
&= \sum_{i=1}^{d_x} 4\lambda_1^{-2}\|\mathsf{P}\mathbf{Z}^\top\mathbf{W}^{(i)}\|^2 + \frac{16\nu^2}{\lambda_1^2\lambda_2^2}\|(I-\mathsf{P})\mathbf{Z}^\top\mathbf{W}^{(i)}\|^2 \\
&= \sum_{i=1}^{d_x} \frac{4}{\lambda_1^2}\left(\sum_{j=1}^{p}\left\langle v_j, \mathbf{Z}^\top\mathbf{W}^{(i)}\right\rangle^2\right) + \frac{16\nu^2}{\lambda_1^2\lambda_2^2}\left(\sum_{j=p+1}^{d_x}\left\langle v_j, \mathbf{Z}^\top\mathbf{W}^{(i)}\right\rangle^2\right)
\end{aligned}$$

For an index j, let $\lambda[j]$ equal λ_1 if $j \leq p$, and λ_2 if $p+1 \leq d \leq d$, and define the vector $\mathbf{Z}_j = \mathbf{Z}v_j$. Then, $\left\langle v_j, \mathbf{Z}^\top\mathbf{W}^{(i)}\right\rangle^2$ can be bounded as as

$$\begin{aligned}
\left\langle v_j, \mathbf{Z}^\top\mathbf{W}^{(i)}\right\rangle^2 &= \|\mathbf{Z}_j^\top\mathbf{W}^{(i)}\|_2^2 \\
&= \|\mathbf{Z}_j\|^2 \cdot \frac{\left\|\mathbf{Z}_j^\top\mathbf{W}^{(i)}\right\|_2^2}{\|\mathbf{Z}_j\|_2^2} \\
&\leq \|\mathbf{Z}_j\|^2 \cdot \frac{3}{2}\frac{\left\|\mathbf{Z}_j^\top\mathbf{W}^{(i)}\right\|_2^2}{\|\mathbf{Z}_j\|_2^2 + \frac{1}{2}\lambda[j]},
\end{aligned}$$

where in the last inequality we use that $\|\mathbf{Z}_j\|_2^2 = v_j^\top \mathbf{\Lambda} v_j \geq v_j^\top(\lambda_1(1-\mathsf{P}) + \lambda_2 \mathsf{P})v_j = \lambda[j]$.

By the scalar-valued self normalized tail inequality, Lemma 3.3, it holds with probability $1-\delta$ that

$$\begin{aligned}
\left\langle v_j, \mathbf{Z}^\top \mathbf{W}^{(i)} \right\rangle^2 &\leq 3\|\mathbf{Z}_j\|_2^2 \log \frac{\frac{1}{2}\lambda[j] + \|\mathbf{Z}_j\|_2^2}{\frac{1}{2}\lambda_j[j]\delta} \\
&\leq 3\lambda[j]\boldsymbol{\kappa}[j] \log \frac{\frac{1}{2}\lambda[j] + \lambda[j]\boldsymbol{\kappa}[j]}{\frac{1}{2}\lambda_j[j]\delta} \\
&\leq 3\lambda[j]\boldsymbol{\kappa}[j] \log \frac{3\boldsymbol{\kappa}[j]}{\delta} \\
&= \begin{cases} 3\lambda_1\boldsymbol{\kappa}_1 \log \frac{3\boldsymbol{\kappa}_1}{\delta} & j \leq d \\ 3\lambda_2\boldsymbol{\kappa}_2 \log \frac{3\boldsymbol{\kappa}_2}{\delta} & j \geq d \end{cases}
\end{aligned}$$

where we set $\boldsymbol{\kappa}[j] := \frac{v_j^\top \mathbf{\Lambda} v_j}{\lambda[j]} \geq 1$, and note that $\boldsymbol{\kappa}_1 := \max_{1\leq j\leq d} \boldsymbol{\kappa}[j]$ and $\boldsymbol{\kappa}_2 := \max_{j>p} \boldsymbol{\kappa}[j]$. Hence, taking a union bound over all dm coordinates

$$\begin{aligned}
&\|\widehat{\theta} - \theta_\star\|_{\mathrm{F}}^2 \\
&\quad\leq \sum_{i=1}^m \frac{4}{\lambda_1^2}\left(\sum_{j=1}^p 3\lambda_1\boldsymbol{\kappa}_1 \log\frac{3dm\boldsymbol{\kappa}_1}{\delta}\right) + \frac{16\nu^2}{\lambda_1^2\lambda_2^2}\left(\sum_{j=p+1}^d 3\lambda_2\boldsymbol{\kappa}_2 \log\frac{3dm\boldsymbol{\kappa}_2}{\delta}\right) \\
&\quad\leq \frac{12mp\boldsymbol{\kappa}_1}{\lambda_1}\log\frac{3md\boldsymbol{\kappa}_1}{\delta} + \left(\frac{\nu}{\lambda_1}\right)^2 \cdot \frac{48m(p-d)\boldsymbol{\kappa}_2}{\lambda_2}\log\frac{3md\boldsymbol{\kappa}_2}{\delta}.
\end{aligned}$$

□

Chapter 6

Nonstochatic Control

While the LQR setting of the previous chapter has received much attention in recent literature it is quite limited: the costs are fixed quadratic functions, the noise i.i.d. Gaussian, and the state fully observed.

In this chapter, we consider a much more general adaptive control setting, and demonstrate that, despite the generality, low regret with respect to a natural benchmark of policies is attainable. This model is *non-stochastic control problem*: a model for dynamics that replaces stochastic noise with adversarial perturbations in the dynamics, and allows for arbitrary changing sequence of convex costs.

In this non-stochastic model, it is impossible to pre-compute an instance-wise optimal controller. Instead, the metric of performance is regret, or total cost compared to the best in hindsight given the realization of the noise. Previous work has introduced new adaptive controllers that are learned using iterative optimization methods, as a function of the noise, and are able to compete with the best controller in hindsight.

This chapter presents a novel approach to non-stochastic control which unifies and generalizes prior results in the literature. Notably, we provide the first sublinear regret guarantees for non-stochastic control with partial observation for both *known* and *unknown* systems.

Organization and Results

In Section 6.1 we introduce the nonstochastic control problem, an online control setting which significantly generalizes the online LQR setting of the previous chapter by 1) considering general and possible adversarially chosen disturbances, 2) considering arbitrary sequences of convex costs not known to the learner in advance, and 3) allowing for partially observed systems. This setting also necessitates a more general notion of regret, defined in terms of a benchmark class of dynamic linear control policies, abbreviated as LDCs.

Subsequently, Section 6.2 introduces a disturbance response control (Drc). This formalism parametrizes control design in nonstochastic control as an *convex* problem, building upon the rich history of convex parameterization in control theory [Youla et al., 1976, Kučera, 1975, Zames, 1981, Wang et al., 2019]. This section also provides further background on Markov

operators needed to formalize the parametrizations. The use of convex parametrization for online control was first proposed by Agarwal et al. [2019b], on which the approach presented in this chapter is based. This section focuses on two simple parametrizations – ones for stable systems, and ones for systems which can be stabilized with static feedback. This helps to simplify presentation; more general parametrizations are introduced at the end of the chapter in Section 6.5.

Section 6.3 leverages the DRC parametrization to obtain low regret in the nonstochastic control problem, assuming the learner knows the system dynamics (but does not disturbances or losses). This section demonstrates a reduction from DRC control with a known system to the problem of online convex optimization with memory (OCOM) [Agarwal et al., 2019a], which admits a simple and efficient solution via online gradient descent due to [Anava et al., 2015].

Section 6.4 extends the reduction to unknown systems by first estimating the relevant system dynamics via least squares, and then applying the reduction from Section 6.3 with the resulting plug-in estimates. The estimation phase is simplified considerably to the formulation of the DRC parametrization solely in terms of Markov operators, and thus permits a direct application of the least-squares algorithm (for partial observation) analysis in Chapter 4.

Finally, Section 6.5 introduces a general DRC formalism, which allows one to consider arbitrary stabilizable and detectable dynamical systems (Definition 6.2). We describe how the parametrization and algorithms extend, and note that the analysis from the previous section carries over as well. This section also presents an approximate certainty-equivalent Youla parametrization which, to our knowledge, is novel in the control literature.

6.1 The Nonstochastic Control problem

In this chapter, we study the problem of nonstochastic control, which generalizes the online LQR setting of the previous chapter. Starting from initial state $\mathbf{x}_1 = 0$, we consider the following dynamics for time steps $1 \le t \le T$:

$$\begin{aligned} \mathbf{x}_{t+1} &= A_\star \mathbf{x}_t + B_\star \mathbf{u}_t + \mathbf{w}_t \\ \mathbf{y}_t &= C_\star \mathbf{x}_t + \mathbf{e}_t, \end{aligned} \tag{6.1}$$

Under partial observation, the system state $\mathbf{x}_t \in \mathbb{R}^{d_y}$ remains hidden, and the learner observes only their chosen inputs $\mathbf{u}_t \in \mathbb{R}^{d_u}$ and the outputs $\mathbf{y}_t \in \mathbb{R}^{d_y}$. Here, we call $\mathbf{w}_t$ the *process noise*, which affects the evolution of the hidden state, and $\mathbf{e}_t$ the *observation noise*, which perturbs the observed outputs. Note that the dynamics coincide with the partially observed identification problem, (4.1), but with $D_\star = 0$ taken for simplicity. We shall frequently refer to the *full observation setting* as a special case, where $\mathbf{x}_t \equiv \mathbf{y}_t$, which is obtained by taking $C_\star = I$ and $\mathbf{e}_t \equiv 0$.

At each time step t, the learner observes $\mathbf{y}_t$, selects an input $\mathbf{u}_t$, and adversary selects a loss $\ell_t(y, u) : \mathbb{R}^{d_y} \times \mathbb{R}^{d_u} \to \mathbb{R}$, and the learner suffers loss $\ell_t(\mathbf{y}_t, \mathbf{u}_t)$. As in Chapter 5, the

learners actions can be specified by a policy π, with actions

$$\mathbf{u}_t = \pi(t, \mathbf{y}_{1:t}, \mathbf{u}_{1:t-1}, \ell_{1:t-1}), \tag{6.2}$$

and where π can itself be randomized. In other words, at time t, the learner observes the current loss ℓ_t and output $\mathbf{y}_t$, but not the system state $\mathbf{x}_t$. The disturbances $\mathbf{w}_t$ and $\mathbf{e}_t$ are never revealed either, thought we shall see that a certain "sufficient statistic" for the disturbances can be computed from inputs and outputs alone.

We let $\mathbf{y}_t^\pi, \mathbf{u}_t^\pi, \mathbf{x}_t^\pi$ denote the outputs, inputs and hidden states that arise from executing policy π to choose the inputs, as in Eq. (6.2) in closed loop with the dynamics (6.1); we refer to these as the *iterates* under π. Note that the disturbance terms $\mathbf{w}_t$ and $\mathbf{e}_t$ are regarded as *open loop*, and do not depend on the choice of policy.

Protocol 6.2 Nonstochastic Control

Intialize: Initial state $\mathbf{x}_1 = 0$, dynamical matrices $(A_\star, B_\star, C_\star)$ (either known to the learner, or unknown).
for each $t = 1, 2, \ldots, T$ **do**
  Learner observes output $\mathbf{y}_t$
  Learner selects input $\mathbf{u}_t$ as a function of $\mathbf{y}_{1:t}$, $\mathbf{u}_{1:t-1}$, and $\ell_{1:t-1}$ (as well as internal randomness)
  Nature reveals loss $\ell_t : \mathbb{R}^{d_y} \times \mathbb{R}^{d_u} \to \mathbb{R}$, and learner suffers $\ell_t(\mathbf{y}_t, \mathbf{u}_t)$.
  Nature selects disturbances $\mathbf{w}_t, \mathbf{e}_t$, and dynamics evolves according to Eq. (6.1).

The superscript-π notation allows us to reason about the iterates under multiple policies π simultaneously, for the given realization of the disturbances. In particular, for each policy π, its cumulative cost is given by

$$\mathbf{J}_T(\pi) := \sum_{t=1}^{T} \ell_t(\mathbf{y}_t^\pi, \mathbf{x}_t^\pi); \tag{6.3}$$

In summary, the nonstochastic control problem generalizes the LQR setting along three axes:

1. It accomodates arbitrarily flexible noise models, rather than restricting consideration to Gaussian noise.

2. It accomodates general and time-varying loss functions, rather than fixed quadratic costs.

3. It accomodates partial observation, rather than full state observation.

This setting was first proposed by Agarwal et al. [2019a], and this chapter generalizes the analysis to partial observation.

Throughout, we assume that the loss functions L are convex, and dominated by quadratic functions. Specifically,

Assumption 6.1. Let $v \in \mathbb{R}^{d_y+d_u}$ denote arguments of the losses ℓ_t. We assume that, for each t, $\ell_t(v)$ is convex, and L-subquadratic. That is, for all $v, v' \in \mathbb{R}^{d_y+d_u}$, $0 \le \ell_t(v) \le L\max\{1, \|v\|^2\}$, and that $|\ell_t(v) - \ell_t(v')\| \le L\max\{1, \|v\|, \|v'\|\}$.

This assumption affords the quadratic growth observed in LQR.

Known v.s. Unknown We shall consider two settings for non-stochastic control. In the *known system setting*, the learner knows the dynamical matrices $(A_\star, B_\star, C_\star)$, but does not know the sequence of disturbances or losses in advance. Thus, the challenge is to select inputs which achieve low cost, despite lack of knowledge of futures losses and disturbances. In the unknown system setting, the learner also does not know $(A_\star, B_\star, C_\star)$, and must learn an appropriate model of the dynamics of the control task.

Output v.s. State Losses For identifiability reasons, we assume the losses act on the observed system outputs $\mathbf{y}_t$, and not on the unobserved states $\mathbf{x}_t$. The algorithms and analysis extend straightforwardly to history-dependence losses of the form $\ell_t(\mathbf{y}_{t:t-k}, \mathbf{u}_{t-k})$ for any fixed $k \in \mathbb{N}$. In certain cases of interest, the history dependence allows for costs which encode dependence on the state not captured by the observation. For example, if $\mathbf{x}_t$ contains the position, velocity, and acceleration of an objective, and $\mathbf{y}_t$ is the projection onto only the position, velocity and acceleration can be recovered by differencing, e.g. $\mathbf{y}_t - \mathbf{y}_{t-1}$.

Examples

Before continuing, we take a moment to describe certain examples that fall under the non-stochastic control setting.

LQR

As may be immediately clear, the LQR setting in Chapter 5 is a special case of the above setup, obtained by taking $\mathbf{y}_t \equiv \mathbf{x}_t$ (and thus $C_\star = I$ and $\mathbf{e}_t \equiv 0$), $\mathbf{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$, and $\ell_t(y, u) = y^\top Q y + u^\top R u$. The total cost Eq. (6.3) is consistent with this special case as well.

LQG

Another classic problem that falls within the scope of nonstochastic control is linear quadratic Gaussian control. Here, the dynamics evolve according to Eq. (6.1), where the process noise and observation noise are i.i.d. Gaussian with certain fixed covariances, and the cost function is a fixed quadratic. The nonstochastic setting can also accomodate generalizations - changing costs, time-varying noise covariances, and even noise that is correlated across time.

Tracking

One motivation for time varying costs are tracking problems Abbasi-Yadkori et al. [2014]. For example, suppose that the observation $\mathbf{y}_t$ corresponds to the physical position of a vehicle in $\mathbb{R}^3$, and the learner wishes to direct their vehicle along a certain trajectory $(\mathbf{a}_t) \subset \mathbb{R}^3$. To elicit such behavior, one can consider the time varying costs $\ell_t(y, u) = \|y - \mathbf{a}_t\|^2 + \gamma\|\mathbf{u}\|^2$, where the second term is a control penalty to ensure well-behaved inputs.

Earlier Online Control

Our setting also captures the online control setting studied by Cohen et al. [2018], which studies a known dynamical system with adversarially chosen, convex quadratic costs $\ell_t(x, u) = x^\top Q_t x + u^\top R_t u$.

A Connection to Robust Control

The non-stochastic problem is also related to a robust control formulation called $\mathcal{H}_\infty$ control [Zhou et al., 1996]. In $\mathcal{H}_\infty$, the goal is to minimize a cumulative cost (e.g., an $\mathcal{L}_2$ cost $\ell_t(y, u) = \|y\|^2 + \|u\|^2$) over the worst case noise sequence satisfying a total bound over the horizon $\sum_{t\geq 1} \|\mathbf{e}_t\|^2 + \|\mathbf{w}_t\|^2$. Regret, in contrast, is an adaptive notion, because it enforces competitive performance on *every sequence* of disturbances, not just the worst case.

Nonstochastic Control and Regret

In the online LQR setting of Chapter 5, we designed an adaptive control policy to compete with the optimal infinite horizon control law. This was made feasible in part by the fact that the optimal law enjoys a closed form expression via the DARE, and that the cost function and noise distribution remained fixed across rounds.

In non-stochastic control, the unpredictability of the noise and losses means that we cannot hope to design an adaptive policy alg such that $\mathbf{J}_T(\mathsf{alg}) - \min_{\text{all policies } \pi} \mathbf{J}_T(\pi)$ grows sublinearly in T [Li et al., 2019]. Moreover, for non-quadratic losses and/or partial observation, even computing the best policy in hindsight may be computationally prohibitive.

Instead, we borrow the *regret* perspective from the online learning community. Rather than attaining performance close to the optimal control law in hindsight, we restrict our attention to attaining comparable performance with all policies in a certain *benchmark class*. Formally, for a set of policies Π, we define the nonstochastic control regret of an algorithm alg with the respect to the benchmark Π as

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi) := \underbrace{\mathbf{J}_T(\mathsf{alg})}_{\text{algorithm cost}} - \underbrace{\inf_{\pi\in\Pi} \mathbf{J}_T(\pi)}_{\text{comparator}}. \tag{6.4}$$

Notices that $\mathbf{J}_T(\cdot)$ and $\textsc{NscReg}_T(\mathsf{alg}; \Pi)$ dependend implicitly on the choice losses ℓ_t, which are not revealed to the learner until the end of the protocol, and one the disturbances $\mathbf{w}_t, \mathbf{e}_t$,

which are never directly revealed, but which influence the dynamics via Eq. (6.1). These quantities also dependent on the dynamics of the system, which may be known or unknown to the learner. We call the term $\inf_{\pi\in\Pi} \mathbf{J}_T(\mathsf{alg})$ the *comparator*. It captures the performance of the best clairvoyant choice of policy $\pi \in \Pi$, given full knowledge of (a) the entire sequence of disturbances $\mathbf{w}_{1:T}$ and $\mathbf{e}_{1:T}$ (b) the sequence of losses $\ell_{1:T}$, and (c) the system dynamics.

Benchmark Policies and LDCs

What are a reasonable class of benchmark policyes Π? In this work, we consider a gold-standard control policies in linear control, which we term *linear dynamic controllers*, or LDCs, whose set is denote Π_{ldc}. An LDC π is a control policy whose form is a linear dynamical system. Formally, π can be represented by the dynamical equations

$$\begin{aligned} \mathring{\mathbf{s}}_t^\pi &= A_\pi \mathring{\mathbf{s}}_t^\pi + B_\pi \mathring{\mathbf{y}}_t^\pi \\ \mathring{\mathbf{u}}_t^\pi &= C_\pi \mathring{\mathbf{s}}_t^\pi + D_\pi \mathring{\mathbf{y}}_t^\pi. \end{aligned} \tag{6.5}$$

The static feedback policies $\mathbf{u}_t = K\mathbf{y}_t$ can be realized by the above by selecting A_π, B_π, C_π to be zero matrices, and $D_\pi = K$. In particular, in the full observation setting, this recovers the static feedback laws $\mathbf{u}_t = K\mathbf{x}_t$ studied in Chapter 5. But the policies in Eq. (6.5) are considerably more general, because they allow for an evolving internal state $\mathbf{z}_t^\pi$. This is necessary to capture the optimal control laws for the linear quadratic gaussian (LGC) control problem, and $\mathcal{H}_\infty$ robust control (under partial observation) [Zhou et al., 1996].

Remark 6.1 (Beyond LDC policies)**.** For concreteness, we focus on LDC policies. However, the techniques in this chapter extend to various more general families. For example, we can compete with policies which have affine terms, or *DC* offsets. That is, policies π with linearly evolving internal state $\mathring{\mathbf{s}}_t$, but outputs are chosen $\mathring{\mathbf{u}}_t^\pi = C_\pi \mathring{\mathbf{s}}_t^\pi + D_\pi \mathring{\mathbf{y}}_t^\pi + \sum_{i=1}^k \alpha_i \psi_i(t)$, where α_i are linear coefficients and $\{\psi_i(t) : 1 \le i \le k\}$ are a fixed set of time varying inputs. This may be useful in tracking problems, where it makes sense to include basis functionals for a tracking problem in addition to pure feedback terms.

Given an LDC of the form (6.5), we refer to the *closed loop* dynamics as the unique dynamical equation satisfying both Eqs. (6.1) and (6.5). This can be expressed as a single dynamical system, with dynamical matrices

$$\begin{aligned} \begin{bmatrix} \mathbf{x}_{t+1}^\pi \\ \mathbf{s}_{t+1}^\pi \end{bmatrix} &= \underbrace{\begin{bmatrix} A_\star + B_\star D_\pi C_\star & B_\star C_\pi \\ C_\star B_\pi & A_\pi \end{bmatrix}}_{A_{\pi,\mathrm{cl}}} \begin{bmatrix} \mathbf{x}_t^\pi \\ \mathbf{s}_t^\pi \end{bmatrix} + \underbrace{\begin{bmatrix} I & B_\star D_\pi \\ 0 & B_\pi \end{bmatrix}}_{B_{\pi,\mathrm{cl}}} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{e}_t \end{bmatrix} \\ \begin{bmatrix} \mathbf{y}_{t+1}^\pi \\ \mathbf{u}_{t+1}^\pi \end{bmatrix} &= \underbrace{\begin{bmatrix} C_\star & 0 \\ D_\pi C_\star & C_\pi \end{bmatrix}}_{C_{\pi,\mathrm{cl}}} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{s}_t \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & I \\ 0 & D_\pi \end{bmatrix}}_{D_{\pi,\mathrm{cl}}} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{e}_t \end{bmatrix}. \end{aligned} \tag{6.6}$$

6.2 Disturbance Response Control

We proceed to introduce a unified framework for nonstochastic control, which we call *disturbance response control*, or DRC. DRC builds upon the rich history of convex parameterization in control theory [Youla et al., 1976, Kučera, 1975, Zames, 1981, Wang et al., 2019] to reduce nonstochastic control to well-studied learning-theoretic framework called online convex optimization, or OCO. The idea of using convex parametrization for online control was developed by Agarwal et al. [2019a], a work which introduced many of the key ideas taken up in this chapter.

Markov Operators, Stability, and Decay

The expression of the closed loop system (6.6) is called an *state space* representation, because it explicitly describes the (co-)evolution of the state $\mathbf{x}_t$ of the dynamical system, and $\mathbf{s}_t$ of the LDC. The *Markov operator*, introduced in Chapter 4, offers a more compact representation.

We let $\mathscr{G}(d, d')$ denote the space of infinite sequences of matrices $G = (G^{[i]})_{i \geq 0}$, with $G^{[i]} \in \mathbb{R}^{d \times d'}$, and endow it with the ℓ_1-operator norm

$$\|G\|_{\ell_1,\mathrm{op}} := \sum_{i \geq 0} \|G^{[i]}\|_{\mathrm{op}}. \tag{6.7}$$

Given dynamical matrices $(A, B, C, D)$with $A \in \mathbb{R}^{d_1 \times d_1}$, $B \in \mathbb{R}^{d_1 \times d_2}$, $C \in \mathbb{R}^{d_3 \times d_1}$, and $D \in \mathbb{R}^{d_3 \times d_2}$ their Markov operator $G = \mathsf{Markov}(A, B, C, D) \in \mathscr{G}(d_3, d_2)$ is the sequence of operators

$$G^{[0]} = D, \quad G^{[i]} = CA^{i-1}B, \quad i > 1. \tag{6.8}$$

In particular, $G_\star = \mathsf{Markov}(A_\star, B_\star, C_\star, 0) \in \mathscr{G}(d_y, d_u)$ is the Markov operator for the dynamics (6.1). Note that the system (A, B, C, D) for which $G = \mathsf{Markov}(A, B, C, D)$ is not unique, because the state space in A can be padded with zeros. We therefore refer to any (A, B, C, D) such that $G = \mathsf{Markov}(A, B, C, D)$ as an *realization* of G.

The *decay function* ψ_G and radius R_G of a Markov operator G are the functions (resp. constant)

$$\psi_G(n) := \sum_{i \geq n} \|G^{[i]}\|_{\mathrm{op}}., \tag{6.9}$$

which sum tail of the operator norms of its components. Note that $\psi_G(n)$ is finite if and only if G is a stable Markov operator, that is, there exists an equivalent realization (A', B', C', D') for which $G = \mathsf{Markov}(A', B', C', D')$, and A' is stable. In this case, there exist constants $c > 0$ and $\rho \in [0, 1)$ such that $\psi_G(n) \leq c\rho^n$.

We will frequently parametrize our approximations in terms of general decay functions ψ. We say that ψ is an *proper decay function* if $\psi(\cdot) : \mathbb{N} \to \mathbb{R}_+$ is non-increasing, and $\psi(0) \in [1, \infty)$.

Nature's Ys

The simplest variant of DRC applies when the dynamical matrix $A_\star$ is already stable, i.e. $\rho(A_\star) < 1$. Recall the Markov operator $G_\star \in \mathscr{G}(d_y, d_u)$, defined by

$$G_\star^{[0]} = 0, \quad G_\star^{[i]} C_\star A_\star^{i-1} B_\star, \quad i > 1. \tag{6.10}$$

As seen in chapter ..., the current system dynamics can be expressed in terms of a sum over the contribution of past inputs, and the *Nature's Y* term $\mathbf{y}_t^{\mathrm{nat}}$, which corresponds to the output of the system in the absence of any input:

$$\mathbf{y}_t = \mathbf{y}_t^{\mathrm{nat}} + \sum_{i=1}^{t-1} G_\star^{[i]} \mathbf{u}_{t-i}, \qquad \text{where } \mathbf{y}_t^{\mathrm{nat}} := \sum_{i=1}^{t-1} C_\star A_\star^{i-1} \mathbf{w}_t + \mathbf{e}_t. \tag{6.11}$$

Note that Nature's Y *does not* dependend on the choice of past inputs, and the actions of the learner. Moreover, given knowledge of the Markov operator $G_\star$, $\mathbf{y}_t^{\mathrm{nat}}$ can be recovered from input output data, via (6.11). This means that, regardless of the chosen sequence of inputs, we can always compute the *counterfactual* output that would have occured had zero input been selected. Moreover, via the same identity, we can compute other counterfactuals, that is, other outputs had a different sequence of controls been selected.

In addition, observe that $\mathbf{y}_t^{\mathrm{nat}}$ can be computed even if the learner does not have direct access to the system states $\mathbf{x}_t$ or disturbances $\mathbf{w}_t, \mathbf{e}_t$. As a consequence, we propose selecting inputs as linear combinations of past Natures Y's, parameterized by sequences of matrices M. Precisely, we consider inputs of the form

$$\mathbf{u}_t^M := \sum_{i=1}^{0} M^{[i-1]} \mathbf{y}_{t-i}^{\mathrm{nat}}, \quad M \in \mathcal{M}(m, R), \tag{6.12}$$

where we define the set $\mathcal{M}(m, R)$ as sequences of m matrices whose cumulative operator norm is bounded by R:

$$\mathcal{M} := \left\{ (M^{[0]}, M^{[1]}, \ldots, M^{[m-1]}) \in (\mathbb{R}^{d_u \times d_y})^m : \sum_{i=0}^{m-1} \|M^{[i]}\|_{\mathrm{op}} \le R \right\}. \tag{6.13}$$

We will call these disturbance response control policies (DRC), because the control is linear in the systems response to the disturbances, namely $\mathbf{y}_t^{\mathrm{nat}}$. We shall use the notation $(\mathbf{u}_t^M, \mathbf{y}_t^M)$ to denote the iterates produced by the DRC with parameter M, Eq. (6.12); we let π^M denote the induced control policy.

Importantly, DRC yields a *convex* controller parameterization.

Lemma 6.1. *The* DRC *parametriztion is convex, in the sense that the function* $M \mapsto (\mathbf{y}_t^M, \mathbf{u}_t^M)$ *is affine, and thus* $M \mapsto \ell_t(\mathbf{y}_t^M, \mathbf{u}_t^M)$. *is convex. In particular,* $M \mapsto \mathbf{J}_T(\pi^M)$ *is convex.*

Proof. We see that $\mathbf{u}_t^M$ is a linear function of M; it's coefficients are determined by Nature's Y's, which are independent of the past choice of player actions. Moreover, $\mathbf{y}_t^M$, in view of Eq. (6.11), is an affine function of $\mathbf{u}_t^M$, which we have established is linear in M. By assumption, ℓ_t is convex. Hence, the function in question is composition of convex function with an affine function of M. □

One consequence of convexity is that one can efficiently compute the best DRC policy in hindsight

$$\arg\min_M \sum_{t=1}^{T} \ell_t(\mathbf{y}_t^M, \mathbf{u}_t^M) = \arg\min_M \mathbf{J}_T(M),$$

provided one has knowledge of $G_\star$. As we shall see, convexity of the parametrization also allows for efficient online control as well.

Note that the previous discussion holds regardless of whether or not $A_\star$ is stable; it is a direct consequence of the open loop (6.1). But if $A_\star$ is unstable, then the Nature's Y terms may be grow very large. We impose the following assumption, we simultaneously captures stability $A_\star$ and boundedness of the noise:

Assumption 6.2. There exists a constant $R_{\text{nat}} \geq 1$ such that, for all t, $\|\mathbf{y}_t^{\text{nat}}\|_2 \leq R_{\text{nat}}$.

The scaling $R_{\text{nat}} \geq 1$ is to simplify the resulting bounds.

The Expressivity of DRC with Nature's Y's

We have just established that it is feasible to optimize over policies π^M. We now show that these policies can also approximate LDC policies in $\pi \in \Pi_{\text{ldc}}$ arbitrarily well.

For an LDC policy, let $G_{\pi,\text{cl},e\to u}$ denote the response of $\mathbf{u}_t^\pi$ to $\mathbf{e}_t$ in (6.6); explicitly,

$$\begin{aligned} G_{\pi,\text{cl},e\to u} &= \mathsf{Markov}(A_{\pi,\text{cl}}, B_{\pi,\text{cl},e}, C_{\pi,\text{cl},u}, D_\pi), \\ \text{where } B_{\pi,\text{cl},e} &= \begin{bmatrix} 0 & B_\pi \end{bmatrix} \text{ and } C_{\pi,\text{cl},u} = \begin{bmatrix} D_\pi C_\star \\ C_\pi \end{bmatrix} \end{aligned} \tag{6.14}$$

In words, $G_{\pi,\text{cl},e\to u}$ captures how the sinputs $\mathbf{u}_t^\pi$ that would arise when executing policy $\pi \in \Pi_{\text{ldc}}$ in closed loop depend on observation noise $\mathbf{e}_t$. The following lemma is essential states that this operator also gives a formula for expressing $\mathbf{u}_t^\pi$ in terms of Nature's Ys.

Lemma 6.2. *The following formula holds for each $t \in \mathbb{N}$:*

$$\mathbf{u}_t^\pi = \sum_{i=0}^{t-1} G_{\pi,\text{cl},e\to u}^{[i]} \cdot \mathbf{y}_{t-i}^{\text{nat}}. \tag{6.15}$$

Lemma 6.2 can be verified by induction on the closed loop dynamics (6.6), and is omitted in the interest of brevity. The important take away is that the inputs selected by LDC policies π can be represented in terms of the quantities Nature's Ys, which, as noted above, can be recovered by the learner via Eq. (6.11). Hence, $\mathbf{y}^{\text{nat}}$ can be regarded as a sufficient statistic for LDC control policies, motivating the DRC parametrization described above.

Still, (6.15) represents $\mathbf{u}_t^\pi$ as a length-t linear combination of Natures Y's, whereas the DRC policies have finite memory m. We now show that, when $G_\star$ and $G_{\pi,\text{cl},e\to u}$ are stable Markov operators, this truncation to finite memory is feasible. To do so, define the class of LDCs with decay at most ψ, where ψ is a proper decay function (non-increasing, non-negative, and $\psi(0) \in [1,\infty)$ on from $\mathbb{N}$ to the reals:

$$\Pi_{e\to u}[\psi] := \{\pi \in \Pi_{\text{ldc}} : \psi_{G_{\pi,\text{cl},e\to u}}(n) \le \psi(n)\} \tag{6.16}$$

To make matters more concrete, we also consider a class of policies with explicit geometric decay. That is,

$$\Pi_{e\to u}(c,\rho) := \{\pi \in \Pi_{\text{ldc}} : \psi_{G_{\pi,\text{cl},e\to u}}(n) \le c\rho^n\} \tag{6.17}$$

Observe that, as long as π is stabilizing, $G_{\pi,\text{cl},e\to u}$ is stable, and therefore, as noted above, there exists some c,ρ such that $\pi \in \Pi_{e\to u}(c,\rho)$. We are now ready to state our main theorem, which describes the fidelity which which LDC policies can be expressed by DRC controllers.

Theorem 6.1. *Suppose Assumptions 6.1 and 6.2 hold. Define* $R_{G_\star} := (1 + \|G_\star\|_{\ell_1,\text{op}})$ *Given a proper decay function* ψ*, a policy* $\pi \in \Pi_{e\to u}[\psi]$*, and* $R_{\mathcal{M}} \ge \psi(0)$ *the following holds for all integers* $m \ge 1$*,*

$$\mathbf{J}_T(\pi) - \inf_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \le 2LTR_{\mathcal{M}}R_{G_\star}^2 R_{\text{nat}}^2 \psi(m). \tag{6.18}$$

More concretely, if $\pi \in \Pi_{e\to u}(c,\rho)$ *for* $c \ge 1$ *and* $\rho \in [0,1)$*, then for* $R_{\mathcal{M}} \ge \frac{c}{1-\rho}$*,*

$$\mathbf{J}_T(\pi) - \inf_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \le c\rho^{-m} \cdot 2LTR_{\mathcal{M}}R_{G_\star}^2 R_{\text{nat}}^2. \tag{6.19}$$

Thus, (6.48) shows that that DRC controllers of length $m = \mathcal{O}(\log T)$ suffices to compete with DRC policies $\pi \in \Pi_{e\to u}(c,\rho)$ ensuring $\mathbf{J}_T(\pi) - \inf_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \le 1/\text{poly}(T)$.

DRC with Static Feedback

The Nature's Y DRC parametrization make sense only when $\rho(A_\star) < 1$, ensuring that $\mathbf{y}_t^{\text{nat}}$ remain bounded with time. As noted above, the more interesting and applicable case is when $A_\star$ is possible unstable, put placed in feedback with a nominal stablilizing controller.

In this section, we study a simple and illustrative intermediate between the stable-$A_\star$ case and the general formalism for stabilized systems. Specifically, we study systems for

which the static feedback law $\mathbf{u}_t = K\mathbf{y}_t$ for a fixed $K \in \mathbb{R}^{d_u \times d_y}$ is stabilizing.[1] When $C_\star$ is full rank - for example, in LQR - there exists such a stabilizing K if there exists *any* LDC that can stabilize the system. However, for general $C_\star$, the existence of a stabilizing K is not guaranteed. More general parametrizations are adressed in Section 6.5.

We let $\mathbf{u}_t^K, \mathbf{y}_t^K$ denote the iterates which arise by placing the control law $\mathbf{u}_t = K\mathbf{y}_t$ in feedback with Eq. (6.1). For stabilizing K, we select inputs

$$\mathbf{u}_t = K\mathbf{y}_t + \mathbf{u}_t^{\mathrm{ex}},$$

where we call $\mathbf{u}_t^{\mathrm{ex}}$ *exogenous input*. Intuitively, the term $K\mathbf{y}_t$ ensures the parametrization is stable, and $\mathbf{u}_t^{\mathrm{ex}}$ is chosen to approximate desired control behavior. One important subtlety is the distinction between the actual input $\mathbf{u}_t$, and exogenous input $\mathbf{u}_t^{\mathrm{ex}}$, which did not arise in the Nature's Ys parametrization (though recovering the parametrization for $K = 0$). The inputs and outputs are now related by

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{u}_t \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t^K \\ \mathbf{u}_t^K \end{bmatrix} + \sum_{i=0}^{t-1} G_K^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}},$$

where $G_K \in \mathscr{G}(d_y + d_u, d_u)$ is the Markov operator

$$G_K = \mathsf{Markov}\left(A_\star + B_\star K C_\star, B_\star, \begin{bmatrix} C_\star \\ KC_\star \end{bmatrix}, \begin{bmatrix} 0 \\ I \end{bmatrix} \right). \tag{6.20}$$

One can check that the following identity holds:

$$\begin{bmatrix} \mathbf{y}_t \\ K\mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t^K \\ \mathbf{u}_t^K \end{bmatrix} + \sum_{i=1}^{t-1} G_K^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}}, , \tag{6.21}$$

which is useful because the right hand side does not depend on $\mathbf{u}_t^{\mathrm{ex}}$, and therefore can be evaluated before the input at time t is selected.

The Drc parametrization with controller K, or Drc-K, we select inputs

$$\mathbf{u}_t^M = K\mathbf{y}_t + \sum_{i=0}^{m-1} M^{[i]} \mathbf{y}_t^K. \tag{6.22}$$

In words, the iterates $\mathbf{y}_t^K$ replace the role of Nature's Ys, and the input includes the additional feedback term $K\mathbf{y}_t$. Since K is stabilizing, $\mathbf{y}_t^K$ can be bounded in time even in $\mathbf{y}_t^{\mathrm{nat}}$ is not. The analogoue of Assumption 6.2 includes a bound on both $\mathbf{y}_t^K$ and $\mathbf{u}_t^K$.

Assumption 6.2b. For all t, $\|(\mathbf{y}_t^K, \mathbf{u}_t^K)\| \le R_{\mathrm{nat}}$ for some $R_{\mathrm{nat}} \ge 1$.

[1] The closed loop dynamics subject to $\mathbf{u}_t = K\mathbf{y}_t$ evolve according to $\mathbf{x}_{t+1} = (A_\star + B_\star K C_\star)\mathbf{x}_t$, and thus K is stabilizing if and only if $\rho(A_\star + B_\star K C_\star) < 1$.

For a policy π, define $G_{K\to\pi} \in \mathscr{G}(d_y + d_u, d_u)$ via

$$G_{K\to\pi} = \mathsf{Markov}\left(A_\star + B_\star K C_\star,\ B_\star(D_\pi - K),\ \begin{bmatrix}(D_\pi - K)C_\star & C_\pi\end{bmatrix},\ \begin{bmatrix}0\\ D_\pi - K\end{bmatrix}\right). \quad (6.23)$$

Again, under the assumption that K is stabilizing, $A_\star + B_\star K C_\star$ is a stable matrix, and thus the $\|G^{[i]}_{K\to\pi}\|_{\mathrm{op}} \le c\rho^i$ for some constants c and ρ. Accordingly, we define the following class of LDC policies for a proper decay function ψ, and constants (c, ρ):

$$\begin{aligned}\Pi_{K\to u}[\psi] &:= \{\pi \in \Pi_{\mathrm{ldc}} : \psi_{G_{K\to\pi}}(n) \le \psi(n)\}\\ \Pi_{K\to u}(c,\rho) &:= \{\pi \in \Pi_{\mathrm{ldc}} : \psi_{G_{K\to\pi}}(n) \le c\rho^n\}\end{aligned} \quad (6.24)$$

An analogue of Lemma 6.2 and Theorem 6.1 holds here, with analogous proofs:

Lemma 6.2b. *The following formula holds for each $t \in \mathbb{N}$:*

$$\mathbf{u}^\pi_t = \sum_{i=0}^{t-1} G^{[i]}_{K\to\pi} \cdot \mathbf{y}^{\mathrm{nat}}_{t-i}. \quad (6.25)$$

Theorem 6.1b. *Suppose Assumptions 6.1 and 6.2 hold. Define $R_G := \|G_K\|_{\ell_1,\mathrm{op}}$. Given a proper decay function ψ, a policy $\pi \in \Pi_{K\to u}[\psi]$, and $R_{\mathcal{M}} \ge \psi(0)$ the following holds for all integers $m \ge 1$,*

$$\mathbf{J}_T(\pi) - \inf_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \le 2LTR_{\mathcal{M}}R_G^2R^2_{\mathrm{nat}}\psi(m). \quad (6.26)$$

More concretely, if $\pi \in \Pi_{K\to u}(c,\rho)$ for $c \ge 1$ and $\rho \in [0,1)$, then for $R_{\mathcal{M}} \ge \frac{c}{1-\rho}$,

$$\mathbf{J}_T(\pi) - \inf_{M\in\mathcal{M}(m,R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \le c\rho^{-m}\cdot 2LTR_{\mathcal{M}}R^2_{G_\star}R^2_{\mathrm{nat}}. \quad (6.27)$$

Proof of Theorem 6.1

We turn to the proof Theorem 6.1; its generalization to static feedback above and general dynamic parametrizations (Section 6.5) is similar. We prove the first statement of the theorem; the second is a direct consequence. To simplify notation, let $\mathbf{v}_t = (\mathbf{y}_t, \mathbf{u}_t) \in \mathbb{R}^{d_y+d_u}$, and appreviate $\mathcal{M}(m, R_{\mathcal{M}})$ by $\mathcal{M}$. Then, for any $M \in \mathcal{M}$, we have

$$|\mathbf{J}_T(\pi) - \mathbf{J}_T(\pi^M)| = |\sum_{t=1}^T \ell_t(\mathbf{v}^\pi_t) - \ell_t(\mathbf{v}^M_t)| \le \sum_{t=1}^T |\ell_t(\mathbf{v}^\pi_t) - \ell_t(\mathbf{v}^M_t)| \quad (6.28)$$

$$\le L\mathcal{C}_{\max}\sum_{t=1}^T \|\mathbf{v}^\pi_t - \mathbf{v}^M_t\|, \quad (6.29)$$

where $\mathcal{C}_{\max} := \max_t \max\{1, \|\mathbf{v}_t^\pi\|_2, \mathbf{v}_t^M\|_2\}$, and where the last line uses Assumption 6.1. To bound $\|\mathbf{v}_t^\pi - \mathbf{v}_t^M\|$, we use Lemma 6.2 and the definition of the DRC control (6.12) to write

$$\mathbf{u}_t^\pi - \mathbf{u}_t^M = \sum_{i=0}^{m-1} (G_{\pi,\mathrm{cl},e\to u}^{[i]} - M^{[i]}) \mathbf{y}_{t-i}^{\mathrm{nat}} + \sum_{i=m}^{t-1} G_{\pi,\mathrm{cl},e\to u}^{[i]} \mathbf{y}_{t-i}^{\mathrm{nat}}.$$

Let us chose M such that

$$M^{[i]} = G_{\pi,\mathrm{cl},e\to u}^{[i]}, \quad \forall i \in \{0, 1, \ldots, m-1\},$$

which lies in $\mathcal{M}$ since $\sum_{i=0}^{m-1} \|M^{[i]}\|_{\mathrm{op}} \le \sum_{i\ge 0} \|G_{\pi,\mathrm{cl},e\to u}^{[i]}\|_{\mathrm{op}} \le R_{\mathcal{M}}$. This selection yields

$$\|\mathbf{u}_t^\pi - \mathbf{u}_t^M\| = \|\sum_{i=m}^{t-1} G_{\pi,\mathrm{cl},e\to u}^{[i]} \mathbf{y}_{t-i}^{\mathrm{nat}}\| \le \psi_{G_{\pi,\mathrm{cl},e\to u}}(m) R_{\mathrm{nat}} \le \psi(m) R_{\mathrm{nat}}.$$

Using the Nature's Y's decomposition (6.11), we find

$$\|\mathbf{y}_t^\pi - \mathbf{y}_t^M\| = \|\sum_{i=0}^{t-1} G_\star^{[i]} (\mathbf{u}_{t-i}^\pi - \mathbf{u}_{t-i}^M)\| \le \|G_\star\|_{\ell_1,\mathrm{op}} \cdot \max_s \|\mathbf{u}_t^\pi - \mathbf{u}_t^M\| \le \|G_\star\|_{\ell_1,\mathrm{op}} \psi(m) R_{\mathrm{nat}}.$$

Thus,

$$\|\mathbf{v}_t^\pi - \mathbf{v}_t^M\| \le (1 + \|G_\star\|_{\ell_1,\mathrm{op}}) \psi(m) R_{\mathrm{nat}} \le R_{G_\star} R_{\mathrm{nat}} . \psi(m)$$

To conclude, it remains to bound the constant $\mathcal{C}_{\max}$.

Claim 6.1. *The constant* $\mathcal{C}_{\max} := \max_t \max\{1, \|\mathbf{v}_t^\pi\|, \mathbf{v}_t^M\|\} \le 2 R_{\mathrm{nat}} R_{G_\star} R_{\mathcal{M}}$

Proof of Claim 6.1. Since $R_{\mathrm{nat}} R_{G_\star} R_{\mathcal{M}} \ge 1$, it suffices to show that, for any t, the bound on $\|\mathbf{v}_t^\pi\|$ and $\|\mathbf{v}_t^M\|$ are both bounded by $2 R_{\mathrm{nat}} R_{G_\star} R_{\mathcal{M}}$. We establish the bound for $\|\mathbf{v}_t^\pi\|$; the bound is analogous for $\|\mathbf{v}_t^M\|$.

$$\|\mathbf{y}_t^\pi\| = \|\mathbf{y}_t^{\mathrm{nat}} + \sum_{i=0}^{t-1} G_\star^{[i]} \mathbf{u}_{t-i}^\pi\|_2 \le \|\mathbf{y}_t^{\mathrm{nat}}\| + \|G_\star\|_{\ell_1,\mathrm{op}} \max_s \|\mathbf{u}_s^\pi\|.$$

Hence,

$$\begin{aligned} \|\mathbf{v}_t^\pi\| &\le \|\mathbf{y}_t^\pi\| + \|\mathbf{u}_t^\pi\| \le \|\mathbf{y}_t^{\mathrm{nat}}\| + (1 + \|G_\star\|_{\ell_1,\mathrm{op}}) \max_s \|\mathbf{u}_s^\pi\| \\ &= \|\mathbf{y}_t^{\mathrm{nat}}\| + R_{G_\star} \max_s \|\mathbf{u}_s^\pi\| \le R_{\mathrm{nat}} + R_{G_\star} \max_s \|\mathbf{u}_s^\pi\| \end{aligned}$$

Finally, using Lemma 6.2, we can bound

$$\|\mathbf{u}_s^\pi\| \le \|G_{\pi,\mathrm{cl},e\to u}\|_{\ell_1,\mathrm{op}} \max_{s'} \|\mathbf{y}_s^{\mathrm{nat}}\| \le R_{\mathcal{M}} R_{\mathrm{nat}}. \tag{6.30}$$

Combining these bounds gives $\|\mathbf{v}_t^\pi\| \le R_{\mathrm{nat}} + R_{G_\star} R_{\mathcal{M}} R_{\mathrm{nat}}$. Since $R_{G_\star}, R_{\mathcal{M}}, R_{\mathrm{nat}} \ge 1$ by assumption, our bound follows. □

In sum, we have shown that

$$\begin{aligned}|\mathbf{J}_T(\pi) - \mathbf{J}_T(\pi^M)| &\le L\mathcal{C}_{\max} \sum_{t=1}^{T} \|\mathbf{v}_t^\pi - \mathbf{v}_t^M\| \\ &\le LT \cdot 2R_{\text{nat}} R_{G_\star} R_{\mathcal{M}} \cdot R_{G_\star} R_{\text{nat}} \psi(m),\end{aligned}$$

as desired. □

6.3 Control of Known Systems via DRC

In this section, we describe how to adaptively update DRC parameterized control policies so as to attain low regret in online control problems. This section begins with a discussion of an online learning framework called Online Convex Optimization with Memory (OCOM), and describes how to reduce to this framework the the online control problem with loss functions induced by the DRC parameterization. We conclude with an end-to-end regret bound leveraging this reduction.

Online Convex Optimization with Memory

Protocol 6.4 Online Convex Optimization (OCO)

Intialize: Compact, convex constraint set $\mathcal{C} \subset \mathbb{R}^d$
for each $t = 1, 2, \dots, T$ **do**
    Learner choses iterate $\mathbf{z}_t \in \mathcal{C}$
    Nature selects convex function $f_t : \mathcal{C} \to \mathbb{R}$
    Learner suffers loss $f_t(\mathbf{z}_t)$

Like nonstochastic control, OCO (Protocoal 6.4) proceeds in rounds. At each round $t = 1, 2, \dots, T$, the learner selects an iterate $\mathbf{z}_t \in \mathcal{C}$, where $\mathcal{C} \subset \mathbb{R}^d$ is a convex and compact set. Subsequently, nature selects a convex function $f_t : \mathcal{C} \to \mathbb{R}$, and the learner suffers loss $f_t(\mathbf{z}_t)$. In OCO, the performance metric is also regret, defined as

$$\text{OCOREG}_T := \sum_{t=1}^{T} f_t(\mathbf{z}_t) - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(\mathbf{z}_t). \tag{6.31}$$

Here, regret compares performance against the best *fixed* $z \in \mathcal{C}$, chosen in hindsight given knowledge of the loss functions f_t.[2] OCO is not quite flexibile enough for our purposes here, because the iterate $\mathbf{z}_t$ is chosen anew at each time t, and the loss incured depends only on that iterate.

[2]Other works have considered more sophisticated comparators, such as sequences of $z_1, z_2, \dots$ with low total movement, or desiderata which enforce low regret on every interval $\{s, s+1, \dots, t\} \subset [T]$

In control, past actions affect future states, and therefore future losses incurred. Thus, the DRC framework is built on a generalization of OCO called OCO *with memory*, abbreviated OCOM, and outlined in Protocol 6.6. In OCOM, Nature selects a $h+1$-argument function $F_t : \mathcal{C}^{h+1} \to \mathbb{R}$, and the learner suffers loss $F_t[\mathbf{z}_{t:t-h}] = F_t[\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-h}]$. To avoid confusion, we use square brackets for $h+1$-ary functions, and parenthesis for unary functions.

The regret is defined with respect to the best constant action. Formally, we let $f_t(z) := F_t[z, \ldots, z]$ denote the *unary specialization* of F_t; that is, the function from $\mathcal{C} \to \mathbb{R}$ obtained by using a fixed z for all $h+1$ arguments. The relevant notion of regret is defined as

$$\text{MEMREG}_T := \sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z). \tag{6.32}$$

Protocol 6.6 Online Convex Optimization with Memory (OCOM)

Intialize: Initial iterates $\mathbf{z}_0, \mathbf{z}_{-1}, \ldots, \mathbf{z}_{h-1}$. Compact, convex constraint set $\mathcal{C} \subset \mathbb{R}^d$
for each $t = 1, 2, \ldots, T$ **do**
    Learner choses iterate $\mathbf{z}_t \in \mathcal{C}$
    Nature selects convex function $F_t : \mathcal{C}^{h+1} \to \mathbb{R}$, with unary loss $f_t(z) = F_t(z, \ldots, z)$.
    Learner suffers loss $F_t(\mathbf{z}_t, \mathbf{z}_{t-1}, \mathbf{z}_{t-h})$

We shall establish a correspondence between policies $\pi \in \Pi_{\text{ldc}}$, and continuous parameters $z \in \mathcal{C}$, for an appropriate convex set $\mathcal{C}$. Before proceeding describe a simple algorithm which attains small MEMREG_T when the show f_t are convex.

Memory-Regret Minimization via OGD

We now introduce online gradient descent (OGD), a classical online learning algorithm which updates interates via gradient updates. In a seminal work, Zinkevich [2003] demonstrated that OGDwith an appropriate selection of step size attains $\text{OCOREG}_T = \mathcal{O}(\sqrt{T})$. An exceptionally elegant argument due to Anava et al. [2015] a decade later demonstrated the same for our criterion of interest, MEMREG_T.

Going forward, we let $\text{Proj}_{\mathcal{C}}$ denote the Euclidean projection onto the convex domain $\mathcal{C}$, and let $\partial f(\cdot)$ denote the subgradient of a convex function f (see e.g. Bubeck [2014, Chapter 3]), which coincides with the gradient $\nabla f(\cdot)$ when f is differentiable. Given a sequence of convex loss functions (f_t), the online gradient descent algorithm selects parameters $\mathbf{z}_t$ which are updated by projected gradient steps $\mathbf{z}_{t+1} = \text{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta \partial f_t(\mathbf{z}_t))$. Pseudocode is given in Algorithm 6.1.

For our OCOM setting of interest, we feed Algorithm 6.1 the sequence (f_t) arising from the unary specializations of the $h+1$-variate, with memory losses F_t. We require the following definition:

Algorithm 6.1 Online Gradient Descent (OGD)

Intialize: Step size η, domain $\mathcal{C}$, arbitrary initial iterate $\mathbf{z}_1 \in \mathcal{C}$
for each $t = 1, 2, \ldots, T$ **do**
    Learner selects iterate $\mathbf{z}_t$,
    Learner recieves function $f_t : \mathcal{C} \to \mathbb{R}$
    Learer updates iterate $\mathbf{z}_{t+1} = \text{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta \partial f_t(\mathbf{z}_t))$.

Definition 6.1. Recall that $f : \mathcal{C} \to \mathbb{R}$ is L_f Lipschitz if, for all $z, z' \in \mathcal{C}'$, $|f(z) - f(z)| \leq L_f \|z - z'\|$. We say that $F : \mathcal{C}^{h+1} \to \mathbb{R}$ is L_s-sup-Lipschitz if, for all $z_{t:t-h}, z'_{t-j} \in \mathcal{C}^{h+1}$,

$$\left|F[z_{t:t-h}] - F[z'_{t:t-h}]\right| \leq L_s \max_{i \in \{0,1,\ldots,h\}} \|z_{t-i} - z'_{t-i}\|.$$

The following lemma provides a regret bound:

Proposition 6.3. *Consider any sequence of $h+1$-variate functions $F_t : \mathcal{C} \to \mathbb{R}$, with unary specialization $f_t(z) = F_t(z, \ldots, z)$. Let L_s be an upper bound on the sup-Lipschitz constant of F_t. Finally, suppose that the domain $\mathcal{C}$ has Euclidean diameter D. Then, the sequence of iterates $(\mathbf{z}_t)$ produced by Algorithm 6.1 when fed the unary functions (f_t) with step size $\eta > 0$ has memory-regret bounded by*

$$\text{MEMREG}_T(\text{OGD}) = \sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - \min_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z) \leq \frac{D^2}{\eta} + \eta T L_s^2 (h+1)$$

In particular, by selecting η to minimze the above expression, the memory regret is at most $DL_s\sqrt{(h+1)T}$.

Notice that the OGD algorithm does not use the $h+1$-variate losses explicitly, but only their unary specialization. Nevertheless, OGD attains low memory-regret against the (F_t) sequence. The key idea is that the memory-regret is related to standard regret, plus a penalty for the overal movement of the iterates. This idea is conceptually important for the following chapter, so we include a brief proof:

Proof. This proof is an adaption of the remarkably elegant argument of Anava et al. [2015]. Begin with the regret decomposition

$$\text{MEMREG}_T(\text{OGD}) = \underbrace{\sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - \sum_{t=1}^{T} f_t(\mathbf{z}_t)}_{(i)} + \underbrace{\sum_{t=1}^{T} f_t(\mathbf{z}_t) - \min_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z)}_{(ii)}.$$

Term (ii) is just the standard OCO regret of the algorithm. We begin with the observation that F_t being L_s-sup Lipschitz implies that f_t is L_s-Lipschitz. Hence, term (ii) isat most $\frac{D^2}{\eta} + \eta T L_s^2$ by the standard analysis of Zinkevich [2003]. The essential difference in OCOM

is term (i), which captures the effect of the memory. Let us bound the term for a fixed time t. Since F_t is L_s-sup-Lipschitz, an application of the triangle inequality yields

$$|F_t[\mathbf{z}_{t:t-h}] - f_t(\mathbf{z}_t)| \le L_s \max_{i\in\{0,1,\dots,h\}} \|\mathbf{z}_t - \mathbf{z}_{t-i}\| \le L_s \sum_{i=1}^{h} \|\mathbf{z}_{t-i+1} - \mathbf{z}_{t-i}\|.$$

For the initial iterates $t - i \le 1$, $\mathbf{z}_{t-i+1} = \mathbf{z}_{t-i}$. Otherwise, $\mathbf{z}_{t-i+1} = \mathrm{Proj}_{\mathcal{C}}(\mathbf{z}_{t-i} - \eta\partial f_t(\mathbf{z}_{t-i}))$. Since the projection operator is a contraction in the Euclidean norm,

$$\|\mathbf{z}_{t-i+1} - \mathbf{z}_{t-i}\| \le \|(\mathbf{z}_{t-i} - \eta\partial f_t(\mathbf{z}_{t-i})) - \mathbf{z}_{t-i}\| \le \eta\|\partial f_t(\mathbf{z}_{t-i})\|.$$

Finally, since f_t is L_s-Lipschitz, $\|\partial f_t(\mathbf{z}_{t-i})\| \le L_s$. Retracing our steps, $|F_t[\mathbf{z}_{t:t-h}] - f_t(\mathbf{z}_t)| \le L_s \sum_{i=1}^{h} \|\mathbf{z}_{t-i+1} - \mathbf{z}_{t-i}\| \le \eta h L_s^2$. Summing this contribution over all times t concludes the proof. □

Reduction to OCO with Memory

Having described the OCOM setting, and having introduced an algorithm which enjoys low regret, we now describe a reduction from nonstochastic control. We describe the reduction in terms of static-feedback DRC parametrization, with feedback matrix K. Recall that, under this parameterization, we select *exogenous inputs* $\mathbf{u}_t^{\mathrm{ex}}$ which are affine functions of the sequence $(\mathbf{y}_t^K)$. Recall also the shorthand $\mathbf{v}_t = (\mathbf{y}_t, \mathbf{u}_t)$. Generalizations to general parameterizations are given in Section 6.5.

Formally, fix a memory length $h \in N$, a control parameter $m \in \mathbb{N}$, and a control radius $R_{\mathcal{M}}$. We consider the convex domain $\mathcal{M} = \mathcal{M}(m, R_{\mathcal{M}})$, which we recall describes all sequences of operators $M = (M^{[0]}, M^{[1]}, \dots, M^{[m-1]}$, where each $M^{[i]} \in \mathbb{R}^{d_u \times d_\omega}$. We define the function

$$u_t^{\mathrm{ex}}(M) := \sum_{i=0}^{m-1} M^{[i]} \cdot \mathbf{y}_{t-i}^K,$$

which is just a finite impulse-reponse to past $(\mathbf{y}_s^K)$. Accordingly, given $M_{t:t-h} \in \mathcal{M}^{h+1}$, define

$$v_t[M_{t:-h}] = \begin{bmatrix} y_t[M_{t:t-h}] \\ u_t[M_{t:t-h}] \end{bmatrix} = \mathbf{v}_t^K + \sum_{i=0}^{h} G_K^{[i]} \cdot u_t^{\mathrm{ex}}(M_{t-i}),$$

where G_K is the Markov operator describing the response from exogenous inputs $\mathbf{u}_t^{\mathrm{ex}}$ to $(\mathbf{y}_t, \mathbf{u}_t)$, defined in Eq. (6.20). Finally, we define the $h+1$-ary loss

$$F_t[M_{t:t-h}] := \ell_t(v_t[M_{t:t-h}]).$$

The unary specializations are

$$f_t(M) := \ell_t(\mathbf{v}_t(M)), \quad \text{where } v_t(M) = v_t[M, \dots, M] = \mathbf{v}_t^{\pi_0} + \sum_{i=0}^{h} G_K^{[i]} u_t^{\mathrm{ex}}(M).$$

In words, $v_t[M_{t:t-h}]$ is the counterfactual (output, input) pair $(\mathbf{y}_t, \mathbf{u}_t)$ which would arise when selecting exogenous inputs $u_s^{\text{ex}}(M_s)$ based on DRC controllers M_s for $s = t-h, t-h+1, \ldots, t$, in an idealized system which had finite memory h. $F_t[M_{t:t-h}]$ then describes the cost that would be suffered in such a system. The unary specialization $v_t(M)$ corresponds to the (ouput,input) pair for a fixed DRC controller M, and again $f_t(M)$ is the corresponding cost. Here, the use of non-bolds is to distinguish between functions instead of iterates.

In addition to the DRC length m and memory h (both supressed in the notation), the definition of the functions above requires access to the Markov operator G_K, and to the counterfactual sequence $\mathbf{v}_t^K = (\mathbf{y}_t^K, \mathbf{u}_t^K)$ which would have occured had K been selected as the control policy. Recall that, in this section, we know the dynamical matrices $(A_\star, B_\star, C_\star)$. Therefore, the operator G_K can be computed explicitly. Then, $\mathbf{v}_t^K$ can be computed by inverting Eq. (6.21):

$$\mathbf{v}_t^K := \begin{bmatrix} \mathbf{y}_t^K \\ \mathbf{u}_t^K \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t \\ K\mathbf{y}_t \end{bmatrix} - \sum_{i=1}^{t} G_K^{[i]} \mathbf{u}_{t-i}^{\text{ex}} \tag{6.33}$$

Hence, at each time t, we can form the loss f_t by recovering $(\mathbf{v}_s^K)$ for times $s \le t$, and using our preccomputed Markov operator $G_{\text{ex}\to v}$.

This gives rise the reduction in Algorithm 6.2. At a high level, Algorithm 6.2 is an online policy iteration procedure, which uses a blackbox algorithm $\mathcal{A}$ to update DRC-parametrized control policiee. More specifically, each time t, we maintain a current DRC parameter $\mathbf{M}_t \in \mathcal{C}$. After observing $\mathbf{y}_t$, we recover the current $\mathbf{y}_t^K$, and select the exogenous input $\mathbf{u}_t^{\text{ex}} = u_t^{\text{ex}}(\mathbf{M}_t)$ using the current parameter. We then choose the total input $\mathbf{u}_t = K\mathbf{y}_t + \mathbf{u}_t^{\text{ex}}$ to include the feedback with the output $\mathbf{y}_t$. Finally, we observe the cost ℓ_t, form the function $f_t(\cdot)$, and feed it to the OCOM learning algorithm $\mathcal{A}$.[3]

Algorithm 6.2 Nonstochastic Control to OCOM reduction

Intialize: OCOM algorithm $\mathcal{A}$, nominal controller K, DRC length m, memory h. inital DRC controller $\mathbf{M}_1 = 0$, set $\mathcal{M} = \mathcal{M}(R_{\mathcal{M}}, m)$.
Precompute: Markov operators G_K
for each $t = 1, 2, \ldots, T$ **do**
    Observe output $\mathbf{y}_t$
    Recover $\mathbf{v}_t^K$ via Eq. (6.33).
    Select exogenous input $\mathbf{u}_t^{\text{ex}} = u_t^{\text{ex}}(\mathbf{M}_t)$
    Select total input $\mathbf{u}_t = \mathbf{u}_t^{\text{ex}} + K\mathbf{y}_t^{\mathsf{alg}}$.
    Receive cost $\ell_t(\cdot, \cdot)$, and form function $f_t(\cdot)$ as above.
    Feed f_t to learning algorithm $\mathcal{A}$, and recieve updated parameter $\mathbf{M}_{t+1} \in \mathcal{M}$.

Notice that the $h+1$-variate functions $F_t[\cdot]$ are not used in the algorithm. This is in part because the OCOM algorithms to which we reduce (e.g. OGD, Algorithm 6.1) only use

[3]For simplicity, we assume the learning algorithm $\mathcal{A}$ only uses unary losses f_t. This is because, to our knowledge, all know algorithms for OCO with memory have this property. However, the reduction still holds even if $\mathcal{A}$ requires the full $h+1$-variate losses F_t.

f_t in their update rules. Nevertheless, the definition of $F_t[\cdot]$ is helpful to understand the guarantees for the reduction. In the following, we assume that Assumption 6.2b holds; that is, $\max_t \|\mathbf{v}_t^K\| \le R_{\text{nat}}$, and let $R_G = \|G_K\|_{\ell_1,\text{op}}$.

Proposition 6.4 (Nonstochastic Control reduces to OCoM)**.** *Consider the reduction described by Algorithm 6.2, with and let ψ be a decay function with $\psi(0) \le R_{\mathcal{M}}$. Then,*

$$\textsc{NscReg}_T(\mathsf{alg};\psi) \le \textsc{MemReg}_T(\mathcal{A}) + 4LTR_{\text{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\frac{\psi_G(h+1)}{R_G} + \frac{\psi(h+1)}{R_{\mathcal{M}}}\right).$$

Moreover, for $L_s = 2LR_{\text{nat}}^2 R_{\mathcal{M}} R_G \sqrt{m}$ and $D = \sqrt{d_u} R_{\mathcal{M}}$, each F_t is L_s-sup-Lipschitz, and the Euclidean diameter is at most D.

An End-to-End Guarantee

Let us recap what we have shown thus far. First, we introduced the online convex optimization (OCO) and OCOwith memory (OCoM) online learning settings, with their associated notions of regret OCoReg (OCO regret) and MemReg (memory regret). We then introduced online gradient descent, OGD, and demonstrated that it enjoys sublinear memory regret (Proposition 6.3). Subsequently, we specified a black box reduction from nonstochastic control to a black box learning procedure $\mathcal{A}$, whose guarantee is given in Proposition 6.4.

To conclude the section, we combine the two guarantees, yielding an end-to-end bound for nonstochastic control with known system dynamics. Again, we assume that Assumption 6.2b holds; that is, $\max_t \|\mathbf{v}_t^K\| \le R_{\text{nat}}$, and let $R_G = \|G_K\|_{\ell_1,\text{op}}$.

Theorem 6.2. *Consider the reduction described by Algorithm 6.2, instantiated with gradient descent with step size η appropriately selected. Set $\bar{R} = R_{\text{nat}} R_{\mathcal{M}} R_G$. Then for a decay function ψ be a decay function with $\psi(0) \le R_{\mathcal{M}}$,*

$$\textsc{NscReg}_T(\mathsf{alg};\psi) \le 4L\frac{\bar{R}^2}{R_G}\sqrt{d_u m(h+1)T} + 4LT\bar{R}^2\left(\frac{\psi_G(h+1)}{R_G} + \frac{\psi(h+1)}{R_{\mathcal{M}}}\right).$$

In particular, if $\psi_G(n) \le c_K\rho^n$ and $\psi(n) \le c_M\rho^n$ for some $\rho \in (0,1)$, then selecting $m = h = \lceil \log_{1/\rho}(T)\rceil$,

$$\textsc{NscReg}_T(\mathsf{alg};\psi) \lesssim L\bar{R}^2 \frac{\log T\sqrt{d_u \log T}}{1-\rho} = \mathcal{O}(\sqrt{T})$$

The above theorem is a direct consequence of the reduction in Proposition 6.4 and of the analysis of OGD afforded Proposition 6.3. [4] The $\sqrt{T}$-rate is shown to be optimal in Simchowitz et al. [2020].

We now turn to the proof of the reduction.

[4] The second estimate uses that $\log_{1/\rho}(x) = \frac{\log x}{\log(1/\rho)} \lesssim \frac{\log x}{1-\rho}$.

Proof of Proposition 6.4

Introduce the shorthand $\Pi_\star = \Pi_{K\to u}(\psi)$ for our policy comparaotr class, and $\mathcal{M} = \mathcal{M}(m, R_{\mathcal{M}})$ the constraint set of DRC controllers. We begin with the following regret decomposition,

$$\begin{aligned}
\textsc{NscReg}_T(\mathsf{alg}; \Pi_\star) &= \sum_{t=1}^T \ell_t(\mathbf{v}_t, \mathbf{u}_t) - \inf_{\pi\in\Pi_\star} \sum_{t=1}^T \ell_t(\mathbf{y}_t^\pi, \mathbf{u}_t^\pi) \\
&\le \underbrace{\sum_{t=1}^T |\ell_t(\mathbf{y}_t, \mathbf{u}_t) - F_t[\mathbf{M}_{t:t-h}]|}_{(i.a)} + \underbrace{\sum_{t=1}^T F_t[\mathbf{M}_{t:t-h}] - \inf_{M\in\mathcal{M}} \sum_{t=1}^T f_t(M)}_{(ii)} \\
&\quad + \underbrace{\max_{M\in\mathcal{M}} \sum_{t=1}^T \left| f_t(M) - \ell_t(\mathbf{y}_t^M, \mathbf{u}_t^M)\right|}_{(i.b)} + \underbrace{\left| \inf_{M\in\mathcal{M}} \sum_{t=1}^T \ell_t(\mathbf{y}_t^M, \mathbf{u}_t^M) - \inf_{\pi\in\Pi_\star} \sum_{t=1}^T \ell_t(\mathbf{y}_t^\pi, \mathbf{u}_t^\pi)\right|}_{(iii)}.
\end{aligned}$$

Observe that term (ii) is precisely the memory regret, $\textsc{MemReg}_T$. Term (iii) captures the policy approximation error, and is at most $2LTR_{\mathcal{M}}R_G^2R_{\text{nat}}^2\psi(m)$ by Theorem 6.1b.

Finally, terms $(i.a)$ and $(i.b)$ capture the effect of truncating the dynamics to have memory h. To bound them, we give some estimates on key quantities, which also play a role in bounding the Lipschitz constants of f_t and F_t. To do so, we define the Frobenius norm of $M \in \mathcal{M}$ in the natural way: $\|M\|_{\text{F}}^2 = \sum_{i=0}^{m-1} \|M^{[i]}\|_{\text{F}}^2$. Our estimates are as follows:

Lemma 6.5. *For all times t, and all $M_{t:t-h}, M'_{t:t-h} \in \mathcal{M}(m, R_{\mathcal{M}})$*

- *$\|\mathbf{v}_t\|$, $\|v_t[M_{t:t-h}]\|$, and $\|v_t(M_t)\|$ are bounded by $2R_GR_{\mathcal{M}}R_{\text{nat}}$.*
- *$\|v_t[M_{t:t-h}] - v_t[M'_{t:t-h}]\| \le R_{\text{nat}}R_G \cdot \max_i \|M_{t-i} - M'_{t-i}\|_{\ell_1,\text{op}} \le \sqrt{m}R_{\text{nat}}R_G \max_i \|M_{t-i} - M'_{t-i}\|_{\text{F}}$*
- *The Euclidean diameter of the set $\mathcal{M}$ is at most $\sqrt{d_u}R_{\mathcal{M}}$.*

Proof of Lemma 6.5. For the first item, let us perform the computation for $\|\mathbf{v}_t\|$; the bounds for the others are similar. First, observe that since our algorithm select $\mathbf{u}_t^{\text{ex}} = \sum_{j=0}^{m-1} \mathbf{M}_t^{[j]}\mathbf{y}_t^K$, and since $\|\mathbf{y}_t^K\| \le \|\mathbf{v}_t^K\| \le R_{\text{nat}}$,

$$\begin{aligned}
\|\mathbf{v}_t\| = \|\mathbf{y}_t^K + \sum_{i=0}^{T-1} G_K^{[i]}\mathbf{u}_t^{\text{ex}}\| &= \|\mathbf{v}_t^{\text{nat}} + \sum_{i=0}^{T-1}\sum_{j=0}^{m-j} G_K^{[i]}\mathbf{M}_t^{[j]}\mathbf{y}_{t-i-j}^K\| \\
&\le \|\mathbf{y}_t^K\| + \sum_{i=0}^{T-1}\sum_{j=0}^{m-j} \|G_K^{[i]}\|_{\text{op}}\|\mathbf{M}_t^{[j]}\|_{\text{op}}\|\mathbf{y}_t^K\| \\
&\le R_{\text{nat}}\left(1 + \|G_K\|_{\ell_1,\text{op}} \max_t \|\mathbf{M}_t\|_{\ell_1,\text{op}}\right) \le 2R_GR_{\mathcal{M}}R_{\text{nat}},
\end{aligned}$$

where we use that $R_G = \|G_K\|_{\ell_1,\mathrm{op}} \geq 1$, and that $\mathbf{M}_t \in \mathcal{M}(m, R_{\mathcal{M}})$.

For the second item, an expansion reveals that

$$\|v_t[M_{t:t-h}] - v_t[M'_{t:t-h}]\| = \|\sum_{i=0}^{h} G_K^{[i]} \sum_{j=0}^{m-1} (M^{[i]} - (M^{[i]})')\mathbf{y}_t^K\|.$$

A similar manipulation to item 1 gives the bound $R_{\mathrm{nat}} R_G \cdot \max_i \|M_{t-i} - M'_{t-i}\|_{\ell_1,\mathrm{op}}$. Moreover, $\|M - M'\|_{\ell_1,\mathrm{op}} \leq \sqrt{m}\|M - M'\|_{\mathrm{F}}$ for all M, M' of length m, yielding the second item. The third item uses the fact that $\|M\|_{\mathrm{F}} \leq \sum_{i=0}^{m-1} \|M^{[i]}\|_{\mathrm{F}} \leq \sqrt{d_u} \sum_{i=0}^{m-1} \|M^{[i]}\|_{\mathrm{op}} \leq \sqrt{d_u} R_{\mathcal{M}}$. □

Using these estimates, we first bound terms $(i.a)$ and $(i.b)$:

Lemma 6.6. *Bound terms $(i.a)$ and $(i.b)$ are at most $2LR_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 R_G \psi_G(h+1)$.*

Proof. Let us bound term $(i.a)$; term $(i.b)$ is similar, wiht shorthand $\mathbf{v}_t = (\mathbf{y}_t, \mathbf{u}_t)$. For any single time t, the L-suquadratic condition, the definition $F_t[\mathbf{M}_{t:t-h}] = \ell_t(v_t[\mathbf{M}_{t:t-h}])$, and the estimates in Lemma 6.5 yield

$$\begin{aligned} |\ell_t(\mathbf{v}_t) - F_t[\mathbf{M}_{t:t-h}]| &\leq L \max\{\|\mathbf{v}_t\|, v_t(\mathbf{M}_{t:t-h}), 1\} \cdot \|v_t[\mathbf{M}_{t:t-h}] - \mathbf{v}_t\| \\ &\leq 2LR_{\mathrm{nat}} R_{\mathcal{M}} R_G \|v_t[\mathbf{M}_{t:t-h}] - \mathbf{v}_t\|. \end{aligned} \tag{6.34}$$

Moreover, we can expand

$$\begin{aligned} v_t[\mathbf{M}_{t:t-h}] &= \mathbf{v}_t^K + \sum_{i=0}^{h} G_K^{[i]} u_t^{\mathrm{ex}}(\mathbf{M}_t) = \mathbf{v}_t^K + \sum_{i=0}^{h} \sum_{j=0}^{m-1} G_K^{[i]} \mathbf{M}_{t-i}^{[j]} \mathbf{y}_{t-i-j}^K \\ \mathbf{v}_t &= \mathbf{v}_t^K + \sum_{i=0}^{t-1} \sum_{j=0}^{m-1} G_K^{[i]} \mathbf{M}_{t-i}^{[j]} \mathbf{y}_{t-i-j}^K. \end{aligned}$$

Hence, with similar simplifications as in Lemma 6.5,

$$\|v_t[\mathbf{M}_{t:t-h}] - \mathbf{v}_t\| \leq \|\sum_{i=h+1}^{t-1} \sum_{j=0}^{m-1} G_K^{[i]} \mathbf{M}_{t-i}^{[j]} \mathbf{y}_{t-i-j}^K\| \leq \sum_{i \geq h+1} \|G_K^{[i]}\| \cdot R_{\mathrm{nat}} \cdot R_{\mathcal{M}} = R_{\mathrm{nat}} \cdot R_{\mathcal{M}} \psi_K(h+1).$$

Hence, $|\ell_t(\mathbf{v}_t) - F_t[\mathbf{M}_{t:t-h}]| \leq 2LR_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 R_G \psi_G(h+1)$. Summing over times T yields the bound. □

Summing the bounds on terms $(i.a)$ and $(i.b)$, and the bound on term (iii) due to Theorem 6.1b, we find

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_\star) \leq \textsc{MemReg}_T(\mathcal{A}) + 4LTR_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\frac{\psi_G(h+1)}{R_G} + \frac{\psi(h+1)}{R_{\mathcal{M}}} \right).$$

Finally, it remains to characterize the sup-Lipschitz constain of L_f. With a similar computation in Eq. (6.34) followed by the second estimate in Lemma 6.5,

$$\begin{aligned}\|F_t[M_{t:t-h}] - F_t[M'_{t:t-h}]\| &\le 2LR_{\text{nat}}R_{\mathcal{M}}R_G\|v_t[M_{t:t-h}] - v_t[M'_{t:t-h}]\| \\ &\le 2LR_{\text{nat}}^2R_{\mathcal{M}}R_G\sqrt{m}\max_i\|M_{t-i} - M'_{t-i}\|_{\text{F}}.\end{aligned}$$

Hence, F_t is L_s-sup-Lipschitz for $L_s = 2LR_{\text{nat}}^2R_{\mathcal{M}}R_G\sqrt{m}$.

6.4 Systems with Unknown Dynamics

In many cases of interest, the system dynamics determined by $(A_\star, B_\star, C_\star)$ are not known a prior. Indeed, the first part of this thesis are concerned solely with estimating system dynamics from data.

To extend to unknown system dynamics, we notice that the above reduction described by Algorithm 6.2 only requires knowledge of the Markov operator G_K, but not knowledge of the state-space representation of the system matrices $(A_\star, B_\star, C_\star)$. Moreover, by assumption, K stabilizes our system, so G_K is a stable Markov operator. Hence, we propose a two-stage algorithm: first, we estimate G_K directly via least squares, as in Chapter 4. We then replace all quantities in Algorithm 6.2 with approximates based on our least squares estimate of G_K, and only the reduction to OCO with memory. This section elaborates on this approach.

A plug-in reduction

In this section, we generalize the reduction in Algorithm 6.2 by replacing exact knowlege of G_K and of $\mathbf{y}^K$ with inexact estimates. Let $\widehat{G} \in \mathscr{G}(d_y + d_u, d_u)$ be an estimate of G_K. We shall assume that our estimate has finite memory h, so that $\widehat{G}^{[i]} = 0$ for all $i > h$.

A natural estimate of $\mathbf{v}_t^K = (\mathbf{y}_t^K, \mathbf{u}_t^K)$ is to use $\widehat{G}$ as a plugin for Eq. (6.33):

$$\hat{\mathbf{v}}_t^K := \begin{bmatrix}\hat{\mathbf{y}}_t^K \\ \hat{\mathbf{u}}_t^K\end{bmatrix} = \begin{bmatrix}\mathbf{y}_t \\ K\mathbf{y}_t\end{bmatrix} - \sum_{i=1}^{h} G_K^{[i]}\mathbf{u}_{t-i}^{\text{ex}} \tag{6.35}$$

Using these estimates, we define an empirical approximation to the DRC input with parameter M, replacing $\mathbf{y}_{1:t}^K$ with $\hat{\mathbf{y}}_{1:t}^K$. In the interest of clarity, we define these approximations in terms of free non-bold parameter $\hat{y}_{1:t}^K, \hat{v}_t^K$; bold will be reserved for iterates recovered according to Eq. (6.35).

$$u_t^{\text{ex}}(M \mid \hat{y}_{1:t}^K) := \sum_{i=0}^{m-1} M^{[i]} \cdot \hat{y}_{t-i}^K,$$

To define the remaining quantities, we require as well $\hat{\mathbf{v}}_t^K$ for the affine term, and an explicit estimate $\widehat{G}$ of G_K:

$$v_t[M_{t:t-h} \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}] := \hat{v}_t^K + \sum_{i=0}^{h} \widehat{G}^{[i]} \cdot u_t^{\mathrm{ex}}(M_{t-i} \mid \hat{y}_{1:t-i}^K),$$
$$F_t[M_{t:t-h} \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}] := \ell_t(v_t[M_{t:t-h} \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}]).$$

The unary specializations are as follows:

$$v_t(M_t \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}) = v_t[M, \ldots, M \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}]$$
$$f_t(M \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}) = F_t[M_{t:t-h} \mid \hat{y}_{1:t}^K, \hat{v}_t^K, \widehat{G}],$$

and our reduction for unknown systems, Algorithm 6.2, generalizes to Algorithm 6.3 as follows.

Algorithm 6.3 Nonstochastic Control to OCOM reduction, estimated system

Intialize: OCOM algorithm $\mathcal{A}$, nominal controller K, DRC length m, memory h, estimate $\widehat{G}$ of G_K.
inital DRC controller $\mathbf{M}_1 = 0$.
for each $t = 1, 2, \ldots, T$ **do**
  Observe output $\mathbf{y}_t$
  Recover approximation $\hat{\mathbf{v}}_t^K$ via Eq. (6.35).
  Select exogenous input $\mathbf{u}_t^{\mathrm{ex}} = u_t^{\mathrm{ex}}(\mathbf{M}_t \mid \hat{\mathbf{y}}_{1:t}^K)$
  Select total input $\mathbf{u}_t = \mathbf{u}_t^{\mathrm{ex}} + K\mathbf{y}_t^{\mathsf{alg}}$.
  Receive cost $\ell_t(\cdot, \cdot)$, and form function $\hat{f}_t(\cdot) = f_t(\cdot \mid \hat{\mathbf{y}}_{1:t}^K, \hat{\mathbf{v}}_t^K, \widehat{G})$ as above.
  Feed f_t to learning algorithm $\mathcal{A}$, and recieve updated parameter $\mathbf{M}_{t+1} \in \mathcal{C}$.

It remains to specify how to estimate $\widehat{G}_K$. We apply the ordinary least squares estimator with i.i.d. Rademacher inputs, as analyzed in Chapter 4. We provide pseudocode for this estimation phase in Algorithm 6.4. The algorithms can be combined into a single reduction:

Guarantees for Unknown Systems

Combining the two procedures yields another reduction, which whose guarantees we now state. The following discussion assume that Assumption 6.2b holds, so that $\max_t \|\mathbf{v}_t^K\| \le R_{\mathrm{nat}}$. We also assume recall that $\|G_K\|_{\ell_1,\mathrm{op}} \le R_G$ and G_K has decay function ψ_G. Recall the memory parameter h and DRC length and radius m and $R_{\mathcal{M}}$. We $\delta > 0$ denote a confidence parameter.

Algorithm 6.4 Estimation of G_K

Intialize: nominal controller K, estimation length N, memory h:
for each $t = 1, 2, \ldots, N$ **do**
  Observe output $\mathbf{y}_t$
  Draw $\mathbf{u}_t^{\text{ex}} \overset{\text{unif}}{\sim} \{\frac{-1}{\sqrt{d_u}}, \frac{1}{\sqrt{d_u}}\}^{d_u}$
  Select total input $\mathbf{u}_t = \mathbf{u}_t^{\text{ex}} + K\mathbf{y}_t^{\mathsf{alg}}$.
Compute least squares estimate

$$\widehat{G}^{[1:h]} \in \sum_{t=1}^{N} \left\| \begin{bmatrix} \mathbf{y}_t \\ K\mathbf{y}_t \end{bmatrix} - \sum_{i=1}^{h} G^{[i]} \cdot \mathbf{u}_{t-i}^{\text{ex}} \right\|_2^2 \tag{6.36}$$

Set $\widehat{G}^{[0]} = \begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix}$ and $\widehat{G}^{[i]} = 0$ for $i > h$.
Return $\widehat{G}$.

Algorithm 6.5 Full Reduction for Unknown Systems

Intialize: nominal controller K, estimation length N, memory h, DRC parameters $m, R_{\mathcal{M}}$, online learning algorithm $\mathcal{A}$
Execute estimation phase Algorithm 6.4 for steps $t = 1, 2, \ldots, N$ to obtain $\widehat{G}$.
Execute Algorithm 6.3 for remaining times $t = N+1, N+2, \ldots, T$ with online learning algorithm $\mathcal{A}$ and estimator Markov operator $\widehat{G}$.

The reduction is stated in terms of the memory regret of the online learning subroutine $\mathcal{A}$ against the sequence of losses induced by our plug-in estimates. Formally, we consider

$$\textsc{MemReg}_T(\mathcal{A}; \hat{f}_{N+1:T}) = \sum_{t=N+1}^{T} \hat{F}_t[\mathbf{M}_{t:t-h}] - \inf_{M \in \mathcal{M}} \sum_{t=N+1}^{T} \hat{f}_t(M)$$
$$\text{where } \hat{F}_t[\cdot] = F_t[\cdot \mid \hat{\mathbf{y}}_{1:t}^K, \hat{\mathbf{v}}_t^K, \widehat{G}],\ \ \hat{f}_t(M) = \hat{F}_t[M, \ldots, M].$$

Our bound is stated in terms of the following constants:

$$\mathcal{C}_{G,\delta} := chR_{\text{nat}}^2 R_{\mathcal{M}}^2 R_G \sqrt{d_u^2 h(\log(1/\delta) + d_u d_y)},$$
$$\mathcal{C}_{\text{aprx}} := 60LR_{\text{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\tfrac{\psi_G(h+1)}{R_G} + \tfrac{\psi(h+1)}{R_{\mathcal{M}}}\right)$$
$$\mathcal{C}_{\text{burn}} := 4(R_G^2 + R_{\text{nat}}^2).$$

Here, $\mathcal{C}_{G,\delta}$ is the error term in our estimation bound, and $c > 0$ denotes an unspecified though universal constant. $\mathcal{C}_{\text{aprx}}$ addresses truncation to finite Markov operator memory h and finite DRC size m. Finally, $\mathcal{C}_{\text{burn}}$ addresses the growth of the state in the burn-in phase

of the learning procedure. We also define the minimal N and h of which these bounds hold:

$$N_\delta := \max\left\{2m, 8d_u h^2 \log(h^2 d_u/\delta), \mathcal{C}_{G,\delta}^2\right\}$$
$$h_0 := \inf\{h : \psi(h+1) \le \tfrac{1}{4R_{\mathcal{M}}}\}.$$

Finally, We can now state our main reduction.

Proposition 6.7. *Given a decacy function $\Pi_\star = \Pi_{K\to u}(\psi)$ denote our benchmark class. Assume $N \ge N_\delta$, $h \ge h_0$, and $R_{\mathcal{M}} \ge \psi(0)$, and that Assumption 6.2b holds. Then, if Algorithm 6.4 is run for N stops, followed by Algorithm 6.3 with the resulting estimate $\widehat{G}$ with online learning algorithm $\mathcal{A}$, then the following bound holds with probability at least $1-\delta$:*

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_{K\to u}(\psi)) \le N\mathcal{C}_{\mathrm{burn}} + \frac{T\mathcal{C}_{G,\delta}}{\sqrt{N}} + T\mathcal{C}_{\mathrm{aprx}} + \textsc{MemReg}_T(\mathcal{A}; \hat{f}_{N+1:T}),$$

Morover, $\hat{F}_t[\cdot]$ is $16LR_{\mathrm{nat}}^2 R_{\mathcal{M}} R_G\sqrt{m}$-sup-Lipschitz. In particular, by tuning N appropriately, the estimation phase contributes regret on the order of $T^{2/3}$:

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_{K\to u}(\psi)) \le (T\cdot \mathcal{C}_{\mathrm{burn}}\mathcal{C}_{G,\delta})^{2/3} + T\mathcal{C}_{\mathrm{aprx}} + \textsc{MemReg}_T(\mathcal{A}; \hat{f}_{N+1:T}).$$

An end-to-end guarantee follows from the above reduction and Proposition 6.3, much in the same way as did Theorem 6.2. In the interest of brevity, we state a bound where all terms have geometric decay, and suppress dependence on problem parameters:

Theorem 6.3. *Consider the reduction in Proposition 6.7, instantiated with online gradient descent with an appropriately tuned step size η. In addition, suppose the benchmark decay function ψ and decay of G_K, ψ_G are upper bounded by $\psi(n) \vee \psi_G(n) \le c\rho^n$. Then, by appropriately setting m, h and the estimation length N,*

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_{K\to u}(\psi)) \le \widetilde{\mathcal{O}}\left(T^{2/3}\right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides terms polynomial in $c, \frac{1}{1-\rho}, R_{\mathrm{nat}}, d_u, d_y, \log(1/\delta)$, and $\log T$.

Proof of Proposition 6.7

All ommited proofs are deferred to the following subsection. Throughout, we assume that

$$\begin{aligned}\|\widehat{G}^{[0:h]} - G_K^{[0:h]}\|_{\ell_1,\mathrm{op}} \le \epsilon_G &:= c_1 h R_{\mathrm{nat}}\sqrt{\frac{d_u^2 h(\log(1/\delta) + d_u d_y)}{N}} + \psi_G(h+1),\\ &\le \frac{\mathcal{C}_{G,\delta}}{48R_{\mathrm{nat}}^2 R_G R_{G_\star}\sqrt{N}} + \psi_G(h+1),\end{aligned}$$

which holds with high probability as a consequence of Lemma 6.10 below, stated formally at the end of the proof. We introduce our regret decomposition. To do so, let us introduce the shorthand:

$$\hat{F}_t[M_{t:t-h}] = F_t[M_{t:t-h} \mid \hat{\mathbf{y}}^K_{1:t}, \hat{\mathbf{v}}^K_t, \widehat{G}], \quad \hat{f}_t(M) = \hat{F}_t[M, \dots, M].$$

Here, $(\hat{F}, \hat{f})$ are the sequences used by the algorithm: they involve empirical estimates of $\hat{\mathbf{y}}^K_{1:t}$ to define the exogenous control input $\mathbf{u}^{\mathrm{ex}}_t = u_t(\mathbf{M}_t \mid \hat{\mathbf{y}}^K_{1:t})$, as well as empirical estimates of $\hat{\mathbf{v}}^K_t$ for the affine term and $\widehat{G}$ of G_K to estimate the Markov operator. However, the actual response to the control input is determined by the true Markov operator G_K and the true affine term $\mathbf{v}^K_t$. To capture this, we introduce

$$F^\star_t[M_{t:t-h}] = F_t[M_{t:t-h} \mid \hat{\mathbf{y}}^K_{1:t}, \mathbf{v}^K_t, G_K], \quad f^\star_t(M) = F^\star_t[M, \dots, M].$$

Finally, we let un-starred f_t denote the "clean" sequences with exact estimates:

$$F_t[M_{t:t-h}] = F_t[M_{t:t-h} \mid \mathbf{y}^K_{1:t}, \mathbf{v}^K_t, G_K], \quad f_t(M) = F_t[M, \dots, M].$$

We now adopt the following regret decomposition, using the nonnegative of the loss functions:

$$\begin{aligned}
\textsc{NscReg}_T(\mathsf{alg}) \le & \underbrace{\sum_{t=1}^{N+m+h} \ell_t(\mathbf{v}_t)}_{(i)} + \underbrace{\sum_{t=N+m+h+1}^{T} \ell_t(\mathbf{v}_t) - F^\star_t[\mathbf{M}_{t:t-h}]}_{(ii.a)} \\
& + \underbrace{\sum_{t=N+1}^{T} F^\star_t[\mathbf{M}_{t:t-h}] - \inf_{M \in \mathcal{M}} \sum_{t=N+1}^{T} f_t(M)}_{(iii)} \\
& \underbrace{\max_{M \in \mathcal{M}} \sum_{t=N+m+1+h}^{T} f_t(M) - \ell_t(\mathbf{v}^M_t)}_{(ii.b)} + \underbrace{\inf_{M \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(\mathbf{v}^M_t) - \inf_{\pi \in \Pi_\star} \sum_{t=1}^{T} \ell_t(\mathbf{v}^\pi_t)}_{(iv)}.
\end{aligned}$$

Term (i) corresponds to the regret lost during the forced exploration phase, and is at most $4N(R^2_{\mathrm{nat}} + R^2_G)$ by Lemma 6.10. Term (iii) is the approximation error induced by the DRC parameterization, and is bounded as in the known system case. Terms $(ii.a)$ and $(ii.b)$ address truncation to memory h; Term $(ii.b)$ is exactly as in the known system case.

Term $(ii.a)$ may be bounded similarly, but because it depends on the estiamtes $\widehat{\mathbf{v}}^K_t$, there is a catch: Because we no longer exactly recover G_K, we do not exactly recover $\mathbf{v}^K_t$. This introduces feedback into our algorithm, and care must be taken to ensure that instability does not ensue:

Lemma 6.8. *Suppose that $\epsilon_G \le \frac{1}{2R_{\mathcal{M}}}$ (which holds under the assumptions of our the proposition). Then, for all t, for all t, $\|\widehat{\mathbf{v}}^K_t\| \le 2R_{\mathrm{nat}}$. Moreover, $\|\widehat{\mathbf{v}}^K_t - \mathbf{v}^K_t\| \le 2\epsilon_G R_{\mathcal{M}} R_{\mathrm{nat}}$.*

Due to this uniform bound, on $\|\widehat{\mathbf{v}}_t^K\|$, we can essentially double every appearance of R_{nat} in the bound. This yields

$$(ii.a) + (ii.b) + (iv) \le 8LTR_{\text{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\frac{\psi_G(h+1)}{R_G} + \frac{\psi(m)}{R_{\mathcal{M}}} \right),$$

and thus, bounding the regret amounts to controlling term (iii)

$$\textsc{NscReg}_T(\mathsf{alg}) \le (iii) + 4N(R_G^2 + R_{\text{nat}}^2) + 8LTR_{\text{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\tfrac{\psi_G(h+1)}{R_G} + \tfrac{\psi(m)}{R_{\mathcal{M}}}\right). \tag{6.37}$$

Let us decompose term (iii) as follows:

$$\begin{aligned}
&\sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \inf_{M\in\mathcal{M}} \sum_{t=N+1}^{T} f_t(M) \\
&= \underbrace{\sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \sum_{t=N+1}^{T} \hat{F}_t[\mathbf{M}_{t:t-h}]}_{(iii.a1)} + \underbrace{\sup_{M\in\mathcal{M}} \left| \sum_{t=N+1}^{T} \hat{f}_t(M) - \sum_{t=N+1}^{T} f_t^\star(M) \right|}_{(iii.a2)} \\
&+ \underbrace{\sum_{t=N+1}^{T} \hat{F}_t[\mathbf{M}_{t:t-h}] - \inf_{M\in\mathcal{M}} \sum_{t=N+1}^{T} \hat{f}_t(M)}_{(iii.b)} \\
&+ \underbrace{\sup_{M\in\mathcal{M}} \left| \sum_{t=N+1}^{T} f_t^\star(M) - \sum_{t=N+1}^{T} f_t(M) \right|}_{(iii.c)}.
\end{aligned}$$

Recall that our reduction runs the online learning algorithm $\mathcal{A}$ on the $\hat{f}_t$ losses. Hence, the reduction guarantees regret on term $(iii.b)$. That is,

$$(iii.b) = \textsc{MemReg}_T(\mathcal{A}; \hat{f}_{N+1:T}).$$

The terms $(iii.a1)$ and $(iii.a2)$ are the error terms which arise because, even those the learning *algorithm* uses $\hat{f}_t$ for the losses, the learners *actual* losses are approximate by $f_t^\star$. Finally, the term $(iii.c)$ emerges because we are using a Drc parameterization based on estimates of $\mathbf{y}_t^K$, rather than the true sequence.

To control all these terms, we introduce some estimates analogous to Lemma 6.5. Before continuing, let us introduce the following notation, consistent with the definitions of $\hat{F}_t, F_t^\star, F_t$:

$$\begin{aligned}
\hat{v}_t[M_{t:t-h}] &= v_t[M_{t:t-h} \mid \hat{\mathbf{y}}_{1:t}^K, \hat{\mathbf{v}}_t^K, \widehat{G}], \quad \hat{v}_t(M) = \hat{v}_t[M,\dots,M] \\
v_t^\star[M_{t:t-h}] &= v_t[M_{t:t-h} \mid \hat{\mathbf{y}}_{1:t}^K, \mathbf{v}_t^K, G_K], \quad v_t^\star(M) = v_t^\star[M,\dots,M] \\
v_t[M_{t:t-h}] &= v_t[M_{t:t-h} \mid \mathbf{y}_{1:t}^K, \mathbf{v}_t^K, G_K], \quad v_t(M) = v_t[M,\dots,M].
\end{aligned}$$

Written out more explicitly,

$$\hat{v}_t[M_{t:t-h}] = \hat{\mathbf{v}}_t^K + \sum_{i=0}^{h}\sum_{j=0}^{m-1} \widehat{G}^{[i]} M_{t-i}^{[j]} \hat{\mathbf{y}}_{t-i-j}^K$$

$$v_t^\star[M_{t:t-h}] = \mathbf{v}_t^K + \sum_{i=0}^{h}\sum_{j=0}^{m-1} G_K^{[i]} M_{t-i}^{[j]} \hat{\mathbf{y}}_{t-i-j}^K$$

$$v_t[M_{t:t-h}] = \mathbf{v}_t^K + \sum_{i=0}^{h}\sum_{j=0}^{m-1} G_K^{[i]} M_{t-i}^{[j]} \mathbf{y}_{t-i-j}^K$$

We state the following lemma without proof; it can be verified by leveraging Lemma 6.8 to inflating R_{nat} and R_G by factors of 2 in the statement of Lemma 6.5:

Lemma 6.9. *For all times t, and all $M_{t:t-h}, M'_{t:t-h} \in \mathcal{M}(m, R_{\mathcal{M}})$*

- *$\|\hat{v}_t\|$, $\|\hat{v}_t[M_{t:t-h}]\|$, and $\|\hat{v}_t(M_t)\|$ are bounded by $8R_G R_{\mathcal{M}} R_{\text{nat}}$. The same is true of $v_t^\star$ and v_t. In particular, $f_t(M), f_t^\star(M), \hat{f}_t(M)$ are all at most $64R_G^2 R_{\mathcal{M}}^2 R_{\text{nat}}^2$.*
- *The functions $\hat{F}_t$ are $8L_s$-sup-Lipschitz, where $L_s = 2LR_{\text{nat}}^2 R_{\mathcal{M}} R_G \sqrt{m}$ was the sup-Lipschitz constant established for the known system in Proposition 6.4.*
- *Let v, v' be any two vectors in the image of $\mathcal{M}$ under $v_t[\cdot]$ $\hat{v}_t[\cdot]$, or $v^\star[\cdot]$ for some time t. Then, $|\ell_t(v) - \ell_t(v')| \le 4R_G R_{\mathcal{M}} R_{\text{nat}} L \|v - v'\|$.*

In light of part 3 of the above lemma, we have that

$$(iii.a1) + (iii.a2) + (iii.c) \le 8R_G R_{\mathcal{M}} R_{\text{nat}} L \times \tag{6.38}$$

$$\sum_{t=N+1}^{T} \|\hat{v}_t[\mathbf{M}_{t:t-h}] - v^\star[\mathbf{M}_{t:t-h}]\| + \sup_{M\in\mathcal{M}} \|\hat{v}_t(M) - v_t^\star(M)\| + \sup_{M\in\mathcal{M}} \|v_t^\star(M) - v_t(M)\|. \tag{6.39}$$

By expanding the definitions and invoking Lemma 6.8, we can bound

$$\begin{aligned}\|\hat{v}_t[\mathbf{M}_{t:t-h}] - v_t^\star[\mathbf{M}_{t:t-h}]\| &\le \|\widehat{\mathbf{v}}_t^K - \mathbf{v}_t^K\| + \|G_K - \widehat{G}\|_{\ell_1,\text{op}} R_{\mathcal{M}} \max_t \|\widehat{\mathbf{y}}_t^K\| \\ &\le 2\epsilon_G R_{\mathcal{M}} R_{\text{nat}} + 2\|G_\star - \widehat{G}\|_{\ell_1,\text{op}} \epsilon_G R_{\text{nat}} = 4\epsilon_G R_{\mathcal{M}} R_{\text{nat}}.\end{aligned}$$

A similar bound holds for $\sup_{M\in\mathcal{M}} \|\hat{v}_t(M) - v_t(M)\|$. Finally, again using Lemma 6.8

$$\|\hat{v}_t[\mathbf{M}_{t:t-h}] - v_t[\mathbf{M}_{t:t-h}]\| \le R_G R_{\mathcal{M}} \max_t \|\widehat{\mathbf{y}}_t^K - \widehat{\mathbf{y}}_t\| \le 2R_G R_{\mathcal{M}}^2 R_{\text{nat}}^2 \epsilon_G.$$

Hence, using $R_G \ge 1$, we conclude

$$\begin{aligned}(iii.a1) + (iii.a2) + (iii.c) &\le 48 R_G R_{\mathcal{M}}^2 R_{\text{nat}}^2 \epsilon_G T \\ &\le \frac{C_{G,\delta}}{\sqrt{N}} + 48 R_G^2 R_{\mathcal{M}}^2 R_{\text{nat}}^2 \cdot \frac{\psi(h+1)}{R_G}.\end{aligned}$$

Hence,

$$(iii) \le \textsc{MemReg}_T(\mathcal{A}; \hat{f}_{N+1:T}) + \frac{C_{G,\delta}}{\sqrt{N}} + 48 R_G^2 R_{\mathcal{M}}^2 R_{\mathrm{nat}}^2 \cdot \frac{\psi(h+1)}{R_G}.$$

Combining this bound with Eq. (6.37) concludes the proof. □

Omitted Proofs

Proof of Lemma 6.8. Introduce the norm $\|\mathbf{u}_{1:t}^{\mathrm{ex}}\|_{2,\infty} := \max_{1\le s\le t} \|\mathbf{u}_s^{\mathrm{ex}}\|$, and similarly for other sequences of vectors. Then,

$$\begin{aligned}
\|\widehat{\mathbf{v}}_t^K - \mathbf{v}_t^K\| &= \|\mathbf{v}_t - \sum_{i=0}^{t-1} G_K^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}} - (\mathbf{v}_t - \sum_{i=0}^{t-1} \widehat{G}^{[i]} \mathbf{u}_t^{\mathrm{ex}})\| \\
&\overset{(i)}{=} \|\sum_{i=1}^{t-1} (G_K^{[i-1]} - \widehat{G}^{[i]}) \mathbf{u}_{t-i}^{\mathrm{ex}}\| \\
&\le \|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}} \|\mathbf{u}_{1:t-1}^{\mathrm{ex}}\|_{2,\infty} \le \epsilon_G \|\mathbf{u}_{1:t-1}^{\mathrm{ex}}\|_{2,\infty},
\end{aligned}$$

where equality (i) uses that $\widehat{G}^{[0]} = G_K^{[0]}$. In particular,

$$\|\widehat{\mathbf{v}}_{1:t}^K\|_{2,\infty} \le \|\mathbf{v}_t^K\|_{2,\infty} + \epsilon_G \|\mathbf{u}_{1:t-1}^{\mathrm{ex}}\|_{2,\infty} \le R_{\mathrm{nat}} + \epsilon_G \|\mathbf{u}_{1:t-1}^{\mathrm{ex}}\|_{2,\infty}. \tag{6.40}$$

On the other hand, for any time t,

$$\|\mathbf{u}_t^{\mathrm{ex}}\| \le \begin{cases} 1 & t \le N \\ \sum_{i=0}^{m-1} \|\mathbf{M}_t^{[i]} \widehat{\mathbf{y}}_t^K\| & t > N \end{cases}$$

In the second case,

$$\|\mathbf{M}_t^{[i]} \widehat{\mathbf{y}}_t^K\| \le R_{\mathcal{M}} \|\widehat{\mathbf{y}}_{1:t}^K\|_{2,\infty} \le R_{\mathcal{M}} \|\widehat{\mathbf{v}}_{1:t}^K\|_{2,\infty}$$

Hence, $\|\mathbf{u}_{1:t-1}^{\mathrm{ex}}\|_{2,\infty} \le \max\{1, R_{\mathcal{M}} \|\widehat{\mathbf{v}}_{1:t-1}^K\|_{2,\infty}\}$. Combing with Eq. (6.40),

$$\|\widehat{\mathbf{v}}_{1:t}^K\|_{2,\infty} \le R_{\mathrm{nat}} + \epsilon_G \max\{1, R_{\mathcal{M}} \|\widehat{\mathbf{v}}_{1:t-1}^K\|_{2,\infty}\}.$$

Since $R_{\mathcal{M}}, R_{\mathrm{nat}} \ge 1$ by assumption, we mind that if $\epsilon_G \le \frac{1}{2R_{\mathcal{M}}}$, the above recursion has $\|\widehat{\mathbf{v}}_{1:t}^K\|_{2,\infty} \le 2R_{\mathrm{nat}}$ for all t. □

Lemma 6.10. *Suppose that $h \le h_0 = \frac{1}{4R_{\mathcal{M}}}$. After $N \ge N_\delta$ steps, the following guarantee holds with probability $1-\delta$:*

$$\|\widehat{G}^{[0:h]} - G_K^{[0:h]}\|_{\ell_1,\mathrm{op}} \le \epsilon_G := c_1 h R_{\mathrm{nat}} \sqrt{\frac{d_u^2 h(\log(1/\delta) + d_u d_y)}{N}} + \psi_G(h+1)$$

Moreover, $\sum_{t=1}^{N+h+m} \ell_t(\mathbf{y}_t, \mathbf{v}_t) \le 4NL(R_{\mathrm{nat}}^2 + R_G^2)$.

Proof of Lemma 6.10. For $i \le j$, we let $G^{[i:j]} = (G^{[i]}, G^{[i+1]}, \dots, G^{[j]})$. Note that $\widehat{G}^{[0]} = G_K^{[0]}$, and that $\widehat{G}^{[i]} = 0$ for $i \ge 1$. We begin by relating $\|\cdot\|_{\ell_1,\mathrm{op}}$ to the operator norm. The following can be established via a variational characterization of both norms:

$$\|\widehat{G}^{[0:h]} - G_K^{[0:h]}\|_{\ell_1,\mathrm{op}} = \|\widehat{G}^{[1:h]} - G_K^{[1:h]}\|_{\ell_1,\mathrm{op}} \le \sqrt{h}\|\widehat{G}^{[1:h]} - G_K^{[1:h]}\|_{\mathrm{op}}.$$

From the same analysis as in Corollary 4.1, with a renormalization of the inputs, we have for a universal constant c_0 and $N \ge N_0(\delta) := 8d_u h^2 \log(h^2 d_u/\delta)$

$$\|\widehat{G}^{[1:h]} - G_K^{[1:h]}\|_{\mathrm{op}} \le c_0 R\sqrt{\frac{d_u h(\log(1/\delta) + hd_u + d_y)}{N}},$$

where, where R is an upper bound on $\max_{t\in[N]} \|\boldsymbol{\delta}_t\|$, where $\boldsymbol{\delta}_t = \mathbf{v}_t - \sum_{i=0}^{[h]} G_k^{[i]}\mathbf{u}_{t-i}^{\mathrm{ex}}$. Using the decomposition $\mathbf{v}_t = \mathbf{v}_t^K + \sum_{i=0}^{t-1} G_k^{[i]}\mathbf{u}_{t-i}^{\mathrm{ex}}$, we can bound $\|\boldsymbol{\delta}_t\| \le R_{\mathrm{nat}} + \psi_G(h+1) \le 2R_{\mathrm{nat}}$, where we use that $\|\mathbf{u}_t^{\mathrm{ex}}\| = 1$, and that h is chosen such that $\psi_G(h+1) \le 1$, and that $R_{\mathrm{nat}} \ge 1$ by assumption. Hence,

$$\|\widehat{G}^{[1:h]} - G_K^{[1:h]}\|_{\mathrm{op}} \le c_0 R_{\mathrm{nat}}\sqrt{\frac{d_u h(\log(1/\delta) + hd_u + d_y)}{N}}.$$

Hence, for $c_1 = 2c_0$

$$\begin{aligned}\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}} &= \|\widehat{G}^{[1:h]} - G_K^{[1:h]}\|_{\ell_1,\mathrm{op}} + \psi_G(h) \\ &\le c_1 h R_{\mathrm{nat}}\sqrt{\frac{d_u^2 h(\log(1/\delta) + d_u d_y)}{N}} + \psi_G(h+1) := \epsilon_G(N,\delta).\end{aligned}$$

For the second bound, the $\|\mathbf{v}_t\| \le \|\mathbf{v}_K\| + \sum_{t=1}^{T-1} \|G_K^{[i]}\|\|\mathbf{u}_{t-i}^{\mathrm{ex}}\| \le R_{\mathrm{nat}} + R_G$. Hence, $\ell_t(\mathbf{v}_t) \le L(R_{\mathrm{nat}} + R_G)^2 \le 2L(R_{\mathrm{nat}}^2 + R_G^2)$ by Assumption 6.1. There are $N + m + h$ terms in the sum in question, and for $N \ge N_\delta$, $m + h \le N$, yielding a total bound of $2N \cdot 2L(R_{\mathrm{nat}}^2 + R_G^2)$. □

6.5 A General form of DRC

In general, a partially observed system can not be able to be stabilized by static feedback. To circumvent this, we describe stabilizing the system with an *dynamic feedback controller*. The following exposition mirrors Simchowitz et al. [2020], but is abridged considerably. We focus on explaining the various paramterizations. It can be verified that all theoretical results in this chapter — importantly, the reductions — extend directly to these parametrizations as well.

Precisely, the following analysis extends to all linear systems which can be stabilized given observations of outputs.

Definition 6.2 (Stabilizable and Detectable)**.** Given a triple $(A_\star, B_\star, C_\star) \in \mathbb{R}^{d_x^2} \times \mathbb{R}^{d_x \times d_u} \times \mathbb{R}^{d_y \times d_y}$, we say that

- $(A_\star, B_\star)$ is *stabilizable* if there exist a feedback matrix K such that $(A_\star + B_\star K)$ is stable.
- $(A_\star, C_\star)$ is *detectable* if there a matrix L such that $A_\star + C_\star L$ is stable.[5]

Going forward, we will abbreviate "detectable and stabilizable" systems as S&D systems.

It is well-known that a system is S&D if and only there exists any control policy π which ensures that the system is stable Zhou et al. [1996]. Hence, Definition 6.2 is essentially the most general definition to which we could hope our results apply.[6]

DRC with dynamic nominal controller

In the general parametrization, we maintain an internal state $\mathbf{s}_t$, which evolves according to the dynamical equations

$$\mathbf{s}_{t+1} = A_{\pi_0}\mathbf{s}_t + B_{\pi_0}\mathbf{y}_t + B_{\pi_0,u}\mathbf{u}_t^{\mathrm{ex}}, \tag{6.41}$$

and selects inputs as a combination of an exogenous input $\mathbf{u}_t^{\mathrm{ex}}$, and an endogenous input determined by the system:

$$\mathbf{u}_t = \mathbf{u}_t^{\mathrm{ex}} + (C_{\pi_0}\mathbf{s}_t + D_{\pi_0}\mathbf{y}_t). \tag{6.42}$$

Lastly, the algorithmic prescribes an *control output*, denoted by $\boldsymbol{\omega}_t \in \mathbb{R}^{d_\omega}$, given by

$$\boldsymbol{\omega}_t = C_{\pi_0,\omega}\mathbf{s}_t + D_{\pi_0,\omega}\mathbf{y}_t \in \mathbb{R}^{d_\omega}, \tag{6.43}$$

which we use to parameterize the controller. In both the internall stable parametrization (Nature's Ys), and the static feedback parametrization) we take $\boldsymbol{\omega}_t = \mathbf{y}_t$. However, for more sophisticated parameterizations described below, other choices of $\boldsymbol{\omega}_t$ are desirable.

We assume that π_0 is *stabilizing*, meaning that, if we have $\max_t \|\mathbf{e}_t\|, \|\mathbf{w}_t\|, \|\mathbf{u}_t^{\mathrm{ex,alg}}\| < \infty$ are bounded, then with $\max_t \|\mathbf{u}_t^{\mathsf{alg}}\|, \|\mathbf{y}_t^{\mathsf{alg}}\|, \|\boldsymbol{\omega}_t^{\mathsf{alg}}\| < \infty$. As a consequence of the Youla parametrization [Youla et al., 1976], one can always construct a controller π_0 which has this property for sufficiently non-pathological systems. Analogous to the sequence $\mathbf{y}_t^K, \mathbf{u}_t^K$, we consider a sequence that arises under no exogenous inputs:

Definition 6.3. We define the 'natural' sequence $\mathbf{y}_t^{\pi_0}, \mathbf{u}_t^{\pi_0}, \boldsymbol{\omega}_t^{\pi_0}$ as the sequence obtained by executing the stabilizing policy π_0 in the absence of $\mathbf{u}_t^{\mathrm{ex}} = 0$; we set $\mathbf{v}_t^{\pi_0} = (\mathbf{y}_t^{\pi_0}, \mathbf{u}_t^{\pi_0}) \in \mathbb{R}^{d_y+d_u}$. Each such sequence is determined uniquely by the disturbances $\mathbf{w}_t, \mathbf{e}_t$.

[5] Equivalently, if $(A_\star^\top, C_\star^\top)$ is controllable.

[6] Indeed, if a system is not S&D, then even *nonlinear* policies π fail to make the system stable in an input output sense: that is, bounded noise into the system can yield an unbounded response. Because we do not see states directly in this model, and only consider output costs, we can generalize the Definition 6.2 condition further to all systems $(A_\star, B_\star, C_\star)$ for which there exists an equivalent realization $(\tilde{A}, \tilde{B}, \tilde{C})$ with the same Markov operator (i.e. $\tilde{C}\tilde{A}^i\tilde{B} = C_\star A_\star^i B_\star$ for all $i \geq 0$) which is S&D. In this case, we can just "pretend" that the system is given by $(\tilde{A}, \tilde{B}, \tilde{C})$.

Moreover, the 'natural' sequences can be related to the sequences visited by the algorithm via linear Markov operators

Definition 6.4. We define the Markov operators $G_{\mathrm{ex}\to v} \in \mathscr{G}(d_y + d_u, d_u)$ and $G_{\mathrm{ex}\to\omega} \in \mathscr{G}(d_\omega, d_u)$ as the operators for which

$$\boldsymbol{\omega}_t^{\mathsf{alg}} = \boldsymbol{\omega}_t^{\pi_0} + \sum_{i=1}^{t} G_{\mathrm{ex}\to\omega}^{[t-i]} \mathbf{u}_i^{\mathrm{ex}}, \quad \mathbf{v}_t = \mathbf{v}_t^{\pi_0} + \sum_{i=1}^{t} G_{\mathrm{ex}\to v}^{[t-i]} \mathbf{u}_i^{\mathsf{alg}}. \tag{6.44}$$

We define the resulting decay function as

$$\psi_{\pi_0}(n) := \max\left\{ \sum_{i\geq n} \|G_{\mathrm{ex}\to\omega}^{[i]}\|_{\mathrm{op}}, \ \sum_{i\geq n} \|G_{\mathrm{ex}\to v}^{[i]}\|_{\mathrm{op}} \right\}.$$

A couple remarks are in order:

- The existence of these Markov operators in Definition 6.4 a direct consequence of the linear dynamics.
- We have that $G_{\mathrm{ex}\to\omega}^{[0]} = 0$, since $\mathbf{u}_t^{\mathrm{ex}}$ does not influence $\boldsymbol{\omega}_t^{\mathsf{alg}}$, and that $G_{\mathrm{ex}\to v}^{[0]} = \begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix}$, since $\mathbf{u}_t^{\mathsf{alg}} = \mathbf{u}_t^{\pi_0} + \mathbf{u}_t^{\mathrm{ex}}$, but $\mathbf{y}_t^{\mathsf{alg}}$ does not depend on $\mathbf{u}_t^{\mathrm{ex}}$.
- As noted about, $\boldsymbol{\omega}_t = \mathbf{y}_t$ for the internally stable and static feedback parametrizations above. Hence, for all i, $G_{\mathrm{ex}\to\omega}^{[i]}$ is just the top $d_y \times d_u$ block of the matrix $G_{\mathrm{ex}\to v}^{[i]}$. In particular, $\psi_{\pi_0}(n) = \sum_{i\geq n} \|G_{\mathrm{ex}\to v}^{[i]}\|_{\mathrm{op}} = \psi_{G_{\mathrm{ex}\to v}}(n)$.

Finally, given LDC policies π, we define a *conversion* Markov operator so that the conclusion of Lemma 6.2 holds tautologically:

Definition 6.5. Given nominal policy π_0 and target policy $\pi \in \Pi_{\mathrm{ldc}}$, the $\pi_0 \to \pi$ conversion operator $G_{\pi_0\to\pi}$ is the element of $\mathscr{G}(d_u, d_\omega)$ satisfying

$$\mathbf{u}_t^{\pi} = \mathbf{u}_t^{\pi_0} + \sum_{i=0}^{t-1} G_{\pi_0\to\pi}^{[i]} \cdot \boldsymbol{\omega}_{t-i}^{\pi_0}. \tag{6.45}$$

For the Natures Y parametrization, the conversion operator is $G_{\pi,\mathrm{cl},e\to u}$ given in Eq. (6.14), and for the static feedback parametrization, the operator is $G_{K\to\pi}$, given in Eq. (6.23).

Again, we define the induced classes of LDC policies: f

$$\begin{aligned} \Pi_{\pi_0\to u}[\psi] &:= \{\pi \in \Pi_{\mathrm{ldc}} : \psi_{G_{\pi_0\to\pi}}(n) \leq \psi(n)\} \\ \Pi_{\pi_0\to u}(c,\rho) &:= \{\pi \in \Pi_{\mathrm{ldc}} : \psi_{G_{\pi_0\to\pi}}(n) \leq c\rho^n\} \end{aligned} \tag{6.46}$$

Theorem 6.1c. *Suppose Assumptions 6.1 and 6.2 hold. Define* $R_{G_\star} := \psi_{\pi_0}(0)$. *Given a proper decay function* ψ, *a policy* $\pi \in \Pi_{\pi_0 \to u}[\psi]$, *and* $R_{\mathcal{M}} \geq \psi(0)$ *the following holds for all integers* $m \geq 1$,

$$\mathbf{J}_T(\pi) - \inf_{M \in \mathcal{M}(m, R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \leq 2LTR_{\mathcal{M}} R_{G_\star}^2 R_{\mathrm{nat}}^2 \psi(m). \tag{6.47}$$

More concretely, if $\pi \in \Pi_{\pi_0 \to u}(c, \rho)$ *for* $c \geq 1$ *and* $\rho \in [0, 1)$, *then for* $R_{\mathcal{M}} \geq \frac{c}{1-\rho}$,

$$\mathbf{J}_T(\pi) - \inf_{M \in \mathcal{M}(m, R_{\mathcal{M}})} \mathbf{J}_T(\pi^M) \leq c\rho^{-m} \cdot 2LTR_{\mathcal{M}} R_{G_\star}^2 R_{\mathrm{nat}}^2. \tag{6.48}$$

One can verify that Theorem 6.1 and Theorem 6.1b are special cases of the theorem above.

Example 1: IO Parametrization

We can select a stabilizing controller π_0 such that $A_{\pi_0}, B_{\pi_0}, C_{\pi_0}, D_{\pi_0}$ need not be zero, but both the internal controller dynamics, and the closed-loop dynamics are stable. That is, $\rho(A_{\pi_0}) < 1$ and $\rho(A_{\pi_0,\mathrm{cl}}) < 1$. Yet again, we set $\boldsymbol{\omega}_t^{\mathsf{alg}} = \mathbf{y}_t^{\mathsf{alg}}$, corresponding to $C_{\pi_0,\omega} = 0$ and $D_{\pi_0,\omega} = I$. This gives rise to the input-ouput, or IO, parametrization [Zames, 1981, Furieri et al., 2019]. Though more general than static feedback, this parametrization does not stable all posible stabilizable sytems [Halevi, 1994].

We define a closed form expression for the $\pi_0 \to \pi$ operator that arises under internally stable feedback:

Definition 6.6 (Definition C.6 in Simchowitz et al. [2020])**.** Given a nominal controller π_0 given by $(A_{\pi_0}, B_{\pi_0}, C_{\pi_0}, D_{\pi_0})$, and a target controller π given by $(A_\pi, B_\pi, C_\pi, D_\pi)$, and recalling the closed loop matrix $A_{\pi,\mathrm{cl}}$ from Eq. (6.6), define the matrices $A_{\pi_0 \to \pi}, B_{\pi_0 \to \pi}, C_{\pi_0 \to \pi}$ by

$$A_{\pi_0 \to \pi} := \left[\begin{array}{cc|c} & & 0 \\ & & 0 \\ \hline B_{\pi_0} C_\star & 0 & A_{\pi_0} \end{array}\right], \quad B_{\pi_0 \to \pi} := \begin{bmatrix} B_\star D_\pi & -B_\star \\ B_\pi & 0 \\ 0 & 0 \end{bmatrix},$$
$$C_{\pi_0 \to \pi} := \begin{bmatrix} (D_\pi - D_{\pi_0}) C_\star & C_\pi & -C_{\pi_0} \end{bmatrix}$$

and $D_{\pi_0 \to \pi} = \begin{bmatrix} D_\pi & 0 \end{bmatrix}$ Define $\bar{G}_{\pi_0 \to \pi} := \mathsf{Markov}(A_{\pi_0 \to \pi}, B_{\pi_0 \to \pi}, C_{\pi_0 \to \pi}, D_{\pi_0 \to \pi})$, and define:

$$G^{[i]}_{\pi_0, y \to (y,u)} = \mathbb{I}_{i=0} \begin{bmatrix} I \\ D_{\pi_0} \end{bmatrix} + \mathbb{I}_{i \geq 1} \begin{bmatrix} 0 \\ C_{\pi_0} A_{\pi_0}^{i-1} B_{\pi_0} \end{bmatrix}.$$

Finally, we the $\pi_0 \to \pi$ coversion operator takes the following form

$$G^{[i]}_{\pi_0 \to \pi} = \sum_{j=0}^{i} \bar{G}^{[i-j]}_{\pi_0 \to \pi} G^{[j]}_{\pi_0, y \to (y,u)}.$$

Proposition 6.11 (Proposition C.1 in Simchowitz et al. [2020]). *For any stabilizing π and internally stable π_0, the Markov operator $G_{\pi_0\to\pi}$ defined in Definition 6.6 is the convolution of two stable Markov operators, and is a $\pi_0 \to \pi$ conversion operator. That is, for all t, the exogenous inputs*

$$\mathbf{u}_t^{\mathrm{ex}} = \sum_{i=0}^{t-1} \bar{G}^{[i]}_{\pi_0\to\pi} \mathbf{y}^{\mathrm{nat}}_{t-i}.$$

produce the input-output pairs $(\mathbf{y}_t^\pi, \mathbf{u}_t^\pi)$ via

$$\begin{bmatrix} \mathbf{y}_t^\pi \\ \mathbf{u}_t^\pi \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t^{\mathrm{nat}} \\ \mathbf{u}_t^{\mathrm{nat}} \end{bmatrix} + \sum_{i=0}^{t-1} G^{[i]}_{\mathrm{ex}\to v} \mathbf{u}^{\mathrm{ex}}_{t-i}.$$

In particular, for the conversion operator to be a stable Markov operator, we require that both the open-loop dynamics nominal controller dynamics (i.e. with dynamical matrix A_{π_0}), as well as the target dynamics of the closed loop system under π (with dynamical matrix $A_{\pi,\mathrm{cl}}$) are both stable. Unfortunately, not all S&D systems can be stabilized in this fashion [Zames, 1981].

Example 2: Youla Parametrization

As described above, certain pathological systems may not admit any stabilizing controller π_0 which is internally stable. However, all S&D *do* admit stabilizing controllers of the following form:

Definition 6.7 (Exact Youla DRC). Consider a stabilizable and detectable system, and fix matrices L, F that satisfy $\rho(A_\star + B_\star F) < 1$ and $\rho(A_\star + LC_\star) < 1$. Exact Observer Feedback with Exogenous inputs denotes the internal state $\mathring{\mathbf{s}}_t$ via $\tilde{\mathbf{x}}_t \in \mathbb{R}^{d_x}$, and has the dynamics

$$\begin{aligned} \tilde{\mathbf{x}}_{t+1} &= (A_\star + LC_\star)\tilde{\mathbf{x}}_t - L\mathbf{y}_t + B_\star \mathbf{u}_t^{\mathsf{alg}} \\ \boldsymbol{\omega}_t &= C_\star \tilde{\mathbf{x}}_t - \mathbf{y}_t^{\mathsf{alg}}, \quad \mathbf{u}_t^{\mathsf{alg}} = \mathbf{u}_t^{\mathrm{ex}} + F\tilde{\mathbf{x}}_t, \end{aligned}$$

with $\tilde{\mathbf{x}}_1 = 0$. This yields an LDC-ex $d_\omega = d_y$, with $A_{\pi_0} = (A_\star + LC_\star + B_\star F)$, $B_{\pi_0} = -L$, $C_{\pi_0} = F$, $D_{\pi_0} = 0$, $C_{\pi_0,\omega} = C_\star$, and $D_{\pi_0,\omega} = -I$.

Note that the optimal LQG controller is an observer-feedback controller. However, for this parametrization, we don't need to know this optimal LQG controller. Rather, *any* observer-feedback controller will suffice.

Lemma 6.12 (Lemma C.2 Simchowitz et al. [2020]). *Under Definition 6.7, following identities hold:*

1. *$G^{[i]}_{\mathrm{ex}\to\omega} = 0$ for all $i > 0$. In other words, $\boldsymbol{\omega}_t^{\mathsf{alg}} = \boldsymbol{\omega}_t = \boldsymbol{\omega}_t^{\pi_0}$ for all t, regardless of exogenous inputs.*

2. *Letting $G_{(w,e)\to\omega}$ denote the Markov operator giving the response from $(\mathbf{w}_t, \mathbf{e}_t)$ to $\boldsymbol{\omega}_t$, we have*

$$G^{[i]}_{(w,e)\to\omega} = \mathbb{I}_{i=0}\begin{bmatrix}0 & I_{d_y}\end{bmatrix} + \mathbb{I}_{i>0}C_\star(A_\star + LC_\star)^{i-1}\begin{bmatrix}I_{d_x} & F\end{bmatrix}.$$

3. *We have the identity*

$$G^{[i]}_{\mathrm{ex}\to v} = \mathbb{I}_{i=0}\begin{bmatrix}0\\ I_{d_u}\end{bmatrix} + \mathbb{I}_{i>0}\begin{bmatrix}C_\star\\ F\end{bmatrix}(A_\star + B_\star F)^{i-1}B_\star.$$

4. *Letting $G_{(w,e)\to v}$ denote the Markov operator giving the response from $(\mathbf{w}_t, \mathbf{e}_t)$ to $\mathbf{v}_t^{\pi_0}$, we have the identity*

$$G^{[i]}_{(w,e)\to v} = \mathbb{I}_{i=0}\begin{bmatrix}0 & I\\ 0 & 0\end{bmatrix} + \mathbb{I}_{i>0}\begin{bmatrix}I & 0\\ 0 & -L\end{bmatrix}\begin{bmatrix}A_\star & B_\star F\\ -LC_\star & A_\star + BF + LC_\star\end{bmatrix}^{i-1}\begin{bmatrix}C_\star & 0\\ 0 & F\end{bmatrix}$$

Moreover, via a change of basis, we can write

$$G^{[i]}_{(w,e)\to v} = \mathbb{I}_{i=0}\begin{bmatrix}0 & I\\ 0 & 0\end{bmatrix} + \mathbb{I}_{i>0}\begin{bmatrix}I & 0\\ -L & -L\end{bmatrix}\begin{bmatrix}A_\star + B_\star F & B_\star F\\ 0 & A_\star + LC_\star\end{bmatrix}^{i-1}\begin{bmatrix}C_\star & -F\\ 0 & F\end{bmatrix}$$

In particular, since $\rho(A_\star+B_\star F), \rho(A_\star+LC_\star) < 1$ hold by assumption due to stabilizability and detectability, all of the above systems are guaranteed to be stable.

The conversion operator is given by the following proposition:

Proposition 6.13 (Proposition C.2 in Simchowitz et al. [2020])**.** *Define the matrices*

$$A_{\mathrm{yla},\pi} := A_{\pi,\mathrm{cl}}, \quad B_{\mathrm{yla},\pi} := \begin{bmatrix}B_\star D_\pi - L\\ B_\star\end{bmatrix}, \quad C_{\mathrm{yla},\pi} = \begin{bmatrix}D_\pi C_\star - F\end{bmatrix}, \quad D_{\mathrm{yla},\pi} = D_\pi$$

Then,

$$G_{\mathrm{yla},\pi} = \mathsf{Markov}(A_{\mathrm{yla},\pi}, B_{\mathrm{yla},\pi}, C_{\mathrm{yla},\pi}, D_{\mathrm{yla},\pi})$$

is a $\pi_0 \to \pi$ conversion operator for the exact Youla DRC *parametrization of Definition 6.7. That is*

$$\mathbf{v}_t^\pi = \mathbf{v}_t^{\pi_0} + \sum_{i=0}^{t-1} G^{[i]}_{\mathrm{ex}\to v}\mathbf{u}^{\mathrm{ex}}_{t-i} \quad \textit{for} \quad \mathbf{u}^{\mathrm{ex}}_s = \sum_{j=0}^{s-1} G^{[i]}_{\mathrm{yla},\pi}\boldsymbol{\omega}^{\pi_0}_{s-i}.$$

The statement of the Youla parametrization is standard, though varies source-to-source. We use the expression in cite Megretski [2004, Theorem 10.1].

Example 3: Certainty-Equivalent Youla

The previously suggested parameterization requires exact specification of the system parameters $(A_\star, B_\star, C_\star)$. However, for an unknown system, one can only hope to estimate parameters approximately. This section details the effects of executing a Youla controller with approximate estimates of the system parameters.

Definition 6.8 (Certainty Equivalent Youla)**.** Given parameter estimates $\hat{A}, \hat{B}, \widehat{C}$, we the following dynamical process produces total inputs $\mathbf{u}_t$ from exogenous inputs $\mathbf{u}_t^{\mathrm{ex}}$:

$$\begin{aligned} \widehat{\mathbf{x}}_{t+1} &= (\hat{A} + L\widehat{C})\widehat{\mathbf{x}}_t - L\mathbf{y}_t + \hat{B}\mathbf{u}_t \\ \hat{\boldsymbol{\omega}}_t &= \widehat{C}\widehat{\mathbf{x}}_t - \mathbf{y}_t \\ \mathbf{u}_t &= \mathbf{u}_t^{\mathrm{ex}} + F\widehat{\mathbf{x}}_t. \end{aligned} \tag{6.49}$$

Note that $\hat{\boldsymbol{\omega}}_t$ depends on the history of exogenous inputs $\mathbf{u}_t^{\mathrm{ex}}$. Still, we can give a closed form representation of the overall system dynamics, and the map from exogenous inputs to outputs/controls:

Lemma 6.14. *Set $\boldsymbol{\delta}_t := \widehat{\mathbf{x}}_t - \mathbf{x}_t$ and $\Delta_{\mathrm{youl}} := \hat{A} - A_\star + L(\widehat{C} - C_\star)$. Then, the dynamics induced by Definition 6.8 satisfy that*

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \boldsymbol{\delta}_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} A_\star + B_\star F & B_\star F \\ \Delta_{\mathrm{youl}} & \hat{A} + L\widehat{C} \end{bmatrix}}_{:=A_{\mathrm{cl,in}}} \begin{bmatrix} \mathbf{x}_t \\ \boldsymbol{\delta}_t \end{bmatrix} + \underbrace{\begin{bmatrix} B_\star \\ \hat{B} - B_\star \end{bmatrix}}_{B_{\mathrm{cl,in}}} \mathbf{u}_t^{\mathrm{ex}} + \begin{bmatrix} I & 0 \\ -I & -L \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{e}_t \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbf{y}_t \\ \hat{\boldsymbol{\omega}}_t \\ \mathbf{u}_t \end{bmatrix} = \begin{bmatrix} C_\star & 0 \\ \widehat{C} - C_\star & \widehat{C} \\ F & F \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \boldsymbol{\delta}_t \end{bmatrix} + \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} \mathbf{u}_t^{\mathrm{ex}} + \begin{bmatrix} I \\ -I \\ 0 \end{bmatrix} \mathbf{e}_t.$$

Denoting by $\hat{G}_{\mathrm{in}}$ the Markov operator describing the map from $\mathbf{u}_t^{\mathrm{ex}} \to (\mathbf{y}_t, \mathbf{u}_t)$, we then have the identity that

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{u}_t \end{bmatrix} = \begin{bmatrix} \mathbf{y}_t^{\mathrm{nat}} \\ \mathbf{u}_t^{\mathrm{nat}} \end{bmatrix} + \sum_{i=0}^{t-1} \hat{G}_{\mathrm{in}}^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}}.$$

Proof. Let's change variables.

$$\begin{aligned} \mathbf{x}_{t+1} &= A_\star \mathbf{x}_t + B_\star F\widehat{\mathbf{x}}_t + B_\star \mathbf{u}_t^{\mathrm{ex}} + \mathbf{w}_t \\ &= (A_\star + B_\star F)\mathbf{x}_t + B_\star F\boldsymbol{\delta}_t + B_\star \mathbf{u}_t^{\mathrm{ex}} + \mathbf{w}_t \\ \widehat{\mathbf{x}}_{t+1} &= (\hat{A} + L\widehat{C})\widehat{\mathbf{x}}_t - L\mathbf{y}_t + \hat{B}F\widehat{\mathbf{x}}_t + \hat{B}\mathbf{u}_t^{\mathrm{ex}} \\ &= (\hat{A} + \hat{B}F)\widehat{\mathbf{x}}_t + L(\widehat{C}\widehat{\mathbf{x}} - C_\star \mathbf{x}_t) + \hat{B}\mathbf{u}_t^{\mathrm{ex}} - L\mathbf{e}_t \\ \boldsymbol{\delta}_{t+1} = \widehat{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1} &= (\hat{A} + L\widehat{C})\widehat{\mathbf{x}}_t - (A_\star + LC_\star)\mathbf{x}_t - L\mathbf{e}_t - \mathbf{w}_t + (\hat{B} - B_\star)\mathbf{u}_t^{\mathrm{ex}} \\ &= (\hat{A} + L\widehat{C})\boldsymbol{\delta}_t + (\hat{A} - A_\star + L(\widehat{C} - C_\star))\mathbf{x}_t - L\mathbf{e}_t - \mathbf{w}_t + (\hat{B} - B_\star)\mathbf{u}_t^{\mathrm{ex}}. \end{aligned}$$

Once again, changing variables, we have

$$\begin{aligned}\hat{\boldsymbol{\omega}}_t &= \widehat{C}\boldsymbol{\delta}_t + (\widehat{C} - C_\star)\mathbf{x}_t - \mathbf{e}_t \\ \mathbf{u}_t &= \mathbf{u}_t^{\mathrm{ex}} + F\boldsymbol{\delta}_t + F\mathbf{x}_t.\end{aligned}$$

□

The conversion operators can be specified as follows:

Proposition 6.15 (Proposition C.3 in Simchowitz et al. [2020])**.** *Let $G_{\mathrm{yla},\pi}$ be as in Proposition 6.13, and define $\bar{G}_{\mathrm{yla},\pi} \in \mathscr{G}^{d_u+d_y \times d_y}$ via*

$$\bar{G}_{\mathrm{yla},\pi}^{[i]} = \begin{bmatrix} G_{\mathrm{yla},\pi}^{[i]} \\ I_{d_y} \cdot \mathbb{I}_{i=0} \end{bmatrix}.$$

Further, define the operators

$$\begin{aligned} G_{:\to\star} &:= \mathsf{Markov}\left(\begin{bmatrix} A_\star + B_\star F & 0 \\ \hat{B}F - LC_\star & \hat{A} + L\widehat{C} \end{bmatrix}, \begin{bmatrix} B_\star & \hat{B} \\ L & L \end{bmatrix}, \begin{bmatrix} F & -F \end{bmatrix}, \begin{bmatrix} I & 0 \end{bmatrix} \right) \\ G_{\star\to:} &:= \mathsf{Markov}\left(\begin{bmatrix} A_\star + LC_\star & B_\star F - L\widehat{C} \\ 0 & \hat{A} + \hat{B}F \end{bmatrix}, \begin{bmatrix} L \\ L \end{bmatrix}, \begin{bmatrix} C_\star & -\widehat{C} \end{bmatrix}, I \right). \end{aligned}$$

Then, the transfer operator $G_{\pi_0\to\pi} := G_{:\to\star} \odot \bar{G}_{\mathrm{yla},\pi} \odot G_{\star\to:}$ is a $\pi_0 \to \pi$ conversion operator for the Approximate Youla LCD-Ex of Definition 6.8.

Again, since $(A_\star, B_\star, C_\star)$ is assumed to be S&D, there exists an L and F such that $A_\star + B_\star F$ and $A_\star + LC_\star$ are stable, Hence, if $\hat{A}, \hat{B}, \widehat{C}$ are sufficiently close to $(A_\star, B_\star, C_\star)$, the above conversion operators are stable as well.

Algorithms for General Parametrizations: Known System

Again, fix memory length $h \in N$, a control parameter $m \in \mathbb{N}$, and a control radius $R_{\mathcal{M}}$. Now the choice of DRC-parametrized inputs depends on

$$u_t^{\mathrm{ex}}(M) := \sum_{i=0}^{m-1} M^{[i]} \cdot \boldsymbol{\omega}_{t-i}^{\pi_0},$$

Given $M_{t:t-h} \in \mathcal{M}^{h+1}$, we define

$$v_t[M_{t:-h}] = \begin{bmatrix} y_t[M_{t:t-h}] \\ u_t[M_{t:t-h}] \end{bmatrix} = \mathbf{v}_t^{\pi_0} + \sum_{i=0}^{h} G_K^{[i]} \cdot u_t^{\mathrm{ex}}(M_{t-i}),$$

where $G_{\mathrm{ex}\to v}$ is the Markov operator describing the response from exogenous inputs $\mathbf{u}_t^{\mathrm{ex}}$ to $(\mathbf{y}_t, \mathbf{u}_t)$, defined in Eq. (6.20). Finally, we define the $h+1$-ary loss and unary specializtion

$$F_t[M_{t:t-h}] := \ell_t(v_t[M_{t:t-h}]), \quad f_t(M) = F_T[M, M, \ldots, M].$$

Again, when the system is known, the operators $G_{\mathrm{ex}\to v}$ and $G_{\mathrm{ex}\to\omega}$ can be computed directly. Therefore, $\mathbf{v}^{\pi_0}$ and $\boldsymbol{\omega}^{\pi_0}$ can be computed by inverting Eq. (6.44)

$$\boldsymbol{\omega}_t^{\pi_0} = \boldsymbol{\omega}_t^{\mathsf{alg}} - \sum_{i=1}^{t} G_{\mathrm{ex}\to\omega}^{[t-i]} \mathbf{u}_i^{\mathrm{ex}}, \quad \mathbf{v}_t^{\pi_0} = \mathbf{v}_t - \sum_{i=1}^{t} G_{\mathrm{ex}\to v}^{[t-i]} \mathbf{u}_i^{\mathsf{alg}}. \tag{6.50}$$

This gives rise the reduction in Algorithm 6.6. Note that this reduction involves updating the internal state of the $\mathbf{s}_t$ of the nominal control policy (denoted by algorithm comments).

Algorithm 6.6 Nonstochastic Control to OCoM reduction with general parametrization

Intialize: OCoM algorithm $\mathcal{A}$, nominal controller π_0, DRC length m, memory h.
inital DRC controller $\mathbf{M}_1 = 0$, set $\mathcal{M} = \mathcal{M}(R_{\mathcal{M}}, m)$.
Intialize: internal state $\mathbf{s}_1 = 0$ `// controller dynamics`
Precompute: Markov operators G_K
for each $t = 1, 2, \ldots, T$ **do**
  Observe output $\mathbf{y}_t$
  Recover $\boldsymbol{\omega}_t^{\pi_0}$ and $\mathbf{v}_t^{\pi_0}$ via Eq. (6.50).
  Select exogenous input $\mathbf{u}_t^{\mathrm{ex}} = u_t^{\mathrm{ex}}(\mathbf{M}_t)$
  Select total input $\mathbf{u}_t = \mathbf{u}_t^{\mathrm{ex}} + C_{\pi_0}\mathbf{s}_t D_{\pi_0}\mathbf{y}_t$.
  Receive cost $\ell_t(\cdot,\cdot)$, and form function $f_t(\cdot)$ as above.
  Feed f_t to learning algorithm $\mathcal{A}$, and recieve updated parameter $\mathbf{M}_{t+1} \in \mathcal{M}$.
  Update internal state $\mathbf{s}_{t+1} = A_{\pi_0}\mathbf{s}_t + B_{\pi_0}\mathbf{y}_t$. `// controller dynamics`

Algorithms for General Parametrizations: Unknown System

As in Section 6.4, we begin with access to plugin estimates $\widehat{G}_{\mathrm{ex}\to v}$ and $\widehat{G}_{\mathrm{ex}\to\omega}$ of $G_{\mathrm{ex}\to v}$ and $G_{\mathrm{ex}\to\omega}$, respectively:

$$\begin{aligned} \hat{\mathbf{v}}_t^{\pi_0} &:= \begin{bmatrix} \mathbf{y}_t \\ D_{\pi_0}\mathbf{y}_t \end{bmatrix} - \sum_{i=1}^{h} \widehat{G}_{\mathrm{ex}\to v}^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}} \\ \hat{\boldsymbol{\omega}}_t^{\pi_0} &:= \sum_{i=1}^{h} \widehat{G}_{\mathrm{ex}\to\omega}^{[i]} \mathbf{u}_{t-i}^{\mathrm{ex}} \end{aligned} \tag{6.51}$$

To condense notation, we let $\widehat{G} = (\widehat{G}_{\mathrm{ex}\to v}, \widehat{G}_{\mathrm{ex}\to\omega}) \in \mathscr{G}(d_y + d_u + d_\omega, \eta)$ denote the two operators stacked vertically.

Using these estimates, we define an empirical approximation to the DRC input with parameter M, replacing $\mathbf{y}_{1:t}^K$ with $\hat{\mathbf{y}}_{1:t}^K$:

$$u_t^{\mathrm{ex}}(M \mid \hat{\eta}_{1:t}^{\pi_0}) := \sum_{i=0}^{m-1} M^{[i]} \cdot \hat{\eta}_{t-i}^{\pi_0},$$

The remaining quanties are analogous to Section 6.4:

$$v_t[M_{t:t-h} \mid \hat{\eta}_{1:t}^{\pi_0}, \hat{v}_t^{\pi_0}, \widehat{G}] := \mathbf{v}_t^K + \sum_{i=0}^{h} \widehat{G}^{[i]} \cdot u_t^{\mathrm{ex}}(M_{t-i} \mid \hat{\eta}_{1:t-i}^{\pi_0}),$$
$$F_t[M_{t:t-h} \mid \hat{\eta}_{1:t}^{\pi_0}, \hat{v}_t^{\pi_0}, \widehat{G}] := \ell_t(v_t[M_{t:t-h} \mid \hat{\eta}_{1:t}^{\pi_0}, \hat{v}_t^{\pi_0}, \widehat{G}]).$$

The unary specializations are as follows:

$$v_t(M_t \mid \hat{\eta}_{1:t}^{\pi_0}, \hat{v}_t^{\pi_0}, \widehat{G}) = v_t[M, \ldots, M \mid \hat{\eta}_{1:t}^{\pi_0}, \widehat{G}]$$
$$f_t(M \mid \hat{\eta}_{1:t}^{\pi_0}, \hat{v}_t^{\pi_0}, \widehat{G}) = F_t[M_{t:t-h} \mid \hat{\eta}_{1:t}^{\pi_0}, \widehat{G}],$$

and our reduction for unknown systems generalizes as follows:

Algorithm 6.7 Nonstochastic Control to OCOM reduction, estimated system

Intialize: OCOM algorithm $\mathcal{A}$, nominal controller π_0, DRC length m, memory h, estimate $\widehat{G} = (\widehat{G}_{\mathrm{ex}\to v}, \widehat{G}_{\mathrm{ex}\to\omega})$
inital DRC controller $\mathbf{M}_1 = 0$.
Recieve: starting state $\mathbf{s}_1$
for each $t = 1, 2, \ldots, T$ **do**
    Observe output $\mathbf{y}_t$
    Recover approximation $\hat{\mathbf{v}}_t^{\pi_0}$ and $\hat{\boldsymbol{\omega}}^{\pi_0}$ via Eq. (6.51).
    Select exogenous input $\mathbf{u}_t^{\mathrm{ex}} = u_t^{\mathrm{ex}}(\mathbf{M}_t \mid \hat{\boldsymbol{\omega}}_{1:t}^{\pi_0})$
    Select total input $\mathbf{u}_t = \mathbf{u}_t^{\mathrm{ex}} + C_{\pi_0}\mathbf{s}_t D_{\pi_0}\mathbf{y}_t$.
    Receive cost $\ell_t(\cdot,\cdot)$, and form function $\hat{f}_t(\cdot) = f_t(\cdot \mid \hat{\boldsymbol{\omega}}_{1:t}^{\pi_0}, \hat{\mathbf{v}}_t^{\pi_0}, \widehat{G})$ as above.
    Feed f_t to learning algorithm $\mathcal{A}$, and recieve updated parameter $\mathbf{M}_{t+1} \in \mathcal{C}$.
    Update state $\mathbf{s}_{t+1} = A_{\pi_0}\mathbf{s}_t + B_{\pi_0}\mathbf{y}_t$.

Algorithm 6.8 Estimation of $\widehat{G}$

Intialize: nominal controller π_0, estimation length N, memory h:
Initial State: $\mathbf{s}_1 = 0$
for each $t = 1, 2, \ldots, N$ **do**
    Observe output $\mathbf{y}_t, \boldsymbol{\omega}_t$
    Draw $\mathbf{u}_t^{\mathrm{ex}} \overset{\mathrm{unif}}{\sim} \{\frac{-1}{\sqrt{d_u}}, \frac{1}{\sqrt{d_u}}\}^{d_u}$
    Select total input $\mathbf{u}_t = \mathbf{u}_t^{\mathrm{ex}} + C_{\pi_0}\mathbf{s}_t + D_{\pi_0}\mathbf{y}_t$.
Compute least squares estimate

$$\widehat{G}^{[1:h]} \in \sum_{t=1}^{N} \left\| \begin{bmatrix} \mathbf{y}_t \\ \mathbf{u}_t - \mathbf{u}_t^{\mathrm{ex}} \\ \boldsymbol{\omega}_t \end{bmatrix} - \sum_{i=1}^{h} G^{[i]} \cdot \mathbf{u}_{t-i}^{\mathrm{ex}} \right\|_2^2 \tag{6.52}$$

Set $\widehat{G}^{[0]} = \begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix}$ and $\widehat{G}^{[i]} = 0$ for $i > h$.
Return $\widehat{G} = (\widehat{G}_{\mathrm{ex}\to v}, \widehat{G}_{\mathrm{ex}\to\omega})$.

Algorithm 6.9 Full Reduction for Unknown Systems

Intialize: nominal controller π_0, estimation length N, memory h, DRC parameters $m, R_{\mathcal{M}}$, online learning algorithm $\mathcal{A}$
Execute estimation phase Algorithm 6.4 for steps $t = 1, 2, \ldots, N$.
Execute Algorithm 6.3 for remaining times $t = N+1, N+2, \ldots, T$ with online learning algorithm $\mathcal{A}$, initialized with estimate $\widehat{G}$ and starting internal state $\mathbf{s}_{N+1}$ where estimation phase $\mathbf{s}_{N+1}$ left off.

Chapter 7

Fast Rates for Nonstochastic Control

In the previous chapter of this thesis, we introduced the rather general setting of nonstochastic control, and derived two regret bounds: a scaling of $\sqrt{T}$ when the dynamics are known, and a scaling of $T^{2/3}$ when the dynamics are unknown.

The bounds for the online LQR problem presented in Chapter 5 were superior: $\sqrt{T}$-regret for *unknown* dynamics. Of course, if the system is known, then we have *zero* regret, because we can synthesize the optimal control policy in advance[1] In this chapter we ask:

> *Can attain nonstochastic regret bounds which are* ***comparable*** *to those attainable in online LQR?*

More quantitatively, is it possible to attain $\sqrt{T}$-regret scaling for unknown system dynamics? And for known dynamics, is it possible to obtain regret which grows very slowly in the problem horizon?

This chapter provides an affirmative answer to the above question: $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for unknown systems, and $\mathrm{poly}(\log T)$ regret for known systems. We refer to these improved regret bounds as *fast rates*. It does so by implementing the same control-to-online learning algorithmic reductions described in the previous chapter, but two key differences:

- A novel, second-order optimization subroutine we call Semi-Ons.
- A more refined analytic framework we call Oco-with-affine memory.

Importantly this implies that the complexity of online control is, under appropriate assumptions, determined primarily by whether or not the system dynamics are known in advance. The bells-and-whistles present in the nonstochastic setting are immaterial to the regret scaling.

[1] This hides the subtle issue of random-fluctuation in the realized regret due to the noise. However, the expected regret is zero with respect to the stipulated benchmark is 0. In fact, it is slightly negative due to finite horizon effects.

Organization of the Chapter

Section 7.1 details the assumptions adopted to obtain fast rates, introduces our Semi-Ons algorithm, and states its regret guarantees guarantees: $\sqrt{T}$ regret for unknown systens, and polylogarithmic for known. It also sketches the key technical challenges and techniques.

Section 7.2 describes the limitations of prior the Oco-with-memory framework (as introduced in the previous chapter) in the absence of strong convexity. Notably, it establishes a lower bound on a popular analysis technique which controls with-memory regret by bounding total Euclidean movement of the optimization iterates.

Section 7.3 details a more tailored framework-Oco-with-affine-memory (OcoAM), of which it demonstrates that online control is a special case. It then shows that Semi-Ons enjoys logarithmic regret against OcoAM losses, circumventing the limitations of the less-specialized OcoM with memory; these bounds directly translate logarithmic regret for online control via the reduction of Section 6.3. This section exposes a key "input-recoverability property" (Definition 7.3) in the analysis, which admits a Fourier-theoretic interpration (Lemma 7.1).

Section 7.4 provides the algorithm and analysis for unknown systems. The key challenge in obtaining fast rates is demonstrating low sensitivity to estimation error in the Markov operator. For didactic purposes and in the interests of brevity, the exposition focuses on demonstrating the robustness of gradient descent to *strongly convex* functions.

This section pays attention to two key technical ingredients: a more subtle regret decomposition, and

The majority of this chapter focuses on systems which can be stabilized via static feedback, and thus are ammenable to the static-feedback Drc parametrization described in Section 6.2. More general systems pose a challenge, because input-recoverability can falter. Section 7.2 elaborates on this point further. It describes a more benign "semi-adversarial" noise regime which affords fast rates for general Drc parametrizations (those described in Section 6.5). It also pinpoints a special case when input-recoverability *does* hold for a certain Youla (non-static) Drc parametrization.

7.1 Main Results

We consider the nonstochastic control setting, detailed in Section 6.1 in the previous chapter. The reader may wish to consult this section, as well as Section 6.2 for definition of relevant quantities. We begin by stating our assumptions.

Assumptions

Assumptions on the Losses

Recall the subquadratic growth assumption (Assumption 6.1), which states that for $v = (y, u)$ the losses $\ell_t(v)$ satisfy $\ell_t(v) \le L \max\{1, \|v\|^2\}$, and $\|\nabla \ell_t(v)\| \le L\|v\|$. We also assume the

losses are α-strongly convex:

Assumption 7.1. Given a parameter $\alpha > 0$, we assume that $\ell_t(v)$ is twice-continuously differentiable and α-strongly convex. That is, the function $\nabla^2 \ell_t(v) \succeq \alpha I$ uniformly. In addition, we assume ℓ_t are L-subquadratic.

This assumption holds in the LQR setting, because the cost function is a positive-definite quadratic form in the state and input. The differentiability assumptions can be relaxed if desired, but are imposed for simplicity.

Strong convexity suffices for attaining low regret when the dynamics are known. When unknown, we require an additional smoothness assumption:

Assumption 7.2. We assume that the losses ℓ_t are L-smooth. That is, ℓ_t are twice continuously differentiable, and $\|\nabla^2 \ell_t(v)\|_{\mathrm{op}} \le L$ for all $v \in \mathbb{R}^{d_y + d_u}$.

Assumption 7.2 also holds in the LQR setting, because quadratic functions have globally bounded second derivative. For simplicity, we assume the subquadratic parameter and smoothness parameters are the same number L (this can be enforced by enlarging L if necessary). This is motivated by the fact that quadratic costs $\ell_t(v) = v^\top Q v$ are both L-smooth and L-subquadratic for the choice $L = \lambda_{\max}(Q)$.

Assumptions on the parametrization

The more restrictive assumption in this chapter is that we can stabilize the system with static feedback. We refer the reader back to Section 6.2 for a refresher on the DRC parametrization, and in particular, DRC with static feedback.

Assumption 7.3. We assume that we can apply the static feedback DRC parameterization outlined in Section 6.2. Specifically, we have access to a static feedback matrix $K \in \mathbb{R}^{d_y \times d_u}$ such that the feedback law $\mathbf{y}_t = K \mathbf{u}_t$ stabilizes the system.

As discussed in the previous chapter, Assumption 7.3 is not restrictive for fully observed systems, or more generally when $\mathbf{y}_t = C\mathbf{x}_t + \mathbf{e}_t$ where $d_y = d_x$ and C is invertible, but is restrictive in general. Unfortunately, as noted in the previous chapter, many partially observed systems of interest cannot be stabilized by static feedback.

Our reliance on static feedback is due to a certain technical condition which requires the Markov operator induced by the parametrization to be "invertible" in a certain sense (Definition 7.3). The condition has a Fourier-theoretic interpretation that the Z-transform of the Markov operator is well-conditioned on the unit circle (Lemma 7.1), which can be verified for the Markov operators G_K which arise from static feedback (Lemma 7.2).

Our results extend to any of the more general parametrizations of Section 6.5 which share this property, though for brevity, we restrict our exposition to static feedback. For example, one can show that this property holds for Youla-parametrized DRC

(Definition 6.7one of the examples sketched in), provide that the eigenvalues of the true system matrix $A_\star$ do not lie on the unit circle (Lemma 7.16).

Under a slightly different noise model, it is possible to obtain fast rates for *all* parametrizations described in Section 6.5, which are general enough to stabilize any stabilizable and detectable system. We call this noise model *semi-adversarial*, where adversarial disturbances are perturbed by a small amount of i.i.d. stochastic noise. This setting is inspired by the smoothed analysis paradigm which popular in the theoretical computer science community Spielman [2005]. We discuss this noise model further and its implications for online control in Section 7.5.

We stress that while the main results in this chapter require access to a fixed stabilizing controller, ***our regret benchmark still competes with arbitrary, dynamical LDC control policies.***

The Semi-Ons algorithm

To take advantage of strongly convex losses, we apply the OCoM-to-Drc reductions detailed in the previous chapter:Algorithm 6.2 for known systems, and Algorithm 6.5 for known. Whereas Chapter 6 instantiates those reduces with online gradient descent, this chapter instantiates both with novel second-order online learning algorithm tailored to our setting.

A key observation is that the losses which arise in the aforementioned reduction may be expressed compactly in the following form

$$f_t(z) := \ell_t(\mathbf{v}_t + \mathbf{H}_t z), \tag{7.1}$$

where ℓ_t are convex costs satisfying Assumption 7.1, $z \in \mathbb{R}^d$ is a convex parameter lying in a constraint set $\mathcal{C}$, $\mathbf{v}_t \in \mathbb{R}^{\mathsf{d_v}}$ is an affine offset term, and $\mathbf{H}_t \in \mathbb{R}^{\mathsf{d_v} \times d}$ is a matrix. Formal derivation of of the structure (7.1) is described by Definition 7.2 for known systems, and at the start of Section 7.4 for unknown.

Importantly, even if $\ell_t(\cdot)$ is strongly convex, the losses in Eq. (7.1) need not be. Moreover, as described in Section 7.2, common generalizations of strong convexity such as exp-concavity do not directly translate to fast regret rates when system dynamics are taken into account.

Our proposed algorithm, Online Semi-Newton Step (Semi-Ons), aims to adress these limitations. Inspired by the landmark Online Newton Step (Ons) algorithm introduced by Hazan et al. [2007], Semi-Ons executes Newton-like updates using a preconditioner based on a running covariance of the matrices $\mathbf{H}_t$:

$$\mathbf{\Lambda}_t = \lambda I + \sum_{s=1}^{t} \mathbf{H}_t^\top \mathbf{H}_t.$$

In contrast, the classic Ons algorithm preconditions based on a running sum of outer products gradients of the losses $\nabla f_t(z)$. Since the gradients of losses of the form (7.1) lie in

Algorithm 7.1 Online Semi-Newton Step (Semi-Ons)

Input: Step size η, regularization parameter $\lambda > 0$, domain $\mathcal{C} \subset \mathbb{R}^d$, arbitrary initial iterate $\mathbf{z}_1 \in \mathcal{C}$
Initialize: $\Lambda_0 = \lambda \cdot I_d$, $\mathbf{z}_1 \leftarrow \mathbf{0}_d$
for each $t = 1, 2, \ldots, T$ **do**
    Learner recieves f_t of the form $f_t(z) = \ell_t(\mathbf{v}_t + \mathbf{H}_t z)$,
    $\nabla_t \leftarrow \nabla f_t(\mathbf{z}_t)$, where $f_t(z) = \ell_t(\mathbf{v}_t + \mathbf{H}_t z)$.
    $\Lambda_t \leftarrow \Lambda_{t-1} + \mathbf{H}_t^\top \mathbf{H}_t$
    $\widetilde{\mathbf{z}}_{t+1} \leftarrow \mathbf{z}_t - \eta \Lambda_t^{-1} \nabla_t$. `// Newton update`
    $\mathbf{z}_{t+1} \leftarrow \arg\min_{z \in \mathcal{C}} \|\Lambda_t^{1/2}(z - \widetilde{\mathbf{z}}_{t+1})\|_2^2$. `// projection`

the rowspace of $\mathbf{H}_t$, it holds that $\mathbf{H}_t^\top \mathbf{H}_t$ roughly dominates $\nabla f_t(z) \nabla f_t(z)^\top$ in a PSD sense. This domination means that Semi-Ons has a considerably larger preconditioner (again, in the PSD sense) than online Newton. The larger preconditioner forces the algorithm take smaller steps which, when translated into online control, helps to achieve low with-memory regret, and low sensivity to estimation error when the system dynamics are unknown.

Guarantees for Drc with Semi-Ons

We now state the key performance guarantees for the Semi-Ons algorithm, in both the known- and unknown-system nonstochastic control setting. These results are instantiated by the analysis in the subsequent sections of this chapter, though proofs are abridged for brevity. Full proofs may be found in Simchowitz [2020].

Our bounds are stated in terms of the subquadratic, strong convexity, and smoothness parameters defined above. In addition, we assume that the static-feedback parametrization enjoys the following properties:

Assumption 7.4. There are parameters $c \geq 1$, $R_{\mathrm{nat}} \geq 1$, and $\rho \in (0, 1)$ such that

- For all times t, the iterates produced by controller K have Euclidean norm $\|\mathbf{v}_t^K\| = \|(\mathbf{y}_t^K, \mathbf{u}_t^K)\|$ uniformly bounded by R_{nat}.
- The decay function of G_K, denoted here as ψ_G, has decay $\psi_G(n) \leq c\rho^n$. In particular $\|G_K\|_{\ell_1,\mathrm{op}} \leq c$.

In addition, we assume that we compete with the benchmark class $\Pi_\star = \Pi_{K \to u}(c, \rho)$; that is, the set of policies π for which the conversion operator $G_{K \to \pi}$ has decay $\psi_{G_{K \to \pi}}(n) \leq c\rho^n$.

We encourage the reader to revisit Section 6.2 for the definitions of relevant terms in Assumption 7.4. The salient point is that, if K is stabilizing (i.e. Assumption 7.3), then there always exists parameters $c, \rho, R_{\mathrm{nat}}$ satisfying the first two items. Moreover, for any stabilizing π, there exists $c \geq 1$ and $\rho \in (0, 1)$ such that $\pi \in \Pi_{K \to u}(c, \rho)$.

We are now prepared to state our bounds. First, we demonstrate polylogarithmic regret for systems with known dynamics:

Theorem 7.1 (Guarantee for Known System)**.** *Suppose Assumptions 7.1, 7.3 and 7.4 holds. Then, for a suitable choice of parameters, the* Semi-Ons *(Algorithm 7.1) implemented as described as in Section 7.3 enjoys the following regret guarantee:*

$$\mathrm{ControlReg}_T(\mathsf{alg}; \Pi_\star) \le \log^4(1+T) \cdot \frac{c_\star^5(1+\|K\|_{\mathrm{op}})^3}{(1-\rho_\star)^5} \cdot d_u d_y R_{\mathrm{nat}}^2 \cdot \frac{L^2}{\alpha}$$

More specifically, we use Semi-Ons in tandem with the Drc-to-Oco-with-memory reduction spelled out in Section 6.3 of the previous chapter. The only subtle point is setting up the correspondence between the losses of the form (7.1) with those that arise in the control reduction. Section 7.3 details this correspondence, and describes the essential elements of the proof. In also provides a generic for Semi-Ons when applied to any sequence losses of the form (7.1), stated in Theorem 7.4.

Analogously, to adress unknown system dynamics, we apply Semi-Ons within the estimate-then-control reduction described in Section 6.4 of the previous chapter. Again, Section 7.4 details this correspondence, and describes the essential elements of the proof. This leads us to the following bound: $\widetilde{\mathcal{O}}(\sqrt{T})$-regret for nonstochastic control, matching the optimal regret $\sqrt{T}$-scaling achievable for online LQR.

Theorem 7.2 (Guarantee for Unknown System)**.** *Suppose Assumptions 7.1 to 7.4 all hold. Then, for a suitable choice of parameters, the* Semi-Ons *combined with an initial exploraton phase, as described as in Section 7.4 enjoys the following regret guarantee with probability* $1-\delta$*:*

$$\textsc{NscReg}_T(\mathsf{alg}; \Pi_\star) \lesssim \sqrt{T}\log^3(1+T)\log\frac{1}{\delta} \cdot \frac{c_\star^8(1+\|K\|_{\mathrm{op}})^5}{(1-\rho_\star)^{10}} \cdot d_y(d_u+d_y)R_{\mathrm{nat}}^5 \cdot \frac{L^2}{\alpha}$$

The proof of this theorem relies on a rather subtle analysis of the sensitivity of Semi-Ons to perturbations in the losses, as well as a more careful regret decomposition. Section 7.4 provides details.

Challenges and Techniques

As described above, the key technical obstacle in providing fast rates is the subtle observaion that

> Even though $\ell_t(\cdot)$ are assumed to be strongly convex, the losses in Eq. (7.1) need not be.

The losses in Eq. (7.1) do exhibit a generalization of strong concavity - exp concavity (Definition 7.1) - which suffices to enjoy logarithmic regret in the standard (i.e., non-dynamic)

OCO problem setting. This does not appear to be sufficient for online control: A lower bound Section 7.2 demonstrates that black-box black-box application of the standard OCO-with-*memory* analysis cannot allow for better that $T^{1/3}$ regret, even when the learner faces simple losses of the form (7.1).

This necessitates a different approproach. We leverage the particular structure of the losses in Eq. (7.1) we called OCOwith *affine* memory. We show that the with-memory and unary losses which arise take the form

$$F_t(z_{t:t-h}) = \ell_t(\mathbf{v}_t + \sum_{i=0}^{h} G^{[i]}\mathbf{Y}_{t-i}z_{t-i}), \quad f_t(z) := \ell_t(\mathbf{v}_t + \mathbf{H}_t z), \quad \mathbf{H}_t = \sum_{i=0}^{h} G^{[i]}\mathbf{Y}_{t-i}.$$

Here the Markov operator G is fixed across times t and, when specialized to control, coincides with G_K. We take advantage of the identity $\mathbf{H}_t = \sum_{i=0}^{h} G^{[i]}\mathbf{Y}_{t-i}$ and the input-recoverability property of G (alluded to above and defined formall in Definition 7.3) to establish a well-conditioned linear bijection between the sequence $(\mathbf{H}_t)$ and $(\mathbf{Y}_t)$. This allows us to demonstrate that the Semi-ONS updates, which depend only on $f_t(\cdot)$, and thus on $\mathbf{H}_t$, nevertheless are slow-moving in a certain geometry induced by the $(\mathbf{Y}_t)$ sequence. This argument is made formal in the analysis of known systems Section 7.3.

For unknown systems, the argument is considerably more subtle, and ommitted in the interest in brevity. To build intuition, we consider what would occur *if* we could assume that the losses f_t were all strongly convex. We then show that OGD would display lower sensitivity to error (quadratic in an error of ϵ rather than linear) than in the non-strongly convex setting of Section 6.4. The extension to the actual losses f_t leverages the same geometric ideas Section 7.3; the full argument can be found in [Simchowitz, 2020, Section 5].

7.2 The Limitations of OCO with Memory

To motivate our algorithm and analysis, we begin by describing the insufficiency of the OCO with memory (OCOM) reduction outlined in the previous chapter, specifically Section 6.3. We focus our discussion on the reduction for known systems.

In Proposition 6.4, we demonstrated that the regret of the reduction in Algorithm 6.2, instantiated by a learning algorithm $\mathcal{A}$, was essentially bounded by

$$\textsc{MemReg}_T(\mathcal{A}; F_{1:T}) = \sum_{t=1}^{T} F_t[\mathbf{M}^{\mathcal{A}}_{t:t-h}] - \inf_{M\in\mathcal{M}} \sum_{t=1}^{T} f_t(M),$$

where $\mathbf{M}^{\mathcal{A}}_t \in \mathcal{M}$ are the iterates produced by the subroutine $\mathcal{A}$, and we had constructed the losses $F_t : \mathcal{M}^{h+1} \to \mathbb{R}$ and $f_t : \mathcal{M} \to \mathbb{R}$ via

$$F_t[M_{t:t-h}] = \ell_t(\mathbf{v}_t + \sum_{i=0}^{h}\sum_{j=0}^{m-1} G^{[i]}_K M^{[j]}_{t-i}\mathbf{y}^K_{t-i-j}), \quad f_t(M) = F_t[M, \dots, M]. \tag{7.2}$$

This movitated us to derive Proposition 6.3, a regret bound for the generic OCOM setting: at each time time, the learner produces an iterate $\mathbf{z}_t \in \mathcal{C}$ for a convex domain $\mathcal{C}$, and suffers cost $F_t[\mathbf{z}_{t:t-h}]$, where $F_t : \mathcal{C}^{h+1} \to \mathbb{R}$[2], and suffer regret

$$\textsc{MemReg}_T = \sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - \inf_{M \in \mathcal{C}} \sum_{t=1}^{T} f_t(z).$$

The analysis in that proposition argued that we decompose

$$\textsc{MemReg}_T = \underbrace{\sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - \sum_{t=1}^{T} f_t(\mathbf{z}_t)}_{(i)} + \underbrace{\sum_{t=1}^{T} f_t(\mathbf{z}_t) - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z)}_{(ii)}, \tag{7.3}$$

where term (ii) is the standard OCO regret (without memory), and (i) represents the relative cost of the losses $F_t[\mathbf{z}_{t:t-h}]$ with memory to the unary losses $f_t(\mathbf{z}_t)$ without. We analyzed online gradient descent (OGD), and leveraged key property we then leverage was that the iterates $\mathbf{z}_t$ were *slow-moving*, that is, $\|\mathbf{z}_t - \mathbf{z}_{t-1}\| \le L_s \eta$, where L_s was a Lipschitz constant, and η was the step-size. This ultimately led to a bound of $(i) \le L_s^2 h T \eta$, which scales as $\sqrt{T}$ for the optimal η selection.

If the losses f_t were α-*strongly convex*, a similar argument can be applied [Anava et al., 2015]. For strongly convex losses, online gradient descent can take much more conservative updates $\mathbf{z}_{t+1} = \mathrm{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta_t \nabla f_t(\mathbf{z}_t))$, where η_t is a time-varying learning rate that decays as $\frac{1}{\alpha t}$. It is well known that this approach yields OCO regret (term (ii)) which scales logarithmically in the time horizon.

In addition this slower learning rate, term (i) can be bounded by roughly

$$L_s^2 h \cdot \sum_{t=1}^{T} \eta_t \lesssim L_s^2 h \cdot \frac{\log T}{\alpha}.$$

Thus, the contribution of the memory term is also logarithmic, and hence so is the total bound on $\textsc{MemReg}_T$.

Unfortunately, the loss functions f_t that arise in our control reduction are *not* strongly convex. Yes, the cost functions ℓ_t are (Assumption 7.1), but the functions which arise from Eq. (7.2) are not. Indeed, for the Hessian of f_t to be full rank, it would requite that the mapping

$$M \mapsto \sum_{i=0}^{h} \sum_{j=0}^{m-1} G_K^{[i]} M^{[j]} \mathbf{y}_{t-i-j}^K$$

be full rank, which may fail in general. One consequence is that we cannot blindly adopt the more conservative $\eta_t = \frac{1}{\alpha t}$ step sizes.

[2]Recall that we use brackets for the $h+1$-ary functions, and parantheses for unary functions

Exp-Concavity and Memory

Though our losses f_t in (7.1) are not strongly convex, they can be shown to satisfy a popular generalization of strong convexity known as exp-concavity.

Definition 7.1 (Exp-Concavity [Hazan et al., 2007])**.** Given $\gamma > 0$, a twice-differentiable convex function f on a convex domain $\mathcal{C}$ is said to be γ-exp-concave if $\gamma \cdot \nabla f(z)(\nabla f(z))^\top \preceq \nabla^2 f(z)$ for all $z \in \mathcal{C}$.

The landmark online Newton step (ONS) algorithm of Hazan et al. [2007] established that it is possible to obtain logarithmic standard OCO regret (term *ii* in Eq. (7.3)) against sequences of exp-concave functions. However, it remains unknown whether the same is true for the OCO-with-memory setting.

Intuitively, obtaining logarithmic regret against exp-concave losses requires an algorithm which adapts to the geometry induced by problem gradients along the iterates: $\nabla_t := \nabla f_t(\mathbf{z}_t)$. Precisely, if the algorithm produces a gradient ∇_t which is roughly orthogonal to the previous ones, such an algorithm needs to be prepared to take an aggressive step in that direction. Unfortunately, these large steps mean large jumps in $\mathbf{z}_{t+1} - \mathbf{z}_t$, which pose a challenge for bounding the movement costs (term i in Eq. (7.3)) that arise in OCO with memory.

A Lower Bound on Euclidean Movement

The analysis from Anava et al. [2015] sketched above bounds the total euclidean movement of the iterates $\mathbf{z}_t$. Specifically, define the euclidean cost of an algorithm,

$$\textsc{EucCost}_T := \sum_{t=1}^{T} \|\mathbf{z}_t - \mathbf{z}_{t-1}\|,$$

and recall the standard OCO regret $\textsc{OcoReg}_T := \sum_{t=1}^{T} f_t(\mathbf{z}_t) - \inf_{z\in\mathcal{C}} \sum_{t=1}^{T} f_t(z)$. For a parameter $\mu > 0$ depending on the Lipschitz parameters and memory length h, the Anava et al. [2015]'s argument described above bounds the μ-regret,

$$\begin{aligned}\mu\text{-}\textsc{Reg}_T &:= \textsc{OcoReg}_T + \mu\textsc{EucCost}_T \\ &= \left(\sum_{t=1}^{T} f_t(\mathbf{z}_t) - \inf_{z\in\mathcal{C}} \sum_{t=1}^{T} f_t(z)\right) + \mu \sum_{t=1}^{T} \|\mathbf{z}_t - \mathbf{z}_{t-1}\|,\end{aligned}$$

In words, μ-$\textsc{Reg}_T$ captures the tradeoff between the attaining low OCO regret, and doing so with iterates which move slowly in the Euclidean distance.

It turns out that *no algorithm* can attain logarithmc μ-$\textsc{Reg}_T$ for the sorts of losses encountered in the DRC reduction, *even* if ℓ_t are strongly convex.

Precisely, let us consider a construction in dimension $d = 1$, domain $\mathcal{C} = [-1, 1]$, $\ell(v) = v^2$, with losses $f_t = \ell(\mathbf{v}_t - \epsilon z) = (\mathbf{v}_t - \epsilon z)^2$, where $\epsilon \in (0, 1]$ is fixed, $\mathbf{v}_t \in \{-1, 1\}$ are chosen by an oblivious adversary[3].

[3]that is, selected in a manner that is not adaptive to the learners chose of parameters

Note that ℓ satisfies Assumption 7.1 with $\alpha = L = 1$. They are also $\mathcal{O}(1)$ exp-concave (Definition 7.1), since both $\|\nabla f_t\| \leq 2\epsilon$ and $\nabla^2 f_t = \epsilon^2$. In particular, f_t are only ϵ^2-strong convex, which is necessary for this construction to evade known upper bounds.

The following lower bound is established in Simchowitz [2020], based off the construction in [Altschuler and Talwar, 2018, Theorem 13]:

Theorem 7.3 (Theorem 2.3 in Simchowitz [2020])**.** *Let $c_1, \dots, c_4$ be constants. For $T \geq 1$ and $\mu \leq c_1 T$, there exists $\epsilon = \epsilon(\mu, T)$ and a joint distribution $\mathcal{D}$ over $\mathbf{v}_1, \dots, \mathbf{v}_T \in \{-1, 1\}^T$ such that any proper learning algorithm (i.e. $\mathbf{z}_t \in \mathcal{C}$ for all t),*

$$\mathbb{E}[\mu\text{-}\textsc{Reg}_T] \geq c_2 (T\mu^2)^{1/3}.$$

In particular, for $\mu = 1$, $\mathbb{E}[\mu\text{-}\textsc{Reg}_T]] \geq c_2 T^{1/3}$, and if $\mathbb{E}[\textsc{OcoReg}_T] \leq R \leq c_3 T$, then, $\mathbb{E}[\text{EucCost}_T] \geq c_4 \sqrt{T/R}$.

Hence, analyses based on Euclidean movement cannot ensure better than $T^{1/3}$ regret in the OcoMsetting, and thus no better than $\Omega(T^{1/3})$ regret for online control of a known system with strongly convex losses. Interestingly, these tradeoffs can be matched by the standard online Newton step algorithm, as verified by Simchowitz [2020, Theorem G.1].

7.3 Fast Rates for Known Systems

In this section, we circumvent the limitations of the OcoM to establish fast rates for regret when the system dynamics are known. We begin by describing the general OcoAM setting, and explain how the losses which arise in the known-system reduction (Algorithm 6.2) arise as a special case. We then turn to the input-recoverability property, which underlies our analysis. We proceed to state, and then prove, a regret guarantee for the general OcoAM setting. Finally, we specialize our findings to nonstochastic control with known system dynamics.

Oco with Affine Memory

The key idea is to introduce a more careful reduction we call Oco with affine memory, or OcoAM. The protocol and notion of regret in OcoAM is identical to that of Oco with memory: one faces a sequences of with-memory losses $F_t[z_{t:t-h}]$ with domain $\mathcal{C}^{h+1}$, and suffers regret compared to optimal performance on unary losses $f_t(z) = F_t[z, \dots, z] : \mathcal{C} \to \mathbb{R}$:

$$\text{MemoryReg}_T = \sum_{t=1}^{T} F_t[z_{t:t-h}] - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z).$$

The key difference is that the losses F_t have a particular form:

$$F_t[z_{t:t-h}] = \ell_t(\mathbf{v}_t + \sum_{i=0}^{h} G^{[i]} \mathbf{Y}_{t-i} z_{t-i}). \tag{7.4}$$

Here, the affine term $\mathbf{v}_t$, the matrix $\mathbf{Y}_t$ and the loss ℓ_t are revealed to the learner at time t. The Markov operator G is known to the learner in advance. Note then that the resulting unary losses exhibit the form of Eq. (7.1):

$$f_t(z) = \ell_t(\mathbf{v}_t + \mathbf{H}_t z_t), \quad \text{where } \mathbf{H}_t := \sum_{i=0}^{h} G^{[i]} \mathbf{Y}_{t-i} \tag{7.5}$$

Notably, DRC-parametrized online control natural embeds into OCOAM:

Definition 7.2 (DRC-to-OCOAM embedding)**.** We represent DRC controllers $M \in \mathcal{M} = \mathcal{M}(m, R_{\mathcal{M}})$ as a vector $z \in \mathbb{R}^d$, where $d = m \cdot d_y \cdot d_u$. We call this representation an *embedding*, let $z = \mathfrak{e}(M)$ denote the embedding of a particular $M \in \mathcal{M}$, and let $\mathcal{C} = \mathfrak{e}(\mathcal{M})$ denote the convex domain under the embedding map. Given the paraters $\mathbf{y}_t^K$, we let $\mathbf{Y}_t$ denote the matrix so that multiplication by $z = \mathfrak{e}(M)$ recover $\mathbf{u}_t^M$:

$$\mathbf{Y}_t z = \sum_{i=0}^{m-1} M^{[i]} \mathbf{y}_{t-i}^K = \mathbf{u}_t^M, \quad z = \mathfrak{e}(M).$$

We let

$$\mathbf{H}_t = \sum_{i=0}^{h} G_K^{[i]} \mathbf{Y}_{t-i}.$$

Thus, $f_t(M)$ and $F_t[M_{t:t-h}]$ in Algorithm 6.2 can be expressed as

$$f_t(z) = \ell_t(\mathbf{v}_t^K + \mathbf{H}_t z), \quad F_t[z_{t:t-h}] = \ell_t(\mathbf{v}_t^K + \sum_{i=0}^{h} G_K^{[i]} \mathbf{Y}_{t-i} s_{t-i})$$

where z and $z_{t:t-h}$ correspond to the embedding of M and $M_{t:t-h}$, respectively.

Hence, our proposed algorithm executes the reduction Algorithm 6.4 with the losses f_t given in Definition 7.2 and Semi-ONS as base learner.

The Input Recoverability Property

The losses f_t fail to be strong convex when the matrix $\mathbf{H}_t$ is rank deficient (or ill-conditioned), and this is possible for adversarial sequences of $\mathbf{y}_t^K$. Luckily, for the DRC parameterization with state feedback, there is a sort of "hidden" strong convexity of which we can take advantage.

Recall $\mathscr{G}(n, m)$ as the set of all Markov operators $G = (G^{[i]})_{i \geq 0}$, where $G^{[i]} \in \mathbb{R}^{n \times m}$.

Definition 7.3. We say that the Markov operator $G = (G^{[i]})_{i\geq 0}$ is κ-*input recoverable* if the following holds for any sequence of inputs $(u_0, u_1, u_2, \dots) \in \mathbb{R}^{d_u}$ normalized so that $\sum_{t=1}^{\infty} \|u_t\|^2 = 1$:

$$\sum_{t=0}^{\infty} \left\| \sum_{i=0}^{t} G^{[i]} u_{t-i} \right\|^2 \geq \kappa.$$

We let $\kappa(G)$ denote the smallest value of $\kappa \in [0, 1]$ such that G is κ-input recoverable.

In other words, the input recoverably of a Markov operator is determined by the the following: Let $(y_0, y_1, y_2, \dots)$ denote the sequence obtained by feeding a sequence $(u_0, u_1, u_2, \dots)$ through the operator G: $y_t = \sum_{i=0}^{t} G^{[i]} u_{t-i}$. Then, $\kappa(G)$ bounds the minimal quotient of the ℓ_2-norm of the sequence of y to that of u:

$$\kappa(G) = \min \left\{ 1, \max_{u_0, u_2, \dots} \frac{\sum_{t\geq 0} \|y_t\|^2}{\sum_{t\geq 0} \|u_t\|^2} \right\} : y_t = \sum_{i=0}^{t} G^{[i]} u_{t-i}.$$

Thus, when $\kappa(G) > 0$, G is nondegenerate in the sense that every non-zero sequence of inputs results in a non-zero sequence of outputs. Input recoverability can be lower bounded via a Fourier-theoretic characterization. To do so, we introduce the notion of the Z-transform:

Definition 7.4. Given $G \in \mathscr{G}(n, m)$, it's Z-transform is the power series $\check{G}(z) :== \sum_{i\geq 0} G^{[i]} z^{-i}$, defined for all complex scalars $z \in \mathbb{C} - \{0\}$.

The input-recovarability parameter of G is then lower bounded by the square of the minimum singular value of itz Z-transform over the complex circle:

Lemma 7.1. *Let $G \in \mathscr{G}(n, m)$ be such that $\|G\|_{\ell_1, \mathrm{op}} < \infty$. Then,*

$$\kappa(G) \geq \min_{z \in C : |z| = 1} \sigma_m(\check{G}(z))^2,$$

where $\sigma_m(\cdot)$ denotes minimal complex singular value - that is, letting H denote Hermitian adjoint,

$$\sigma_m(\check{G}(z))^2 = \sqrt{\lambda_m(\check{G}(z)^{\mathsf{H}} \check{G}(z)}.$$

Proof. The proof relies on two facts: Parseval's identity and the convolution theorem for the Z-transform. Fix $u_0, u_1, \dots$ with $\sum_{n=0}^{\infty} \|u_n\|^2 = 1$, and define a Markov-shaped vector $U = (U^{[i]})$, with $U^{[i]} = u_i$, and its Z-transform $\check{U}(z) := \sum_{i=0}^{n} U^{[i]} z^{-i}$. We have that

$$\sum_{n\geq 0} \left\| \sum_{i=0}^{n} G^{[i]} u_{n-i} \right\|_2^2 = \sum_{n\geq 0} \|(G * U)^{[n]}\|^2$$

where $*$ denotes the convolution operator.

Since $\|G\|_{\ell_1,\mathrm{op}} = \sum_{i\geq 0}\|G^{[i]}\|_{\mathrm{op}} < \infty$, Cauchy-Swchwarz implies that G is square-summable: $\sum_{i\geq 0}\|G^{[i]}\|_{\mathrm{op}}^2 < \infty$. In addition, (u_t) is square summable by assumption. Therefore, we may invoke Parseval's identity, which implies that

$$\sum_{n\geq 0}\left\|(G*U)^{[n]}\right\|_2^2 = \frac{1}{2\pi}\int_0^{2\pi}\|\widetilde{(G*U)}(e^{\iota\theta})\|_2^2\mathrm{d}\theta,$$

where $\widetilde{(G*U)}(z) = \sum_{i\geq 0}(G*U)^{[i]}z^{-i}$ is the Z-transform of $G*U$. Because convolutions become multiplications under the Z-transformation, we have that for the Z-transform of U,

$$\frac{1}{2\pi}\int_0^{2\pi}\|\widetilde{(G*U)}(e^{\iota\theta})\|_2^2\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi}\|\check{G}(e^{\iota\theta})\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta.$$

This establishes the first equality of the claim. For the inequality, we have

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi}\|\check{G}(e^{\iota\theta})\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta &\geq \frac{1}{2\pi}\int_0^{2\pi}\sigma_{\min}(\check{G}(e^{\iota\theta}))^2\|\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta \\ &\geq \min_{z:|z|=1}\sigma_{\min}(\check{G}(z))^2\cdot\frac{1}{2\pi}\int_0^{2\pi}\|\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta.\end{aligned}$$

To conclude, we note that by Parsevals identity, $\frac{1}{2\pi}\int_0^{2\pi}\|\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta. = \sum_{n\geq 0}\|U^{[n]}\| = \sum_{n\geq 0}\|u_n\|^2 = 1$, giving

$$\sum_{n\geq 0}\left\|\sum_{i=0}^{n}G^{[i]}u_{n-i}\right\|_2^2 = \frac{1}{2\pi}\int_0^{2\pi}\|\check{G}(e^{\iota\theta})\check{U}(e^{\iota\theta})\|_2^2\mathrm{d}\theta \geq \min_{z:|z|=1}\sigma_{\min}\check{G}(z)^2.$$

□

This condition may seem quite strong, but holds when $G = G_K$ results from a stabilizing feedback parametrization:

Lemma 7.2. *Let G_K denote a Markov operator that arises from a stabilizing, static-K* DRC *parametrization, described at length in Section 6.2. Then, we have $\kappa(G_K) \geq \frac{1}{4}\min\{1, \|K\|_{\mathrm{op}}^{-2}\}$.*

Proof Sketch. We provide a brief sketch; a full proof can be found in Simchowitz [2020, Appendix D.3]. Using Lemma 7.1,

$$\kappa(G) \geq \min_{z\in\mathcal{C}:|z|=1}\sigma_{\min}(\check{G}(z))^2,$$

where $\check{G}(z)$ is the Z-transform of G, that is $\check{G}(z) = \sum_{i\geq 0}G^{[i]}z^{-i}$. Given the Markov operator G_K which arises from the stabilizing controller parameterization, its Z-transform takes the following form:

$$\check{G}_K(z) = \begin{bmatrix} C_\star\check{A}(z)B_\star \\ I + KC_\star\check{A}(z)B_\star\end{bmatrix},$$

where $\check{A}(z) = (zI - (A_\star + B_\star K C_\star))^{-1}$is the Z-transform of the closed-loop matrix. Using a linear algebraic argument due to Agarwal et al. [2019b], we show that any matrix of the form $\begin{bmatrix} X \\ I + KX \end{bmatrix}$ must have minimal singular value at least $\frac{1}{2}\min\{1, \|K\|_{\mathrm{op}}\}$. The bound follows. □

Unfortunately, it is not clear if just a computation holds for more general controller parametrizations that do not rely on static feedback. A discussion of strong convexity for general parametrizations is deferred to Section 6.2.

Regret bounds for Semi-Ons

We now turn to stating a regret guarantee for the regret of Semi-Ons in a generic OcoAM setting. That is, she is given a known Markov operator G, and at each time t, a triple $(\mathbf{v}_t, \mathbf{Y}_t, \ell_t)$ is revealed to the learner, and she faces with-memory loss $F_t[\cdot]$ of the form (7.4), and executes updates on the unary losses $f_t(\cdot)$ of the form (7.5). The notion of regret is

$$\mathrm{MemoryReg}_T = \sum_{t=1}^{T} F_t[z_{t:t-h}] - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z).$$

Our regret guarantee requires the following bound on relevant parameters:

Definition 7.5 (Bounds on Relevant Parameters)**.** definitiondefnPolPars

- The diameter $D := \max\{\|z - z'\| : z, z' \in \mathcal{C}\}$, Y-radius $R_Y := \max_{t\in[T]} \|\mathbf{Y}_t\|_{\mathrm{op}}$, and $R_{Y,\mathcal{C}} := \max_t \max_{z\in\mathcal{C}} \|\mathbf{Y}_t z\|$.
- We define the radii $R_v := \max_{t\in[T]} \max\{\|\mathbf{v}_t\|_2\}$ and $R_G := \max\{1, \|G\|_{\ell_1,\mathrm{op}}\}$.
- We define the H-radius $R_H = R_G R_Y$, and define the effective Lipschitz constant $L_{\mathrm{eff}} := L \max\{1, R_v + R_G R_{Y,\mathcal{C}}\}$.

Finally, we recall the decay function $\psi_G(n) := \sum_{i\geq n} \|G^{[i]}\|_{\mathrm{op}}$.

Theorem 7.4 (Semi-Ons regret, exact case)**.** *Suppose $\kappa = \kappa(G) > 0$, Assumption 7.1 holds, and consider the running Algorithm 7.1 on the unary losses of the form Eq. (7.5), with parameters $\eta = \frac{1}{\alpha}$, $\lambda := 6hR_Y^2 R_G^2$. Suppose in addition that h is large enough to satisfy $\psi_G(h+1)^2 \leq R_G^2/T$. Then,*

$$\mathrm{MemoryReg}_T \leq 3\alpha h D^2 R_H^2 + \frac{3dh^2 L_{\mathrm{eff}}^2 R_G}{\alpha \kappa^{1/2}} \log(1+T).$$

Proof of Theorem 7.4

Here, we present a proof of Theorem 7.4. Most proofs are ommitted, and a full proof can be found in Section 4 of Simchowitz [2020].

As in the analysis of OCO with memory, we decompose the with-memory regret into standard OCO regret and a term containing the effect of memory:

$$\textsc{MemReg}_T = \underbrace{\sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - f_t(\mathbf{z}_t)}_{(i)} + \underbrace{\sum_{t=1}^{T} f_t(\mathbf{z}_t) - \inf_{z \in \mathcal{C}} \sum_{t=1}^{T} f_t(z)}_{(ii)}, \tag{7.6}$$

Term (ii), the standard OCO regret against the f_t sequence, can be upper bounded by arguments which mirror the online Newtop step algorithm [Hazan et al., 2007]; see Simchowitz [2020, Section 4.1] for proof details and a formal statement.

To bound term (i), the central idea is to replace a bound in terms of Euclidean movement (which is insufficient, in view of Theorem 7.3) with an upper bound in terms of a more geometry away quantity. Specifically, we show that

Lemma 7.3. *For all $t \geq 1$, we have*

$$|F_t[\mathbf{z}_{t:t-h}] - f_t(\mathbf{z}_t)| \leq L_{\text{eff}} R_G \sum_{i=1}^{h} \|\mathbf{Y}_{t-i}(\mathbf{z}_t - \mathbf{z}_{t-i})\|$$

Therefore, by the triangle inequality, rearranging summations, and the assumption $\mathbf{z}_s = \mathbf{z}_1$ for $s \leq 1$,

$$\sum_{t=1}^{T} F_t[\mathbf{z}_{t:t-h}] - f_t(\mathbf{z}_t) \leq h L_{\text{eff}} R_G \sum_{s=1-h}^{T} \sum_{i=1}^{h-1} \|\mathbf{Y}_s(\mathbf{z}_{s+i+1} - \mathbf{z}_{s+i})\| \cdot \mathbb{I}_{1 \leq s+i \leq t-1}.$$

The important point here is that we need only control iterate of movement in the norms weighted by the sequence of matrices $(\mathbf{Y}_t)$. In particular, if the matrices $\mathbf{Y}_t$ tend not be be large in certain directions, then we can tolerate more iterate movement in those directions as well. We remark that Lemma 7.3 is a consequence only of the OCOAM loss structure, and does depend on our selection of Semi-ONS as the learning algorithm.

Our use of Semi-ONS becomes important in the proof of the next lemma. Here, we use the Semi-ONS update rule to upper bound the terms $\|\mathbf{Y}_s(\mathbf{z}_{t+1} - \mathbf{z}_t)\|_2$:

Lemma 7.4. *Adopt the convention $\Lambda_s = \Lambda_1$ for $s \leq 0$. Further, consider $s \leq t$, with $t \geq 1$ and s possibly negative. Then,*

$$\|\mathbf{Y}_s(\mathbf{z}_{t+1} - \mathbf{z}_t)\|_2 \leq \eta L_{\text{eff}} \operatorname{tr}(\mathbf{Y}_s \Lambda_s^{-1} \mathbf{Y}_s)^{1/2} \operatorname{tr}(\mathbf{H}_t^\top \Lambda_t^{-1} \mathbf{H}_t))^{1/2}.$$

Therefore, by Cauchy Schwartz (recalling the shorthand $\nabla_t = \nabla f_t(\mathbf{z}_t)$)

$$\text{MoveDiff}_T \le \eta h^2 L_{\text{eff}} R_G \cdot \sqrt{\underbrace{\sum_{t=1-h}^{T} \text{tr}(\mathbf{Y}_t \Lambda_t^{-1} \mathbf{Y}_t)}_{(iii.a)}} \cdot \sqrt{\underbrace{\sum_{t=1}^{T} \text{tr}(\nabla_t^\top \Lambda_t^{-1} \nabla_t)}_{(iii.b)}} .$$

In words, we pay for the movement of the gradients ∇_t in the norm induced by the preconditioner Λ_t^{-1}, and the norm of the matrices $\mathbf{Y}_t$ in the same norm.

Term $(iii.b)$ is quite standard in the analysis of second order methods, and in fact it appears in the analysis of term (i). It relies on two observations. First, the structure of the OcoAM losses implies that

$$\nabla f_t(z) \nabla f_t(z)^\top \preceq L_{\text{eff}}^2 \mathbf{H}_t^\top \mathbf{H}_t,$$

where L_{eff} was the effective Lipschitz constant defined in Definition 7.5. The second is a popular argument knows as the log-det potential lemma. We state (without proof) a strengthening of a more common statement:

Lemma 7.5 (Log-det potential, Lemma 4.5 in Simchowitz [2020])**.** *Consider a general sequence of matrices $(\widetilde{\Lambda}_t)$ satisfying $\widetilde{\Lambda}_t \succeq c \sum_{t=1}^T \mathbf{H}_t^\top \mathbf{H}_t + \lambda_0$. Then,*

$$\sum_{t=1}^{T} \text{tr}(\mathbf{H}_t \widetilde{\Lambda}_t^{-1} \mathbf{H}_t^\top) \le \frac{d}{c} \log\left(1 + \frac{cTR_H^2}{\lambda_0}\right) \tag{7.7}$$

In particular, the above holds with $\mathbf{H}_t$ replaced by ∇_t, provided the RHS is scaled by a factor of L_{eff}^2.

Thus taking $\widetilde{\Lambda}_t = \Lambda_t = \lambda I + \sum_{t=1}^T \mathbf{H}_t^\top \mathbf{H}_t$ allows us to bound term $(iii.b)$ by a term which grows at most logarithmically in the horizon.

We now turn to term $(iii.a)$, and we adopt an arugment of the same flavor. The problem is that term $(iii.a)$ considers movement of the $\mathbf{Y}_t$-matrices, whereas Eq. (7.7) is consider $\mathbf{H}_t$. Recall the relationship between the two:

$$\mathbf{H}_t = \sum_{i=0}^{h} G^{[i]} \mathbf{Y}_{t-i}.$$

In other words, $\mathbf{H}_t$ arises by convolving $\mathbf{Y}_t$ with the Markov operator G.

Recall the input-recoverability definition (Definition 7.3), which states that convolution with G is well conditioned map. In particular, this implies that sequence of inputs can be recovered uniquely from the sequence of outputs under G. By extension, it should be the case that the sequence $(\mathbf{Y}_t)$ can be recovered by the sequence of convolutions $(\mathbf{H}_t)$ (taking into account finite memory h, among other things). This does indeed turn out to be true, and leads to the following bound:

Lemma 7.6. *Suppose that $\kappa = \kappa(G) > 0$, and define $c_{\psi;t} := 1 \vee \frac{t\psi_G(h+1)^2}{hR_G^2}$. Then, for any $\mathbf{Y}_{1-h}, \mathbf{Y}_{2-h}, \dots, \mathbf{Y}_t$, the matrices $\mathbf{H}_s = \sum_{i=0}^{[h]} G^{[i]}\mathbf{Y}_{s-i}$ satisfy*

$$\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{H}_s \succeq \frac{\kappa}{2} \cdot \left(\sum_{s=1-h}^{t} \mathbf{Y}_s^\top \mathbf{Y}_s \right) - 5hR_H^2 c_{\psi;t} I.$$

In particular, for a sufficiently large λ so as to dominate the remainder term, we can show that

$$\mathbf{\Lambda}_t = \lambda I + \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{H}_s \gtrsim \lambda + \kappa \sum_{s=1}^{t} \mathbf{Y}_s^\top \mathbf{Y}_s.$$

Therefore, we can apply the log-det potential argument, Lemma 7.5 with $\mathbf{Y}$ replacing $\mathbf{H}_t$ to obtain:

$$\sum_{t=1-h}^{T} \operatorname{tr}(\mathbf{Y}_t \Lambda_t^{-1} \mathbf{Y}_t) \lesssim \frac{d}{\kappa} \log\left(1 + \frac{\kappa T R_Y^2}{\lambda}\right).$$

Thus, we have established that both terms (*iii.a*) and (*iii.b*) are at most logarithmic in the problem horizon, completing the challenging portion of the proof. □

Specializing to the control setting

To conclude, we specialize Theorem 7.4 to the control setting. To do, we instantiate the bounds in Definition 7.5 as follows.

Lemma 7.7 (Parameter Bounds)**.** *Consider the embedding outlined in Definition 7.2. Recall $R_{\text{nat}} = \max_t \|\mathbf{v}_t^K\|$ and $\mathcal{M} = (m, R_{\mathcal{M}})$; assume for normalization that $R_{\text{nat}}, R_{\mathcal{M}} \geq 1$, and ψ_G, R. Then, the following bounds hold:*

(a) We have $D \leftarrow \max\{\|z - z'\| : z, z' \in \mathcal{M}_{\mathfrak{e}}\} \leq 2\sqrt{m}R_{\mathcal{M}}$.

(b) We have $R_Y \leftarrow \sqrt{m}R_{\text{nat}}$.

(c) We have $R_{Y,\mathcal{C}} \leftarrow R_{\mathcal{M}}R_{\text{nat}}$.

(d) $R_G \leftarrow R_G$, $\psi_G \leftarrow \psi_G$, and $R_H \leftarrow \sqrt{m}R_G R_{\text{nat}}$

(e) We have $R_v \leftarrow R_{\text{nat}}$, and $L_{\text{eff}} \leq 2LR_G R_{\mathcal{M}} R_{\text{nat}}$.

Moreover, $d = md_u d_\omega$

Proof. We proceed item by item:

(a) We have $D \le 2\max\{\|z\| : z \in \mathcal{M}_{\mathfrak{e}}\}$. For $z = \mathfrak{e}(M)$, have that $\|z\| = \|M\|_{\mathrm{F}} \le \sqrt{m}\|M\|_{\ell_1,\mathrm{op}} \le \sqrt{m}R_{\mathcal{M}}$.

(b) Let $z = \mathfrak{e}(M)$. Then, $\|\mathbf{Y}_t z\| = \|\sum_{i=0}^{m-1} \boldsymbol{\omega}_{t-i}^{\mathrm{nat}} M^{[i]}\| \le R_{\mathrm{nat}}\|M\|_{\ell_1,\mathrm{op}} \le \sqrt{m}R_{\mathrm{nat}}\|M\|_{\mathrm{F}}$. Since $\|z\|_2 = \|M\|_{\mathrm{F}}$, the bound follows.

(c) As argued above, $\|\mathbf{Y}_t z\|_{\mathrm{op}} \le R_{\mathrm{nat}}\|M\|_{\ell_1,\mathrm{op}} \le R_{\mathcal{M}}R_{\mathrm{nat}}$ for $M \in \mathcal{M}$.

Items (d) and (e) follow directly from the assumptions. □

With the above subsitutions, we obtain that

$$\mathrm{MemoryReg}_T \lesssim (\alpha\sqrt{\kappa})^{-1} \max\{m,h\}^3 d_u d_y R_G^3 R_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 L^2 \log(1+T)).$$

Theorem 7.1 then follows by combining the reduction Algorithm 6.4, and taking $\kappa \gtrsim 1/(1+\|K\|_{\mathrm{op}})$ due to Lemma 7.2.

7.4 Fast Rates for Unknown Dynamics

Next, we turn to systems with unknown dynamics. Again, we apply the reduction from the previous chapter, this time for unknown systems (Algorithm 6.5). And, as in the previous section, we instantiate the reduction by invoking Semi-Ons as the learning subroutine. Specifically, we instantiate Algorithm 6.5, feeding in the the losses

$$\hat{f}_t(M) := \ell_t\left(\widehat{\mathbf{v}}_t + \sum_{i=0}^{h} \widehat{G}^{[i]} \sum_{j=0}^{m-1} M^{[j]} \hat{y}_{t-i-j}^K\right).$$

into the Semi-Ons algorithm Algorithm 7.1. Note that these loss functions have the requisite OcoAM form. Specificially, given the embedded $z = \mathfrak{e}(M)$ parameter, we set $\mathbf{Y}_t$ to be matrix so that

$$\mathbf{Y}_t z = \sum_{j=0}^{m-1} M^{[j]} \hat{y}_{t-j}^K.$$

Note that, with this substitution, we can express $\hat{f}_t$ as functions of the parameter $z \in \mathcal{C}$ with

$$\hat{f}_t(z) = \ell_t\left(\widehat{\mathbf{v}}_t + \widehat{\mathbf{H}}_t z\right), \quad \text{where } \widehat{\mathbf{H}}_t = \sum_{i=0}^{h} \widehat{G}^{[i]} \mathbf{Y}_{t-i}.$$

The remainder of the section analysis of the above proposal, providing a proof sketch of Theorem 7.2.

A more careful regret decomposition

To establish a $\sqrt{T}$-regret scaling,we establish an ϵ^2-error sensitivity to ϵ-approximate estimates of the Markov operator G_K. Precisely, letting ϵ_G denote the upper bound on the estimation error $\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}}$ derived in Section 6.4, we show that the regret scales like

$$N + T\epsilon_G^2, \text{ where } \epsilon_G \sim 1/\sqrt{N}.$$

Tuning N $\sqrt{T}$ yields $\sqrt{T}$ regret. Note that, as in Section 6.4, the "regret term" turns out to be second order relative to the sensitivity to error ϵ_G.

Implementation of this argument necessitates a more careful regret decomposition. Recall from the proof of Proposition 6.7 in Section 6.4 the loss functions

$$\begin{aligned}
\hat{F}_t[M_{t:t-h}] &= F_t[M_{t:t-h} \mid \hat{\mathbf{y}}_{1:t}^K, \hat{\mathbf{v}}_t^K, \widehat{G}], \quad \hat{f}_t(M) = \hat{F}_t[M, \dots, M] \\
F_t^\star[M_{t:t-h}] &= F_t[M_{t:t-h} \mid \hat{\mathbf{y}}_{1:t}^K, \mathbf{v}_t^K, G_K], \quad f_t^\star(M) = F_t^\star[M, \dots, M] \\
F_t[M_{t:t-h}] &= F_t[M_{t:t-h} \mid \mathbf{y}_{1:t}^K, \mathbf{v}_t^K, G_K], \quad f_t(M) = F_t[M, \dots, M].
\end{aligned}$$

Again, the "hat" sequence are the losses based on empirical estimates used by the learning algorithm, the "starred" sequence is a finite-memory approximation of the loss the algorithm actual incurs (using the correct Markov parameter and affine term), and the F_t, f_t sequence captures the loss the algorithm *would* incur if it had used the correct values of $\mathbf{y}_{1:t}^K$.

In Eq. (6.37) in the previous chapter, it is shown that

$$\begin{aligned}
\textsc{NscReg}_T(\mathsf{alg}) &\le \mathcal{R}_T(\mathcal{M}) + 4N(R_G^2 + R_{\mathrm{nat}}^2) + 8LTR_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\tfrac{\psi_G(h+1)}{R_G} + \tfrac{\psi(m)}{R_{\mathcal{M}}}\right), \\
&\text{where } \mathcal{R}_T(\mathcal{M}) := \sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \inf_{M \in \mathcal{M}} \sum_{t=N+1}^{T} f_t(M).
\end{aligned} \tag{7.8}$$

An important step in obtaining fast rates is *overparametrization*. The algorithm still updates parameters in the DRC set $\mathcal{M} = \mathcal{M}(m, R_{\mathcal{M}})$, but we introduce for the purpose of analysis a slightly smaller set $\mathcal{M}_0 = \mathcal{M}(m_0, R_0)$, with $m_0 \le m$ and $R_0 \le R_{\mathcal{M}}$. Subsituting in $\mathcal{M}_0$ in for $\mathcal{M}$ in Eq. (7.8) and using $R_{\mathcal{M}} \ge R_0$ gives:

$$\textsc{NscReg}_T(\mathsf{alg}) \le \mathcal{R}_T(\mathcal{M}_0) + 4N(R_G^2 + R_{\mathrm{nat}}^2) + 8LTR_{\mathrm{nat}}^2 R_{\mathcal{M}}^2 R_G^2 \left(\tfrac{\psi_G(h+1)}{R_G} + \tfrac{\psi(m_0)}{R_{\mathcal{M}}}\right) \tag{7.9}$$

We then fix a comparator $M_\star \in \mathcal{M}$, to be chosen carefully, and decompose:

$$\begin{aligned}
\mathcal{R}_T(\mathcal{M}_0) &= \underbrace{\sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \sum_{t=N+1}^{T} f_t^\star(M_\star)}_{(i)} \\
&= \underbrace{\sum_{t=N+1}^{T} f_t(M_\star) - \inf_{M_0 \in \mathcal{M}_0} \sum_{t=N+1}^{T} f_t^\star(M_0)}_{(ii)}
\end{aligned}$$

There are two key differences between this regret decomposition and the one considered in the previous chapter:

- Term (ii): Rather than considering the worst case difference between the starred and unstarred sequence $\sup_{M\in\mathcal{M}} |\sum_{t=N+1}^{T} f_t^\star(M) - f_t(M)|$, we consider the difference between the best in hindsight comparator $M_0 \in \mathcal{M}_0$, and the carefully chosen comparator $M_\star \in \mathcal{M}$. $M_\star$ is chosen as a function of the optimal M_0 so that the performance of $M_\star$ with inputs $\hat{\mathbf{y}}_t^K$ is ϵ_G^2-away from the performance of using controller M_0 with the true inputs $\mathbf{y}_t^K$. We will call this the *comparator error*.
- Term (i): Rather than relating the regret on the "star" sequence to the regret on the "hat" sequence, we establish a bound directly on the "star" sequence. This involves accounting for the error incurred by running an online learning algorithm on the $\hat{f}$ sequence, even though the loss is suffered on the $f^\star$ sequence. We call this term *regret sensitivity.* Importantly, the comparator chosen in $M_\star \in \mathcal{M}$.

We now state upper bounds on terms (i) and (ii), starting with term (ii); proofs are given in the subsequent subsections. Throughout, let $\epsilon_G \geq \sqrt{T}$ be an upper bound on $\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}}$. We also define, for $M \in \mathcal{M}$,

$$\hat{u}_t(M) := \sum_{i=0}^{m-1} M^{[i]} \hat{\mathbf{y}}_t^K,$$

which is the input the learning algorithm selects, based on estimates of $\mathbf{y}_t^K$, with DRC parameter M.

Proposition 7.8. *Suppose the set $\mathcal{M} = \mathcal{M}(m, R)$ over parameterizes $\mathcal{M}_0 = \mathcal{M}(m_0, R_0)$ such that $m \geq 2m_0 + h$ and $R_{\mathcal{M}} \geq 2R_0$. Finally, suppose that $\psi_G(h+1) \leq \epsilon_G^2$, and that, as in Section 6.4, $\epsilon_G \leq \frac{1}{R_{\mathcal{M}}}$. Then,*

$$\sum_{t=N+1}^{T} f_t^\star(M_\star) - f_t(M_0) \leq \inf_{\tau\in(0,1]} \frac{16LR_{\mathcal{M}}^3 R_G^2}{\tau} \cdot T\epsilon_G^2 + L\tau \sum_{t=N+1}^{T} \|\mathbf{u}_t - \hat{u}_t(M_\star)\|^2 + 64LR_G^2 R_{\mathcal{M}}^2 R_{\mathrm{nat}}^2 (m+h).$$

The propositions takes advantage of the overparametrization to select $M_\star$ to capture the first-order effects of the errors between $\hat{\mathbf{y}}_t^K$ and $\mathbf{y}_t^K$, leaving a second-order error $T\epsilon_G^2$ as a remainder. The challenge is that errors $\hat{\mathbf{y}}_t^K - \mathbf{y}_t^K$ depend on feedback from the chosen inputs $\mathbf{u}_t$, chosen as $\mathbf{u}_t = \hat{u}_t(\mathbf{M}_t)$ from changing DRC parameters. This makes it challenging to capture all the first order effects in a single parameter. Thus, we incur a penalty for the *square* differences $\sum_{t=N+1}^{T} \|\mathbf{u}_t - \hat{u}_t(M_\star)\|^2$ between the inputs $\mathbf{u}_t$ selected, and those that would have been chosen with fixed parameter $M_\star$. Our final bound leaves a free parameter $\tau \in (0, 1]$ to trade off between the $T\epsilon_G^2$-sensitivity term and the cumulative movement penalty.

We now turn our attention to the term (i):

Proposition 7.9. *Let $\kappa = \kappa(G_K)$ be the input recoverability parameter of G_K (Definition 7.3), which we know is bounded below by $\frac{1}{4}\min\{1, \|K\|_{\mathrm{op}}^2\}$, in view of Lemma 7.2. Moreover, suppose that ϵ_G is selected so that $\epsilon_G \geq \sqrt{T}$. Then, for an appropriate choice of λ, η, the* Semi-Ons *subroutine enjoys*

$$\sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \sum_{t=N+1}^{T} f_t^\star(M_\star) \leq C\log(T)\frac{d_\omega d_y m^{3/2} L^2 R_G^3 R_{\mathrm{nat}}^3 R_{\mathcal{M}}^2}{\alpha\kappa^{1/2}}\left(T\epsilon_G^2 + h^2(R_G^2 + R_{\mathrm{nat}})\right) - \nu_\star \sum_{t=N+1}^{T} \|\mathbf{u}_t - \hat{u}(M_\star)\|^2,$$

where above, C is a universal constant, and $\nu_\star \gtrsim \alpha\sqrt{\kappa}/\sqrt{m}R_{\mathrm{nat}}$.

The above bound attains the desired $T\epsilon_G^2$ sensitivity, and also contains the negative regret term necessary to offset the movement cost from the previous proposition. We can then bound $\mathcal{R}_T(\mathcal{M}_0)$ by combining the above two propositions, choosing the free paramaeter τ in Proposition 7.8 to cancel out $\nu_\star$ in Proposition 7.9. [4] Theorem 7.2 follows by combining the above two propositions with the unknown-system reduction given by Eq. (7.9), using the arguments in that proposition to bound the magnitude of ϵ_G in terms of the sample size N. See Simchowitz [2020, Appendix D.5] for details.

Bounding Regret Sensitivity (Proposition 7.9)

We begin by adressing term (i), the regret sensitivity. To obtain our desired regret bound, we must establish that an error of $\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}} \leq \epsilon_G$ translates into a regret of $T\epsilon_G^2$.

The proof of this sensitivity bound for Semi-Ons is quite involved; we state the bound and its setup at the end of the section, along with the necessary specializations to the control setting.

We begin with a statement and proof a simple bound for a toy setup that illustrates the key ideas. The key observation is that we can view the term

$$\sum_{t=N+1}^{T} F_t^\star[\mathbf{M}_{t:t-h}] - \inf_{M\in\mathcal{M}} \sum_{t=N+1}^{T} f_t^\star(M).$$

as if we are running want to attain low regret on a sequence $(F_t^\star, f_t^\star)$, but observe perturbed losses $(\hat{f}_t)$. In particular, we can view the gradient $\nabla\hat{f}_t(M)$ of the "hat" sequence as a corrupted gradient of the gradient $\nabla\hat{f}_t(M)$ of the star sequence; Leveraging smoothness (Assumption 7.2), the difference in the gradients scales linearly in ϵ_G.

Thus, we aim to understand the sensitivity of online learning to ϵ-corruped gradients. For simplicity, we study the robustness of online gradient descent (Algorithm 6.1) when

[4] We use that $\alpha \leq L$ so that our choice of τ lies in $(0, 1]$

the losses $f^\star$ are strongly convex — the extension to general OCOAM losses follows from the same sorts of arguments detailed in the previous section (again, see Simchowitz [2020, Section 5] for details).

It is already known that gradient descent enjoys quadratic error sensitivity in the stochastic optimization setting [Devolder et al., 2014]. The following result extends the guarantee to online learning:

Proposition 7.10 (Robustness of Strongly Convex OGD). *Let $\mathcal{C} \subset \mathbb{R}^d$ be convex with diameter D, and let $(f_t^\star)$ denote a sequence of α-strongly convex and L-Lipschitz functions on $\mathcal{C}$. Consider the gradient update rules $z_{t+1} = \mathrm{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta_{t+1}(\nabla f_t^\star(\mathbf{z}_t) + \boldsymbol{\epsilon}_t))$, where $\boldsymbol{\epsilon}_t$ is an arbitrary error sequence. Then, for step size $\eta_t = \frac{3}{\alpha t}$,*

$$\sum_{t=1}^T f_t^\star(\mathbf{z}_t) - \inf_{z\in\mathcal{C}} \sum_{t=1}^T f_t^\star(z) \le \frac{6L^2}{\alpha}\log(T+1) + \alpha D^2 + \frac{6}{\alpha}\sum_{t=1}^T \|\boldsymbol{\epsilon}_t\|_2^2 .$$

In fact, the following, stronger "negative regret" bound holds for any comparator $z \in \mathcal{C}$:

$$\sum_{t=1}^T f_t^\star(\mathbf{z}_t) - f_t^\star(z) \le \frac{6L^2}{\alpha}\log(T+1) + \alpha D^2 + \frac{6}{\alpha}\sum_{t=1}^T \|\boldsymbol{\epsilon}_t\|_2^2 - \frac{\alpha}{6}\sum_{t=1}^T \|\mathbf{z}_t - z\|_2^2.$$

In particular, we observe that if the errors satisfy $\|\boldsymbol{\epsilon}_t\| \le \epsilon$, the, the regret scales like $\log T + T\epsilon^2$, yielding the desired quadratic error sensitivity. Interestingly, as in the "slow-rate" regime, the regret term ($\log T$) is dominated by the sensitivity term ($T\epsilon^2$) in the natural regime where $\epsilon \sim 1/\sqrt{T}$.

The second bound incorporates a negative term, which penalizes the overall *quadratic* movement of the iterates $\sum_{t=1}^T \|\mathbf{z}_t - z\|_2^2$. This term will prove useful in the following subsection.

Proof of Proposition 7.10. The proof is "by the book", following the standard analysis of online gradient descent for strongly convex losses. Our proof follows the presentation of Hazan [2019]. Let $\hat\nabla_t := \nabla f_t(\mathbf{z}_t) + \boldsymbol{\epsilon}_t$ denote the corrupted gradients, and let $\nabla_t := \nabla f_t(\mathbf{z}_t)$ the uncorrupted. Fix the comparator point $z \in \mathcal{C}$. From [Hazan, 2019, Eq. 3.4], strong convexity of $f_t^\star$ implies that

$$2(f_t^\star(\mathbf{z}_t) - f^\star(z)) \le 2\nabla_t^\top(\mathbf{z}_t - z) - \alpha\|\mathbf{z}_t - z\|_2^2. \tag{7.10}$$

Recall that the corrupted gradient updates are $\mathrm{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta_{t+1}(\nabla f_t^\star(\mathbf{z}_t) + \boldsymbol{\epsilon}_t)) = \mathrm{Proj}_{\mathcal{C}}(\mathbf{z}_t - \eta_{t+1}\hat\nabla_t)$. The Pythagorean Theorem implies

$$\|\mathbf{z}_{t+1} - z\|_2^2 \le \|\mathbf{z} - z - \eta_{t+1}\hat\nabla_t\|_2^2 = \|\mathbf{z}_t - z\|_2^2 + \eta_{t+1}^2\|\hat\nabla_t\|^2 - 2\eta_{t+1}\hat\nabla_t^\top(\mathbf{z}_t - z), \tag{7.11}$$

which can be re-expressed as

$$-2\hat\nabla_t^\top(\mathbf{z}_t - z) \ge \frac{\|\mathbf{z}_{t+1} - z\|_2^2 - \|\mathbf{z}_t - z\|_2^2}{\eta_{t+1}} - \eta_{t+1}\|\hat\nabla_t\|^2 \tag{7.12}$$

Furthermore, using the elementary inequality $ab \le \frac{a^2}{2\tau} + \frac{\tau}{2}b^2$ for any a, b and $\tau > 0$, we have that for any $\tau > 0$

$$\begin{aligned} -\langle \hat{\nabla}_t, \mathbf{z}_t - z\rangle &= -\langle \nabla_t, \mathbf{z}_t - z\rangle - \langle \boldsymbol{\epsilon}_t, \mathbf{z}_t - z\rangle \\ &\le -\langle \nabla_t, \mathbf{z}_t - z\rangle + \frac{\alpha\tau}{2}\|\mathbf{z}_t - z\|_2^2 + \frac{1}{2\alpha\tau}\|\boldsymbol{\epsilon}_t\|_2^2 \end{aligned} \tag{7.13}$$

Combining Equations (7.12) and (7.13), and rearranging,

$$\begin{aligned} 2\nabla_t^\top(\mathbf{z}_t - z) &\le \frac{\|\mathbf{z}_t - z\|^2 - \|\mathbf{z}_{t+1} - z\|^2}{\eta_{t+1}} + \eta_{t+1}\|\hat{\nabla}_t\|^2 + \frac{1}{\tau\alpha}\|\boldsymbol{\epsilon}_t\|_2^2 + \tau\alpha\|\mathbf{z}_t - z\|^2 \\ &\le \frac{\|\mathbf{z}_t - z\|^2 - \|\mathbf{z}_{t+1} - z\|^2}{\eta_{t+1}} + 2\eta_{t+1}L^2 + (2\eta_{t+1} + \frac{1}{\tau\alpha})\|\boldsymbol{\epsilon}_t\|_2^2 + \tau\alpha\|\mathbf{z}_t - z\|^2. \end{aligned}$$

where we used $\|\hat{\nabla}_t\|_2^2 \le 2(\|\nabla f(z_t)\|_2^2 + \|\boldsymbol{\epsilon}_t\|_2^2) \le 2(L_f^2 + \|\boldsymbol{\epsilon}_t\|_2^2)$. Combining with (7.10), we have

$$\begin{aligned} \sum_{t=1}^T f^\star(\mathbf{z}_t) - f^\star(z) \le \frac{1}{2}\sum_{t=1}^T(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - (1-\tau)\alpha)\|\mathbf{z}_t - z\|_2^2 \\ + \sum_{t=1}^T 2\eta_{t+1}L^2 + (\frac{1}{\tau\alpha} + 2\eta_{t+1})\|\boldsymbol{\epsilon}_t\|_2^2 + \frac{\|\mathbf{z}_1\|_2^2}{\eta_1} \end{aligned}$$

Finally, let us set $\eta_t = \frac{3}{\alpha t}$, $\tau = \frac{1}{3}$, and recall $D = \mathrm{Diam}(\mathcal{C})$. Then, we have that

- $(\frac{1}{\tau\alpha} + 2\eta_{t+1}) \le \frac{6}{\alpha}$, and $2\sum_{t=k+1}^T \eta_{t+1}L_{\mathrm{f}}^2 \le 2\cdot 3L_{\mathrm{f}}^2\log(T+1)/\alpha$
- $\frac{\|\mathbf{z}_1\|_2^2}{\eta_1} \le \alpha D^2/3$.
- Finally,

$$\frac{1}{2}\sum_{t=1}^T(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - (1-\tau)\alpha)\|\mathbf{z}_t - z\|_2^2 = \frac{1}{2}\sum_{t=1}^T(\alpha/3 - 2\alpha/3)\|\mathbf{z}_t - z\|_2^2,$$

  which is equal to $\frac{-\alpha}{6}\sum_{t=1}^T\|z_t - z_\star\|^2$.

The bound follows by tying up loose ends. □

Formal Statements for Semi-Ons

We begin by a formal statement of the regret for Semi-Ons in a general setup. Here, we run Semi-Ons on approximate losses of the form

$$\hat{f}_t(z) = \ell_t\left(\widehat{\mathbf{v}}_t + \widehat{\mathbf{H}}_t z\right), \quad \text{where } \widehat{\mathbf{H}}_t = \sum_{i=0}^h \widehat{G}^{[i]}\mathbf{Y}_{t-i},$$

However, we are interested in regret on "exact" losses

$$F^\star[z_{t:t-h}] = \ell_t\left(\mathbf{v}_t + \sum_{i=0}^{h} G^{[i]}\mathbf{Y}_{t-i}z_{t-i}\right)$$
$$f_t^\star(z) = \ell_t\left(\mathbf{v}_t + \widehat{\mathbf{H}}_t z\right), \quad \text{where } \mathbf{H}_t = \sum_{i=0}^{h} G^{[i]}\mathbf{Y}_{t-i}.$$

Under the correspondence $z = \mathfrak{e}(M)$, the above recovers the true losses $F_t^\star, f_t^\star$ in our control regret decomposition with the subsitutions $\mathbf{Y}_t z = \sum_{j=0}^{m-1} M^{[j]}\hat{y}_{t-j}^K$, $\widehat{\mathbf{v}}_t \leftarrow \hat{\mathbf{v}}_t^K$, $\mathbf{v}_t \leftarrow \mathbf{v}_t^K$, $G \leftarrow G_K$, and $\widehat{G} \leftarrow \widehat{G}$. To state the general bound, we introduce the following parameters:

Definition 7.6 (Bounds on Relevant Parameters)**.** We assume $\mathcal{C}$ contains the oirgon.

- The diameter $D := \max\{\|z - z'\| : z, z' \in \mathcal{C}\}$, Y-radius $R_Y := \max_{t\in[T]} \|\mathbf{Y}_t\|_{\mathrm{op}}$, and $R_{Y,\mathcal{C}} := \max_t \max_{z\in\mathcal{C}} \|\mathbf{Y}_t z\|$.
- We define $R_v := \max_{t\in[T]} \max\{\|\mathbf{v}_t\|_2, \|\widehat{\mathbf{v}}_t\|_2\}$, $R_G := \max\{1, \|G\|_{\ell_1,\mathrm{op}}, \|\widehat{G}\|_{\ell_1,\mathrm{op}}\}$.
- Define the the H-radius $R_H = R_G R_Y$, and define the effective Lipschitz constant $L_{\mathrm{eff}} := L\max\{1, R_v + R_G R_{Y,\mathcal{C}}\}$.
- We let ϵ_G be an upper bound so that $\|\widehat{G} - G_\star\|_{\ell_1,\mathrm{op}} \le \epsilon_G$, $\max_{t\ge1} \|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|_2 \le c_v\epsilon_G$ for some $c_v > 0$, and that $\widehat{G}^{[i]} = 0$ for all $i > h$.

Our main guarantee is as follows:

Theorem 7.5 (Theorem 2.2a in Simchowitz [2020])**.** *Suppose that the losses ℓ_t satisfy the strong convexity and moothness assumptions Assumptions 7.1 and 7.2. Then for any comparator $z_\star \in \mathcal{C}$, the following regret bound holds*

$$\sum_{t=1}^{T} F_t^\star(\mathbf{z}_{t:t-h}) - f_t^\star(z_\star) + \nu_\star \sum_{t=1}^{T} \|\mathbf{Y}_t(\mathbf{z}_t - z_\star)\|^2$$
$$\lesssim \log(1+T)\left(\frac{C_1}{\alpha\kappa^{1/2}} + C_2\right)\left(T\epsilon_G^2 + h^2(R_G^2 + R_Y)\right),$$

where $C_1 := (1+R_Y)R_G(h+d)L_{\mathrm{eff}}^2$, $C_2 := (L^2c_v^2/\alpha + \alpha D^2)$, and $\nu_\star = \frac{\alpha\sqrt{\kappa}}{48(1+R_Y)}$.

Note that, in the control setting $\mathbf{Y}_t(z) = \hat{u}_t(M)$ for $z = \mathfrak{e}(M)$. Hence, the negative regret term becomes $\sum_{t=1}^T \|\mathbf{Y}_t(\mathbf{z}_t - z_\star)\|^2 = \sum_{t=1}^T \|\hat{u}(\mathbf{M}_t) - \hat{u}(M_\star)\|^2 = \sum_{t=1}^T \|\mathbf{u}_t - \hat{u}(M_\star)\|^2$. Proposition 7.9 follows by instantiating the above bound for times $t = N+1, N+2, \ldots, T$ with the following lemma:

Lemma 7.11 (Parameter Bounds for Unknown Setting)**.** *Assume $R_{\mathrm{nat}} \geq 1$, and that the assumptions of the previous proposition are satisfied. Then, for $t_0 := N + m + h + 1$, the following hold*

(a) We have $D = \max\{\|z - z'\| : z, z' \in \mathcal{M}_{\mathfrak{e}}\} \leq \sqrt{m} R_{\mathcal{M}}$.

(b) We have $R_Y := \max_{t \geq t_0} \|\mathbf{Y}_t\|_{\mathrm{op}} \leq 2\sqrt{m} R_{\mathrm{nat}}$.

(c) We have $R_{Y,\mathcal{C}} = \max_{t \geq t_0} \max_{z \in \mathcal{C}} \|\mathbf{Y}_t z\| \leq 2 R_{\mathcal{M}} R_{\mathrm{nat}}$.

(d) For $G = G_{\mathrm{ex} \to v}$, we have $R_G = \|\widehat{G}_{\mathrm{ex} \to (y,u)}\|_{\ell_1,\mathrm{op}} \vee \|G_{\mathrm{ex} \to v}\|_{\ell_1,\mathrm{op}} \leq 2R_{\pi_0}$, $\psi_G \leq \psi_{\pi_0}$, and $R_H \leq 2\sqrt{m} R_{\pi_0} R_{\mathrm{nat}}$

(e) We have $R_v := \max_{t \geq t_0} \|\mathbf{v}_t^K\| \vee \|\hat{\mathbf{v}}_t^K\| \leq 2R_{\mathrm{nat}}$, and $L_{\mathrm{eff}} := 8 L R_{\pi_0} R_{\mathcal{M}} R_{\mathrm{nat}}$.

(f) We can take c_v to be $3 R_{\mathcal{M}} R_{\mathrm{nat}}$.

Moreover, $d = d_\omega d_y m$.

Sketch. The proof is analogous to that of Lemma 7.7, but inflating parameters as needed due to approximate quantities, as in Lemma 6.9. □

Bounding comparator error (Proposition 7.8)

Next, we describe how to bound the comparator error

$$(ii) = \inf_{M \in \mathcal{M}} \sum_{t=N+1}^{T} f_t^\star(M) - \inf_{M_0 \in \mathcal{M}_0} \sum_{t=N+1}^{T} f_t(M_0),$$

which accounts for using $\hat{\mathbf{y}}_t^K$ instead of $\mathbf{y}_t^K$ in the DRC parameterization. Going forward, let M_0 witness the infininum of $M_0 \in \mathcal{M}_0$ on the sum on the unstarred sequence $\sum_{t=N+1}^{T} f_t(M_0)$. We shall construct a carefully taylored $M_\star \in \mathcal{M}$ which attains comparable performance; that is, we bound

$$\sum_{t=N+1}^{T} f_t^\star(M_\star) - f_t(M_0).$$

Define the DRC inputs based on the estimated and the true $\hat{\mathbf{y}}^K$'s,

$$\hat{u}_s(M) = \sum_{i=0}^{m-1} M^{[i]} \hat{\mathbf{y}}_{s-i}^K, \quad u_s(M) = \sum_{i=0}^{m-1} M^{[i]} \mathbf{y}_{s-i}^K.$$

The following lemma shows that the comparator error is bounded by the cumulative difference in these inputs:

Lemma 7.12. *For any $N_0 \geq N$*

$$\|\sum_{t=N_0+1}^{T} f_t^\star(M_\star) - f_t(M_0)\| \leq 8LR_G^2 R_{\mathcal{M}} R_{\mathrm{nat}} \sum_{t=N_0+1-h}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\|.$$

Proof. Arguing as in Eq. (6.39), we have

$$\|\sum_{t=N_0+1}^{T} f_t^\star(M_\star) - f_t(M_0)\| \leq 8R_G R_{\mathcal{M}} R_{\mathrm{nat}} L \sum_{t=N+1}^{T} \|v_t^\star(M_\star) - v_t(M_0)\|,$$

where $v_t^\star(M) = \mathbf{v}_t^K + \sum_{i=0}^{h} \hat{u}_{t-i}(M)$ and $v_t(M) = \mathbf{v}_t^K + \sum_{i=0}^{h} G_K^{[i]} u_{t-i}(M)$. The bound follows by summing and using $\|G_K\|_{\ell_1,\mathrm{op}} \leq R_G$. □

The following lemma controls the terms $\sum_{t=N_0+1-h}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\|$:

Lemma 7.13. *Set $N_0 = N + m + 2h \geq N$. Moreover, suppose the set $\mathcal{M} = \mathcal{M}(m, R)$ over parameterizes $\mathcal{M}_0 = \mathcal{M}(m_0, R_0)$ such that $m \geq 2m_0 + h$ and $R_{\mathcal{M}} \geq 2R_0$. Finally fix an error bound ϵ_G such that $\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}} \leq \epsilon_G \leq \frac{1}{R_{\mathcal{M}}}$, and $\psi_G(h+1) \leq \epsilon_G^2$. Then, the following bound holds for any $\tau \leq 1/2R_{\mathrm{nat}}R_{\mathcal{M}}$:*

$$\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \leq \frac{R_{\mathcal{M}}^2 T \epsilon_G^2}{\tau} + \tau \sum_{t=N_0+1}^{T} \|\hat{u}_t(\mathbf{M}_t - M_\star)\|_{\mathrm{F}}^2.$$

We prove Lemma 7.13 just below. The proof carefully expands the erros between $\hat{\mathbf{y}}_t^K$ and $\mathbf{y}_t^K$ in terms of past inputs $\mathbf{u}_s, s \leq t$, which take the form $\mathbf{u}_s = \hat{u}_s(\mathbf{M}_s)$. We then show that, if instead the inputs took the form $\hat{u}_s(M_0)$ for the *fixed* controller $M_0 \in \mathcal{M}_0$, then we can find an overparametrized $M_\star \in \mathcal{M}$ for which $\hat{u}_t(M_\star) \approx u_t(M_0)$ for all t, giving $\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \approx= 0$. Accouting for the real inputs $\mathbf{u}_s = \hat{u}_s(\mathbf{M}_s)$, we suffer the differences between $\mathbf{M}_s - M_0 \approx \mathbf{M}_s - M_\star$. A key subtle point is to maintain the penalty for these differences as arguments of the selected inputs $\hat{u}_t(\cdot)$.

Combining the two lemmas with the reparametrization $\tau \leftarrow \tau/8R_G^2 R_{\mathcal{M}} R_{\mathrm{nat}}$[5] yields

$$\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \leq \inf_{\tau \leq 1} \frac{16LR_{\mathrm{nat}}R_{\mathcal{M}}^3 R_G^2 T \epsilon_G^2}{\tau} + L\tau \sum_{t=N+1}^{T} \|\mathbf{M}_t - M_\star\|_{\mathrm{F}}^2.$$

Using nonnegativity of the losses f_t and using an upper bound on $f_t^\star(M_\star)$ from Lemma 6.9,

$$\sum_{t=N+1}^{T} f_t^\star(M_\star) - f_t(M_0) \leq \inf_{\tau \leq 1} \frac{16LR_{\mathrm{nat}}R_{\mathcal{M}}^3 R_G^2 T \epsilon_G^2}{\tau} + L\tau \sum_{t=N+1}^{T} \|\mathbf{M}_t - M_\star\|_{\mathrm{F}}^2 + 64LR_G^2 R_{\mathcal{M}}^2 R_{\mathrm{nat}}^2 (m+h).$$

This completes the proof of Proposition 7.8.

[5]For which it is enough that the reparametrized τ is less than or equal to one

Proof of Lemma 7.13

We begin by developing

$$\begin{aligned}\hat{u}_t(M_\star) - u_t(M_0) &= \sum_{i=0}^{m-1} M_\star^{[i]} \hat{\mathbf{y}}_{t-i}^K - M_0^{[i]} \mathbf{y}_{t-i}^K \\ &= \sum_{i=0}^{m-1} (M_\star - M_0)^{[i]} \hat{\mathbf{y}}_{t-i}^K + M_0^{[i]} (\hat{\mathbf{y}}_{t-i}^K - \mathbf{y}_{t-i}^K).\end{aligned}$$

Let us star examining the difference between the $\mathbf{y}^K$ and $\hat{\mathbf{y}}^K$ terms. Introducing $\Delta_G = \widehat{G} - G_K$,

$$\begin{aligned}\hat{\mathbf{y}}_t^K - \mathbf{y}_t^K &= \sum_{i=0}^{h} (\widehat{G}^{[i]} - G_K^{[i]}) \mathbf{u}_{t-i} + \sum_{i>h} (\widehat{G}^{[i]} - G_K^{[i]}) \mathbf{u}_t \\ &= \sum_{i=0}^{h} \Delta_G^{[i]} \mathbf{u}_{t-i} \pm \psi_G(h+1) R_u,\end{aligned}$$

where $\pm c$ denotes a term with Euclidean norm bounded by c. Here, we take R_u to be an upper bound on $\max_t \|\mathbf{u}_t\|$ (to be instantiated later), and recall $\psi_G(n) = \sum_{i>n} \|G_K^h\|_{\mathrm{op}}$.

For $t \geq N + m$, $\mathbf{u}_t$ is precisely equal to $\hat{u}_t(\mathbf{M}_t)$, since inputs are selected based on the online learning procedure with parameter $\mathbf{M}_t$. Hence, for $t \geq N + m + h$,

$$\hat{\mathbf{y}}_t^K - \mathbf{y}_t^K = \sum_{i=0}^{h} \Delta_G^{[i]} \hat{u}_t(\mathbf{M}_t) \pm \psi_G(h+1) R_u,$$

and therefore, for $t \geq N + m + 2h$,

$$\hat{u}_t(M_\star) - u_t(M_0) = \sum_{i=0}^{m-1} (M_\star - M_0)^{[i]} \hat{\mathbf{y}}_{t-i}^K + \sum_{i=0}^{m-1} \sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[j]} \hat{u}_{t-i-j}(\mathbf{M}_{t-i-j}) \pm R_{\mathcal{M}} R_u \psi_G(h+1)$$

This expression may seem daunting, but the essential idea is to try to "fold" most of the second term into the first. To do so, let us add and subtract $\hat{u}_t(M_0)$:

$$\begin{aligned}\hat{u}_t(M_\star) - u_t(M_0) = &\sum_{i=0}^{m-1} (M_\star - M_0)^{[i]} \hat{\mathbf{y}}_{t-i}^K + \sum_{i=0}^{m-1} \sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[j]} \hat{u}_{t-i-j}(M_0) \\ &+ \sum_{i=0}^{m-1} \sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[j]} \hat{u}_{t-i-j}(\mathbf{M}_{t-i-j} - M_0) + R_{\mathcal{M}} R_u \psi_G(h+1).\end{aligned}$$

The above display uses the fact that $\hat{u}_t(\cdot)$ is a linear function. The most important prt of the argument is the following lemma:

Lemma 7.14. *Recall the overparametrized* DRC *set $\mathcal{M} = \mathcal{M}(m, R)$, and underparametrized* DRC *set $\mathcal{M}_0 = \mathcal{M}(m_0, R_0)$, such that $m \geq 2m_0 + h$ and $R_{\mathcal{M}} \geq 2R_0$, and $\|\widehat{G} - G_K\|_{\ell_1,\mathrm{op}} \leq \epsilon_G \leq \frac{1}{R_{\mathcal{M}}}$. Then, there is a choice of $M_\star \in \mathcal{M}$ such that*

$$\sum_{i=0}^{m-1} (M_\star - M_0)^{[i]} \hat{\mathbf{y}}^K_{t-i} + \sum_{i=0}^{m-1}\sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[i]} \hat{u}_{t-i-j}(M_0)$$

is identically zero. *Moreover, the choice of $M_\star$ satisfies*

$$\|M_\star - M_0\|_{\ell_1,\mathrm{op}} \leq R_{\mathcal{M}}^2 \epsilon_G + (8LR_G^2 R_{\mathcal{M}} R_{\mathrm{nat}})$$

Proof. Expanding the term $\hat{u}_{t-i-j}(M_0)$, it suffices to find an $M_\star$ such that

$$\sum_{i=0}^{m-1} M_\star^{[i]} \hat{\mathbf{y}}^K_{t-i} = \sum_{i=0}^{m-1} M_0^{[i]} \hat{\mathbf{y}}^K_{t-i} + \sum_{i=0}^{m-1}\sum_{j=0}^{h}\sum_{\ell=0}^{m-1} M_0^{[i]} \Delta_G^{[j]} M_0^{[\ell]} \hat{\mathbf{y}}^K_{t-i-j-\ell}.$$

Since $M_0 \in \mathcal{M}_0 = \mathcal{M}(m_0, R_0)$ is of a smaller length, the i and ℓ indices on the RHS can be taken to run from $0, 1, \ldots, m_0 - 1$. Rearranging that sum, it suffices to find $M_\star$ such that

$$\sum_{i=0}^{m-1} M_\star^{[i]} \hat{\mathbf{y}}^K_{t-i} = \sum_{i=0}^{2m_0+h-1} \left(M_0^{[i]} + \sum_{j_1+j_2+j_3=i} \mathbb{I}_{j_2 \leq h} M_0^{[j_1]} \Delta_G^{[j_2]} M_0^{[j_3]} \right) \hat{\mathbf{y}}^K_{t-i}.$$

Hence, we can simply select

$$M_\star^{[i]} = M_0^{[i]} + \sum_{j_1+j_2+j_3=i} \mathbb{I}_{j_2 \leq h} M_0^{[j_1]} \Delta_G^{[j_2]} M_0^{[j_3]}, i > 0.$$

Then, $M_\star^{[i]} = 0$ for all $i \geq 2m_0 + h$. Moreover,

$$\begin{aligned}\|M_\star - M_0\|_{\ell_1,\mathrm{op}} &= \sum_i \sum_{j_1+j_2+j_3=i} \mathbb{I}_{j_2 \leq h} \|M_0^{[j_1]} \Delta_G^{[j_2]} M_0^{[j_3]}\| \\ &\leq \|M_0\|_{\ell_1,\mathrm{op}}^2 \|\Delta_G\|_{\ell_1,\mathrm{op}} \leq R_0^2 \epsilon_G \leq R_{\mathcal{M}} R_0 \leq R_{\mathcal{M}}^2 \epsilon_G,\end{aligned}$$

where we use that $R_0 \leq R_{\mathcal{M}}$, and $\|M_0\|_{\ell_1,\mathrm{op}} \leq R_0$. In particular, if $\epsilon_G \leq 1/R_{\mathcal{M}}$, then $\|M_\star\|_{\ell_1,\mathrm{op}} \leq R + \|M_\star - M_0\|_{\ell_1,\mathrm{op}} \leq 2R_0$. □

Selecting $M_\star$ as in the above lemma, we find that for $t \geq N + m + 2h$, we find that

$$\begin{aligned}\hat{u}_t(M_\star) - u_t(M_0) &= \sum_{i=0}^{m-1}\sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[j]} \hat{u}_{t-i-j}(\mathbf{M}_{t-i-j} - M_0) \pm R_{\mathcal{M}} R_u \psi_G(h+1) \\ &\overset{(a)}{=} \sum_{i=0}^{m-1}\sum_{j=0}^{h} M_0^{[i]} \Delta_G^{[j]} \hat{u}_{t-i-j}(\mathbf{M}_{t-i-j} - M_\star) \pm (R_{\mathcal{M}}^3 \epsilon_G^2 R_y + R_{\mathcal{M}} R_u \psi_G(h+1)),\end{aligned}$$

where R_y is an upper bound on $\max_t \|\hat{\mathbf{y}}_t^K\|$, and where inequality (a) adds and substracts $\hat{u}_{t-i-j}(M_\star - M_0)$, and invokes Lemma 7.14 to upper bound its contribution by

$$\|M_0\|_{\ell_1,\mathrm{op}} \cdot \|\Delta_G\|_{\ell_1,\mathrm{op}} \cdot R_y \|M_\star - M_0\|_{\ell_1,\mathrm{op}} \le R_{\mathcal{M}}^3 \epsilon_G^2 R_y.$$

Carefully[6] summing the bound over $t = N_0 + 1, N_0 + 2, \ldots, T$ where $N_0 = N + m + 2h \ge N$ yields:

$$\begin{aligned}
&\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \\
&\le \|M_0\|_{\ell_1,\mathrm{op}} \|\Delta_G\|_{\ell_1,\mathrm{op}} \left(\sum_{t=N+1}^{T} \|\hat{u}_t(\mathbf{M}_t - M_\star)\| \right) + T(R_{\mathcal{M}}^3 \epsilon_G^2 R_y + R_{\mathcal{M}} R_u \psi_G(h+1)) \\
&\le R_{\mathcal{M}} \varepsilon_G \left(\sum_{t=N+1}^{T} \|\hat{u}_t(\mathbf{M}_t - M_\star)\| \right) + T(R_{\mathcal{M}}^3 \epsilon_G^2 R_y + R_{\mathcal{M}} R_u \psi_G(h+1)).
\end{aligned}$$

Using the elementary inequality that $xy \le \frac{x^2}{2\tau} + \frac{\tau y^2}{2}$ for any $\tau > 0$, we find

$$\begin{aligned}
\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \le \frac{T R_{\mathcal{M}}^2 \epsilon_G^2}{2\tau} + \tau \sum_{t=N+1}^{T} \|\hat{u}_t(\mathbf{M}_t - M_\star)\|^2 \\
+ T(R_{\mathcal{M}}^3 \epsilon_G^2 R_y + R_{\mathcal{M}} R_u \psi_G(h+1)).
\end{aligned}$$

To conclude, we introduce a number of numerical simplifications: the arguments of Section 6.4 to bound $R_u \le 2R_{\mathrm{nat}} R_{\mathcal{M}}$ and $R_y \le 2R_{\mathrm{nat}}$, invoke the assumption that $\tau \le 1/4 R_Y R_{\mathcal{M}} \le 1$, $\psi_G(h+1) \le \epsilon_G^2$, and $R_{\mathrm{nat}} \ge 1$. This lets us bound

$$\sum_{t=N_0+1}^{T} \|\hat{u}_t(M_\star) - u_t(M_0)\| \le \frac{2R_{\mathcal{M}}^2 T \epsilon_G^2}{\tau} + \tau \sum_{t=N_0+1}^{T} \|\mathbf{M}_t - M_\star\|_{\mathrm{F}}^2.$$

□

7.5 Fast Rates Beyond Static Feedback

It is natural to understand what occurs for the general DRC parametrizations, detailed in Section 6.5, which rely on more than static feedback. In this section, we sketch two avenues to extend our results to that section. The reader may wish to reread Section 6.5 for a review of notation.

[6] The careful summation groups together all appearance of $\mathbf{M}_t$ for a given t in the double sum, avoiding additional factos of m and h and picking up ℓ_1, op-norms of relevant operators.

One possibly is to apply the same Semi-Ons algorithm and analysis, but with the more general Drc parametrization. All that must be checked is that the suitable generalization of the G_K Markov operator - $G_{\mathrm{ex}\to v}$ - also satisfies the input-recoverability property $\kappa(G_{\mathrm{ex}\to v}) > 0$. As shown in Lemma 7.1, it is sufficient that the Z-transform (Definition 7.4) of $G_{\mathrm{ex}\to v}$ be well conditioned on the unit circle. At the end of the chapter, this is verified for the special case of the exact Youla parametrization, when no eignevalues of the dynamic matrix $A_\star$ lie on the complex unit circle.

It is unclear whether the condition of Lemma 7.1 can be verified more generally. However, fast rates *can* be attained for arbitrary systems in a slightly more restrictive noise model: semi-adversarial noise.

Fast Rates under semi-adversarial noise

Here, we (very concisely) describe the semi-stochastic noise model, which permits fast learning rates for general stabilizing Drc parametrizations. Our discussion summarizes Appendices E and F of Simchowitz et al. [2020].

Under semi-adversarial noise, we assume that the noise variables $\mathbf{e}_t$ and $\mathbf{w}_t$ have a stochastic, and nonstochastic component. Previously, all assumptions places on $\mathbf{e}_t$ and $\mathbf{e}_t$ were through the system responses: $\mathbf{y}_t^{\mathrm{nat}}$ for stable sytems, $\mathbf{y}_t^K$ under static feedback, and $\boldsymbol{\omega}_t^{\pi_0}$ under more general parametrizations. Here, the assumption is placed on the noise process itself:

Assumption 7.5. We assume that the noise terms decompose into adversarial and stochastic sequences $\mathbf{w}_t = \mathbf{w}_t^{\mathrm{st}} + \mathbf{w}_t^{\mathrm{st}}$ and $\mathbf{e}_t = \mathbf{e}_t^{\mathrm{st}} + \mathbf{e}_t^{\mathrm{st}}$. We assume that the stochastic noises are $(\mathbf{w}_t^{\mathrm{st}}, \mathbf{e}_t^{\mathrm{st}})$ are independent across t (and of the adversarial components), mean-zero, and satisfy

$$\mathbb{E}\left[\begin{bmatrix}\mathbf{w}_t^{\mathrm{st}}\\ \mathbf{e}_t^{\mathrm{st}}\end{bmatrix}\begin{bmatrix}\mathbf{w}_t^{\mathrm{st}}\\ \mathbf{e}_t^{\mathrm{st}}\end{bmatrix}\right] \succeq \sigma_{\mathrm{st}}^2 I, \quad \forall t.$$

Assumption 7.5 essentially means that the adversary's choice of noise is perturbed by some random and well conditioned component. The independence of the noise can be relaxed considerably (see Simchowitz et al. [2020, Assumption 6b]).

Let $(\mathcal{F}_t)_{t\geq 1}$ denote the filtration generated by the stochastic component of the noise. [7] Then, one can show that one can still achieve fast rates if, for some $k = m + 2h$, it holds that

$$\mathbb{E}[f_t(M) \mid \mathcal{F}_{t-k}] \text{ is } \alpha_f(m,h)\text{-strongly convex}, \tag{7.14}$$

where $f_t(M)$ is the relavant cost function in the Drc-to-Oco-with-memory reduction. It is show in Simchowitz et al. [2020, Appendix E] that this conditional strong convexity suffices

[7] We assume for simplicity that the adversarial components are decided before the game, and thus measurable with respect to $\mathcal{F}_0$.

for fast rates (scaled by $\frac{1}{\alpha_f(m,h)}$), using arguments similar to those earlier in this chapter. As shown in that appendix, thistranslates (roughly) into regret squares of

$$\frac{\text{poly}(\log T)}{\alpha_f(m,h)} \text{for known system dynamics.}$$
$$\sqrt{\frac{T}{\alpha_f(m,h)}} \text{ for unknown dynamics.}$$

To evaluate the strong convex parameter α_f, we require a few preliminaries. Specifically, recall that $G_{\text{ex}\to v}$ denotes the Markov response from exogenous inputs $\mathbf{u}^{\text{ex}}$ to input-outpairs $v = (y, u)$. IIn addition, let $G_{(w,e)\to\omega}$ describe the Markov response from the noise variables $(\mathbf{e}_t, \mathbf{w}_t)$ to the $\boldsymbol{\omega}_t^{\text{nat}}$-variable used for the DRC parametriation. We let $G^\top_{(w,e)\to\omega}$ denote the Markov operator whose terms are the transposes of those in $G_{(w,e)\to\omega}$.

Next, we let $\kappa_s(G)$ idenote input-length analogue of the input-recoverability parameter $\kappa(G)$ defined in Definition 7.3, formally:

$$\kappa_s(G) = \min_{u_0,u_2,\dots} \sum_{t=0}^{\infty} \left\| \sum_{i=0}^{t} G^{[i]} u_{t-i} \right\|^2$$
$$\text{such that } \sum_t \|\mathbf{u}_t\|^2 = 1 \textbf{ and } u_t = 0, \ \forall t > s.$$

In words, $\kappa_s(G)$ gives a lower bound on the minimal response under a Markov parameter G from a finite sequence of implues $u_0, u_1, \dots, u_s$.

With the above definitions in place, the strong convexity can be bounded as follows

Informal Theorem (Informal version of Theorem 11 in Simchowitz et al. [2020])**.** *When ℓ_t are α-strongly convex, we choose lookback $k = m + 2h$, and m and h are sufficiently large (relative to appropriate decay functions), it holds that we may select α_f in Eq. (7.14) as*[8]

$$\alpha_f(m,h) = \frac{1}{2}\alpha \cdot \sigma_{\text{st}}^2 \cdot \kappa_m(G_{\text{ex}\to v}) \cdot \kappa_{m+h}(G^\top_{(w,e)\to\omega}).$$

Essentially, the above result replaces the full notion of input recoverability this this finite-horizon analogues $\kappa_m(\cdot)$ and $\kappa_{m+h}(\cdot)$, at the expense of requiring well-conditioned ($\sigma_{\text{st}}^2 > 0$) semi-stochastic noise. Of course, in our bounds, m and h are selected to be logarithmic in the time horizon T. Thus, to achieve fast rates, it is necessary to understanding how poorly $\kappa_m(G_{\text{ex}\to v})$ and $\kappa_{m+h}(G^\top_{(w,e)\to\omega})$ degrade as a function for m and h.

An important observation which is true of all the example parametrizations of Section 6.5 is that, for $G_{\text{ex}\to v}$, or $G^\top_{(w,e)\to\omega}$ have the form of a Markov operator $G \in \mathscr{G}(n, m)$, where $n \geq m$ (that is, output dimension exceeds input dimension), for which

[8]The version of the theorem stated in Simchowitz et al. [2020] is stated in terms of the Z-transformed definition of κ_s, similar to Lemma 7.1.

- G can be written as the Markov operator of an LTI system of finite order: that is, there exists matrices A, B, C such that $G^{[i]} = CA^{i-1}B$ for all $i \geq 1$.
- The zero-index in the operator is full rank:

$$\sigma_{\min}(G^{[0]}) > 0.$$

Indeed for all parametrizations described in Section 6.5, one can compute a finite order state space represenation of the induced dynamics, and

$$G^{[0]}_{\mathrm{ex}\to v} = \begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix}, \quad \left(G^{[0]}_{(w,e)\to\omega}\right)^\top = \begin{bmatrix} 0 \\ I_{d_y} \end{bmatrix}.$$

The key technical insight of Simchowitz et al. [2020, Appendix F] is that, for such systems, $\kappa_s(G)$ degrades at most polynomially in s:

Proposition 7.15 (Proposition F.1 in Simchowitz et al. [2020])**.** *Let $G \in \mathscr{G}(n, m)$, $n \geq m$ be a Markov operator with $\|G\|_{\ell_1,\mathrm{op}} < \infty$, such that G admits a finite-order state space representation, and and $\sigma_m(G^{[0]}) > 0$. Then, there exist constants c_1 and c_2 dependening only on G such that*

$$\kappa_s(G) \geq c_1 \cdot s^{-c_2}, \quad \forall s > 0.$$

The above proposition is proven by a careful complex-analytic interpolation argument, and the resulting constants c_1 and c_2 that emerge may be quite poor; for example, $\kappa_s(G)$ be degrade exponentially poorly in problem dimension. But, importantly, in terms of dependence on s itself, $\kappa_s(G)$ degrades at most polynomially. In particular,

$$\alpha_f(m, h) = \frac{1}{2}\alpha \cdot \sigma_{\mathrm{st}}^2 \cdot \frac{1}{\mathrm{poly}(m, h)}.$$

Since we end up selecting m and h to be logarithmic in the time horizon T, this degradation of strong convexity suffices for fast rates.

Input-recoverability for exact youla

Lemma 7.16. *Suppose that $A_\star$ has no unit-norm eigenvalues. Then, for the exact Youla* Drc *parametrization detailed in Definition 6.7. Then, for the Markov operator $G_{\mathrm{ex}\to v}$ induced by the general* Drc *parametrization (see Section 6.5), it holds that:*

$$\min_{z\in\mathbb{C}:|z|=1} \sigma_{\min}(\check{G}_{\mathrm{ex}\to v}(z)) > 0$$

Proof. Recall that, in the exact Youla parametrization, there is a feedback matrix F such that $A_\star + B_\star F$ is stable. The Markov operator takes the form

$$G_{\mathrm{ex}\to v} = \mathbb{I}_{i=0}\begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix} + \mathbb{I}_{i>0}\begin{bmatrix} C_\star \\ F \end{bmatrix}(A_\star + B_\star F)^{i-1}B_\star.$$

Hence, the Z-transform can be computed to be

$$\check{G}_{\mathrm{ex}\to v}(z) = \begin{bmatrix} 0 \\ I_{d_u} \end{bmatrix} + \begin{bmatrix} C_\star \\ F \end{bmatrix}\check{A}_{BF}(z)B_\star = \begin{bmatrix} C_\star \check{A}_{BF}(z)B_\star \\ I_{d_u} + F\check{A}_{BF}(z)B_\star \end{bmatrix},$$

where $\check{A}_{BF} = (zI - (A_\star + B_\star F))^{-1}$. In particular,

$$\min_{z\in\mathbb{C}:|z|=1} \sigma_{\min}(\check{G}_{\mathrm{ex}\to v}(z)) \geq \min_{z\in\mathbb{C}:|z|=1} \sigma_{\min}(I_{d_u} + F\check{A}_{BF}B_\star).$$

Define $X(z) := zI - A_\star$. Then, applying the Woodbury identity,

$$\begin{aligned} I_{d_u} + F\check{A}_{BF}(z)B_\star &= I + F(zI - A_\star - B_\star F)^{-1}B_\star = I + F(X(z) - B_\star F)^{-1}B_\star \\ &= I + F(X(z) - B_\star F)^{-1}B_\star \\ &= -((-I) - F(X(z) + B_\star(-I)F)^{-1}B_\star) \\ &= -(-I + FX(z)^{-1}B_\star)^{-1}. \end{aligned}$$

This implies that

$$\sigma_{\min}(I_{d_u} + F\check{A}_{BF}(z)B_\star) = \frac{1}{\|-I + FX(z)^{-1}B_\star\|_{\mathrm{op}}} \leq \frac{1}{1 + \|F\|_{\mathrm{op}}\|B_\star\|_{\mathrm{op}}\|X(z)^{-1}\|_{\mathrm{op}}}.$$

Substituting $X(z) = (zI - A_\star)$, we have

$$\min_{z\in\mathbb{C}:|z|=1} \geq \frac{1}{1 + \|F\|_{\mathrm{op}}\|B_\star\|_{\mathrm{op}} \max_{z\in\mathbb{T}} \|(zI - A_\star)^{-1}\|_{\mathrm{op}}}.$$

In other words, if the eigenvalues of $A_\star$ are bounded away from 1 in magnitude, then

$$\min_{z\in\mathbb{C}:|z|=1} \sigma_{\min}(\check{G}_{\mathrm{ex}\to v}(z)) > 0.$$

□

Concluding Remarks

Chapter 8

Concluding Remarks

This dissertation studies the relative difficulty of linear system identification and adaptive control across range of problem settings and assumptions. We find that certain system properties, e.g. *mixing*, are non-essential for estimation, and that very general adaptive problems, e.g. with arbitrary disturbances and with changing costs, admit adaptive control algorithms whose regret matches the optimal rates in more restrictive settings.

Real-world dynamics, however, are almost always nonlinear. This poses a number of challenges:

- Planning optimal trajectories in nonlinear systems can be computationally prohibitive. In contrast, planning in linear systems often admits convex, computationally efficient formulations.
- In nonlinear systems, dynamical behavior may differ significantly in different regions of the state space. In linear systems, on the other hand, local behavior uniquely determines global behavior.
- Nonlinear systems can exhibit qualitative behavior not observed in linear systems. This is especially true of the non-smooth state transitions induced by contact dynamics.

The author and his collaborators have begun preliminary work carving out nonlinear control formulations which permit rigorous theoretical guarantees. Such settings include LQR with rich nonlinear observations [Mhammedi et al., 2020], and nonlinear receding horizon control with first-order planning oracles [Westenbroek et al., 2021]. We hope the thesis provides helpful scaffolding for future study of more challenging nonlinear control problems to come.

Bibliography

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, and Varun Kanade. Tracking adversarial targets. In *International Conference on Machine Learning*, pages 369–377, 2014.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In *Artificial Intelligence and Statistics*, pages 1246–1254, 2017.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, pages 23–31. PMLR, 2020.

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

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

Jason Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. *arXiv preprint arXiv:1803.01548*, 2018.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning (ICML)*, pages 1747–1754, 2012.

Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

M.C. Campi and Erik Weyer. Finite Sample Properties of System Identification Methods. *IEEE Transactions on Automatic Control*, 47(8), 2002.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Lin Chen, Qian Yu, Hannah Lawrence, and Amin Karbasi. Minimax regret of switching-constrained online convex optimization: No phase transition. *arXiv preprint arXiv:1910.10873*, 2019.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1028–1037, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

Victor de la Pena, Guodong Pang, et al. Exponential inequalities for self-normalized processes with applications. *Electronic Communications in Probability*, 14:372–381, 2009.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator. *arXiv:1710.01688*, 2017.

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

Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic MDP. In *International Conference on Machine Learning*, pages 675–683, 2013.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Feng Ding. Two-stage least squares based iterative estimation algorithm for CARARMA system modeling. *Applied Mathematical Modelling*, 37(7):4798–4808, April 2013. ISSN 0307904X. doi: 10.1016/j.apm.2012.10.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0307904X12006191.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite Time Analysis of Optimal Adaptive Policies for Linear-Quadratic Systems. *arXiv:1711.07230*, 2017.

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

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On optimality of adaptive linear-quadratic regulators. *arXiv preprint arXiv:1806.10749*, 2018c.

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*, pages 1467–1476, 2018.

Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 72–80, 1997.

Dylan J Foster and Max Simchowitz. Logarithmic regret for adversarial online control. *arXiv preprint arXiv:2003.00189*, 2020.

Luca Furieri, Yang Zheng, Antonis Papachristodoulou, and Maryam Kamgarpour. An input–output parametrization of stabilizing controllers: Amidst youla and system level synthesis. *IEEE Control Systems Letters*, 3(4):1014–1019, 2019.

Miguel Galrinho. Least squares methods for system identification of structured models. 2016. URL http://www.diva-portal.org/smash/get/diva2:953835/FULLTEXT01.pdf.

Miguel Galrinho, Cristian Rojas, and Hákan Hjalmarsson. A weighted least-squares method for parameter estimation in structured models. In *53rd IEEE Conference on Decision and Control*, December 2014. doi: 10.1109/CDC.2014.7039903.

Gautam Goel and Babak Hassibi. Regret-optimal control in dynamic environments. *arXiv preprint arXiv:2010.10473*, 2020.

Paul J Goulart, Eric C Kerrigan, and Jan M Maciejowski. Optimization over state feedback policies for robust control with constraints. *Automatica*, 42(4):523–533, 2006.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. *arXiv preprint arXiv:2007.04393*, 2020.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471–1530, 2004.

Lei Guo and Dawei Huang. Least-squares identification for ARMAX models without the positive real condition. *IEEE Transactions on Automatic Control*, 34(10):1094–1098, October 1989. ISSN 0018-9286. doi: 10.1109/9.35285.

Yoram Halevi. Stable lqg controllers. *IEEE Transactions on Automatic Control*, 39(10): 2104–2106, 1994.

Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica: Journal of the Econometric Society*, pages 1269–1286, 1982.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient Descent Learns Linear Dynamical Systems. *arXiv:1609.05191*, 2016.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 6702–6712, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

Elad Hazan, Sham M. Kakade, and Karan Singh. The nonstochastic control problem. *arXiv preprint arXiv:1911.12178*, 2019.

B. L. Ho and R. E. Kalman. Effective construction of linear state-variable models from input/output functions. *Automatisierungs-Technik*, 14(1-12):545–548, 1966.

BL Ho and RE Kalman. Effective construction of linear state-variable models from input/output data. 1965.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

Petros A Ioannou and Jing Sun. *Robust adaptive control*. Courier Corporation, 2012.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Sham Kakade, Mengdi Wang, and Lin F Yang. Variance reduction methods for sublinear reinforcement learning. *arXiv preprint arXiv:1802.09184*, 2018.

Sham M Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London (University of London), 2003.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82.1:35–45, 1960.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR, 2018.

Miroslav Krstic, Petar V Kokotovic, and Ioannis Kanellakopoulos. *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.

Vladimír Kučera. Stability of discrete linear feedback systems. *IFAC Proceedings Volumes*, 8(1):573–578, 1975.

Vitaly Kuznetsov and Mehryar Mohri. Generalization Bounds for Non-Stationary Mixing Processes. *Machine Learning*, 106(1), 2017.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020b.

Yingying Li, Xin Chen, and Na Li. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. In *NeurIPS*, pages 14858–14870, 2019.

Lennart Ljung. *System Identification: Theory for the User*. 1999.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Daniel J. McDonald, Cosma R. Shalizi, and Mark Schervish. Nonparametric Risk Bounds for Time-Series Forecasting. *Journal of Machine Learning Research*, 18, 2017.

Alexander Megretski. Lecture 10: Q-parametrization. *6.245: Multivariable Control Systems*, 2004.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. *Advances in Neural Information Processing Systems*, 33, 2020.

Mehryar Mohri and Afshin Rostamizadeh. Stability Bounds for Non-i.i.d. Processes. In *Neural Information Processing Systems*, 2007a.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher Complexity Bounds for Non-I.I.D. Processes. In *Neural Information Processing Systems*, 2007b.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. 2008.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. *arXiv preprint arXiv:1809.03019*, 2018.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Thrasyvoulos Pappas, Alan Laub, and Nils Sandell. On the numerical solution of the discrete-time algebraic riccati equation. *IEEE Transactions on Automatic Control*, 25(4):631–641, 1980.

Juan C Perdomo, Max Simchowitz, Alekh Agarwal, and Peter Bartlett. Towards a dimension-free understanding of adaptive linear control. *arXiv preprint arXiv:2103.10620*, 2021.

Orestis Plevrakis and Elad Hazan. Geometric exploration for online control. *arXiv preprint arXiv:2010.13178*, 2020.

Jan Willem Polderman. On the necessity of identifying the true parameter in adaptive LQ control. *Systems & control letters*, 8(2):87–91, 1986.

Jan Willem Polderman. Adaptive LQ control: Conflict between identification and control. *Linear algebra and its applications*, 122:219–244, 1989.

S. Joe Qin. An overview of subspace identification. *Computers & Chemical Engineering*, 30 (10-12):1502–1513, September 2006. ISSN 00981354. doi: 10.1016/j.compchemeng.2006.05.045. URL https://linkinghub.elsevier.com/retrieve/pii/S009813540600158X.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Anders Rantzer. Concentration Bounds for Single Parameter Adaptive Control. In *American Control Conference*, 2018.

Paria Rashidinejad, Jiantao Jiao, and Stuart Russell. Slip: Learning to predict in unknown dynamical systems with long-term memory. *arXiv preprint arXiv:2010.05899*, 2020.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019.

Michael Rotkowitz and Sanjay Lall. A characterization of convex problems in decentralized control. *IEEE transactions on Automatic Control*, 50(12):1984–1996, 2005.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Parikshit Shah, Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Linear System Identification via Atomic Norm Regularization. In *Conference on Decision and Control*, 2012.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. Society for Industrial and Applied Mathematics, 2018.

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

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

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*, 2020.

Daniel A Spielman. The smoothed analysis of algorithms. In *International Symposium on Fundamentals of Computation Theory*, pages 17–18. Springer, 2005.

William Spinelli, Luigi Piroddi, and Marco Lovera. On the role of prefiltering in nonlinear system identification. *IEEE Transactions on Automatic Control*, 50(10):1597–1602, October 2005. ISSN 0018-9286. doi: 10.1109/TAC.2005.856655. URL http://ieeexplore.ieee.org/document/1516260/.

Robert F Stengel. *Optimal control and estimation*. Courier Corporation, 1994.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning. 1998.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pages 2627–2636, 2017.

Michel Verhaegen. Subspace model identification part 3. analysis of the ordinary output-error state-space model identification algorithm. *International Journal of Control*, 58(3): 555–586, 1993.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Mats Viberg, Bo Wahlberg, and Björn Ottersten. Analysis of state space system identification methods based on instrumental variables and subspace fitting. *Automatica*, 33(9):1603–1616, 1997.

Mathukumalli Vidyasagar and Rajeeva L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3), 2008.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. *arXiv preprint arXiv:2102.05214*, 2021.

D. Q. Wang. Least squares-based recursive and iterative estimation for output error moving average systems using data filtering. *IET Control Theory Applications*, 5(14):1648–1657, September 2011. ISSN 1751-8644. doi: 10.1049/iet-cta.2010.0416.

Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. A system level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 2019.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial iIntelligence*, pages 538–545. Morgan Kaufmann Publishers Inc., 2001.

Tyler Westenbroek, Max Simchowitz, Michael I Jordan, and S Shankar Sastry. On the stability of nonlinear receding horizon control: a geometric perspective. *arXiv preprint arXiv:2103.15010*, 2021.

Dante Youla, Hamid Jabr, and Jr Bongiorno. Modern wiener-hopf design of optimal controllers–part ii: The multivariable case. *IEEE Transactions on Automatic Control*, 21(3):319–338, 1976.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. The power of predictions in online control. *arXiv preprint arXiv:2006.07569*, 2020.

George Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on automatic control*, 26(2):301–320, 1981.

Yong Zhang. Unbiased identification of a class of multi-input single-output systems with correlated disturbances using bias compensation methods. *Mathematical and Computer Modelling*, 53(9-10):1810–1819, May 2011. ISSN 08957177. doi: 10.1016/j.mcm.2010.12.059. URL https://linkinghub.elsevier.com/retrieve/pii/S0895717711000045.

Wei Xing Zheng. A revisit to least-squares parameter estimation of ARMAX systems. In *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, volume 4, pages 3587–3592 Vol.4, December 2004. doi: 10.1109/CDC.2004.1429269.

Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive lqr with unknown b-matrix. *arXiv preprint arXiv:2011.09288*, 2020.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pages 1583–1591, 2013.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.