Lectures Notes for ESE6180: Learning, Dynamics and Control

Ingvar Ziemann

October 1, 2024

Contents

1. Introduction **4**
- 1.1. Notation . . . 5
- 1.2. Outline . . . 5
  - 1.2.1. Time-Series and Dynamics . . . 6
  - 1.2.2. Learning . . . 6
  - 1.2.3. Control . . . 7
- 1.3. Further Reading . . . 8
- 1.4. Acknowledgements . . . 8

I. Learning **9**

2. Probability and Statistics Preliminaries **10**
- 2.1. Mean Estimation . . . 10
  - 2.1.1. The Central Limit Theorem and the Scale Root-n . . . 11
  - 2.1.2. Markov and Chernoff . . . 11
  - 2.1.3. Further Properties of sub-Gaussian Random Variables . . . 14
- 2.2. A First Look at Random Design Linear Regression . . . 15
  - 2.2.1. The Law of Large Numbers and the Scale n . . . 15
  - 2.2.2. Sub-Exponential Concentration . . . 17
- 2.3. More Concentration Inequalities . . . 21

3. Learning in $\mathbb{R}^d$ **22**
- 3.1. Mean Estimation in $\mathbb{R}^d$ . . . 22
  - 3.1.1. Covering Numbers . . . 23
  - 3.1.2. Controlling the Random Walk in $\mathbb{R}^d$ . . . 25
- 3.2. Random Design Linear Regression in Higher Dimensions . . . 25

4. The Hanson-Wright Inequality: Concentration with Linear Dependence **30**
- 4.1. Proof of The Hanson-Wright Inequality . . . 30
  - 4.1.1. Gaussian Comparison Inequalities for sub-Gaussian Quadratic Forms . . . 31
  - 4.1.2. Finishing the proof of Theorem 4.0.1 . . . 33

5. Linear Regression with Linear Dependence **36**
- 5.1. Instantiating the Hanson-Wright Inequality . . . 37
  - 5.1.1. The Random Walk . . . 37
  - 5.1.2. The Lower Tail . . . 39
  - 5.1.3. Guarantees for Linear Regression with Dependence . . . 41

5.2. Elements of Linear System Identification . . . 41
5.3. Learning from many Trajectories . . . 43
5.4. Notes . . . 44

6. Beyond Stability: The Lower Tail Revisited **45**
6.1. Causal Processes . . . 45
6.1.1. A Decoupling Inequality for sub-Gaussian Quadratic Forms . . . 47
6.1.2. The Lower Tail of the Empirical Covariance of Causal sub-Gaussian Processes 49
6.2. Learning without Stability . . . 52
6.3. Notes . . . 53
6.4. Proof of Theorem 6.1.1 . . . 54

II. Control **57**

7. The Linear Quadratic Regulator **58**
7.1. Dynamic Programming Solution to LQR . . . 58
7.2. Regret . . . 60
7.3. Elements of Linear Control Theory . . . 61
7.4. Deterministic Optimal Control . . . 62
7.4.1. Proof of Theorem 7.3.1 . . . 63
7.5. Notes . . . 65

III. Experiment Design and Statistical Optimality **66**

IV. Further Topics **67**

References **68**

1. Introduction

The use of sequential data in modern machine learning solutions is abundant. Recent success stories range from advances in natural language processing by pre-training large models autoregressively [Brown et al., 2020], walking robots [Yang et al., 2020], to games such as Go and StarCraft [Silver et al., 2017]. Nevertheless, much of our modern theory of supervised learning focuses on learning from independent—and often identically distributed—data streams. Indeed, traditional models of generalization and learnability typically posit that data arrives independently and drawn at random from a fixed distribution. The independence assumption is violated in all of the examples above and it often nontrivial to directly port results from the independent setting to correctly capture the effects of temporal dependency. For instance and by contrast, natural language exhibits strong inter-word dependencies; we predict the next token from the previous context-length many tokens. Similarly, the current position of a robot or state of a game certainly has large bearing on any future position or state.

As optimistic as these developments have been, a host of new challenges present themselves as we proceed to deploy learning algorithms in dynamical systems. The stakes of failure in these emergent applications of learning are typically higher; it does not take much imagination to see that an erroneously learned dynamics or observation model for a self-driving car can result in a crash. These issues are amplified by the fact that we have a relatively poorer theoretical understanding of learning from dependent, temporally correlated, data. Many of the recent advances in high-dimensional probability that have proved immensely useful for learning and statistics chiefly apply to the setting of independent and identically distributed (iid) data.

At their core, these recent successes can in one way or another be regarded as problems of decision-making under uncertainty. Control theory and engineering has long been concerned with exactly this: to use feedback to mitigate dynamic uncertainty. In parallel the related field of system identification has sought to use sampled data from such dynamics to further mitigate this uncertainty. To some extent, over the past few decades or so, these fields have developed in isolation from machine learning, even though they often seek to tackle many of the same problems. In light of these shared ambitions, it is somewhat natural that, historically speaking, the modern incarnations of controls and learning share a common ancestor in Wiener's cybernetics. Bellman's work on dynamic programming is also just as important in controls as it is in reinforcement learning.

This observation—of shared aims and problem formulations—is certainly not novel to these notes. Over the past half-decade or so, a sizeable group of researchers from controls and machine learning have made this observation. This has resulted in a rich body of work offering a fresh perspective on problems classical problems in system identification, reinforcement learning and adaptive controls among others. To date, no cohesive effort has been made to synthesize these developments and make them easily accessible to beginning graduate students. The aim of these notes is to provide a self-contained and streamlined exposition of a select portion of these developments. While we have decided to mainly focus on the settings of linear system identification and learning to control the

linear quadratic regulator, we will also cover certain nonlinear extensions.

It is also worth to point out that the classical literature on system identification has done a formidable job at—often very accurately—characterizing the asymptotic performance of identification algorithms [Ljung, 1999]. Our aim is not to supplant this literature but rather to complement the asymptotic picture with finite sample guarantees by relaying recently developed technical tools drawn from high-dimensional probability, statistics and learning theory [Vershynin, 2018, Wainwright, 2019].

1.1. Notation

For a positive integer $n \in \mathbb{N}$, we define the shorthand $[n] \triangleq \{1, \dots, N\}$. Maxima (resp. minima) of two numbers $a, b \in \mathbb{R}$ are denoted by $a \vee b = \max(a, b)$ ($a \wedge b = \min(a, b)$). For two sequences $\{a_t\}_{t\in\mathbb{N}}$ and $\{b_t\}_{t\in\mathbb{N}}$ we introduce the shorthand $a_t \lesssim b_t$ if there exists a universal constant $C > 0$ and an integer t_0 such that $a_t \leq Cb_t$ for every $t \geq t_0$. If $a_t \lesssim b_t$ and $b_t \lesssim a_t$ we write $a_t \asymp b_t$. When working with sequences in linear spaces, say X, of finite length, say $n \in \mathbb{N}$, it will often be convenient to identify the sequence $\{x_i\}_{i\in[n]}$ with the vector $x_{1:n} \in \mathsf{X}^d$, whose i:th component is $x_i \in \mathsf{X}$.

Differentiation and Integration We use $\mathtt{D}$ for Jacobian, $\mathtt{d}$ for differential and ∇ for the gradient. Expectation (resp. probability) with respect to all the randomness of the underlying probability space is denoted by $\mathbf{E}$ (resp. $\mathbf{P}$).

Linear Algebra The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|_2$, and the unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1}$. The standard inner product on $\mathbb{R}^d$ is denoted $\langle \cdot, \cdot \rangle$. We embed matrices $M \in \mathbb{R}^{d_1\times d_2}$ in Euclidean space by vectorization: $\mathsf{vec}\, M \in \mathbb{R}^{d_1 d_2}$, where vec is the operator that vertically stacks the columns of M (from left to right and from top to bottom). For a matrix M the Euclidean norm is the Frobenius norm, i.e., $\|M\|_F \triangleq \|\mathsf{vec}\, M\|_2$. We similarly define the inner product of two matrices M, N by $\langle M, N \rangle \triangleq \langle \mathsf{vec}\, M, \mathsf{vec}\, N \rangle$. The transpose of a matrix M is denoted by M^T and $\operatorname{tr} M$ denotes its trace. For a matrix $M \in \mathbb{R}^{d_1\times d_2}$, we order its singular values $\sigma_1(M), \dots, \sigma_{d_1\wedge d_2}(M)$ in descending order by magnitude. We also write $\|M\|_{\mathsf{op}}$ for its largest singular value: $\|M\|_{\mathsf{op}} \triangleq \sigma_1(M)$. To not carry dimensional notation, we will also use $\sigma_{\min}(M)$ for the smallest nonzero singular value. For square matrices $M \in \mathbb{R}^{d\times d}$ with real eigenvalues, we similarly order the eigenvalues of M in descending order as $\lambda_1(M), \dots, \lambda_d(M)$. In this case, $\lambda_{\min}(M)$ will also be used to denote the minimum (possibly zero) eigenvalue of M. For two symmetric matrices M, N, we write $M \succ N$ ($M \succeq N$) if $M - N$ is positive (semi-)definite.

1.2. Outline

The core of this class consists of understanding the non-asymptotic behavior of learning algorithms when they interact with temporally dependent data and the use of such learning algorithms in decision-making (control).

1.2.1. Time-Series and Dynamics

To fix ideas, let us consider a time-series of the form:

$$Y_i = f_\star(X_i) + W_i, \qquad i = 1, \dots, n. \tag{1.2.1}$$

We call (1.2.1) a time-series to highlight the fact that the variables (Y_i, X_i) are allowed to depend on past (Y_j, X_j) for $j < i$. The variables Y_i (with values in Y) are called the output (or target), the X_i (with values in X) are typically called covariates and the W_i (with values in Y) are the noise variables. The function $f_\star$ is called the regression functions and is typically the object we want to learn from data. In other situations, noteably when $Y_i = X_{i+1}$, the function $f_\star$ is called the dynamics function or map. This is because in this particular case (1.2.1) reads $X_{i+1} = f_\star(X_i) + W_i$ and so we may think of this as a discrete time dynamical system. While we will look at general, possibly nonlinear $f_\star$ in **??**, much of our focus will be on the linear situation:

$$Y_i = \theta_\star X_i + W_i, \qquad i = 1, \dots, n. \tag{1.2.2}$$

where $\mathsf{X} = \mathbb{R}^{d_\mathsf{X}}$, $\mathsf{Y} = \mathbb{R}^{d_\mathsf{Y}}$ and $\theta_\star \in \mathbb{R}^{d_\mathsf{Y} \times d_\mathsf{X}}$ is a matrix. The simplest dynamical system falling into our model is thus a linear dynamical system:

$$X_{i+1} = \theta_\star X_i + W_i, \qquad i = 1, \dots, n. \tag{1.2.3}$$

1.2.2. Learning

To learn the function $f_\star$, just as in ordinary iid supervised learning, we can often resort to empirical risk minimization. If $\mathsf{L} : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$ is a loss function and $\mathscr{F}$ a hypothesis class, we simply pick $\widehat{f}$ that minimizes the empirical risk

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \mathsf{L}(f(X_i), Y_i). \tag{1.2.4}$$

if Y is a normed space, say $\mathbb{R}^{d_\mathsf{Y}}$ equipped with the Euclidean norm, $\|\cdot\|$, a very natural criterion, dating back to at least Gauss, is the square loss function $\mathsf{L}_{\mathrm{sq}}(y', y) \triangleq \|y' - y\|^2$. If further the map $f_\star$ is linear, say represented by a matrix $\theta_\star$, and the search space (hypothesis class) is taken to be all such matrices, this becomes the Ordinary Least Squares (OLS) estimator:

$$\widehat{\theta} \in \operatorname*{argmin}_{\mathbb{R}^{d_\mathsf{Y} \times d_\mathsf{X}}} \frac{1}{n} \sum_{i=1}^{n} \|Y_i - \theta_\star X_i\|^2. \tag{1.2.5}$$

We will spend a great deal of time understanding the linear model (1.2.2) in conjunction with the estimator (1.2.5). It turns out that, despite its apparent simplicity, analyzing (1.2.5) in conjunction with (1.2.3) is already a formidable task; linear dynamics can already display a rich set of behaviors that are not present when the samples $(X, Y)_{1:n}$ are drawn iid. In the first few weeks of this class, we will first draw up a first principles analysis of linear regression with dependent data. In fact, to ease into it, and as there are a few preliminaries from probability we would like to cover first, we will begin with the situation when $(X, Y)_{1:n}$ are drawn iid. We will then see how our analysis must change once we remove the iid assumption.

1.2.3. Control

Once we have a thorough understanding of learning in linear time-series and dynamics, we will introduce control into the mix. Namely, we will generalize (1.2.3) to include a sequence of control inputs $U_{1:n}$:

$$X_{i+1} = A_\star X_i + B_\star U_i + W_{i+1}, \qquad X_1 = W_0, \qquad i = 1, \dots, n. \tag{1.2.6}$$

The U_i play a special role in that they serve as optimization variables. Namely, they are chosen such as to minimize a cost criterion. In these notes, we will focus on the quadratic cost criterion:

$$\mathtt{V}_n^\pi \triangleq \mathbf{E}^\pi \left[X_n^\mathsf{T} Q_n X_n + \sum_{i=1}^{n-1} X_i^\mathsf{T} Q X_i + U_i^\mathsf{T} R U_i \right], \qquad Q, Q_n \succeq 0, R \succ 0, \tag{1.2.7}$$

and where $\pi = \pi_{1:n}$ is the policy, a sequence of conditional distributions, with π_i dictating the choice of $U_i \mid X_{1:i}$. To be precise then, it is not the U_i but π_i that are the optimization variables. The combination of (1.2.6) and (1.2.7) is called the linear quadratic regulator (LQR). Characterzing the optimal policy π is a classical problem in control theory and we will review the basics of its solution in Chapter 7. Besides being a classical problem in control theory, it is also an excellent candidate for beginning to understand reinforcement learning in large state spaces.

Namely, we will typically assume that $A_\star$ and $B_\star$ are unknown and that we are only given access to the dynamics through the collection of some dataset $X_{1:n+1}, U_{1:n}$. One approach that we will study in the sequel, illustrated in Figure 1.1 below, is to first find estimates $\widehat{\theta} = (\widehat{A}, \widehat{B})$ of $(A_\star, B_\star)$ and then use these estimates to synthesize a control policy.

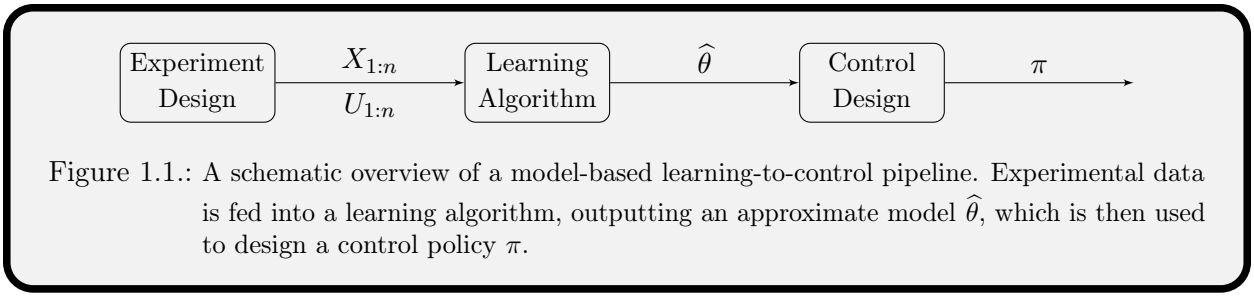

Figure 1.1.: A schematic overview of a model-based learning-to-control pipeline. Experimental data is fed into a learning algorithm, outputting an approximate model $\widehat{\theta}$, which is then used to design a control policy π.

One may also ask what the complementary question of what the optimal way to design the data is for downstream control use. Such questions fall within the purview of experiment design and the notion of an optimal experiment. This requires us to understand statistical optimality: what is the best we can do with a given dataset? We provide a crash course on information-theoretic lower bounds to answer this question in **??**. Once equipped with a refined understanding of data efficiency, we can turn to finding the policy which gives the best dataset, in terms of that dataset consisting of the most informative sample. In other words, in this class we will cover all three of the boxes in Figure 1.1. Finally, we will also get to see how to extend these ideas to more challenging situations such as partially observed (hidden Markov) models and nonlinear models.

1.3. Further Reading

In preparation of this manuscript the author has relied heavily, but not exclusively, on the following sources.

Part I: The first two chapters, mainly dealing with independent data, draw from parts of Wainwright [2019] and Vershynin [2018]. The subsequent development for dependent data builds an alternative approach to the results in Simchowitz et al. [2018], Jedra and Proutiere [2022], Tu et al. [2024] and is based on an expanded form of Ziemann [2023].

Part II: The main references for this part are Mania et al. [2019] and Fazel et al. [2018].

Part III: The main references for this part are Bobrovsky et al. [1987], Wagenmaker et al. [2021] and Lee et al. [2024].

Part IV: The main references for this part are Ziemann and Tu [2022] and Ziemann et al. [2024].

There have also been a few earlier surveys and tutorials on this topic that you may find useful, see in particular Recht [2019], Matni et al. [2019], Tsiamis et al. [2023], Ziemann et al. [2023].

1.4. Acknowledgements

The author thanks Bruce Lee and Thomas Zhang for providing valuable feedback on an earlier draft of this manuscript. Any (and there will be some) remaining mistakes are entirely the fault of the author. The author is also grateful to Nikolai Matni for making available his course material from an earlier year.

Part I.

Learning

2. Probability and Statistics Preliminaries

2.1. Mean Estimation

Let us consider the simplest parametric estimation problem: that of mean estimation. Here, we are simply given n independently drawn observations from a distribution P over $\mathbb{R}^d$. This distribution has a fixed mean *parameter*, say $\theta_\star \in \mathbb{R}^d$, and the learning objective is to use the samples $Y_{1:n} \sim \mathsf{P}^{\otimes n}$, to estimate the mean as well as possible. If we define the *noise variables* $W_i = Y_i - \theta_\star$, this observation model can conveniently be written as

$$Y_i = \theta_\star + W_i, \quad i = 1, \dots, n. \tag{2.1.1}$$

We will see statistical models—regression models—of the more general form $Y_i = f(\theta_\star, X_i) + W_i$ time and time again in this manuscript (where the X_i are random variables known as inputs or covariates). Of course, (2.1.1) is just a particularly simple form of such a regression model in which the *regression function* f is the constant $\theta_\star$.

What is a natural *estimator* to estimate the mean parameter $\theta_\star$? Of course, a reasonable thing to try is just the empirical mean

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{2.1.2}$$

Analyzing the performance of estimators (or more generally decisions made under uncertainty!) such as (2.1.2) will be our main endeavour in the sequel. It turns out that (2.1.2) is particularly simple to analyze. Indeed, (2.1.2) is a (relatively) rare example in which the population level mean square error is analytically available:

$$\begin{aligned}
\mathbf{E}\|\widehat{\theta} - \theta_\star\|^2 &= \mathbf{E} \left\| \frac{1}{n} \sum_{i=1}^{n} Y_i - \theta_\star \right\|^2 \\
&= \mathbf{E} \left\| \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta_\star) \right\|^2 \\
&= \mathbf{E} \left\| \frac{1}{n} \sum_{i=1}^{n} W_i \right\|^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E} \|W_i\|^2 \qquad (\mathbf{E}\langle W_i, W_j \rangle = 0, i \neq j) \\
&= \frac{\operatorname{tr} \mathbf{V}(W)}{n}.
\end{aligned} \tag{2.1.3}$$

We will later see that (2.1.3) is not improvable in a certain sense—a learner essentially needs a priori knowledge of the parameter θ to do better.[1] Statements of the form that no learner can to better than a certain performance level are often called *information-theoretic lower bounds*. This is a more advanced topic that will be covered in part in **??**.

2.1.1. The Central Limit Theorem and the Scale Root-n

While (2.1.3) informs us that the scale of the error $\widehat{\theta} - \theta_\star$ is $n^{-1/2}$, it says relatively little about the distribution of these errors. Moreover, the above calculation (2.1.3) is not really replicable for more advanced models. We would also like a more "instructive" approach.

The key to recognize is that objects of the form

$$\sum_{i=1}^{n}(Y_i - \theta_\star) = \sum_{i=1}^{n} W_i \tag{2.1.4}$$

are well studied in probability theory, and called *random walks*. The perspective here is that a (possibly drunk) person takes a random, mean zero step at time $i = 1$ in the random direction $Y_1 - \theta_\star$ and then at time $i = 2$ takes another step in the direction $Y_2 - \theta_\star$. The walker proceeds and their position at time $i = n$ is given by the sum (2.1.4).

The Lindeberg-Lévy Central Limit Theorem allows us to reason about such random walks. Let us for simplicity assume that $d = 1$, then we have:

$$n^{-1/2}\sum_{i=1}^{n}(Y_i - \theta_\star) \to N\left(0, \mathbf{V}\left(W\right)\right) \qquad \text{in distribution as } n \to \infty. \tag{2.1.5}$$

In other words, the typical "random walker" will tend to move away from the origin at a rate $\sqrt{n}$. This directly translates to a rate of convergence of $1/\sqrt{n}$ for our mean estimator. In particular, the limiting probability that $\widehat{\theta} - \theta$ falls outside of a certain interval exhibits asymptotically Gaussian tails:

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sqrt{\frac{n}{\mathbf{V}(W)}}[\widehat{\theta} - \theta_\star]\right| > s\right) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-s}^{s} e^{-u^2/2}du. \qquad (s \in \mathbb{R}_+) \tag{2.1.6}$$

In other words, the typical deviations of $\widehat{\theta} - \theta_\star$ are of the order $\sqrt{\mathbf{V}(W)/n}$. In the limit, deviations of larger order of magnitude are exponentially rare.

2.1.2. Markov and Chernoff

Let us now try to replicate (2.1.6) *non-asymptotically*; we want sub-Gaussian (super-exponentially decaying) tails for the probability that $\widehat{\theta} - \theta_\star$ falls outside of an interval of length s without explicitly looking at the limiting scale $\sqrt{n}$—without taking limits as in (2.1.6). To this end, we now discuss a few preliminary inequalities that control the tail of a random variable. Our first inequality is Markov's.

[1] This statement presupposes that the mean and variance of P exist. When they do not the situation is rather more subtle.

Lemma 2.1.1 (Markov). *Let X be a nonnegative random variable. For every $s > 0$ we have that*

$$\mathbf{P}(X \geq s) \leq s^{-1}\mathbf{E}[X]. \tag{2.1.7}$$

Proof. We have that $\mathbf{E}[X] \geq \mathbf{E}[\mathbf{1}_{X\geq s}X] \geq s\mathbf{E}[\mathbf{1}_{X\geq s}]$. Since $\mathbf{E}[\mathbf{1}_{X\geq s}] = \mathbf{P}(X \geq s)$ the result follows by rearranging. ■

Typically, Markov's inequality itself is insufficient for our goals: we seek deviation inequalities that taper of exponentially fast in s and not as s^{-1}. Such scaling is for instance predicted asymptotically by the central limit theorem by the asymptotic normality of renormalized sums of square integrable iid random variables; that is, sums of the form $S_n/\sqrt{n} = (X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ where the $X_i, i \in [n]$ are independent and square integrable—see the discussion immediately above and (2.1.6). For random variables possessing a moment generating function, Markov's inequality can be "boosted" by the so-called "Chernoff trick". Namely, we apply Markov's inequality to the moment generating function of the random variable instead of applying it directly to the random variable itself.

Corollary 2.1.1 (Chernoff). *Fix $s > 0$ and suppose $\mathbf{E}\exp(\lambda X)$ exists for $\lambda \in \Lambda \subset \mathbb{R}_+$. Then:*

$$\mathbf{P}(X \geq s) \leq \min_{\lambda\in\Lambda} e^{-\lambda s}\mathbf{E}\exp(\lambda X). \tag{2.1.8}$$

Proof. Fix $\lambda \geq 0$. We have:

$$\begin{aligned}\mathbf{P}(X \geq s) &= \mathbf{P}(\exp(\lambda X) \geq \exp(\lambda s)) && \text{(monotonicity of } x \mapsto e^{\lambda x}\text{)}\\ &\leq e^{-\lambda s}\mathbf{E}\exp(\lambda X) && \text{(Markov's inequality).}\end{aligned}$$

The result follows by optimizing. ■

Recall that the function $\psi_X(\lambda) \triangleq \mathbf{E}\exp(\lambda X)$ is the moment generating function of X. For instance, if X has univariate Gaussian distribution with mean zero and variance σ^2, the moment generating function appearing in (2.1.8) is just $\mathbf{E}\exp(\lambda X) = \exp\left(\lambda^2\sigma^2/2\right)$. Hence the probability that said Gaussian exceeds s is upper-bounded:

$$\mathbf{P}(X > s) \leq \min_{\lambda\geq 0} e^{-\lambda s}\exp\left(\lambda^2\sigma^2/2\right) = \exp\left(\frac{-s^2}{2\sigma^2}\right) \tag{2.1.9}$$

which (almost) exhibits the correct Gaussian tails as compared to (2.1.7). We write almost because $\exp(-s^2/2\sigma^2) \approx \mathbf{P}(V > s)$ where $V \sim N(0, \sigma^2)$ but the expression is not exact—cf. (2.1.6). It should be pointed out that assumptions stronger than those of the Central Limit Theorem (finite variance) are indeed needed for a non-asymptotic theory with sub-Gaussian tails as in (2.1.9). An assumption of this kind which is relatively standard in the literature is introduced next.

In the sequel, we will not want to impose the Gaussian assumption. Instead, we define a class of random variables that admit reasoning analogous to (2.1.9).

Definition 2.1.1. *We say that a centered random vector W taking values in $\mathbb{R}^d$ is σ^2-sub-Gaussian (σ^2-subG) if for every $v \in \mathbb{R}^d$ we have that:*

$$\mathbf{E}\exp(\langle v, W\rangle) \leq \exp\left(\frac{\sigma^2\|v\|^2}{2}\right). \tag{2.1.10}$$

Similarly, we say that W is σ^2-conditionally sub-Gaussian with respect to a σ-field $\mathcal{F}$ if (2.1.10) holds with $\mathbf{E}[\cdot]$ replaced by $\mathbf{E}[\cdot|\mathcal{F}]$ and the conditional mean of W given $\mathcal{F}$ is zero.

The term σ^2 appearing in (2.1.10) is called the variance proxy of a sub-Gaussian random variable. The significance of this definition is that the one-dimensional projections $X = \langle v, W \rangle$ (with $\|v\| = 1$) satisfy the tail inequality (2.1.9). While obviously Gaussian random variables are sub-Gaussian with their variance as variance-proxy, there are many examples beyond Gaussians that fit into this framework. It is for instance straightforward to show that bounded random variables have variance proxy proportional to the square of their width [see eg. Wainwright, 2019, Examples 2.3 and 2.4]. Moreover, it is readily verified that the normalized sum mentioned above—$S_n/\sqrt{n} = (X_1 + \cdots + X_n)/\sqrt{n}$—satisfies the same bound (2.1.9) provided that the entries of $X_{1:n}$ are independent, mean zero and σ^2-sub-Gaussian. To see this, notice that the moment generating function "tensorizes" across products. Namely, for every $\lambda \in \mathbb{R}$:

$$\mathbf{E} \exp\left(\frac{\lambda}{\sqrt{n}} \sum_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbf{E} \exp\left(\frac{\lambda}{\sqrt{n}} X_i\right) \leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 \sigma^2}{2n}\right) = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{2.1.11}$$

Hence, by the exact same reasoning leading up to (2.1.9) such normalized sub-Gaussian sums satisfy the same tail bound (2.1.9).

Indeed returning to our mean estimator in (2.1.6), if we impose the further restriction that $Y_i - \theta$ is σ^2-sub-Gaussian we can now instantiate the above bounds with $X_i = Y_i - \theta_\star$. We obtain:

$$\begin{aligned} \mathbf{P}\left(\sqrt{n}(\widehat{\theta} - \theta_\star) > s\right) &= \mathbf{P}\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \theta_\star\right) > s\right) \\ &= \mathbf{P}\left(\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (Y_i - \theta_\star)\right) > s\right) \\ &\leq \exp\left(\frac{-s^2}{2\sigma^2}\right) \qquad \text{((2.1.11))} \end{aligned} \tag{2.1.12}$$

which is consistent with our earlier asymptotic expression (2.1.6) . The fact that (2.1.12) holds for all finite $n \in \mathbb{N}$ precisely cost us the sub-Gaussian assumption of Definition 2.1.1 and a weakening of the bound in that we have replaced $\mathbf{V}(W)$ by the variance proxy σ^2. Is this assumption reasonable? The answer is of course that it depends. However, at least all bounded random variables are sub-Gaussian.

Exercise 2.1.1. *As an intermediate step to proving that all bounded random variables are sub-Gaussian, let us consider the case of* Rademacher *random variables. Namely, let R be uniformly distributed over $\{-1, -1\}$. Show R is sub-Gaussian with variance proxy* 1. hint: series expand $\mathbf{E}e^{\lambda R} = \frac{e^{-\lambda} + e^{\lambda}}{2}$.

Example 2.1.1 (Hoeffding)**.** *Suppose $|X| \leq b$ for $b \in \mathbb{R}_+$. We will show that X is sub-Gaussian. We will prove this by a technique called* symmetrization*. To this end, let us introduce an auxiliary random variable R which is uniformly distributed over $\{-1, 1\}$. If we further denote by X' an independent copy of X, we have that:*

$$\begin{aligned} \mathbf{E}_X \exp(\lambda(X - \mathbf{E}_X X)) &\leq \mathbf{E}_{X,X'} \exp(\lambda(X - X')) && \text{(Jensen's)} \\ &= \mathbf{E}_{X,X',R} \exp(\lambda R(X - X')) && \text{(symmetry)} \\ &\leq \mathbf{E}_{X,X'} \exp(\lambda^2 (X - X')^2/2) && \text{(Exercise 2.1.1)} \\ &\leq \exp\left(\frac{4\lambda^2 b^2}{2}\right). \end{aligned} \tag{2.1.13}$$

In light of Example 2.1.1 one might think that boundedness and sub-Gaussian concentration is all we need. However, the next example demonstrates that such bounds can be quite far from optimal.

Example 2.1.2. *Fix $p \in (0,1)$ and consider a Bernoulli random variable B which is 1 with probability p and 0 with probability $1-p$. the sub-Gaussian variance proxy obtained from the previous example (of order 1) can be arbitrarily worse than the variance parameter $p(1-p)$ (of order p).*[2]

2.1.3. Further Properties of sub-Gaussian Random Variables

This section collects a few properties of sub-Gaussians that will useful in the sequel.

We saw in (2.1.11) that sums of independent sub-Gaussian random variables are sub-Gaussian. The next lemma shows that this remains true even without independence at the cost of a worsening in the sub-Gaussian constant.

Lemma 2.1.2. *Let X and Y be centered sub-Gaussian random variables with variance proxies σ_X^2 and σ_Y^2. $X+Y$ is centered sub-Gaussian with variance proxy at most $2\sigma_X^2 + 2\sigma_Y^2$.*

Proof. We prove this by Cauchy-Schwarz. Let $\lambda \in R$, then:

$$\begin{aligned}
\mathbf{E}\exp\left(\lambda(X+Y)\right) &\leq \sqrt{\mathbf{E}\exp\left(2\lambda X\right)\mathbf{E}\exp\left(2\lambda Y\right)} \\
&\leq \sqrt{\exp\left(\frac{4\lambda^2\sigma_X^2}{2}\right)\exp\left(\frac{4\lambda^2\sigma_Y^2}{2}\right)} \\
&= \exp\left(\frac{\lambda^2(2\sigma_X^2+2\sigma_Y^2)}{2}\right)
\end{aligned} \tag{2.1.14}$$

establishing the result. ■

Naturally, control of the moment generating function also yields control of the individual moments.

Lemma 2.1.3. *Let $X \in \mathrm{subG}(\sigma^2)$. We have that $\mathbf{E}|X|^p \leq 2p(\sigma^2)^{p/2}\Gamma(p/2) \leq (4e)^{1/p}\sigma^p\sqrt{p}^p$ for all $p \in \mathbb{N}$.*

Proof. We write:

$$\begin{aligned}
\mathbf{E}[|X|^p] &= \int_0^\infty \mathbf{P}(|X|^p > s)ds \\
&= \int_0^\infty \mathbf{P}(|X| > s^{1/p})ds \\
&\leq 2\int_0^\infty \exp\left(\frac{s^{2/p}\sigma^2}{2}\right)ds \\
&= 2p(\sigma^2)^{p/2}\int_0^\infty u^{p/2-1}e^{-u}du \\
&= 2p(\sigma^2)^{p/2}\Gamma(p/2) \\
&\leq 4e\sigma^p p^{p/2}
\end{aligned} \tag{2.1.15}$$

as was required. ■

[2] Note that the sub-Gaussian variance proxy of a Bernoulli random variable is actually smaller than suggested by Example 2.1.1.

Exercise 2.1.2. *Fill in the details for the proof of Lemma 2.1.3.*

1. *For every positive random variable X, show that $\mathbf{E}X = \int_0^\infty \mathbf{P}(X > s)ds$.*
2. *for every integer n: $\Gamma(n) = (n)! \leq n^n$.*
3. *For every integer n : $\Gamma(n+1/2) \leq 1 + \Gamma(n+1) \leq 2(n+1)^{n+1}$*
4. *For every odd integer p: $2p\left(\frac{p+1}{2}\right)^{(p+1)/2} \leq 4ep^{p/2}$. Hint: maximize $2p^{3/2}\left(\frac{p+1}{2p}\right)^{(p+1)/2}$.*

The factor $(4e)^{1/p}$ is approximately 1 for large p, but can still be improved—can you come up with a tighter proof for small p?

2.2. A First Look at Random Design Linear Regression

Let us now go beyond mean estimation and consider a more complicated model. In the most general setting we consider, we will let the parameter $\theta_\star$ be a matrix in $\mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$. We define distributions P_X and P_W over $\mathbb{R}^{d_{\mathsf{X}}}$ and $\mathbb{R}^{d_{\mathsf{Y}}}$ respectively and draw $(X_{1:n}, W_{1:n}) \sim \mathsf{P}_X^{\otimes n} \otimes \mathsf{P}_W^{\otimes n}$. We will again assume that all variables have finite variance. This allows us to define observations as noise corrupted versions of $\theta_\star X_i$ via:

$$Y_i = \theta_\star X_i + W_i, \qquad i = 1, \ldots, n. \tag{2.2.1}$$

The learner now has access to tuples $(X_i, Y_i), i = 1 \ldots, n$ to recover the parameter. In this case, the natural algorithm is a little (but not much) more complicated than the empirical mean. Namely, we will look at the ordinary least squares (OLS) estimator:

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}} \left\{ \frac{1}{n} \sum_{i=1}^n \|Y_i - \theta X_i\|^2 \right\}. \tag{2.2.2}$$

Exercise 2.2.1. *Show that on the event that $\sum_{i=1}^n X_i X_i^\mathsf{T}$ is invertible, the solution to (2.2.2) is given by*

$$\widehat{\theta} - \theta_\star = \left(\sum_{i=1}^n W_i X_i^\mathsf{T} \right) \left(\sum_{i=1}^n X_i X_i^\mathsf{T} \right)^{-1}. \tag{2.2.3}$$

2.2.1. The Law of Large Numbers and the Scale n

Let us consider the one-dimensional setting, in which $d_{\mathsf{X}} = d_{\mathsf{Y}} = 1$. In this case, the OLS error equation (2.2.3) takes the form

$$\widehat{\theta} - \theta_\star = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n X_i^2} = n^{-1/2} \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i X_i}{\frac{1}{n} \sum_{i=1}^n X_i^2}. \tag{2.2.4}$$

How do we analyze the error (2.2.4) or (2.2.3) more generally? Equation (2.2.4) is already quite instructive of a general approach. Namely, notice that the denominator in the rightmost expression

is the empirical average of the iid variables X_i^2. The law of large numbers informs us that averaging random variables of a fixed mean preserves their scale:

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \to \mathbf{E}X^2 \qquad \text{almost surely as } n \to \infty. \tag{2.2.5}$$

Put differently, the sum $\sum_{i=1}^n X_i^2$ is of order n.

Turning to the numerator of (2.2.4), this is just as in our analysis of mean estimation a random walk. Indeed, the W_iX_i are iid mean zero random variables and so

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n W_iX_i \to N(0, \mathbf{V}(WX)) \qquad \text{in distribution as } n \to \infty. \tag{2.2.6}$$

Combining (2.2.5) with (2.2.6) and applying Slutsky's Theorem yields that:

$$\sqrt{n}(\widehat{\theta} - \theta_\star) \to N\left(0, \frac{\mathbf{V}(WX)}{\mathbf{E}X^2}\right) \qquad \text{in distribution as } n \to \infty. \tag{2.2.7}$$

Equation (2.2.7) thus shows that, just as in the case of mean estimation, the typical rate of convergence for linear regression is $n^{-1/2}$. Let us now establish a nonasymptotic version of (2.2.7). We will proceed in a sequence of steps. Let us assume that both $X_{1:n}$ and $W_{1:n}$ are drawn from sub-Gaussian distributions with variance proxies σ_X^2 and σ_W^2 respectively. The astute reader will notice that the assumption of sub-Gaussianity is not sufficient to establish the an analogue of (2.2.7). The reason for this is that while we have assumed that individual X_i and W_i are sub-Gaussian, their products (and squares) are not necessarily. Fortunately, they satisfy a weaker notion, known as the class of *sub-exponential* random variables. We introduce these in Section 2.2.2 below. For now, we will instead assume that $XW \in \text{subG}(\sigma_{XW}^2)$ and $X^2 \in \text{subG}(\sigma_{X^2}^2)$. In this case, we first obtain, just as before, that:

$$\mathbf{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n W_iX_i \geq \sqrt{2\sigma_{WX}^2 s^2}\right) \leq e^{-s^2}. \tag{2.2.8}$$

Moreover, under our assumptions $\frac{1}{n}\sum_{i=1}^n X_i^2 \in \text{subG}(n^{-1}\sigma_{X^2}^2)$ and so:

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 \leq \frac{1}{n}\sum_{i=1}^n \mathbf{E}X_i^2 - \sqrt{\frac{2\sigma_{X^2}^2 s^2}{n}}\right) \leq e^{-s^2}. \tag{2.2.9}$$

We will often want to make sure a statement holds with a fixed failure probability, say δ. To this end, in this particular case we set $\delta = \exp(-s^2)$ or equivalently, $s = \sqrt{\log(1/\delta)}$. To clean up our result, let us further require that for some $\varepsilon > 0$:

$$\sqrt{\frac{2\sigma_{X^2}^2 \log(1/\delta)}{n}} \leq \frac{\varepsilon}{n}\sum_{i=1}^n \mathbf{E}X_i^2 \Leftrightarrow n \geq \frac{2\sigma_{X^2}^2 \log(1/\delta)}{\varepsilon^2(\mathbf{E}X^2)^2}. \tag{2.2.10}$$

Putting everything together (with two union bounds) yields that for $\delta \in (0, 1/3)$ and with probability $1 - 3\delta$ we have that:

$$|\widehat{\theta} - \theta_\star| \leq \frac{\sqrt{2\sigma_{XY}^2 \log(1/\delta)}}{\sqrt{n}(1-\varepsilon)\mathbf{E}X^2} \tag{2.2.11}$$

as long as (2.2.10) holds. Conditions such as (2.2.10)—*burn-in conditions*—will appear repeatedly in the analysis of various estimators. More generally in the case of the least squares estimator (2.2.2), they correspond to the minimal sample size such that the matrix $\sum_{i=1}^n X_i X_i^\mathsf{T}$ is sufficiently removed from being ill-conditioned.

2.2.2. Sub-Exponential Concentration

As noted above, the assumptions that $XW \in \mathrm{subG}(\sigma_{XW}^2)$ and $X^2 \in \mathrm{subG}(\sigma_{X^2}^2)$ are not entirely natural—they for instance rule out the classical situation in which X and W are jointly Gaussian (why?). The following slightly weaker notion however turns out be sufficient to not exclude the Gaussian situation.

Definition 2.2.1. *We say that a centered random vector Z taking values in $\mathbb{R}^d$ is sub-exponential with variance proxy σ^2 and domain range α if for all $v \in \mathbb{S}^{d-1}$*

$$\mathbf{E} \exp\left(\lambda \langle v, Z\rangle\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \qquad \text{for all } \lambda \in \left[\frac{-1}{\alpha}, \frac{1}{\alpha}\right]. \tag{2.2.12}$$

Our first result regarding sub-exponential random variables is that they admit tail bounds that are almost as good as sub-Gaussian random variables.

Lemma 2.2.1. *Let $Z \in \mathrm{subE}(\sigma^2, \alpha)$. We have that*

$$\mathbf{P}(Z \geq s) \leq \begin{cases} \exp\left(-\frac{s^2}{2\sigma^2}\right) & s \leq \sigma^2/\alpha, \\ \exp\left(-\frac{s}{2\alpha}\right) & \text{else.} \end{cases} \tag{2.2.13}$$

A concise way to think of Lemma 2.2.1 can be obtained by noticing that if we set $s = \sqrt{2\sigma^2 \log(1/\delta)}$ we have with probability $1-\delta$ that $Z \leq \sqrt{2\sigma^2 \log(1/\delta)}$ in the first range and if we instead set $s = 2\alpha^{-1}\log(1/\delta)$ we have with probability $1-\delta$ that $Z \leq 2\alpha \log(1/\delta)$. In particular, it always holds that with probability $1-\delta$:

$$Z \leq \sqrt{2\sigma^2 \log(1/\delta)} + 2\alpha \log(1/\delta). \tag{2.2.14}$$

Proof. We have that

$$\mathbf{P}(Z \geq s) \leq \min_{\lambda \in [0, \alpha^{-1}]} \exp\left(-\lambda s + \frac{\lambda^2\sigma^2}{2}\right) \tag{2.2.15}$$

If $s \leq \sigma^2/\alpha$ this becomes $\exp(-s^2/2\sigma^2)$. Otherwise, if $s \geq \sigma^2/\alpha$ the minimum is achieved at the boundary $\lambda = \alpha^{-1}$ (why?), in which case we may verify that

$$\exp\left(-\lambda s + \frac{\lambda^2\sigma^2}{2}\right) \leq \exp\left(-\alpha^{-1}s + \frac{\alpha^{-1}s}{2}\right) = \exp\left(-\frac{s}{2\alpha}\right) \tag{2.2.16}$$

as was required. ∎

Let us next show that sums of sub-exponential random variables are again sub-exponential, irrespectively of their dependence structure.

Lemma 2.2.2. *Let $X \in \mathrm{subE}(\sigma_X^2, \alpha_X), Y \in \mathrm{subE}(\sigma_Y^2, \alpha_Y)$. We then have that $X+Y \in \mathrm{subE}(2\sigma_X^2 + 2\sigma_Y^2, \alpha_X \wedge \alpha_Y)$*

Exercise 2.2.2. *Prove Lemma 2.2.2.*

Before we proceed with investigating the behavior of squares of sub-Gaussian variables, let us note that bounded random variables are sub-exponential. The following is often attributed to Bernstein.

Lemma 2.2.3 (Bernstein's Inequality)**.** *Fix $b > 0$ and let X be a bounded random variable, such that $|X - \mathbf{E}X| \leq b$. Then for every $\lambda \in (-b^{-1}, b^{-1})$:*

$$\mathbf{E} \exp\left(\lambda(X - \mathbf{E}X)\right) \leq \exp\left(\frac{\lambda^2 \mathbf{V}(X)}{2(1 - |\lambda| b)}\right). \tag{2.2.17}$$

Proof.

$$\begin{aligned}
\mathbf{E} \exp\left(\lambda X - \mathbf{E}X\right) &= 1 + \frac{\lambda^2 \mathbf{V}(X)}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k \mathbf{E}(X - \mathbf{E}X)^k}{k!} \\
&\leq 1 + \frac{\lambda^2 \mathbf{V}(X)}{2} + \frac{\lambda^2 \mathbf{V}(X)}{2} \sum_{k=3}^{\infty} \frac{2(|\lambda| b)^{k-2}}{k!} \quad (|X - \mathbf{E}X| \leq b) \\
&\leq 1 + \frac{\lambda^2 \mathbf{V}(X)}{2} + \frac{\lambda^2 \mathbf{V}(X)}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2} \qquad (k \geq 3 \geq 2) \\
&= 1 + \frac{\lambda^2 \mathbf{V}(X)}{2(1 - |\lambda| b)}.
\end{aligned} \tag{2.2.18}$$

The result follows by invoking the inequality $1 + x \leq e^x, x \in \mathbb{R}$. ■

The following lemma is key to relaxing the sub-Gaussianity assumption made in the previous section.

Lemma 2.2.4. *Let X and Y be real random variables, jointly distributed according to $\mathsf{P}_{X,Y}$. Suppose that the centered marginal distributions P_X and P_Y are sub-Gaussian with variance proxies σ_X^2 and σ_Y^2 respectively.*

1. *The random variable $Z = X^2 - \mathbf{E}X^2$ is sub-exponential with variance proxy $32\sigma_X^4$ and domain range $4\sigma_X^2$.*
2. *The random variable $Z = XY - \mathbf{E}XY$ is sub-exponential with variance proxy $64(\sigma_X^2 + \sigma_Y^2)^2$ and domain range $2(\sigma_X^2 + \sigma_Y^2)$.*
3. *If furthermore X and Y are independent with Y mean zero, then the random variable $Z = XY$ is sub-exponential with variance proxy $8\sigma_X^2\sigma_Y^2$ and domain range $4(\sigma_X)(1 + \sigma_Y)$.*

Proof. To prove 1, we expand the moment generating function of Z:

$$\begin{aligned}
\mathbf{E}\exp(\lambda Z) &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbf{E}(X^2 - (\mathbf{E}X^2))^k}{k!} && (Z \text{ is centered}) \\
&= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^k \mathbf{E}\left(\frac{X^2}{2} - \frac{(\mathbf{E}X^2)}{2}\right)^k}{k!} \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^{k-1}[\mathbf{E}(X^2)^k + (\mathbf{E}X^2))^k]}{k!} && (\text{Jensen's}) \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^k \mathbf{E}X^{2k}}{k!} && (\text{Jensen's}) \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^k (4k\sigma^{2k}(k-1)!)}{k!} && (\text{Lemma 2.1.3}) \\
&\leq 1 + 4\sum_{k=2}^{\infty} \lambda^k 2^k \sigma^{2k} \\
&= 1 + 4(2\lambda\sigma^2)^2 \sum_{k=0}^{\infty} (2\lambda\sigma^2)^k \\
&= 1 + 8(\lambda\sigma^2)^2 \frac{1}{1 - 2\lambda\sigma^2} && (|\lambda| < (2\sigma^2)^{-1} \\
&= 1 + 16\lambda^2\sigma^4 && (|\lambda| \leq (4\sigma^2)^{-1}) \\
&\leq \exp\left(\frac{32\lambda^2\sigma^4}{2}\right) && (x \in \mathbb{R} \Rightarrow 1 + x \leq e^x)
\end{aligned} \tag{2.2.19}$$

which proves the first claim.

To prove the second claim, observe that $XY = \left(\frac{X+Y}{2}\right)^2 - \left(\frac{X-Y}{2}\right)^2$. We may assume without loss of generality that XY is centered. By Lemma 2.1.2 it is clear that both $\frac{X+Y}{2}$ and $\frac{X-Y}{2}$ are sub-Gaussian with variance proxy $\frac{\sigma_X^2+\sigma_Y^2}{2}$ and thus by the first point both $\left(\frac{X+Y}{2}\right)^2$ and $\left(\frac{X-Y}{2}\right)^2$ are sub-exponential with variance proxy $16(\sigma_X^2 + \sigma_Y^2)^2$ and domain range $(2(\sigma_X^2 + \sigma_Y^2))^{-1}$. The second claim now follows by a final application of Lemma 2.2.2.

As for the the third claim, notice that by independence of X and Y:

$$\begin{aligned}
\mathbf{E}\exp(\lambda XY) &\le \mathbf{E}\exp\left(\frac{\lambda^2 X^2\sigma_Y^2}{2}\right) && \text{(centered } Y\in \mathrm{subG}(\sigma_Y^2)\text{)}\\
&\le \mathbf{E}\exp\left(\frac{\lambda^2(X^2-\mathbf{E}X^2)\sigma_Y^2}{2}+\frac{\lambda^2\mathbf{E}X^2\sigma_Y^2}{2}\right)\\
&\le \exp\left(\frac{64\lambda^4\sigma_X^4\sigma_Y^4}{8}+\frac{\lambda^2\mathbf{E}X^2\sigma_Y^2}{2}\right) && \text{(part 1. and } |\lambda|\le(4\sigma_X)^{-1}\text{)}\\
&\le \exp\left(\frac{4\lambda^2\sigma_X^2\sigma_Y^2}{2}+\frac{\lambda^2\mathbf{E}X^2\sigma_Y^2}{2}\right) && \text{(}|\lambda|\le(4\sigma_X(1+\sigma_Y))^{-1}\text{)}\\
&\le \exp\left(\frac{8\lambda^2\sigma_X^2\sigma_Y^2}{2}\right) && \text{(}\mathbf{E}X^2\le 4\sigma_X^2 \text{ by Lemma 2.1.3}\text{)}
\end{aligned} \tag{2.2.20}$$

as was required for the the third part. ■

Exercise 2.2.3. *This exercise asks you to prove an analogue of (2.2.11) without imposing sub-Gaussianity of the W_iX_i and X_i^2. As before, assume that $X_i\in\mathrm{subG}(\sigma_X^2), i=1,\dots,n$ and $W_i\in\mathrm{subG}(\sigma_W^2), i=1,\dots,n$.*

1. *Show there exist universal positive constant $c,c'>0$ that for $\delta\in(0,1)$ it holds with probability at least $1-\delta$ that:*

$$\frac{1}{n}\sum_{i=1}^n W_iX_i \le \sqrt{\frac{c\sigma_X^2\sigma_Y^2\log(1/\delta)}{n}}+\frac{c'(\sigma_X(1+\sigma_W))\log(1/\delta)}{n}. \tag{2.2.21}$$

2. *Show that there exist universal positive constants $c,c'>0$ such that with probability at least $1-\delta$*

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \ge c\mathbf{E}X^2 \tag{2.2.22}$$

as long as $n\ge\frac{c'\sigma_X^2\log(1/\delta)}{\mathbf{E}X^2}$.

3. *Conclude that there exist universal positive constants c,c' such that with probability at least $1-2\delta$ we have that*

$$|\widehat{\theta}-\theta_\star| \le \sqrt{\frac{c(\sigma_X\sigma_W)^2\log(1/\delta)}{n(\mathbf{E}X^2)^2}} \tag{2.2.23}$$

as long as $n\ge c'\left(\frac{\sigma_X(1+\sigma_W)}{\sigma_X\sigma_W}]\vee\frac{\sigma_X^2}{\mathbf{E}X^2}\right)\log(1/\delta)$.

4. *Note that (2.2.23) is still unsatisfactory in that the leading term depends on $\sigma_X\sigma_W$ which is qualitatively larger than the variance of XY. Show that if X and Y are bounded by B_X and B_Y then $\sigma_X\sigma_W$ can be replaced by $\mathbf{V}(XY)$ in (2.2.23) at the cost of inflating the burn-in ($n\ge\dots$).*

2.3. More Concentration Inequalities

Let us consider a non-negative random variable Z, e.g., $Z = \sum_{i=1}^n X_i^2$ for some sequence of random variables $X_{1:n}$. In Exercise 2.2.3 we used concentration inequalities to establish that such random variables do not become "too small". In particular, we used a sub-exponential concentration inequality, requiring the existence of all moments, to establish such control. The following lemma shows that the lower tail exhibits sub-Gaussian behavior even if only the first two moments exist.

Lemma 2.3.1 (Better Anti-Concentration)**.** *Let Z be a non-negative random variable. For all $\lambda \in [0, \infty)$ we have that:*

$$\mathbf{E} \exp(-\lambda Z) \leq \exp\left(-\lambda \mathbf{E} Z + \frac{\lambda^2}{2} \mathbf{E} Z^2\right) \tag{2.3.1}$$

Therefore, for every $t \in \mathbb{R}_+$ and every sequence $Z_{1:n}$ of independent copies of Z:

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \leq \mathbf{E} Z - t\right) \leq \exp\left(\frac{-nt^2}{2\mathbf{E} Z^2}\right). \tag{2.3.2}$$

Contrast this with our earlier Hoeffding style bound which for $Z \in \mathrm{subG}(\sigma_Z^2)$ would have given

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \leq \mathbf{E} Z - t\right) \leq \exp\left(\frac{-nt^2}{2\sigma_Z^2}\right). \tag{2.3.3}$$

In other words, Lemma 2.3.1 is an improvement of a factor $\sigma_Z^2/\mathbf{E} Z^2$ on the naive bound.

Proof. The first part follows by estimating the exponential function twice:

$$\begin{aligned} \mathbf{E} \exp(-\lambda Z) &\leq \mathbf{E}\left[1 - Z + \frac{\lambda^2}{2} Z^2\right] && \left(x \geq 0 \Rightarrow e^{-x} \leq 1 - x + \frac{x^2}{2}\right) \\ &\leq \exp\left(-\lambda \mathbf{E} Z + \frac{\lambda^2}{2} \mathbf{E} Z^2\right) && \left(x \in \mathbb{R} \Rightarrow 1 + x \leq e^x\right). \end{aligned} \tag{2.3.4}$$

The second part follows from the first and a Chernoff bound:

$$\begin{aligned} \mathbf{P}\left(\sum_{i=1}^n Z_i \leq \sum_{i=1}^n \mathbf{E} Z_i - nt\right) &= \mathbf{P}\left(\sum_{i=1}^n [-Z_i + \mathbf{E} Z_i] \geq nt\right) \\ &= \min_{\lambda \geq 0} \mathbf{P}\left(\exp(-\lambda \sum_{i=1}^n [-Z_i + \mathbf{E} Z_i]) \geq e^{\lambda nt}\right) \\ &\leq \min_{\lambda \geq 0} \exp\left(-\lambda nt + \frac{n\lambda^2}{2} \mathbf{E} Z^2\right) && (Z_i \text{ iid and Chernoff}) \\ &= \exp\left(\frac{-nt^2}{2\mathbf{E} Z^2}\right) \end{aligned} \tag{2.3.5}$$

as was required. ∎

Exercise 2.3.1. *Improve the burn-in requirement in Exercise 2.2.3, by showing that (2.2.22) holds while only requiring the existence of $\mathbf{E} X^4$ (i.e., provide a refined analysis of the lower tail of the scalar empirical covariance).*

3. Learning in $\mathbb{R}^d$

In the previous chapter we saw how elementary concentration inequalities allowed us to control the performance of the empirical mean estimator and the least squares estimator in 1-dimensional parameter space. In this chapter, we will see how to extend these ideas to various higher-dimensional problems in $\mathbb{R}^d$. Our main technique to achieve this is called *covering* or the *ε-net argument*. Let us for instance suppose we wanted to control the operator norm of a random matrix $M \in \mathbb{R}^{d\times d'}$. This is for instance pertinent if we wish to analyze the general least squares estimator

$$\widehat{\theta} - \theta_\star = \left(\sum_{i=1}^n W_i X_i^\mathsf{T}\right)\left(\sum_{i=1}^n X_i X_i^\mathsf{T}\right)^{-1}. \tag{3.0.1}$$

Both $\sum_{i=1}^n W_i X_i^\mathsf{T}$ and $\sum_{i=1}^n X_i X_i^\mathsf{T}$ are random matrices and controlling their spectral information is a natural step in proving guarantees on $\widehat{\theta} - \theta_\star$.

3.1. Mean Estimation in $\mathbb{R}^d$

Before we tackle the question of analyzing (3.0.1), let us return to the (arguably simpler) mean estimation problem discussed (2.1). Recall that we were given observations

$$Y_i = \theta_\star + W_i, \quad i = 1, \dots, n. \tag{3.1.1}$$

where we now assume that $Y_i, \theta_\star, W_i \in \mathbb{R}^d$. Just as before, we assume that the samples Y_i are drawn iid. However, let us also impose that the random variables $W_i, i = 1 \dots, n$ are mean zero σ^2-sub-Gaussian random vectors. We seek to control the estimation error:

$$\left\|\widehat{\theta} - \theta_\star\right\| = \left\|\frac{1}{n}\sum_{i=1}^n Y_i - \theta_\star\right\| = \frac{1}{n}\left\|\sum_{i=1}^n W_i\right\|. \tag{3.1.2}$$

Our previous success in controlling the scalar version of (3.1.2) hinged on the fact that the moment generating function of the random walk (sum of the independent random variables) $\sum_{i=1}^n W_i$ simply is their product. However, the appearance of the norm in (3.1.2) complicates matters. It is not an obvious task to compute moment generating functions of norms of random variables based only on information about the random variables themselves. The key observation turns out to be that norms admit variational characterizations. For instance, the Euclidean norm of a vector $x \in \mathbb{R}^d$ can be written as

$$\|x\| = \max_{v\in\mathbb{S}^{d-1}} \langle v, x\rangle = \max_{v\in\mathbb{S}^{d-1}} \sum_{i=1}^d v_i x_i \qquad (v_i, x_i \text{ coordinate projections of } v, x) \tag{3.1.3}$$

and sums of independent random variables we know how to control.

3.1.1. Covering Numbers

In other words, we will often find ourselves in a situation where it is possible to obtain a scalar concentration bound but need this to hold uniformly (the maximum can set-theoretically be thought of as a union) for many random variables at once. The ε-net argument, which proceeds via the notion of covering numbers, is a relatively straightforward way of converting concentration inequalities for scalars into their counterparts for vectors, matrices and functions more generally.

Definition 3.1.1. *Let (X, d) be a metric space and fix $\varepsilon > 0$. A subset $\mathcal{N}$ of X is called an ε-net of X if every point of X is within radius ε of a point of $\mathcal{N}$:*

$$\sup_{x \in \mathsf{X}} \inf_{x' \in \mathcal{N}} d(x, x') \leq \varepsilon. \tag{3.1.4}$$

Moreover, the minimal cardinality of $\mathcal{N}$ necessary such that (3.1.4) holds is called the covering number at resolution ε of (X, d) and is denoted $\mathcal{N}(\varepsilon, \mathsf{X}, d)$.

To understand why the notion of a covering number is important, notice that the maximum in (3.1.3) is taken over an infinite (even uncountably so) set, $\mathbb{S}^{d-1}$. Recall now that the union bound states that the probability that the maximum of a *finite collection* ($|S| < \infty$) $\{X_i\}_{i \in S}$ of random variables exceeds a certain threshhold can be bounded by the sum of their probabilities:

$$\mathbf{P}\left(\max_{i \in S} X_i > t\right) = \mathbf{P}\left(\bigcup_{i \in S} \{X_i > t\}\right) \leq \sum_{i \in S} \mathbf{P}\left(X_i > t\right). \tag{3.1.5}$$

Unfortunately as we noted, the unit sphere appearing (3.1.3) is not a finite set and so the union bound (3.1.5) cannot be directly applied. However, when the domain of optimization has geometric structure, one can often exploit this to leverage the union bound not directly but rather in combination with a discretization argument. We now provide a discretized version of the variational form in (3.1.3).

Lemma 3.1.1. *Let $x \in \mathbb{R}^d$ and suppose that $\mathcal{N}$ is an ε-net of $\mathbb{S}^{d-1}$. Then*

$$\|x\| \leq \frac{1}{1-\varepsilon} \max_{v_0 \in \mathcal{N}} \langle v_0, x \rangle. \tag{3.1.6}$$

Proof. Fix $x \in \mathbb{R}^d$ and let $v \in \mathbb{S}^{d-1}$ be such that $\langle v, x \rangle = \|x\|$ (i.e. $v = x/\|x\|$). Let further $v_0 \in \mathcal{N}$ be such that $\|v - v_0\| \leq \varepsilon$. We have that:

$$\begin{aligned} \|x\| &= \langle v, x \rangle \\ &= \langle v - v_0 + v_0, x \rangle \\ &\leq |\langle v - v_0, x \rangle| + \langle v_0, x \rangle \\ &\leq \|v - v_0\| \|x\| + \langle v_0, x \rangle \quad \text{(Cauchy-Schwarz)} \\ &\leq \varepsilon \|x\| + \langle v_0, x \rangle. \qquad (\|v - v_0\| \leq \varepsilon) \end{aligned} \tag{3.1.7}$$

Re-arranging yields that $(1-\varepsilon)\|x\| \leq \max_{v_0 \in \mathcal{N}} \langle v_0, x \rangle$ for every $x \in \mathbb{R}^d$. ∎

The reader should now ask: but exactly how many points do we need to cover, say, the sphere? We will make frequent use of the following fact, which quantifies the covering numbers of the unit ball $\mathbb{B}^d$ and unit sphere $\mathbb{S}^{d-1}$ in the Euclidean metric.

Lemma 3.1.2 (Volumetric Argument)**.** *The covering number of the unit ball is exponential in d:*

$$\frac{1}{\varepsilon^d} \leq \mathcal{N}(\varepsilon, \mathbb{B}^d, \|\cdot\|) \leq \left(1 + \frac{2}{\varepsilon}\right)^d. \tag{3.1.8}$$

Moreover, the upper bound remains true for the unit sphere: $\mathcal{N}(\varepsilon, \mathbb{S}^{d-1}, \|\cdot\|) \leq \left(1 + \frac{2}{\varepsilon}\right)^d$.

Before we proceed with the proof of Lemma 3.1.2 we will need to introduce the auxiliary notion of a packing number.

Definition 3.1.2. *An ε-packing of a metric space (X, d) is a set $\mathcal{M}$ such that $d(x, y) > \varepsilon$ for all $x, y \in \mathcal{M}$. The packing number of (X, d) at resoultion ε, $\mathcal{M}(\varepsilon, \mathsf{X}, d)$ is the maximal cardinality of any ε-packing of (X, d).*

The proof of Lemma 3.1.2 relies on the following observation, relating covering numbers to packing numbers.

Lemma 3.1.3. *For any metric space (X, d) and all $\varepsilon > 0$ the covering and packing numbers satisfy the following:*

$$\mathcal{M}(2\varepsilon, \mathsf{X}, d) \leq \mathcal{N}(\varepsilon, \mathsf{X}, d) \leq \mathcal{M}(\varepsilon, \mathsf{X}, d). \tag{3.1.9}$$

Proof. We only prove the second inequality in (3.1.9). To this end, let $\mathcal{M} = \{x_1, \ldots, x_m\}$ be an optimal ε-packing of (X, d). Observe that we necessarily have for any $x \in \mathsf{X}$ $\|x - x_i\| \leq \varepsilon$ for at least one index $i \in [m]$ by virtue of the fact our packing is optimal (any point $x \in \mathsf{X}$ not satisfying this could otherwise be added to the packing, contradicting optimality).

The proof of the first inequality is not central to the main development of this text and left as an exercise. ■

We are now ready to prove Lemma 3.1.2.

Proof of Lemma 3.1.2. Let us denote Lebesque measure in $\mathbb{R}^d$ by $\lambda(\cdot)$. Suppose that $x_1, \ldots, x_n$ is an ε cover of $(\mathbb{B}^d, \|\cdot\|)$. Clearly,

$$\mathbb{B}^d \subseteq \sum_{i=1}^{n} x_i + \varepsilon \mathbb{B}^d \tag{3.1.10}$$

which yields that $\lambda(\mathbb{B}^d) \leq n\lambda(\varepsilon\mathbb{B}^d) = n\varepsilon^d\lambda(\mathbb{B}^d)$. Hence: $|\mathcal{N}| \geq \varepsilon^{-d}$ for any ε-covering $\mathcal{N}$ and in particular this is true for the (any) optimal covering. This establishes the first inequality in (3.1.8).

To prove the second inequality in (3.1.8), let instead $x_1, \ldots, x_m$ be an optimal ε-packing of $(\mathbb{B}^d, \|\cdot\|)$. By optimality, $x_1, \ldots, x_m$ is also an ε-cover of $(\mathbb{B}^d, \|\cdot\|)$. Next, observe that by the packing property the balls $x_i + (\varepsilon/2)\mathbb{B}^d, i \in [m]$ are all disjoint and their union is contained in $\mathbb{B}^d + (\varepsilon/2)\mathbb{B}^d$. Hence:

$$\begin{aligned}
& m\lambda((\varepsilon/2)\mathbb{B}^d) \leq \lambda\left((\varepsilon/2)\mathbb{B}^d + \mathbb{B}^d\right) \\
\Leftrightarrow \\
& m(\varepsilon/2)^d\lambda(\mathbb{B}^d) \leq (\varepsilon/2)^d\lambda((1 + 2/\varepsilon)\mathbb{B}^d) \\
\Leftrightarrow \\
& m \leq (1 + 2/\varepsilon)^d.
\end{aligned} \tag{3.1.11}$$

The result for the unit ball follows since $x_1, \ldots, x_m$ is, as we previously noted, an admissible ε-covering.

Finally, the upper bound for the unit sphere follows almost exactly in the same way as the second inequality above since $m\lambda((\varepsilon/2)\mathbb{B}^d) \leq \lambda\left((\varepsilon/2)\mathbb{B}^d + \mathbb{S}^{d-1}\right) \leq \lambda\left((\varepsilon/2)\mathbb{B}^d + \mathbb{B}^d\right)$ whenever m is the cardinality of an optimal packing of $\mathbb{S}^{d-1}$. ■

3.1.2. Controlling the Random Walk in $\mathbb{R}^d$

We have now established all the preliminaries to prove a non-asymptotic guarantee on the d-dimensional mean estimation error in (3.1.2).

Fix $\varepsilon > 0$ and let $\mathcal{N}_\varepsilon$ be an optimal ε-cover of $\mathbb{S}^{d-1}$. As per Lemma 3.1.1 we have that for every $\varepsilon \in (0, 1)$:

$$\frac{1}{n}\left\|\sum_{i=1}^n W_i\right\| \leq \frac{1}{\sqrt{n}(1-\varepsilon)} \max_{v \in \mathcal{N}_\varepsilon} \left\langle v, \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \right\rangle = \frac{1}{\sqrt{n}(1-\varepsilon)} \max_{v \in \mathcal{N}_\varepsilon} \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle v, W_i \rangle . \tag{3.1.12}$$

Each of the variables $\langle v, W_i \rangle$ are 1-dimensional mean zero independent σ^2-sub-Gaussian and consequently so is the normalized sum $\frac{1}{\sqrt{n}} \sum_{i=1}^n \langle v, W_i \rangle$. Consequently for every fixed v with probabiltiy at least $1 - \delta$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \langle v, W_i \rangle \leq \sqrt{2\sigma^2 \log(1/\delta)}. \tag{3.1.13}$$

Let us set $\varepsilon = 2/3$. We now apply the union bound (3.1.5) to the (at most) $(1 + 2/\varepsilon)^d = 4^d$ elements of $\mathcal{N}_\varepsilon$ (invoke Lemma 3.1.2) to conclude that with probability at least $1 - 4^d\delta$ that we have:

$$\frac{1}{n}\left\|\sum_{i=1}^n W_i\right\| \leq \frac{3\sqrt{2\sigma^2 \log(1/\delta)}}{\sqrt{n}}. \tag{3.1.14}$$

A change of variables from δ to $\delta' = 4^d\delta$ yields that with probability at least $1 - \delta'$:

$$\frac{1}{n}\left\|\sum_{i=1}^n W_i\right\| \leq \frac{3\sqrt{2\sigma^2 (d \log(4) + \log(1/\delta'))}}{\sqrt{n}}. \tag{3.1.15}$$

Comparing with $\sqrt{\operatorname{tr} \mathbf{V}(W)/n}$ from (2.1.3) we see that this is indeed the correct scaling with n and cannot be uniformly improved in the class of sub-Gaussian random variables (up to universal constants at least). Indeed, the factor $d\sigma^2$ is exactly equal to $\operatorname{tr} \mathbf{V}(W)$ if $W_i, i \in [n]$ are independent Gaussian. It is also worth to point out that the *deviation term*, $\log(1/\delta)$, in principle remains unaffected by the increased dimensionality—the effect is additive and contributes to the mean but not directly to the deviation from the mean.

3.2. Random Design Linear Regression in Higher Dimensions

We now turn our attention to analyzing the $(d_{\mathsf{X}}, d_{\mathsf{Y}})$-dimensional ordinary least squares estimator (3.0.1). The learner obtains access to observations $(X, Y)_{1:n}$ which satisfy

$$Y_i = \theta_\star X_i + W_i \tag{3.2.1}$$

where $\theta_\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{Y}}}$ is unknown and where the W_i are iid σ_W^2-sub-Gaussian. We also assume that the X_i are drawn iid σ_X^2-sub-Gaussian.

With these preliminary estimates in place, let us now proceed to analyze (3.0.1). We write (3.0.1) in prefiltered form as

$$\begin{aligned}(\widehat{\theta} - \theta_\star)\sqrt{\Sigma_X} &= \left(\sum_{i=1}^n W_i(\Sigma_X^{-1/2}X_i)^{\mathsf{T}}\right)\left(\sum_{i=1}^n \Sigma_X^{-1/2}X_i(\Sigma_X^{-1/2}X_i^{\mathsf{T}})\right)^{-1} \\ &= \left(\sum_{i=1}^n W_i\tilde{X}_i^{\mathsf{T}}\right)\left(\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right)^{-1}\end{aligned} \tag{3.2.2}$$

where $\tilde{X}_i \triangleq \Sigma^{-1/2}X_i, i \in [n]$ are the whitened covariates. Our main new challenge, as noted above, is that the $\sum_{i=1}^n W_i\tilde{X}_i^{\mathsf{T}}$ and $\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^{\mathsf{T}}$ are random matrices instead of scalars or vectors. To set the stage for the analysis to be done in the sequel, notice that for $\circ \in \{F, \mathsf{op}\}$:

$$\left\|(\widehat{\theta} - \theta_\star)\sqrt{\Sigma_X}\right\|_\circ \le \left\|\sum_{i=1}^n W_i\tilde{X}_i^{\mathsf{T}}\right\|_\circ \left\|\left(\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right)^{-1}\right\|_{\mathsf{op}}. \tag{3.2.3}$$

Thankfully, the Frobenius and operator norms admit discretization results analogous to Lemma 3.1.1. The analysis of $\sum_{i=1}^n W_i\tilde{X}_i^{\mathsf{T}}$ is indeed similar to what we have previously seen and proceeds for the Frobenius norm by the simple observation that $\|M\|_F^2 = \operatorname{tr} M^{\mathsf{T}}M = (\mathsf{vec}\, M)^{\mathsf{T}}\,\mathsf{vec}\, M$. Hence to control the random walk in Frobenius norm, Lemma 3.1.1 can invoked with $x = \mathsf{vec}\, M$. As for the matrix operator norm, we have the following result.

Lemma 3.2.1. *Fix $\varepsilon \in (0, 1/2)$, let $M \in \mathbb{R}^{d' \times d}$ and let $\mathcal{N}, \mathcal{M}$ be ε-nets of $\mathbb{S}^{d-1}$ and $\mathbb{S}^{d'-1}$. We have that*

$$\sup_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Mx, y\rangle \le \|M\|_{\mathsf{op}} \le \frac{1}{1-2\varepsilon}\sup_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Mx, y\rangle. \tag{3.2.4}$$

Moreover if $d = d'$ and M is symmetric:

$$\sup_{x\in\mathcal{N}} \langle Mx, x\rangle \le \|M\|_{\mathsf{op}} \le \frac{1}{1-2\varepsilon}\sup_{x\in\mathcal{N}} \langle Mx, x\rangle. \tag{3.2.5}$$

The proof proceeds analogously to that of (3.1.1).

Exercise 3.2.1. *Prove Lemma 3.2.1.*

However, the analysis of the whitened empirical covariance matrix $\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^{\mathsf{T}}$ is a little different. Inspecting (3.2.2) and (3.2.3)—since operator norm of the inverse of a positive semidefinite matrix is the smallest eigenvalue value of that matrix—it seems that we require a guarantee that with high probability

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right) \ge c \tag{3.2.6}$$

for some universal positive constant c. Equivalently this can be stated as:

$$\frac{1}{n}\sum_{i=1}^n \langle v, \tilde{X}_i\rangle^2 \ge c \quad \text{for all } v \in \mathbb{S}^{d_{\mathsf{X}}-1}. \tag{3.2.7}$$

In other words, we wish to show that the smallest eigenvalue of the whitened empirical covariance matrix is lower bounded by some absolute constant independent of the data generating mechanism.[1] While Lemma 3.2.1 controls the largest eigenvalue of positive semidefinite matrices, we now require control at the other end of the spectrum. The following variation of Lemma 3.2.1 does the trick.

Lemma 3.2.2. *Fix $\varepsilon \in (0, 1/2)$, let $M_i \in \mathbb{R}^{d\times d}, i \in [n]$ be symmetric positive semidefinite matrices and let $\mathcal{N}$ be an ε-net of $\mathbb{S}^{d-1}$. For any sequence of square roots $\sqrt{M_i}, i \in [n]$ of the M_i and any $\alpha \in \mathbb{R}_+$ we have that:*

$$\lambda_{\min}\left(\sum_{i=1}^n M_i\right) \geq \frac{1}{1+\alpha}\min_{u\in\mathcal{N}}\sum_{i=1}^n \|\sqrt{M_i}u\|^2 - \frac{\varepsilon^2}{\alpha}\left\|\sum_{i=1}^n M_i\right\|_{\mathsf{op}}. \tag{3.2.8}$$

Proof. We write

$$\lambda_{\min}\left(\sum_{i=1}^n M_i\right) = \min_{v\in\mathbb{S}^{d-1}}\sum_{i=1}^n \langle M_i v, v\rangle = \min_{v\in\mathbb{S}^{d-1}}\sum_{i=1}^n \langle \sqrt{M_i}v, \sqrt{M_i}v\rangle = \min_{v\in\mathbb{S}^{d-1}}\sum_{i=1}^n \|\sqrt{M_i}v\|^2. \tag{3.2.9}$$

Let now $v \in \mathbb{S}^{d-1}$ and choose $u \in \mathcal{N}$ to be determined later. For any $\alpha \in \mathbb{R}_+$, using Young's inequality we have that:

$$\begin{aligned}
\sum_{i=1}^n \|\sqrt{M_i}u\|^2 &= \sum_{i=1}^n \|\sqrt{M_i}(u-v+v)\|^2 \\
&= \sum_{i=1}^n \|\sqrt{M_i}v\|^2 + \|\sqrt{M_i}(u-v)\|^2 + 2\langle\sqrt{M_i}(u-v), \sqrt{M_i}v\rangle \\
&\leq \sum_{i=1}^n \|\sqrt{M_i}v\|^2 + \|\sqrt{M_i}(u-v)\|^2 + \alpha\|\sqrt{M_i}v\|^2 + \frac{1}{\alpha}\|\sqrt{M_i}(u-v)\|^2 \\
&\leq \sum_{i=1}^n (1+\alpha)\|\sqrt{M_i}v\|^2 + \left(1+\frac{1}{\alpha}\right)\left\|\sum_{i=1}^n M_i\right\|_{\mathsf{op}}\varepsilon^2
\end{aligned} \tag{3.2.10}$$

where in the last line we chose $u = u(v)$ to be within distance ε of v. Upon re-arranging and then minimizing over first v and then u:

$$\min_{v\in\mathbb{S}^{d-1}}\sum_{i=1}^n \|\sqrt{M_i}v\|^2 \geq \frac{1}{1+\alpha}\min_{u\in\mathcal{N}}\sum_{i=1}^n \|\sqrt{M_i}u\|^2 - \left(\frac{1}{\alpha}\right)\left\|\sum_{i=1}^n M_i\right\|_{\mathsf{op}}\varepsilon^2 \tag{3.2.11}$$

as was required (use $1 + 1/\alpha = (1+\alpha)/\alpha$). ■

The path forward to establishing (3.2.7) is now clear. We first show that $\frac{1}{n}\sum_{i=1}^n \langle v, \tilde{X}_i\rangle^2 \geq c$ pointwise for each v with high probability and then make this estimate uniform via (3.2.2). Instantiating Lemma 2.3.1 we have that:

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n \langle v, \tilde{X}_i\rangle^2 \leq \frac{1}{n}\sum_{i=1}^n \mathbf{E}\langle v, \tilde{X}_i\rangle^2 - t\right) \leq \exp\left(-\frac{nt^2}{2\mathbf{E}\langle v, \tilde{X}_i\rangle^4}\right) \tag{3.2.12}$$

[1] The whitening step is not necessary but simplifies the bookkeeping in the sequel as it ensures that applying the operator norm bound in (3.2.3) is not too wasteful (why?).

Since the $\tilde{X}_i$ are whitened, it becomes convenient to define the hypercontractivity constant $\kappa_X \triangleq \max_{v\in\mathbb{S}^{d_{\mathsf{X}}-1}} \sqrt{\mathbf{E}\langle v, \tilde{X}_i\rangle^4}$.[2] The above then simplifies to:

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\langle v, \tilde{X}_i\rangle^2 \leq 1-t\right) \leq \exp\left(-\frac{nt^2}{2\kappa_X^2}\right) \tag{3.2.13}$$

We may use our just-established discretization argument for the smallest eigenvalue in Lemma 3.2.2 to conclude that for every $\varepsilon > 0$ and with probability at least $1 - \left(1+\frac{2}{\varepsilon}\right)^{d_{\mathsf{X}}} \exp\left(-\frac{nt^2}{2\kappa_X^2}\right)$ we have that

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right) \geq \frac{1-t}{1+\varepsilon} - \varepsilon\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right\|_{\mathsf{op}} \tag{3.2.14}$$

Equation (3.2.14) is almost what we want, but there still is an operator norm term we are yet to bound. The argument for this is standard; for $\mathcal{M}$ a (2/3)-net of $\mathbb{S}^{d_{\mathsf{X}}-1}$ we write

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right\|_{\mathsf{op}} \leq \sup_{v\in\mathcal{M}}\frac{3}{n}\sum_{i=1}^{n}\langle \tilde{X}_i, v\rangle^2 \tag{3.2.15}$$

If $\tilde{X} \in \mathrm{subG}(\sigma_{\tilde{X}}^2)$ we have that (pointwise) for every v, for some universal positive constant $c > 0$ and every $s > 0$:

$$\sum_{i=1}^{n}\langle \tilde{X}_i, v\rangle^2 \leq \sum_{i=1}^{n}\mathbf{E}\langle \tilde{X}_i, v\rangle^2 + s \tag{3.2.16}$$

with probability at least $1 - \exp\left(-\frac{cs}{n\sigma_{\tilde{X}}^2}\right)$ (in the large s regime—which we are in fact interested in). In particular, (3.2.16) holds uniformly for every $v \in \mathcal{M}$ with probability at least $1 - 4^{d_{\mathsf{X}}}\exp\left(-\frac{cs}{n\sigma_{\tilde{X}}^2}\right)$. Let us select $s = n^2$, this gives

$$\left\|\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right\|_{\mathsf{op}} \leq n + n^2 \tag{3.2.17}$$

with probability at least $1 - 4^{d_{\mathsf{X}}}\exp\left(-\frac{cn}{\sigma_{\tilde{X}}^2}\right) = 1 - \exp\left(c' d_{\mathsf{X}} - \frac{cn}{\sigma_{\tilde{X}}^2}\right)$. A further union and combining the above with (3.2.14) yields that:

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right) \geq \frac{1-t}{1+\varepsilon} - 3\varepsilon(1+n) \tag{3.2.18}$$

with probability at least $1 - \exp\left(c' d_{\mathsf{X}} - \frac{cn}{\sigma_{\tilde{X}}^2}\right) - \exp\left(1+\frac{2}{\varepsilon}\right)^{d_{\mathsf{X}}}\exp\left(-\frac{nt^2}{2\kappa_X^2}\right)$. In order to render the second term small, we set $\varepsilon = \frac{t}{3(1+n)}$. Cleaning up the details, and noticing that $\kappa_X^2 \lesssim \sigma_{\tilde{X}}^2$, we have proven the following.

[2] Note that since $\tilde{X}_i = \Sigma_X^{-1/2}X_i$, κ_X actually controls the ratio between the fourth and second moments of X.

Proposition 3.2.1. *There exists a universal positive constant $c > 0$ and with $\tilde{X}_i$ as above, we have for every $t \in (0,1)$ that with probability at least $1 - \delta$*

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^{\mathsf{T}}\right) \geq 1 - t \tag{3.2.19}$$

as long as

$$n \geq \frac{c\sigma_{\tilde{X}}^2}{t^2}\left(d_{\mathsf{X}}\log n + \log(1/\delta)\right). \tag{3.2.20}$$

Combining the above with an argument analogous to that leading up to (3.1.15) but leveraging sub-exponential concentration instead of sub-Gaussian concentration yields the main result of this chapter.

Theorem 3.2.1. *There exist universal positive constants $c, c' > 0$ such that with probability $1 - 2\delta$ we have that:*

$$\left\|(\widehat{\theta} - \theta_\star)\sqrt{\Sigma_X}\right\|_{\mathsf{op}} \leq c\sqrt{\sigma_{\tilde{X}}^2\sigma_W^2(d_{\mathsf{X}} + d_{\mathsf{Y}} + \log(1/\delta))} \tag{3.2.21}$$

as long as

$$n \geq c'\left(\sigma_{\tilde{X}}^2 d_{\mathsf{X}}\log n + \left[\sigma_{\tilde{X}}^2 + 1 + \sigma_W\right]\log(1/\delta)\right) \tag{3.2.22}$$

4. The Hanson-Wright Inequality: Concentration with Linear Dependence

Suppose we wanted to "identify" the following linear dynamical system:

$$X_{t+1} = A_\star X_t + W_{t+1} \qquad X_1 = W_1 \qquad t = 1, \dots, n. \tag{4.0.1}$$

Here, the variables (entries of) $X_{1:n+1}, W_{1:n+1}$ are each elements of $\mathbb{R}^{d_{\mathsf{X}}}$ so that $A_\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$. Our goal will be to estimate the matrix $A_\star$. If it were not for the fact that the variables $X_{1:n+1}$ were highly correlated, this would fit perfectly into our previous model and analysis of the least squares estimator (with $Y_{1:n} = X_{2:n+1}$). Namely, we can still write

$$\widehat{A} - A_\star = \left(\sum_{i=1}^n W_i X_i^\mathsf{T} \right) \left(\sum_{i=1}^n X_i X_i^\mathsf{T} \right)^\dagger. \tag{4.0.2}$$

Unfortunately temporal dependence rules out use of our previous results. In order to salvage the situation, it turns out that the key observation here is that any dependence in the model (4.0.1) is linear: the variables $X_{1:n}$ are linear in the noise $W_{1:n}$. Consequently both the numerator and denominator of (4.0.2) are quadratic in $W_{1:n}$. Hence, we also need to understand how squares of sub-Gaussian random variables behave. Fortunately, quadratic forms are a well-studied topic in the literature. The main result of this chapter shows that sub-Gaussian quadratic forms exhibit similar tail behavior to the Chi-squared distribution (often in the literature referred to as sub-exponential tails). It is known as the Hanson-Wright Inequality.

Theorem 4.0.1 (Hanson and Wright [1971], Rudelson and Vershynin [2013])**.** *Let $M \in \mathbb{R}^{d \times d}$. Fix a random variable $W = W_{1:d}$ where each $W_i, i \in [d]$ is a scalar, mean zero and independent σ^2-sub-Gaussian random variable. Then for every $s \in [0, \infty)$:*

$$\mathbf{P}\left(|W^\mathsf{T} M W - \mathbf{E} W^\mathsf{T} M W| > s \right) \leq 2 \exp\left(- \min\left(\frac{s^2}{144 \sigma^4 \|M\|_F^2}, \frac{s}{16\sqrt{2}\sigma^2 \|M\|_{\mathsf{op}}} \right) \right). \tag{4.0.3}$$

4.1. Proof of The Hanson-Wright Inequality

Let us now proceed with the proof of the Hanson-Wright Inequality (Theorem 4.0.1). We will need a few intermediate exponential inequalities relating sub-Gaussian variables to their Gaussian counterparts.

4.1.1. Gaussian Comparison Inequalities for sub-Gaussian Quadratic Forms

Lemma 4.1.1 (Quadratic Comparison to Gaussian)**.** *Fix symmetric positive semidefinite $M \in \mathbb{R}^{d\times d}$ and let W be a σ^2-sub-Gaussian assuming values in $\mathbb{R}^d$. Let further G be an independent $N(0, I_d)$-distributed random variable. For every $\lambda \in \mathbb{R}$ with $|\lambda| \leq 1/(\sqrt{2}\sigma^2\|M\|_{\mathsf{op}})$:*

$$\mathbf{E}\exp\left(\lambda W^\mathsf{T} M W\right) \leq \mathbf{E}\exp\left(\lambda\sigma^2 G^\mathsf{T} M G\right) \leq \exp\left(\lambda^2\sigma^4\|M\|_F^2\right) \tag{4.1.1}$$

Proof. For the first inequality note that:

$$\begin{aligned}\mathbf{E}\exp\left(\lambda W^\mathsf{T} M W\right) &= \mathbf{E}\exp\left(\sqrt{2\lambda}W^\mathsf{T}\sqrt{M}G\right) && \text{(Gaussian MGF)}\\ &\leq \mathbf{E}\exp\left(\lambda\sigma^2 G^\mathsf{T} M G\right). && \text{(sub-Gaussian MGF)}\end{aligned} \tag{4.1.2}$$

To obtain the second inequality, let VSV^T be the eigendecomposition of M with orthonormal V. Let $v_i, i \in [d]$ be the columns of V. Then

$$G^\mathsf{T} M G = \sum_{i=1}^d s_i^2 (v_i^\mathsf{T} G)^2 \tag{4.1.3}$$

Note by Gaussian rotational invariance that the vector G' with entries $v_i^\mathsf{T} G$ is equal in distribution to G. Hence:

$$\begin{aligned}\mathbf{E}\exp\left(\lambda\sigma^2 G^\mathsf{T} M G\right) &= \prod_{i=1}^d \mathbf{E}\exp\left(\lambda\sigma^2 s_i^2 (v_i^\mathsf{T} G)^2\right)\\ &= \prod_{i=1}^d \mathbf{E}\exp\left(\lambda\sigma^2 s_i^2 G_i^2\right) && \text{(Gaussian Rotation Invariance)}\\ &= \prod_{i=1}^d \left(1-\lambda^2\sigma^4 s_i^2\right)^{-1/2} && \text{(Gaussian-Squared MGF; } \lambda \text{ in our range)}\\ &= \prod_{i=1}^d \exp\left(\frac{-1}{2}\log(1-\lambda^2\sigma^4 s_i^2)\right)\\ &\leq \prod_{i=1}^d \exp\left(\lambda^2\sigma^4 s_i^2\right) && (-\log(1-x) \leq 2x \text{ if } x \in [0, 1/2])\\ &= \exp\left(\lambda^2\sigma^4\|M\|_F^2\right) && \left(\sum_{i=1}^d s_i^2 = \|M\|_F^2\right)\end{aligned} \tag{4.1.4}$$

as per requirement. ■

Lemma 4.1.2 (Decoupled Comparison to Gaussians)**.** *Fix $M \in \mathbb{R}^{d\times d}$ and let W, W' be independent σ^2-sub-Gaussian assuming values in $\mathbb{R}^d$. Let further G, G' be two independent $N(0, I_d)$-distributed random variables. For every $\lambda \in \mathbb{R}$:*

$$\mathbf{E}\exp\left(\lambda W^\mathsf{T} M W'\right) \leq \mathbf{E}\exp\left(\lambda\sigma^2 G^\mathsf{T} M G'\right).$$

Proof. The result follows by straightforward computation, alternatingly using the sub-Gaussian property and the closed form of the Gaussian MGF. Namely:

$$\begin{aligned}
\mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW'\right) &\leq \mathbf{E}\exp\left(\frac{\sigma^2\lambda^2\|MW'\|_2^2}{2}\right) && (W \text{ is } \sigma^2-\text{subG})\\
&= \mathbf{E}\exp\left(\lambda\sigma G^{\mathsf{T}}MW'\right) && (\text{MGF of } G)\\
&\leq \mathbf{E}\exp\left(\frac{\lambda^2\sigma^4\|MG\|_2^2}{2}\right) && (W' \text{ is } \sigma^2-\text{subG})\\
&= \mathbf{E}\exp\left(\lambda\sigma^2 G^{\mathsf{T}}MG'\right). && (\text{MGF of } G')
\end{aligned} \tag{4.1.5}$$

The desired result has been established. ■

Lemma 4.1.3 (MGF of Gaussian Chaos)**.** *Let $W, W' \sim N(0, I_d)$ be independent and let $M \in \mathbb{R}^{d\times d}$. Then for every $\lambda \in \mathbb{R}$ with $|\lambda| \leq 1/(\sqrt{2}\|M\|_{\mathsf{op}})$:*

$$\mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW'\right) \leq \exp\left(\lambda^2\|M\|_F^2\right) \tag{4.1.6}$$

Proof. By the singular value decomposition we may write $M = USV^{\mathsf{T}}$ where the columns $u_{1:d}$ ($v_{1:d}$) of U (of V) form orthonormal bases of $\mathbb{R}^d$. Hence

$$W^{\mathsf{T}}MW = \sum_{i=1}^{d} s_i(W^{\mathsf{T}}u_i)((W')^{\mathsf{T}}v_i) \tag{4.1.7}$$

where $s_{1:d}$ are the singular values of M. Note next that $G_i \triangleq (W^{\mathsf{T}}u_i)$ and $G_i' \triangleq ((W')^{\mathsf{T}}v_i)$ are again independent and standard normal by Gaussian rotational invariance. Hence:

$$\begin{aligned}
\mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW\right) &= \mathbf{E}\exp\left(\lambda\sum_{i=1}^{d} s_iG_iG_i'\right)\\
&= \prod_{i=1}^{d}\mathbf{E}\exp\left(\lambda s_iG_iG_i'\right) && (\text{independence})\\
&= \prod_{i=1}^{d}\mathbf{E}\exp\left(\lambda^2 s_i^2G_i^2/2\right) && (\text{Gaussian MGF})\\
&= \prod_{i=1}^{d}\left(1-\lambda^2 s_i^2\right)^{-1/2} && (\text{Gaussian-Squared MGF; } \lambda \text{ in our range})\\
&= \prod_{i=1}^{d}\exp\left(\frac{-1}{2}\log(1-\lambda^2 s_i^2)\right) && (\exp\circ\log \text{ is the identity function})\\
&\leq \prod_{i=1}^{d}\exp\left(\lambda^2 s_i^2\right) && (-\log(1-x)\leq 2x \text{ if } x\in[0,1/2])\\
&= \exp\left(\lambda^2\|M\|_F^2\right) && \left(\sum_{i=1}^{d} s_i^2 = \|M\|_F^2\right)
\end{aligned} \tag{4.1.8}$$

as per requirement. ■

4.1.2. Finishing the proof of Theorem 4.0.1

We need one more preliminary tool before we arrive at the proof of Theorem 4.0.1. The next Theorem constitutes a useful decoupling inequality which allows us to treat the mixed terms in the quadratic form $W^{\mathsf{T}}MW$ as independent.

Theorem 4.1.1 (Theorem 6.1.1 in of Vershynin [2018]). *Let W be a d-dimensional random vector with mean zero and independent entries. For every convex function f and every $M = (m_{ij})_{i,j=1}^{d} \in \mathbb{R}^{d\times d}$ it holds true that:*

$$\mathbf{E}f\left(\sum_{i,j=1,i\neq j}^{d} m_{ij}W_iW_j\right) \leq \mathbf{E}f\left(4\sum_{i,j=1}^{d} m_{ij}W_iW'_j\right) \tag{4.1.9}$$

where W' is an independent copy of W (i.e., equal to W in distribution but independent of W).

Proposition 4.1.1 (Hanson-Wright Exponential Inequality Form). *Let $M \in \mathbb{R}^{d\times d}$. Fix a random variable $W = W_{1:d}$ where each $W_i, i \in [d]$ is an independent scalar σ^2-sub-Gaussian random variable. For every $\lambda \in \mathbb{R}$ with $|\lambda| \leq \frac{1}{8\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}$ we have that:*

$$\max\left\{\mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW - \lambda\mathbf{E}W^{\mathsf{T}}MW\right), \mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW\right)\right\} \leq \exp\left(36\lambda^2\sigma^4\|M\|_F^2\right) \tag{4.1.10}$$

Proof. By rescaling λ we may assume without loss of generality that $\sigma^2 = 1$. Let now m_{ij} with $i, j \in [d]$ denote the entries of M. We begin by writing

$$W^{\mathsf{T}}MW - \mathbf{E}W^{\mathsf{T}}MW = \sum_{i=1}^{d} m_{ii}(W_i^2 - \mathbf{E}W_i^2) + \sum_{i\neq j} m_{ij}W_iW_j. \tag{4.1.11}$$

Hence by the Cauchy-Schwarz inequality:

$$\begin{aligned}&\mathbf{E}\exp\left(\lambda W^{\mathsf{T}}MW - \lambda\mathbf{E}W^{\mathsf{T}}MW\right)\\ &\qquad\leq \sqrt{\mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}(W_i^2 - \mathbf{E}W_i^2)\right)} \times \sqrt{\mathbf{E}\exp\left(2\lambda\sum_{i\neq j} m_{ij}W_iW_j\right)}.\end{aligned} \tag{4.1.12}$$

We proceed to analyze both terms appearing on the RHS of (4.1.12) separately. We begin by

analyzing the diagonal term. Let W' be an independent copy of W. Then:

$$\begin{aligned}
&\mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}(W_i^2 - \mathbf{E}W_i^2)\right) \\
&\leq \mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}(W_i^2 - \mathbf{E}(W_i')^2)\right) && (W = W' \text{ in distribution}) \\
&\leq \mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}(W_i^2 - (W_i')^2)\right) && (\text{Jensen's inequality}) \\
&= \mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}W_i^2\right)\mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d}(-m_{ii})(W_i)^2\right). && (\text{independence})
\end{aligned} \tag{4.1.13}$$

Hence, if we combine Lemma 4.1.1 with (4.1.13) we find that:

$$\sqrt{\mathbf{E}\exp\left(2\lambda\sum_{i=1}^{d} m_{ii}(W_i^2 - \mathbf{E}W_i^2)\right)} \leq \exp\left(4\lambda^2\|M\|_F^2\right) \tag{4.1.14}$$

as long as $|\lambda| \leq 1/(2\sqrt{2}\|M\|_{\mathsf{op}})$.

Next, we argue similarly for the off-diagonal term and use Theorem 4.1.1 to control the off-diagonal term in (4.1.12). Again letting W' be an independent copy of W and also letting G, G' be two independent isotropic Gaussians in $\mathbb{R}^d$, we have that:

$$\begin{aligned}
\mathbf{E}\exp\left(2\lambda\sum_{i\neq j} m_{ij}W_iW_j\right) &\leq \mathbf{E}\exp\left(8\lambda\sum_{i,j=1}^{d} m_{ij}W_iW_j'\right) && (\text{Theorem 4.1.1}) \\
&\leq \mathbf{E}\exp\left(8\lambda\sum_{i,j=1}^{d} m_{ij}G_iG_j'\right) && (\text{Lemma 4.1.2}) \\
&\leq \exp\left(64\lambda^2\|M\|_F^2\right) && (\text{Lemma 4.1.3})
\end{aligned} \tag{4.1.15}$$

as long as $|\lambda| \leq 1/(8\sqrt{2}\|M\|_{\mathsf{op}})$. The result follows by combining (4.1.14) and (4.1.15) with (4.1.12) and then finally rescaling λ. We also note that the non-centered result follows analogously by skipping the step (4.1.13). ■

The Hanson-Wright Inequality is usually stated in high probability form as in Theorem 4.0.1. We finish the proof of this result below.

Finishing the proof of Theorem 4.0.1 We employ the the Chernoff trick as in (2.1.8) combined with the MGF bound of Proposition 4.1.1. For $\lambda \in \mathbb{R}$ with $|\lambda| \leq \frac{1}{8\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}$ we have that:

$$\mathbf{P}\left(W^{\mathsf{T}}MW - \mathbf{E}W^{\mathsf{T}}MW > s\right) \leq \exp\left(-\lambda s + 36\lambda^2\sigma^4\|M\|_F^2\right). \tag{4.1.16}$$

An admissible choice of λ is:

$$\lambda = \begin{cases} \frac{s}{72\sigma^4\|M\|_F^2}, & \text{if} \quad \frac{s}{72\sigma^4\|M\|_F^2} \leq \frac{1}{8\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}, \\ \frac{1}{8\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}, & \text{if} \quad \frac{s}{72\sigma^4\|M\|_F^2} > \frac{1}{8\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}. \end{cases} \tag{4.1.17}$$

Note that the second condition in (4.1.17) can be rewritten as $\frac{72\sigma^2\|M\|_F^2}{8\sqrt{2}\|M\|_{\mathsf{op}}} < s$. Hence with the choice (4.1.17) inserted into (4.1.16) we have:

$$\mathbf{P}\left(W^{\mathsf{T}}MW - \mathbf{E}W^{\mathsf{T}}MW > s\right) \leq \exp\left(-\min\left(\frac{s^2}{144\sigma^4\|M\|_F^2}, \frac{s}{16\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}\right)\right).$$

The result follows by applying the same calcuation to $-M$ and using a union bound. ■

5. Linear Regression with Linear Dependence

Let us fix ideas. We are concerned with linear time-series models of the form:

$$Y_i = \theta^\star X_i + V_i \quad i = 1, 2, \dots, n \tag{5.0.1}$$

where $Y_{1:n}$ is a sequence of outputs (or targets) assuming values in $\mathbb{R}^{d_{\mathsf{Y}}}$ and $X_{1:n}$ is a sequence of inputs (or covariates) assuming values in $\mathbb{R}^{d_{\mathsf{X}}}$. The goal of the user (or learner) is to recover the a priori unknown linear map $\theta^\star \in \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$ using only the observations $X_{1:n}$ and $Y_{1:n}$. The linear relationship in the regression model (5.0.1) is perturbed by a stochastic noise sequence $V_{1:n}$ assuming values in $\mathbb{R}^{d_{\mathsf{Y}}}$. We refer to the regression model (5.0.1) as a time-series to emphasize the fact that the observations $X_{1:n}$ and $Y_{1:n}$ may arrive sequentially and in particular that past X_i and Y_i may influence future X_j and Y_j (i.e. with $j > i$). In particular, we are interested in the performance of the least squares estimator:

$$(\widehat{\theta} - \theta_\star)\sqrt{\Gamma} = \left(\sum_{i=1}^n V_i X_i^\mathsf{T} \Gamma^{-1/2}\right)\left(\sum_{k=1}^n \Gamma^{-1/2} X_k X_k^\mathsf{T} \Gamma^{-1/2}\right)^\dagger \tag{5.0.2}$$

where Γ is a positive definite weighting (whitening) matrices. We will typically—and it is helpful to think this way—set $\Gamma = \frac{1}{n}\sum_{i=1}^n \mathbf{E} X_i X_i^\mathsf{T}$. Irrespective of this choice, we will always assume that $\Sigma_X \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{E} X_i X_i^\mathsf{T}$ is full rank (and this is essentially without loss of generality since $X_i \in \operatorname{span} \Sigma_X, i \in [n]$ almost surely).

The Dependence Structure Let us specify the dependence structure. We fix a noise source $W_{1:n+1}$ evolving on $\mathbb{R}^{d_{\mathsf{W}}}$. We assume that each of the $(n+1) \times d_{\mathsf{W}}$-many entries of $W_{1:n+1}$ are iid with variance 1 and K^2-sub-Gaussian. We assume that both $X_{1:n}$ and $V_{1:n}$ are linear transformation of this noise source. Namely, we fix two matrices $\mathbf{L}$ and $\mathbf{H}$ and set $X_{1:n} = \mathbf{L} W_{1:n+1}$ and $V_{1:n} = \mathbf{H} W_{1:n+1}$. We also point out that the choice that $W_{1:n+1}$ have variance 1 is without loss of generality since $W_{1:n+1}$ appears quadratically in both the "numerator" and "denominator" of (5.0.2). Moreover, heteroskedasticity can be captured by the matrices $\mathbf{H}$ and $\mathbf{L}$.

It is helpful to have an example in mind, and our canonical one will be that of a first order auto-regressive linear dynamical system. Namely

$$X_i = A_\star X_{i-1} + W_i \qquad X_1 = W_1, \qquad i = 2, 3, \dots, n, n+1. \tag{5.0.3}$$

In terms of (5.0.1), we thus set $Y_i = X_{i+1}$, $V_i = W_i$ and $\theta_\star = A_\star$. In terms of our linear structure, this holds with $\mathbf{L} = \mathbf{L}_{A_\star, n}$ and $\mathbf{H} = \mathbf{H}_{\mathrm{LDS}}$ where for $k \in \mathbb{N}$:

$$\mathbf{L}_{A_\star, k} \triangleq \begin{bmatrix} I_{d_{\mathsf{X}}} & 0 & 0 & \dots & 0 \\ A_\star & I_{d_{\mathsf{X}}} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \dots & 0 \\ A_\star^{k-1} & A_\star^{k-2} & \dots & \dots & I_{d_{\mathsf{X}}} \end{bmatrix}, \quad \text{and} \quad \mathbf{H}_{\mathrm{LDS}} \triangleq \begin{bmatrix} 0 & I_{d_{\mathsf{X}} n} \end{bmatrix}. \tag{5.0.4}$$

5.1. Instantiating the Hanson-Wright Inequality

The main result of this chapter relies on the following simple observation, which brings one-dimensional projections of the numerator of (5.0.2) into a form amenable to analysis by the Hanson-Wright inequality. We notice as before that:

$$(\widehat{\theta} - \theta_\star)\sqrt{\Gamma} = \left(\sum_{i=1}^{n} V_i X_i^\mathsf{T} \Gamma^{-1/2}\right)\left(\sum_{k=1}^{n} \Gamma^{-1/2} X_k X_k^\mathsf{T} \Gamma^{-1/2}\right)^\dagger. \tag{5.1.1}$$

We will deal with each term inside the two parentheses separately and in order. The analysis follows along the same lines that we have previously seen, but we will now leverage the fact that we have the powerful Hanson-Wright Inequality at our disposal combined with our assumption that both the "numerator" and "denominator" in (5.0.2) are quadratic in the noise variable $W_{1:n+1}$.

5.1.1. The Random Walk

We first turn to the random walk component. The following lemma writes the 1-dimensional projections of this component as a quadratic form.

Lemma 5.1.1. *Suppose that $V_{1:n} = \mathbf{H} W_{1:n+1}$ and $X_{1:n} = \mathbf{L} W_{1:n+1}$. For $v \in \mathbb{R}^{d_\mathsf{Y}}$ and $x \in \mathbb{R}^{d_\mathsf{X}}$, it holds that*

$$\sum_{i=1}^{n} v^\mathsf{T} V_i X_i^\mathsf{T} \Gamma^{-1/2} x = W_{1:n+1}^\mathsf{T} \mathbf{H}^\mathsf{T} \operatorname{blkdiag}(v x^\mathsf{T}) \mathbf{G} \mathbf{L} W_{1:n+1} \tag{5.1.2}$$

where $\mathbf{G} = \operatorname{blkdiag}_{1:n}(\Gamma^{-1/2})$

Proof. Note that (5.1.2) is a sum over scalar multiples $v^\mathsf{T} V_i$ and $X_i^\mathsf{T} \Gamma^{-1/2} x$—it is an inner product in $\mathbb{R}^T$. To use this observation, notice that we may write

$$\begin{bmatrix} x^\mathsf{T} \Gamma^{-1/2} X_1 \\ \vdots \\ x^\mathsf{T} \Gamma^{-1/2} X_i \end{bmatrix} = \operatorname{blkdiag}(x^\mathsf{T}) \mathbf{G} \mathbf{L} W_{1:n+1} \tag{5.1.3}$$

and similarly

$$\begin{bmatrix} v^\mathsf{T} V_1 \\ \vdots \\ v^\mathsf{T} V_i \end{bmatrix} = \operatorname{blkdiag}(v^\mathsf{T}) \mathbf{H} W_{1:n+1}. \tag{5.1.4}$$

Consequently,

$$\sum_{i=1}^{n} v^\mathsf{T} V_i X_i^\mathsf{T} \Gamma^{-1/2} x = W_{1:n+1}^\mathsf{T} \mathbf{H}^\mathsf{T} \operatorname{blkdiag}(v x^\mathsf{T}) \mathbf{G} \mathbf{L} W_{1:n+1} \tag{5.1.5}$$

as was required. ∎

Proposition 5.1.1. *Fix*

$$\begin{aligned} \sigma_G &= \max_{v \in \mathbb{S}^{d_\mathsf{Y}-1}, x \in \mathbb{S}^{d_\mathsf{X}-1}} \|\mathbf{H}^\mathsf{T} \operatorname{blkdiag}(v x^\mathsf{T}) \mathbf{G} \mathbf{L}\|_F \\ \sigma_E &= \max_{v \in \mathbb{S}^{d_\mathsf{Y}-1}, x \in \mathbb{S}^{d_\mathsf{X}-1}} \|\mathbf{H}^\mathsf{T} \operatorname{blkdiag}(v x^\mathsf{T}) \mathbf{G} \mathbf{L}\|_{\mathsf{op}}. \end{aligned} \tag{5.1.6}$$

There exists a universal positive constant c such that with probability $1-\delta$:

$$\left\|\sum_{i=1}^{n} V_i X_i^{\mathsf{T}} \Gamma^{-1/2}\right\|_{\mathsf{op}} \le cK^2\left(\sqrt{\sigma_G^2(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta))}+\sigma_E(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta))\right). \tag{5.1.7}$$

Proof. By Theorem 4.0.1 we have that:

$$\begin{aligned} &W_{1:n+1}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}W_{1:n+1} \\ &\quad\le cK^2\left(\sqrt{\|\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}\|_F^2\log(1/\delta)}+\|\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}\|_{\mathsf{op}}\log(1/\delta)\right). \end{aligned} \tag{5.1.8}$$

Consequently, a union bound and a standard discretization argument yields that:

$$\left\|\sum_{i=1}^{n} V_i X_i^{\mathsf{T}} \Gamma^{-1/2}\right\|_{\mathsf{op}} \le c'K^2\left(\sqrt{\sigma_G^2(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta'))}+\sigma_E(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta'))\right) \tag{5.1.9}$$

where $\delta' = (c'')^{d_{\mathsf{X}}+d_{\mathsf{Y}}}\delta$. Above, c, c', c'' are some fixed positive constants we have not bothered to determine. ∎

The reason for not bounding $\max_{v\in\mathbb{S}^{d_{\mathsf{Y}}-1},x\in\mathbb{S}^{d_{\mathsf{X}}-1}}\|\mathbf{H}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}\|_F^2 \le \|\mathbf{H}\|_{\mathsf{op}}^2\|\mathbf{G}\mathbf{L}\|_F^2$ is made clear by the following proposition. It gives a win in terms of dimensional factors over the naive bound. To build some intuition for why such a win is possible, note that (5.1.7) controls a random walk in W_i weighted by the *whitened covariates* $\tilde{X}_i = \Sigma_X^{-1/2}X_i, i\in[n]$ which have identity covariance.

Proposition 5.1.2. *Let σ_G, σ_E be as in (5.1.6). If $\Gamma = \frac{1}{n}\sum_{i=1}^n \mathbf{E}X_iX_i^{\mathsf{T}}$ we have that*

$$\sigma_G^2 \le \|\mathbf{H}\|_{\mathsf{op}}^2 n \qquad \text{and} \qquad \sigma_E^2 \le \|\mathbf{H}\|_{\mathsf{op}}^2(n\wedge\|\mathbf{G}\mathbf{L}\|_{\mathsf{op}}^2). \tag{5.1.10}$$

Proof. Since $\sigma_E^2 \le \sigma_G^2$ it suffices to prove the result for σ_G (operator norm is less than Frobenius norm). With that said, the proof is mostly linear algebra:

$$\begin{aligned} &\operatorname{tr}\left(\mathbf{L}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\operatorname{blkdiag}(xv^{\mathsf{T}})\mathbf{H}^{\mathsf{T}}\mathbf{H}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}\right) \\ &= \operatorname{tr}\left(\mathbf{E}W_{1:n+1}W_{1:n+1}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\operatorname{blkdiag}(xv^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}\right) && (\mathbf{E}W_{1:n+1}W_{1:n+1}^{\mathsf{T}} = I) \\ &= \mathbf{E}\operatorname{tr}\left(W_{1:n+1}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\operatorname{blkdiag}(xv^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}\mathbf{L}W_{1:n+1}\right) && (\text{trace cyclic property}) \\ &= \mathbf{E}\operatorname{tr}\left(X_{1:n}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\operatorname{blkdiag}(xv^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\mathbf{G}X_{1:n}\right) && (X_{1:n} = \mathbf{L}W_{1:n+1}) \\ &= \mathbf{E}\operatorname{tr}\left(\tilde{X}_{1:n}^{\mathsf{T}}\operatorname{blkdiag}(xv^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(vx^{\mathsf{T}})\tilde{X}_{1:n}\right) && (\tilde{X}_i = \Gamma^{-1/2}X_i, i\in[n]) \\ &\le \|\operatorname{blkdiag}(v^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(v)\|_{\mathsf{op}}\mathbf{E}\operatorname{tr}\left(\tilde{X}_{1:n}^{\mathsf{T}}\operatorname{blkdiag}(x)\operatorname{blkdiag}(x^{\mathsf{T}})\tilde{X}_{1:n}\right) && (*) \\ &= \|\operatorname{blkdiag}(v^{\mathsf{T}})\mathbf{H}\mathbf{H}^{\mathsf{T}}\operatorname{blkdiag}(v)\|_{\mathsf{op}}\mathbf{E}\operatorname{tr}\left((x^{\mathsf{T}}\tilde{X})_{1:n}^{\mathsf{T}}(x^{\mathsf{T}}\tilde{X})_{1:n}\right) \\ &\le \|\mathbf{H}\|_{\mathsf{op}}^2\mathbf{E}\operatorname{tr}\left(\sum_{i=1}^{n}(x^{\mathsf{T}}\tilde{X}_i)^2\right) \\ &= n\|\mathbf{H}\|_{\mathsf{op}}^2 && (**) \end{aligned} \tag{5.1.11}$$

where the last equality follows from the fact that the covariance on average is the identity, since $\Gamma^{1/2}$ is a whitening factor, and where $(*)$ follows from the observation that for any matrix $A : M \succeq 0 \Rightarrow \operatorname{tr}(A^\mathsf{T} M A) \leq \|M\|_{\mathsf{op}} \operatorname{tr}(A^\mathsf{T} A)$. To expand on the last part $(**)$ notice that:

$$\mathbf{E} \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^\mathsf{T} = \Gamma^{-1/2} \left(\mathbf{E} \sum_{i=1}^{n} X_i X_i^\mathsf{T} \right) \Gamma^{-1/2} = \Gamma^{-1/2} (T\Gamma) \Gamma^{-1/2} = nI. \tag{5.1.12}$$

The proof is thus finished by noting that $x \in \mathbb{S}^{d_\mathsf{X}-1}$ meaning that the term in $(**)$ is just $n\|\mathbf{H}\|_{\mathsf{op}}^2 \|x\|^2 = n\|\mathbf{H}\|_{\mathsf{op}}^2$. ■

5.1.2. The Lower Tail

We next turn to analyzing the lower tail. Our first observation is, in complete analogy with Lemma 5.1.1, that also the empirical covariance matrix is a quadratic form in the noise source $W_{1:n+1}$.

Lemma 5.1.2. *For any $x \in \mathbb{R}^{d_\mathsf{X}}$ we have that:*

$$\sum_{i=1}^{n} x^\mathsf{T} \Gamma^{-1/2} X_i X_i^\mathsf{T} \Gamma^{-1/2} x = W_{1:n+1}^\mathsf{T} \mathbf{L}^\mathsf{T} \mathbf{G}^\mathsf{T} \operatorname{blkdiag}(xx^\mathsf{T}) \mathbf{G} \mathbf{L} W_{1:n+1}. \tag{5.1.13}$$

It will be convenient to define $\mathbf{M}_x \triangleq \mathbf{L}^\mathsf{T} \mathbf{G}^\mathsf{T} \operatorname{blkdiag}(xx^\mathsf{T}) \mathbf{G} \mathbf{L}$ as this is the corresponding Hanson-Wright weighting matrix. We also observe that virtually the same proof as that of Proposition 5.1.2 gives that for every $x \in \mathbb{S}^{d_\mathsf{X}-1}$: $\operatorname{tr} \mathbf{M}_x \leq n$ and $\|\mathbf{M}_x\|_{\mathsf{op}} \leq \|\mathbf{G}\mathbf{L}\|_{\mathsf{op}}^2$.

Proposition 5.1.3. *Fix $\Gamma = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} X_i X_i^\mathsf{T}$ and assume that $\Gamma \succ 0$. There exist universal positive constants $c, c' > 0$ such that if*

$$n \geq cK^4 \max_{v \in \mathbb{S}^{d_\mathsf{X}-1}} \|\mathbf{M}_x\|_{\mathsf{op}} (d_\mathsf{X} + \log(1/\delta)) \tag{5.1.14}$$

then with probability at least $1 - \delta$:

$$\lambda_{\min} \left(\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^\mathsf{T} \right) \geq c' n. \tag{5.1.15}$$

Regarding our earlier remark that it is essentially without loss of generality to assume that $\frac{1}{n} \sum_{i=1}^{n} \mathbf{E} X_i X_i^\mathsf{T}$ has full rank, we point out that the proposition remains true with $\lambda_{\min}$ replaced by the smallest nonzero eigenvalue if this matrix is rank deficient. All the subsequent analysis still follows restricting from $\mathbb{R}^{d_\mathsf{X}}$ to the range of $\frac{1}{n} \sum_{i=1}^{n} \mathbf{E} X_i X_i^\mathsf{T}$.

Proof. We begin by noticing that

$$\sum_{i=1}^{n} x^\mathsf{T} \Gamma^{-1/2} X_i X_i^\mathsf{T} \Gamma^{-1/2} x = W_{1:n+1}^\mathsf{T} \mathbf{L}^\mathsf{T} \mathbf{G}^\mathsf{T} \operatorname{blkdiag}(xx^\mathsf{T}) \mathbf{G} \mathbf{L} W_{1:n+1} \tag{5.1.16}$$

For $\mathbf{M}_x \triangleq \mathbf{L}^\mathsf{T}\mathbf{G}^\mathsf{T}\,\mathrm{blkdiag}(xx^\mathsf{T})\mathbf{G}\mathbf{L}$ Proposition 4.1.1 informs us that:

$$\mathbf{E}\exp\left(-\lambda W_{1:n+1}^\mathsf{T}\mathbf{M}_x W_{1:n+1}\right) \leq \mathbf{E}\exp\left(-\lambda\,\mathrm{tr}\,\mathbf{M}_x + \frac{72\lambda^2K^4\,\mathrm{tr}(\mathbf{M}_x^2)}{2}\right) \tag{5.1.17}$$

as long as $|\lambda| \leq \frac{1}{8\sqrt{2}K^2\|\mathbf{M}_x\|_{\mathsf{op}}}$. A Chernoff bound thus yields for $t \geq 0$:

$$\mathbf{P}\left(W_{1:n+1}^\mathsf{T}\mathbf{M}_x W_{1:n+1} \leq \mathrm{tr}\,\mathbf{M}_x - t\right) \leq \exp\left(\frac{-t^2}{144K^4\,\mathrm{tr}(\mathbf{M}_x^2)}\right) \tag{5.1.18}$$

Let now $t = \frac{1}{2}\,\mathrm{tr}\,\mathbf{M}_x$ and note that $\mathrm{tr}(\mathbf{M}_x^2) \leq \|\mathbf{M}_x\|_{\mathsf{op}}\,\mathrm{tr}\,\mathbf{M}_x$. We find:

$$\mathbf{P}\left(W_{1:n+1}^\mathsf{T}\mathbf{M}_x W_{1:n+1} \leq \frac{1}{2}\,\mathrm{tr}\,\mathbf{M}_x\right) \leq \exp\left(\frac{-\,\mathrm{tr}(\mathbf{M}_x)}{576K^4\|\mathbf{M}_x\|_{\mathsf{op}}}\right). \tag{5.1.19}$$

It will be convenient to define $\kappa_{\tilde{X}}(n) \triangleq \inf_v \frac{\mathrm{tr}(\mathbf{M}_x)}{\|\mathbf{M}_x\|_{\mathsf{op}}}$. In fact by our previous whitening argument $\mathrm{tr}\,\mathbf{M}_x = n$ and so $\kappa_{\tilde{X}}(n) = \frac{n}{\max_{v\in\mathbb{S}^{d_\mathsf{X}-1}}\|\mathbf{M}_x\|_{\mathsf{op}}}$. Hence if we invoke our earlier discretization argument from Lemma 3.2.2, we find that for any $\varepsilon > 0$ with probability at least $1 - \left(1+\frac{2}{\varepsilon}\right)^{d_\mathsf{X}}\exp\left(\frac{-\kappa_X(n)}{576K^4}\right)$

$$\lambda_{\min}\left(\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^\mathsf{T}\right) \geq \frac{n}{2(1-\varepsilon)} - \varepsilon\left\|\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^\mathsf{T}\right\|_{\mathsf{op}} \tag{5.1.20}$$

A Chernoff argument with roles of λ and $-\lambda$ reversed further yields that:

$$\mathbf{P}\left(W_{1:n+1}^\mathsf{T}\mathbf{M}_x W_{1:n+1} \geq \frac{3n}{2}\right) \leq \exp\left(\frac{-n}{576K^4\|\mathbf{M}_x\|_{\mathsf{op}}}\right) \tag{5.1.21}$$

Consequently, for $\varepsilon' > 0$ and with probability at least $1 - \left(1+\frac{2}{\varepsilon'}\right)^{d_\mathsf{X}}\exp\left(\frac{-\kappa_X(n)}{576K^4}\right)$ a union bound yields

$$\left\|\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^\mathsf{T}\right\|_{\mathsf{op}} \leq \frac{3n}{2(1-2\varepsilon')}. \tag{5.1.22}$$

Putting everything together, we have that with probability $1-\left(1+\frac{2}{\varepsilon}\right)^{d_\mathsf{X}}\exp\left(\frac{-\kappa_X(n)}{576K^4}\right)-\left(1+\frac{2}{\varepsilon'}\right)^{d_\mathsf{X}}\exp\left(\frac{-\kappa_X(n)}{576K^4}\right)$:

$$\lambda_{\min}\left(\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^\mathsf{T}\right) \geq \frac{n}{2(1-\varepsilon)} - \frac{3n\varepsilon}{2(1-2\varepsilon')} \tag{5.1.23}$$

We can certainly pick $\varepsilon = \varepsilon'$ sufficiently small such for some universal positive constants $c, c' > 0$

$$\lambda_{\min}\left(\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^\mathsf{T}\right) \geq c'T \tag{5.1.24}$$

with probability at least $1 - 2\exp\left(cd_\mathsf{X} - \frac{\kappa_X(n)}{576K^4}\right)$. The result follows by probability inversion. ■

The obvious question now is whether $\max_{x\in\mathbb{S}^{d_\mathsf{X}-1}}\|\mathbf{M}_x\|_{\mathsf{op}}$ is bounded for any reasonable system model, such as, e.g., (5.0.3)-(5.0.4). In the sequel, we will see that this is true if the system (5.0.3) is *stable*: the spectral radius of $A_\star$ satisfies $\rho(A_\star) < 1$.

5.1.3. Guarantees for Linear Regression with Dependence

The main result of this chapter is as follows.

Theorem 5.1.1. *Fix $\Gamma = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E} X_i X_i^{\mathsf{T}}$ and assume that $\Gamma \succ 0$. There exist universal positive constants c, c' such that if*

$$n \geq cK^4(1 + \|\mathbf{H}\|_{\mathsf{op}}^2) \max_{x \in \mathbb{S}^{d_{\mathsf{X}}-1}} \|\mathbf{M}_x\|_{\mathsf{op}}(d_{\mathsf{X}} + \log(1/\delta)) \tag{5.1.25}$$

then it holds with probability at least $1 - \delta$ that

$$\|(\widehat{\theta} - \theta_\star)\sqrt{\Gamma}\|_{\mathsf{op}} \leq \frac{c' K^2 \|\mathbf{H}\|_{\mathsf{op}} \sqrt{(d_{\mathsf{X}} + d_{\mathsf{Y}}) + \log(1/\delta)}}{\sqrt{n}}, \tag{5.1.26}$$

Proof. The proof follows by combining Proposition 5.1.1 and Proposition 5.1.2 with Proposition 5.1.3. That is, we note that as long as the design is invertible:

$$\begin{aligned}
\left\|(\widehat{\theta} - \theta_\star)\sqrt{\Gamma}\right\|_{\mathsf{op}} &= \left\| \left(\sum_{i=1}^{n} V_i X_i^{\mathsf{T}} \Gamma^{-1/2} \right) \left(\sum_{k=1}^{n} \Gamma^{-1/2} X_k X_k^{\mathsf{T}} \Gamma^{-1/2} \right)^{-1} \right\|_{\mathsf{op}} \\
&\leq \frac{1}{\lambda_{\min}\left(\sum_{k=1}^{n} \Gamma^{-1/2} X_k X_k^{\mathsf{T}} \Gamma^{-1/2}\right)} \left\| \left(\sum_{i=1}^{n} V_i X_i^{\mathsf{T}} \Gamma^{-1/2} \right) \right\|_{\mathsf{op}}
\end{aligned} \tag{5.1.27}$$

and we use the above propositions to control each parenthesis individually. To obtain the correct burn-in, notice that $\sigma_E^2 \leq \|\mathbf{H}\|_{\mathsf{op}}^2 \max_{x \in \mathbb{S}^{d_{\mathsf{X}}-1}} \|\mathbf{M}_x\|_{\mathsf{op}}$. Consequently, once the burn-in (5.1.25) is satisfied, we have that 1), the sub-Gaussian term in Proposition 5.1.1 is dominant and 2) the lower bound of Proposition 5.1.3 holds. ∎

5.2. Elements of Linear System Identification

We next prove that the term $\mathbf{GL}$ above is well-behaved when $X_{1:n}$ is a linear auto-regression. In this section we fix $\Gamma = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{t-1} A_\star^k (A_\star^{\mathsf{T}})^k$.

Lemma 5.2.1. *Let $X_{1:n}$ satisfy $X_{i+1} = A_\star X_i + W_{i+1}, i \in [n-1]$ with $X_1 = W_1$. If we define for $A \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$ $\mathrm{stab}_n(A) \triangleq \sum_{i=0}^{n-1} \|\Gamma^{-1/2} A^i\|_{\mathsf{op}}$, we have that:*

$$\|\mathbf{GL}_{A_\star, n}\|_{\mathsf{op}} \leq \mathrm{stab}_n(A_\star) \tag{5.2.1}$$

where $\mathbf{L}$ is as in (5.0.4) and as before $\mathbf{G} = \mathrm{blkdiag}(\Gamma^{-1/2})$.

Proof. Let $\mathbf{Z}$ be the downshift operator:

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ I & 0 & 0 & \dots & 0 \\ 0 & I & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & I & 0 \end{bmatrix} \tag{5.2.2}$$

Note that we may bound $\mathbf{GL}_{A_\star,n}$ in terms of its (sub-)diagonal decomposition:

$$\begin{aligned}\|\mathbf{GL}_{A_\star,n}\|_{\mathsf{op}} &= \left\|\sum_{i=0}^{n-1} \mathbf{GZ}^i \operatorname{blkdiag}(A_\star^i)\right\|_{\mathsf{op}} \\ &\le \sum_{i=0}^{n-1} \left\|\mathbf{GZ}^i \operatorname{blkdiag}(A_\star^i)\right\|_{\mathsf{op}}\end{aligned} \tag{5.2.3}$$

Now, the matrices $\mathbf{GZ}^i \operatorname{blkdiag}(A_\star^i)$ are constant along the block (sub-)diagonal, and hence the operator norm is the operator norm of the entries of the block-sub-diagonal:

$$\left\|\mathbf{GZ}^i \operatorname{blkdiag}(A_\star^i)\right\|_{\mathsf{op}} = \|\Gamma^{-1/2} A_\star^i\|_{\mathsf{op}}. \tag{5.2.4}$$

Hence

$$\|\mathbf{GL}_{A_\star,n}\|_{\mathsf{op}} \le \sum_{i=0}^{n-1} \|\Gamma^{-1/2} A_\star^i\|_{\mathsf{op}} \tag{5.2.5}$$

which gives the result. ■

Observe that by Gel'fand's Formula, whenever $\rho(A_\star) < 1$, we have that $\lim_{n\to\infty} \sum_{i=0}^{n-1} \|\Gamma^{-1/2} A_\star^i\|_{\mathsf{op}} < \infty$. Consequently, we may instantiate our main result of this chapter, Theorem 5.1.1, to obtain guarantees for stable linear systems.

Corollary 5.2.1. *Suppose that $\rho(A_\star) < 1$. There exists a universal positive constant $c > 0$ and a constant $C_{\mathrm{sys}}(A_\star)$ only depending on $A_\star$ such that if*

$$n \ge C_{\mathrm{sys}}(A_\star) K^4 (d_{\mathsf{X}} + \log(1/\delta)) \tag{5.2.6}$$

then

$$\|(\widehat{A} - A_\star)\sqrt{\Gamma}\|_{\mathsf{op}} \le \frac{cK^2\sqrt{d_{\mathsf{X}} + \log(1/\delta)}}{\sqrt{n}} \tag{5.2.7}$$

In particular, as long as $\rho(A_\star) < 1$, we can take $C_{\mathrm{sys}}(A_\star)$ to be polynomial in $\lim_{n\to\infty} \mathrm{stab}_n(A_\star)$.

For a linear dynamical system, the matrix $\Gamma = \frac{1}{n}\sum_{i=1}^n \mathbf{E} X_i X_i^{\mathsf{T}}$ can be interpred as an average over so-called controllability (reachability) gramians. Namely

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E} X_i X_i^{\mathsf{T}} = \frac{1}{n}\sum_{i=1}^n \sum_{j=0}^{i-1} (A^\star)^j \Sigma_W (A^{\star,\mathsf{T}})^j \tag{5.2.8}$$

where we have suggestively reintroduced the covariance $\Sigma_W = I$ of the W_i. The point is that (5.2.8) measures amplification of the noise to the state. In particular, if (5.2.8) is full rank (which we have assumed so far, but can get around), parameter recovery is possible.

Let us also note that the assumption that $A_\star$ is stable is necessary for the above to work. Indeed, if it is not we may find that $\mathrm{stab}_n(A_\star) \to \infty$. We will see in Chapter 6 how to circumvent this

assumption. To understand why this is necessary, consider the scalar case with $A_\star = 1$. Then $\Gamma = cn$ for some constant $c > 0$. Hence

$$\mathrm{stab}_n(1) = \sum_{i=0}^{n-1} (cn)^{-1/2} = \frac{\sqrt{n}}{c} \tag{5.2.9}$$

and hence the burn-in cannot be satisfied in general.

There is a clear conflict here. On the one hand, the matrix Γ, which weights $\|(\widehat{\theta} - \theta_\star)\sqrt{\Gamma}\|_{\mathsf{op}}$ *grows* as we loose stability: the identification error $\widehat{\theta} - \theta_\star$ becomes *smaller* with less stability. On the other hand, as we just noted, our burn-in requirement (5.2.6) *also grows* as we loose stability. From the perspective of our proof, the reason this somewhat contradictory phenomenon occurs is because:

- The lower tail of $\sum_{i=1}^n X_i X_i^\mathsf{T}$ in the "denominator" is larger for less stable systems. Less stability translates to a larger signal.
- Unfortunately, the deviations of $\sum_{i=1}^n X_i X_i^\mathsf{T}$ also grow as stability is lost, and in particular if $\rho(A_\star) \geq 1$, the normalized variable $\frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$ no longer concentrates around its mean. It does not satisfy a law of large numbers at scale T.

The general situation of learning in the marginally stable regime $\rho(A_\star) \approx 1$ is quite subtle and a general sharp understanding is still lacking from the literature. However, there has been some partial progress in two situations: 1) when one has access to several independently drawn trajectories from the same system; and 2) when a certain condition number of the system is not too large. In fact, our current analysis is sufficient to cover the first case of many trajectories. In Chapter 6 we will see how to tackle the second situation by providing a refined analysis of the lower tail of the empirical covariance matrix—it exhibits some, milder, degree of anti-concentration even if we only have access to a single trajectory from a marginally stable system. In either case, it should be made clear that *there is signal in the learning problem even without stability.*

5.3. Learning from many Trajectories

From the above discussion, it is clear that the key issue is the concentration of $\frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$. A simple and rather natural fix to this is simply to assume that we have many independent draws from the same dynamical system—the experiment we try to learn from is being repeated at the trajectory level. To make this precise, suppose we have m independent trajectories, each of length T, so that our total number of data points is $n = mT$. In other words, for fixed $j \in [m]$ our statistical model is described by:

$$X_t^{(j)} = A_\star X_{t-1}^{(j)} + W_t^{(j)}, \quad X_1^{(j)} = W_1^{(j)}, \quad t = 2, \ldots, T+1. \tag{5.3.1}$$

The notation is chosen here such that key vectors $X_{1:n}$, $Y_{1:n}$, and $V_{1:n}$ satisfy

$$X_{1:n} = \begin{bmatrix} X_{1:T}^{(1)} \\ \vdots \\ X_{1:T}^{(m)} \end{bmatrix}, \quad Y_{1:n} = \begin{bmatrix} X_{2:T+1}^{(1)} \\ \vdots \\ X_{2:T+1}^{(m)} \end{bmatrix} \quad \text{and} \quad V_{1:n} = \begin{bmatrix} W_{2:T+1}^{(1)} \\ \vdots \\ W_{2:T+1}^{(m)} \end{bmatrix} \tag{5.3.2}$$

Just as before, this can also concisely be written in our linear form $X_{1:n} = \mathbf{L}W_{1:n+m}$ and $V_{1:n} = \mathbf{H}W_{1:n+m}$. Note that our total number of noise source variables is $(n+m)d_{\mathsf{W}}$ instead of $(n+1)d_{\mathsf{W}}$ for the bookkeeping to work out (this is because we require m initial conditions). The matrices $\mathbf{L}$ and $\mathbf{H}$ actually have a particularly well-structured form for this setup. Namely, we have that

$$\mathbf{L} = \operatorname{blkdiag}_{1:m}(\mathbf{L}_{A_\star,T}) \text{ and } \mathbf{H} = \operatorname{blkdiag}_{1:m}(\mathbf{H}_{\mathrm{LDS}}) \tag{5.3.3}$$

where $\mathbf{L}_{A_\star,T}$ and $\mathbf{H}_{\mathrm{LDS}}$ are given by (5.0.4) (with $k = t$).

The fact that the block-diagonal appears in (5.3.3) is very convenient. The operator norm of $\mathbf{GL}$ for this choice of $\mathbf{L}$ is particularly simple. As before, $\mathbf{G} = \operatorname{blkdiag}(\Gamma^{-1/2})$ where $\Gamma = \frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T} = \frac{1}{T}\sum_{t=1}^T X_t^j (X_t^j)^\mathsf{T}$ for any j. Indeed, since for any matrix M, $\|\operatorname{blkdiag}(M)\|_{\mathsf{op}} = \|M\|_{\mathsf{op}}$ we have for our choice

$$\|\mathbf{GL}\|_{\mathsf{op}} = \|\mathbf{GL}_{A_\star,T}\|_{\mathsf{op}} \le \sqrt{T}. \tag{5.3.4}$$

These observations immediately yield a meaningful guarantee using Theorem 5.1.1. In fact, the key is precisely that $\mathbf{L}$ and $\mathbf{H}$ have block-diagonal structure since this is precisely what is required for (5.3.4) to be true. We have thus established the following corollary to Theorem 5.1.1.

Corollary 5.3.1. *Let $n = mT$ and suppose that there exist matrices $\mathbf{l}$ and $\mathbf{h}$ such that $\mathbf{L} = \operatorname{blkdiag}_{1:m}(\mathbf{l})$ and $\mathbf{H} = \operatorname{blkdiag}_{1:m}(\mathbf{h})$. Set $\Gamma = \frac{1}{n}\sum_{i=1}^n \mathbf{E} X_i X_i^\mathsf{T}$. There exist universal positive constants c, c' such that if*

$$m \ge cK^4(1 + \|\mathbf{h}\|_{\mathsf{op}}^2)(d_{\mathsf{X}} + \log(1/\delta)) \tag{5.3.5}$$

then it holds with probability at least $1-\delta$ that

$$\|(\widehat{\theta} - \theta_\star)\sqrt{\Gamma}\|_{\mathsf{op}} \le \frac{c' K^2 \sqrt{\|\mathbf{h}\|_{\mathsf{op}}^2 (d_{\mathsf{X}} + d_{\mathsf{Y}} + \log(1/\delta))}}{\sqrt{n}}. \tag{5.3.6}$$

Proof. Instantiate Theorem 5.1.1 and note that

$$\|\mathbf{M}_x\|_{\mathsf{op}} = \|\mathbf{L}^\mathsf{T}\mathbf{G}^\mathsf{T} \operatorname{blkdiag}(xx^\mathsf{T})\mathbf{GL}\|_{\mathsf{op}} = \|\mathbf{l}^\mathsf{T}\mathbf{g}^\mathsf{T} \operatorname{blkdiag}(xx^\mathsf{T})\mathbf{gl}\|_{\mathsf{op}} \le T \tag{5.3.7}$$

where $\mathbf{g}$ is defined in the same way as $\mathbf{G}$ but with only as many block-diagonals as $\mathbf{l}$ (instead of matching those of L). Similarly, $\|\mathbf{H}\|_{\mathsf{op}} = \|\mathbf{h}\|_{\mathsf{op}}$. ∎

5.4. Notes

The idea to use the Hanson-Wright inequality to control the empirical covariance matrix for stable linear system identification is due to Jedra and Proutiere [2022]. However, our analysis of the random walk term differs from those found in the literature in that we again use the Hanson-Wright inequality here. The more "standard" argument proceeds via the method of self-normalized martingales [Peña et al., 2009]. As for the situation with many trajectories, we refer to Tu et al. [2024] for further reading. Their proofs are again based on a combination of a martingale argument and the more advanced small-ball technique (to control the lower tail) introduced to learning theory by Mendelson [2014] and popularized in system identification by Simchowitz et al. [2018]. Our Corollary 5.3.1 is similar in spirit to their results but our proof technique differs from theirs and is based entirely on our simplified approach using the Hanson-Wright inequality.

6. Beyond Stability: The Lower Tail Revisited

Recall that our outline of the analysis of the least squares estimator in Equation (5.0.2) consists of two main components, one of which being the lower tail of the empirical covariance matrix $\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}}$. In this section we provide an alternative analysis of this random matrix for a class of "causal" systems. Moreover, we will emphasize only the lower tail of this random matrix as to sidestep issues with bounds degrading with the stability of the system considered, cf. the requirement of $\rho(A_\star) < 1$ in Corollary 5.2.1. This allows us to quantitatively separate the notions of persistence of excitation and stability.

6.1. Causal Processes

Let us now carry out this program. Fix two integers T and k such that $T/k \in \mathbb{N}$. We consider causal processes of the form $X_{1:T} = (X_1^{\mathsf{T}}, \dots, X_T^{\mathsf{T}})^{\mathsf{T}}$ evolving on $\mathbb{R}^{d_{\mathsf{X}}}$. More precisely, we assume the existence of an isotropic sub-Gaussian process evolving on $\mathbb{R}^{d_{\mathsf{W}}}$, $W_{1:T+1}$ with $\mathbf{E} W_{1:T} W_{1:T}^{\mathsf{T}} = I_{d_{\mathsf{W}} T}$ and a (block-) lower-triangular matrix $\mathbf{L} \in \mathbb{R}^{d_{\mathsf{X}} T \times d_{\mathsf{X}} T}$ such that

$$X_{1:T} = \mathbf{L} W_{1:T}. \tag{6.1.1}$$

We will assume that all the pT-many entries of $W_{1:T}$ are independent K^2-sub-Gaussian for some positive $K \in \mathbb{R}$.

We say that $X_{1:T}$ is k-causal if the matrix $\mathbf{L}$ has the block lower-triangular form:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 & 0 \\ \mathbf{L}_{2,1} & \mathbf{L}_{2,2} & 0 & 0 & 0 \\ \mathbf{L}_{3,1} & \mathbf{L}_{3,2} & \mathbf{L}_{3,3} & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{L}_{T/k,1} & \dots & \dots & \dots & \dots \mathbf{L}_{T/k,T/k} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \mathbf{L}_3 \\ \vdots \\ \mathbf{L}_{T/k} \end{bmatrix} \tag{6.1.2}$$

where each $\mathbf{L}_{ij} \in \mathbb{R}^{dk \times pk}, i, j \in [T/k] \triangleq \{1, 2, \dots, T/k\}$. In brief, we say that $X_{1:T}$ satisfying the above construction is k-causal with independent K^2-sub-Gaussian increments.

Obviously, every 1-causal process is k-causal for every $k \in \mathbb{N}$ as long as the divisibility condition holds. To analyze the lower tail of the empirical covariance of $X_{1:T}$ we will also associate a decoupled random process

$$\tilde{X}_{1:T} = \mathrm{blkdiag}(\mathbf{L}_{11}, \dots, \mathbf{L}_{T/k,T/k}) W_{1:T}.$$

Hence, the process $\tilde{X}_{1:T}$ is generated in much the same way as $X_{1:T}$ but by removing the sub-diagonal

entries of $\mathbf{L}$:

$$\tilde{\mathbf{L}} \triangleq \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 \\ 0 & \mathbf{L}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{L}_{T/k,T/k} \end{bmatrix} \implies \tilde{X}_{1:T} = \tilde{\mathbf{L}} W_{1:T}.$$

We emphasize that by our assumptions on $W_{1:T}$ and the block-diagonal structure of $\tilde{\mathbf{L}}$ the variables $\tilde{X}_{1:k}, \tilde{X}_{k+1:2k}, \dots, \tilde{X}_{T-k+1:T}$ are all independent of each other; they have been decoupled. This decoupled process will effectively dictate our lower bound, and we will show under relatively mild assumptions that

$$\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}}\right) \gtrsim \lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t \tilde{X}_t^{\mathsf{T}}\right) \tag{6.1.3}$$

with probability that approaches 1 at an exponential rate in the sample size T. More precisely, the following statement is the main result of this chapter.

Theorem 6.1.1. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process taking values in $\mathbb{R}^{d_{\mathsf{x}}}$ with K^2-sub-Gaussian increments. Suppose further that the diagonal blocks are all equal: $\mathbf{L}_{j,j} = \mathbf{L}_{1,1}$ for all $j \in [T/k]$. Suppose $\lambda_{\min}\left(\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right) > 0$. We have that:*

$$\mathbf{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \not\succeq \frac{1}{8T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right) \leq (C_{\mathsf{sys}})^{d_{\mathsf{x}}} \exp\left(-\frac{T}{576K^2k}\right) \tag{6.1.4}$$

where

$$C_{\mathsf{sys}} \triangleq 1 + 4\sqrt{2}\frac{\left(\frac{T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{18k\lambda_{\min}\left(\sum_{t=1}^{T}\mathbf{E}X_tX_t^{\mathsf{T}}\right)} + 9\right)\lambda_{\max}\left(\sum_{t=1}^{T}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}{\lambda_{\min}\left(\sum_{t=1}^{T}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}. \tag{6.1.5}$$

To parse Theorem 6.1.1, note that it simply informs us that there exist a a system-dependent constant C_{sys}—which itself has no more than polynomial dependence on relevant quantities—such that if

$$T/k \geq 576K^2(d\log C_{\mathsf{sys}} + \log(1/\delta)) \tag{6.1.6}$$

then on an event with probability mass at least $1-\delta$:

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \succeq \frac{1}{8T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}.$$

The proof of Theorem 6.1.1 is quite similar to what we have just seen in the proof Proposition 5.1.3. Indeed, we again rely on the Hanson-Wright inequality. However, the trick is to use a decoupling argument—summarized in Proposition 6.1.1—together with the tower property of conditional expectation to prove an exponential inequality which does not saturate for unstable systems.

Remark 6.1.1. *Since the blocks of $\mathbf{L}$ can be regarded to specify the noise-to-output map, the assumption that the diagonal blocks are constant is for instance satisfied by linear time-invariant (LTI) systems. The assumption can be removed at the cost of a more complicated expression.*

The next example serves as the archetype for the reduction from $\mathbf{L}$ to $\tilde{\mathbf{L}}$.

Example 6.1.1. *Suppose that* (6.1.1) *is specified via*

$$X_t = A_\star X_{t-1} + B_\star W_t \tag{6.1.7}$$

for $t \in [T]$ *and where* $(A_\star, B_\star) \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}} + d_{\mathsf{X}} \times d_{\mathsf{W}}}$. *We set* $d = d_{\mathsf{X}}$ *and* $p = d_{\mathsf{W}}$ *in the theorem above. The reduction from* $X_{1:T} = \mathbf{L} W_{1:T}$ *to* $\tilde{X}_{1:T} = \mathrm{blkdiag}(\mathbf{L}_{11}, \dots, \mathbf{L}_{T/k,T/k}) W_{1:T}$ *corresponds to replacing a single trajectory from the linear system* (6.1.7) *of length* T *by* T/k *trajectories of length* k *each and sampled independently of each other. The price we pay for decoupling these systems is that our lower bound is dictated by the gramians up to range* k*:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t \tilde{X}_t^{\mathsf{T}} = \frac{1}{k}\sum_{t=1}^{k} \mathbf{E}\tilde{X}_t \tilde{X}_t^{\mathsf{T}} = \frac{1}{k}\sum_{t=1}^{k}\sum_{j=0}^{t-1} (A^\star)^j B^\star B^{\star,\mathsf{T}} (A^{\star,\mathsf{T}})^j \tag{6.1.8}$$

instead of the gramians up to range T*:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E} X_t X_t^{\mathsf{T}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{t-1} (A^\star)^j B^\star B^{\star,\mathsf{T}} (A^{\star,\mathsf{T}})^j. \tag{6.1.9}$$

Put differently, the reduction from $\mathbf{L}$ *to* $\tilde{\mathbf{L}}$ *can be thought of as restarting the system every* k *steps.*

Comparing with Corollary 5.2.1, the advantage of Theorem 6.1.1 is that it allows us to provide persistence-of-excitation type guarantees that do not rely strongly on the stability of the underlying system. While the proof strategy yielding Corollary 5.2.1 is able to give two-sided concentration results, it comes at the cost of the guarantees becoming vacuous as the spectral radius of $A^\star$ in Example 6.1.1 tends to marginal stability (tends to 1). By contrast, Theorem 6.1.1 does not exhibit such a blow-up since the dependence on C_{sys} in (6.1.6) is logarithmic (instead of polynomial). The distinction might seem small, but it is qualitatively important as it (almost) decouples the phenomena of stability and persistence of excitation.

6.1.1. A Decoupling Inequality for sub-Gaussian Quadratic Forms

Our proof of Theorem 6.1.1 will make heavy use of Proposition 6.1.1 below. This is the crucial probabilistic inequality that allows us to decouple—or restart as discussed in Example 6.1.1.

Proposition 6.1.1. *Fix* $K \geq 1$, $x \in \mathbb{R}^n$ *and a symmetric positive semidefinite* $Q \in \mathbb{R}^{(n+m)\times(n+m)}$ *of the form* $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ *with* $Q_{22} \succ 0$. *Let* W *be an* m*-dimensional mean zero, isotropic and* K^2*-sub-Gaussian random vector with independent entries. Then for every* $\lambda \in \left[0, \frac{1}{8\sqrt{2}K^2\|Q_{22}\|_{\mathsf{op}}}\right]$ *it holds true that:*

$$\mathbf{E}\exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right) \leq \exp\left(-\lambda \operatorname{tr} Q_{22} + 36K^4\lambda^2 \operatorname{tr} Q_{22}^2\right). \tag{6.1.10}$$

By combining Lemma 6.1.1 below with the exponential form of Hanson-Wright (Proposition 4.1.1) we obtain the exponential inequality (6.1.10), which in the sequel will allow us to control the lower tail of the conditionally random quadratic form

$$\begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}.$$

We point out that (6.1.10) is not the best possible if the entries of are W independent and Gaussian as opposed to just isotropic and sub-Gaussian. In this case, the factor $36K^4\lambda^2(\operatorname{tr} Q_{22})^2$ in (6.1.10) can be improved to $\frac{\lambda^2}{2}\operatorname{tr} Q_{22}^2$ and the inequality can be shown to hold for the entire range of non-negative λ [Ziemann, 2023, Lemma 2.1]. Irrespectively, we will see in the sequel that it captures the correct qualitative behavior.

Lemma 6.1.1 (sub-Gaussian Decoupling)**.** *Fix $K \geq 1$, $x \in \mathbb{R}^n$ and a symmetric positive semidefinite $Q \in \mathbb{R}^{(n+m)\times(n+m)}$ of the form $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$. Let W be an m-dimensional mean zero and K^2-sub-Gaussian random vector. Then for every $\lambda \in \left[0, \frac{1}{4K^2\|Q_{22}\|_{\mathsf{op}}}\right]$ it holds true that:*

$$\mathbf{E}\exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right) \leq \sqrt{\mathbf{E}\exp\left(-2\lambda W^{\mathsf{T}} Q_{22} W\right)}. \tag{6.1.11}$$

Proof. First, we remark that we may prove the lemma under the additional hypothesis that $Q_{22} \succ 0$ without loss of generality by regrouping terms. We now proceed to prove the lemma under this additional hypothesis.

Let us introduce the new variable $\mu = Q_{22}^{-1/2} Q_{12} x$. Rewriting our quadratic form in terms of this new variable yields:

$$\begin{aligned}
&\mathbf{E}\exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right) \\
&= \mathbf{E}\exp\left(-\lambda x^{\mathsf{T}} Q_{11} x - \lambda W^{\mathsf{T}} Q_{22} W + 2\lambda x^{\mathsf{T}} Q_{12}^{\mathsf{T}} W\right) \\
&= \mathbf{E}\exp\left(-\lambda x^{\mathsf{T}} Q_{11} x - \lambda W^{\mathsf{T}} Q_{22} W + 2\lambda \mu^{\mathsf{T}} Q_{22}^{1/2} W - \lambda \mu^{\mathsf{T}}\mu + \lambda x^{\mathsf{T}} Q_{12}^{\mathsf{T}} Q_{22}^{-1} Q_{12} x\right) \\
&= \mathbf{E}\exp\left(-\lambda x^{\mathsf{T}}(Q_{11} - Q_{12}^{\mathsf{T}} Q_{22}^{-1} Q_{12}) x - \lambda W^{\mathsf{T}} Q_{22} W + 2\lambda \mu^{\mathsf{T}} Q_{22}^{1/2} W - \lambda \mu^{\mathsf{T}}\mu\right) \\
&\leq \mathbf{E}\exp\left(-\lambda \begin{bmatrix} \mu \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} I_n & Q_{22}^{1/2} \\ Q_{22}^{1/2} & Q_{22} \end{bmatrix} \begin{bmatrix} \mu \\ W \end{bmatrix}\right).
\end{aligned} \tag{6.1.12}$$

where the last inequality uses the fact that $(Q_{11} - Q_{12}^{\mathsf{T}} Q_{22}^{-1} Q_{12})$ is the Schur complement of Q_{22} in the positive semidefinite matrix Q and hence positive semidefinite.

To finish the proof, we note that

$$\begin{aligned}
&\mathbf{E}\exp\left(-\lambda\begin{bmatrix}\mu\\ W\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}I_n & Q_{22}^{1/2}\\ Q_{22}^{1/2} & Q_{22}\end{bmatrix}\begin{bmatrix}\mu\\ W\end{bmatrix}\right)\\
&= \mathbf{E}\exp\left(-2\lambda\mu^{\mathsf{T}}\sqrt{Q_{22}}W - \lambda\mu^{\mathsf{T}}\mu - \lambda W^{\mathsf{T}}Q_{22}W\right)\\
&\le \sqrt{\mathbf{E}\exp\left(-2\lambda\mu^{\mathsf{T}}\sqrt{Q_{22}}W - \lambda\mu^{\mathsf{T}}\mu\right)}\sqrt{\mathbf{E}\exp\left(-\lambda W^{\mathsf{T}}Q_{22}W\right)} \quad \text{(Cauchy-Schwarz)}\\
&\le \sqrt{\mathbf{E}\exp\left(4\lambda^2K^2\mu^{\mathsf{T}}Q_{22}\mu - \lambda\mu^{\mathsf{T}}\mu\right)}\sqrt{\mathbf{E}\exp\left(-\lambda W^{\mathsf{T}}Q_{22}W\right)}. \quad \text{(sub-Gaussianity)}
\end{aligned} \tag{6.1.13}$$

The result follows by noting that $\sqrt{\mathbf{E}\exp\left(4\lambda^2K^2\mu^{\mathsf{T}}Q_{22}\mu - \lambda\mu^{\mathsf{T}}\mu\right)} \le 1$ for our range of λ. ■

Once equipped with (6.1.11), Proposition 6.1.1 follows immediately by Proposition 4.1.1.

6.1.2. The Lower Tail of the Empirical Covariance of Causal sub-Gaussian Processes

Repeated application of Proposition 6.1.1 to the process $X_{1:T} = \mathbf{L}W_{1:T}$ in combination with the tower property of conditional expectation yields the following exponential inequality that controls the lower tail of the empirical covariance in any fixed direction.

Theorem 6.1.2. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process driven by independent K^2-sub-Gaussian increments as described in Section 6.1. Fix also $v \in \mathbb{R}^{d_X}$. Let $Q_{\max} \triangleq \max_{j\in[T/k]} \|\mathbf{L}_{j,j}^{\mathsf{T}}\mathrm{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{j,j}\|_{\mathsf{op}}$ Then for every $\lambda \in \left[0, \frac{1}{8\sqrt{2}K^2Q_{\max}}\right]$:*

$$\begin{aligned}
&\mathbf{E}\exp\left(-\lambda\sum_{t=1}^{T}\langle v, X_t\rangle^2\right)\\
&\qquad\le \exp\left(-\lambda\sum_{j=1}^{T/k}\mathrm{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}}\mathrm{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{j,j}\right) + 36K^4\lambda^2\sum_{j=1}^{T/k}\mathrm{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}}\mathrm{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{j,j}\right)^2\right).
\end{aligned}$$

Proof of Theorem 6.1.2 By repeated use of the tower property we have that:

$$\mathbf{E}\exp\left(-\lambda\sum_{t=0}^{T-1}\langle v, X_t\rangle^2\right) = \mathbf{E}\exp\left(-\lambda\sum_{t=0}^{k-1}\langle v, X_t\rangle^2\right) \times \cdots \times \mathbf{E}_{T-k-1}\exp\left(-\lambda\sum_{t=T-k}^{T-1}\langle v, X_t\rangle^2\right). \tag{6.1.14}$$

We will bound each conditional expectation in (6.1.14) separately, starting with the outermost. Observe that

$$\sum_{t=T-k}^{T-1}\langle v, X_t\rangle^2 = \begin{bmatrix}\langle v, X_{T-k}\rangle\\ \vdots\\ \langle v, X_{T-1}\rangle\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\langle v, X_{T-k}\rangle\\ \vdots\\ \langle v, X_{T-1}\rangle\end{bmatrix} = W_{0:T-1}^{\mathsf{T}}\mathbf{L}_{T/k}^{\mathsf{T}}\mathrm{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{T/k}W_{0:T-1}$$

We now apply Proposition 6.1.1 conditionally with $x = W_{0:T-k-1}$. We find that:

$$\mathbf{E}_{T-k-1} \exp\left(-\lambda \sum_{t=T-k}^{T-1} \langle v, X_t\rangle^2\right)$$
$$\leq \exp\left(-\lambda \operatorname{tr}\left[\mathbf{L}_{T/k,T/k}^{\mathsf{T}} \operatorname{blkdiag}(vv^{\mathsf{T}}) \mathbf{L}_{T/k,T/k}\right] + 36K^4\lambda^2 \left[\operatorname{tr} \mathbf{L}_{T/k,T/k}^{\mathsf{T}} \operatorname{blkdiag}(vv^{\mathsf{T}}) \mathbf{L}_{T/k,T/k}\right]^2 \right).$$

Repeatedly applying Proposition 6.1.1 as above yields the result. ■

To appreciate the terms appearing in Theorem 6.1.2, it is worth to point out that

$$\sum_{j=1}^{T/k} \operatorname{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}} \operatorname{blkdiag}(vv^{\mathsf{T}}) \mathbf{L}_{j,j}\right) = \sum_{t=1}^{T} \mathbf{E}\langle v, \tilde{X}_t\rangle^2.$$

Hence Theorem 6.1.2 effectively passes the expectation inside the exponential at the cost of working with the possibly less excited process $\tilde{X}_{1:T}$ and a quadratic correction term. Note also that the assumption that T is divisible by k is not particularly important. If not, let T' be the largest integer such that $T'/k \in \mathbb{N}$ and $T' \leq T$ and apply the result with T' in place of T.

The significance of Theorem 6.1.2 is demonstrated by the following simple observation, which is just the Chernoff approach applied to the exponential inequality in Theorem 6.1.2.

Lemma 6.1.2. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process with independent K^2-sub-Gaussian increments. Suppose further that the diagonal blocks are all equal: $\mathbf{L}_{j,j} = \mathbf{L}_{1,1}$ for all $j \in [T/k]$. For every $v \in \mathbb{R}^{d_{\mathsf{X}}}$ we have that:*

$$\mathbf{P}\left(\sum_{t=1}^{T} \mathbf{E}\langle v, X_t\rangle^2 \leq \frac{1}{2}\sum_{t=1}^{T} \mathbf{E}\langle v, \tilde{X}_t\rangle^2\right) \leq \exp\left(-\frac{T}{576K^2k}\right). \tag{6.1.15}$$

Proof of Lemma 6.1.2 For any fixed $v \in \mathbb{R}^{d_{\mathsf{X}}}$ and $\lambda \geq 0$ to be determined below in (6.1.16) we have that:

$$
\begin{aligned}
&\mathbf{P}\left(\sum_{t=1}^{T}\langle v, X_t\rangle^2 \leq \frac{1}{2}\sum_{t=1}^{T}\mathbf{E}\langle v, \tilde{X}_t\rangle^2\right) \\
&\leq \mathbf{E}\exp\left(\frac{\lambda}{2}\sum_{t=1}^{T}\mathbf{E}\langle v, \tilde{X}_t\rangle^2 - \lambda\sum_{t=1}^{T}\langle v, X_t\rangle^2\right) \quad \text{(Chernoff)} \\
&\leq \exp\Bigg(-\frac{\lambda}{2}\sum_{j=1}^{T/k}\operatorname{tr}\left[\mathbf{L}_{j,j}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{j,j}\right] \\
&+ 36\lambda^2K^4\sum_{j=1}^{T/k}\operatorname{tr}\left[\mathbf{L}_{j,j}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{j,j}\right]^2\Bigg) \quad \text{(Theorem 6.1.2)} \\
&= \exp\Bigg(-\frac{\lambda T}{2k}\operatorname{tr}\left[\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right] \\
&+ \frac{36\lambda^2TK^4}{k}\operatorname{tr}\left[\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right]^2\Bigg) \quad (\mathbf{L}_{j,j} = \mathbf{L}_{1,1}) \\
&= \exp\Bigg(-\frac{\lambda T}{2k}\operatorname{tr}\left[\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right] \\
&+ \frac{36\lambda^2TK^4}{k}\left[\operatorname{tr}\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right]^2\Bigg) \quad \text{(Cauchy-Schwarz)} \\
&\leq \exp\left(-\frac{T}{576K^2k}\right) \qquad \left(\lambda = \frac{\operatorname{tr}\left[\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{1,1}\right]}{144K^2\left[\operatorname{tr}\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{1,1}\right]^2}\right)
\end{aligned}
\tag{6.1.16}
$$

by optimizing λ in the last line. We point out that the λ used in the above calculation is admissible since $\left[\operatorname{tr}\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right]^2 \geq \|\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\|_{\mathsf{op}}\operatorname{tr}\left[\mathbf{L}_{1,1}^{\mathsf{T}}\operatorname{blkdiag}(vv^{\mathsf{T}})\mathbf{L}_{1,1}\right]$ and so can be seen to satisfy the constraints of Theorem 6.1.2. ■

Note that Lemma 6.1.2 only yields *pointwise* control of the empirical covariance—i.e. pointwise on the sphere $\mathbb{S}^{d_{\mathsf{X}}-1}$. The result holds for a fixed vector on the sphere, but not uniformly for all such vectors at once. Thus, returning to our over-arching goal of providing control of the smallest eigenvalue of the empirical covariance matrix, we now combine (6.1.15) with a union bound and a discretization argument similar to that of Lemma 3.2.2. This approach yields Theorem 6.1.1, of which the proof—along with that of its supporting lemmas—is given in full in Section 6.4.[1]

[1]Similar results can also be obtained for restricted eigenvalues.

6.2. Learning without Stability

Equipped with Theorem 6.1.1 we have positioned ourselves to revisit our earlier result Corollary 5.2.1 which was valid for stable linear systems of the form $X_{t+1} = A_\star X_t + W_{t+1}$—we required $\rho(A_\star) < 1$. Moreover, the burn-in time for this result became vacuous as $\rho(A_\star) \to 1$. Earlier we argued that there was still signal in the learning problem for marginally stable systems $\rho(A_\star) \approx 1$. Indeed, in Section 5.3 we saw that once we have access to sufficiently many trajectories from the same system, stability becomes irrelevant to learnability. We will now see that this is qualitatively the correct behavior—even when we only have access to a single trajectory of data.

We intend to use Theorem 6.1.1. Similar to before we begin with a decomposition of the estimation error. Let us define $\Gamma_s = \frac{1}{p}\sum_{t=1}^{s} \mathbf{E} X_t X_t^\mathsf{T}$ for $s \in [T]$. We write for some k dividing T:

$$\begin{aligned} \widehat{\theta} - \theta_\star &= \left(\sum_{t=1}^{T} V_t X_t^\mathsf{T} \Gamma_T^{-1/2}\right) \Gamma_T^{1/2}\Gamma_k^{-1/2} \left(\sum_{t=1}^{T} \Gamma_k^{-1/2} X_t X_t^\mathsf{T} \Gamma_k^{-1/2}\right)^{-1} \Gamma_k^{-1/2} \\ &\Rightarrow \\ \|(\widehat{\theta} - \theta_\star)\Gamma_T^{-1/2}\|_{\mathsf{op}} &\le \frac{\left\|\sum_{t=1}^{T} V_t X_t^\mathsf{T} \Gamma_T^{-1/2}\right\|_{\mathsf{op}} \|\Gamma_T^{1/2}\Gamma_k^{-1/2}\|_{\mathsf{op}} \|\Gamma_k^{-1/2}\Gamma_T^{1/2}\|_{\mathsf{op}}}{\lambda_{\min}\left(\sum_{t=1}^{T} \Gamma_k^{-1/2} X_t X_t^\mathsf{T} \Gamma_k^{-1/2}\right)} \end{aligned} \tag{6.2.1}$$

As before we let $\mathbf{L}$ be such that $X_{1:T} = \mathbf{L} W_{1:T}$ and further set $\mathbf{G}_T = \mathrm{blkdiag}(\Gamma_T^{-1/2})$. We may use Proposition 5.1.1 just as before to control the first term in the numerator (together with Proposition 5.1.2), this yields that with probability at least $1-\delta$:

$$\begin{aligned} \left\|\sum_{t=1}^{T} V_t X_t^\mathsf{T} \Gamma_T^{-1/2}\right\|_{\mathsf{op}} &\le cK^2\left(\sqrt{\sigma_G^2(d_\mathsf{X} + d_\mathsf{Y} + \log(1/\delta))} + \sigma_E(d_\mathsf{X} + d_\mathsf{Y} + \log(1/\delta))\right) \\ &\le cK^2\Bigg(\sqrt{T\sigma_V^2(d_\mathsf{X} + d_\mathsf{Y} + \log(1/\delta))} \\ &\qquad + \sqrt{\sigma_V^2(T \wedge \|\mathbf{G}_T\mathbf{L}\|_{\mathsf{op}}^2)(d_\mathsf{X} + d_\mathsf{Y} + \log(1/\delta))}\Bigg) \end{aligned} \tag{6.2.2}$$

Combining this with the main result of this chapter, Theorem 6.1.1, we arrive at the following reuslt.

Theorem 6.2.1. *Fix $\delta \in (0,1)$, an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process taking values in $\mathbb{R}^{d_\mathsf{X}}$ with K^2-sub-Gaussian increments. Suppose further that the diagonal blocks are all equal: $\mathbf{L}_{j,j} = \mathbf{L}_{1,1}$ for all $j \in [T/k]$. Suppose $\lambda_{\min}\left(\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^\mathsf{T}\right) > 0$.*

There exist universal positive constants c, c' such that

$$T/k \ge c'K^2(d \log C_{\mathsf{sys}} + \log(1/\delta)) \tag{6.2.3}$$

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

$$\|(\widehat{\theta}-\theta_\star)\Gamma_T^{-1/2}\|_{\mathsf{op}} \le \frac{cK^2\sigma_V\gamma_{T,k}^2\sqrt{(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta))}}{\sqrt{T}} \times \Big(1+\sqrt{(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta))}\Big(1\wedge T^{-1/2}\|\mathbf{G}_T\mathbf{L}\|_{\mathsf{op}}\Big)\Big) \quad (6.2.4)$$

where $\gamma_{T,k} \triangleq \|\Gamma_T^{1/2}\Gamma_k^{-1/2}\|_{\mathsf{op}} \vee \|\Gamma_k^{-1/2}\Gamma_T^{1/2}\|_{\mathsf{op}}$.

A few remarks are in order. The glaring difference between Theorem 6.2.1 and our earlier result Corollary 5.3.1 is the extra dimensional factor $\sqrt{(d_{\mathsf{X}}+d_{\mathsf{Y}}+\log(1/\delta))}$. It has been shown by Tu et al. [2024] that this is in fact unavoidable in the marginally stable regime when only given access to a single trajectory.[2] Note also that this term disappears in the large T regime when for instance $X_{1:T}$ is given by a stable linear dynamical system, since then $T^{-1/2}\|\mathbf{G}_T\mathbf{L}\|_{\mathsf{op}} \ll 1$. Moreover, there is an extra condition number term $\gamma_{T,k}^2$ appearing. Some dependency on this condition number is probably necessary, but we do not believe the bound above to be optimal. There is a more involved argument reducing the dependency to approximately $\gamma_{T,k}\log(\gamma_{T,k})$. This involves using the method of self-normalized martingales instead of the Hanson-Wright inequality to control the random walk component, thereby avoiding the need for the first multiple of $\Gamma_T^{1/2}\Gamma_k^{-1/2}$ in (6.2.1). We will return to see this improvement later in **??**.

Summarizing, the main message of this chapter is that stability and learnability should be thought of as disjoint phenomena: one does not need (approximate) independence of samples to be able to learn. While these phenomena sometimes interact, there can still be signal in processes that are fundamentally unstable. While this is true, let us nevertheless end this chapter with a note of caution: the condition number $\gamma_{T,k}$ above can very well render the bound (6.2.4) vacuous. Indeed, for a linear dynamical system where the matrix $A_\star$ is chosen to be a single Jordan block of size d_{X} with eigenvalue 1 it is easily seen that this is the case (in dimension $d_{\mathsf{X}} \ge 2$). By contrast, if the matrix $A_\star$ is diagonalizable, (6.2.4) is meaningful.

6.3. Notes

In this manuscript we have chosen a perhaps less well-known but conceptually simpler approach to establishing lower bounds on the empirical covariance matrix Equation (6.1.3). The first proof of a statement similar to Theorem 6.1.1 is due to Simchowitz et al. [2018] which in turn relies on a more advanced notion from probability theory known as the small-ball method, due to Mendelson [2014]. The emphasis therein is on anti-concentration—which can hold under milder moment assumptions—rather than concentration. However, the introduction of this tool is not necessary for Gaussian (or sub-Gaussian) system identification. For instance, Sarkar and Rakhlin [2019] leverage the method of self-normalized martingales introduced in **??** below.

Our motivation for providing a different proof is to streamline the exposition as to fit control of the lower tail into the "standard machinery", which roughly consists of: (1) prove a family of scalar exponential inequalities, (2) invoke the Chernoff method, and (3) conclude by a discretization

[2] They work in expectation, so the term $\log(1/\delta)$ does not appear there and so we cannot say for certain whether this part of the term is unavoidable.

argument and a union bound to port the result from scalars to matrices. Our proof here follows this outline and emphasizes the exponential inequality in Theorem 6.1.2.

6.4. Proof of Theorem 6.1.1

Let $\mathcal{N}_\varepsilon$ be an optimal ε-cover of the unit sphere $\mathbb{S}^{d-1}$ and fix a multiplier $q \in (1, \infty)$. We define the events:

$$\begin{aligned} \mathcal{E}_1 &= \bigcup_{v \in \mathcal{N}_e} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} v^\mathsf{T} X_t X_t^\mathsf{T} v \le \frac{1}{2T} \sum_{t=0}^{T-1} \mathbf{E} v^\mathsf{T} \tilde{X}_t \tilde{X}_t^\mathsf{T} v \right\} \\ \mathcal{E}_2 &= \left\{ \left\| \sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right\|_{\mathsf{op}} \ge q \times \left\| \sum_{t=0}^{T-1} \mathbf{E} X_t X_t^\mathsf{T} \right\|_{\mathsf{op}} \right\}. \end{aligned} \tag{6.4.1}$$

for any v, it is true on the complement of $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$ that for every $v_i \in \mathcal{N}_\varepsilon$:

$$\begin{aligned} &\frac{1}{T} \sum_{t=0}^{T-1} v^\mathsf{T} X_t X_t^\mathsf{T} v \\ &\ge \frac{1}{2T} \sum_{t=0}^{T-1} v_i^\mathsf{T} X_t X_t^\mathsf{T} v_i - \frac{1}{T} \sum_{t=0}^{T-1} (v - v_i)^\mathsf{T} X_t X_t^\mathsf{T} (v - v_i) && \text{(parallellogram)} \\ &\ge \frac{1}{2T} \sum_{t=0}^{T-1} v_i^\mathsf{T} X_t X_t^\mathsf{T} v_i - \frac{\varepsilon^2}{T} \left\| \sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right\|_{\mathsf{op}} && \text{(covering)} \\ &\ge \frac{1}{4T} \sum_{t=0}^{T-1} \mathbf{E} v_i^\mathsf{T} \tilde{X}_t \tilde{X}_t^\mathsf{T} v_i - \frac{q\varepsilon^2}{T} \left\| \sum_{t=0}^{T-1} \mathbf{E}[X_t X_t^\mathsf{T}] \right\|_{\mathsf{op}} && (\mathcal{E}^c) \end{aligned} \tag{6.4.2}$$

where we used that $v - v_i$ has norm at most ε for some choice of v_i by the covering property. For this choice we have that:

$$\frac{1}{T} \sum_{t=0}^{T-1} v^\mathsf{T} X_t X_t^\mathsf{T} v \ge \frac{1}{8T} \sum_{t=0}^{T-1} v_i^\mathsf{T} \mathbf{E}[\tilde{X}_t \tilde{X}_T^\mathsf{T}] v_i$$

as long as:

$$\varepsilon^2 \le \frac{\lambda_{\min} \left(\sum_{t=0}^{T-1} \mathbf{E} \tilde{X}_t \tilde{X}_t^\mathsf{T} \right)}{8q \lambda_{\max} \left(\sum_{t=0}^{T-1} \mathbf{E} X_t X_t \right)}. \tag{6.4.3}$$

To finish the proof, it suffices to estimate the failure probabilities $\mathbf{P}(\mathcal{E}_1)$ and $\mathbf{P}(\mathcal{E}_2)$. By (6.1.16), a volumetric argument [see e.g. Wainwright, 2019, Example 5.8] and our particular choice of ε we have:

$$\mathbf{P}(\mathcal{E}_1) \le \left(1 + \frac{2}{\varepsilon^2}\right)^d \exp\left(-\frac{T}{576K^2k}\right).$$

To estimate $\mathbf{P}(\mathcal{E}_2)$, observe first that for $\mathbb{S}^{d-1}_{1/4}$ a 1/4-net of $\mathbb{S}^{d-1}$, we have by **??**, (**??**):

$$\left\|\sum_{t=0}^{T-1} X_t X_t^\mathsf{T}\right\|_{\mathsf{op}} \leq 2 \sup_{\tilde{v}\in\mathbb{S}^{d-1}_{1/4}} \sum_{t=0}^{T-1} \tilde{v}^\mathsf{T} X_t X_t^\mathsf{T} \tilde{v} \tag{6.4.4}$$

We now invoke Proposition 4.1.1 with $M = \mathbf{L}_{\tilde{v}}\mathbf{L}_{\tilde{v}}^\mathsf{T}$ where $\mathbf{L}_{\tilde{v}} = (I_T \otimes \tilde{v}^\mathsf{T})\mathbf{L}$ for fixed $\tilde{v}$ of the form $(v - v_i)/\varepsilon$ as before and finish by a union bound. For fixed $\tilde{v}$ and $\lambda \geq 0$ to be determined below, Proposition 4.1.1 yields via a Chernoff argument:

$$\begin{aligned}
&\mathbf{P}\left(\sum_{t=0}^{T-1} \tilde{v}^\mathsf{T} X_t X_t^\mathsf{T} \tilde{v} \geq q \times \sum_{t=0}^{T-1} \tilde{v}^\mathsf{T} \mathbf{E}[X_t X_t^\mathsf{T}] \tilde{v}\right) \\
&= \mathbf{P}\left(W_{0:T-1}^\mathsf{T} \mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T} W_{0:T-1}^\mathsf{T} \geq q \operatorname{tr} \mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\right) \\
&\leq \exp\left(-\lambda q \operatorname{tr} \mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}} + 36\lambda^2 K^4 \operatorname{tr}(\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T})^2\right) && \left(\lambda \leq \frac{1}{8\sqrt{2} K^2 \|\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\|_{\mathsf{op}}}\right) \\
&\leq \exp\left(-\lambda q \operatorname{tr} \mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}} + 36\lambda^2 K^4 \|\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\|_{\mathsf{op}} \operatorname{tr}(\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T})\right) \\
&= \exp\left(-\frac{\operatorname{tr}(\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T})}{K^2 \|\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\|_{\mathsf{op}}} \left(\frac{q}{8\sqrt{2}} - \frac{9}{32}\right)\right) && \left(\lambda = \frac{1}{8\sqrt{2} K^2 \|\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\|_{\mathsf{op}}}\right) \\
&\leq \exp\left(-\frac{T}{576 K^2 k}\right) && \left(q \geq \frac{8\sqrt{2} T \|\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T}\|_{\mathsf{op}}}{576 k \operatorname{tr}(\mathbf{L}_{\tilde{v}} \mathbf{L}_{\tilde{v}}^\mathsf{T})} + \frac{9 \times 8\sqrt{2}}{32}\right).
\end{aligned} \tag{6.4.5}$$

Observe that

$$q = \frac{8\sqrt{2} T \|\mathbf{L}\mathbf{L}^\mathsf{T}\|_{\mathsf{op}}}{576 k \lambda_{\min}\left(\sum_{t=0}^{T-1} \mathbf{E} X_t X_t^\mathsf{T}\right)} + \frac{9 \times 8\sqrt{2}}{32} \tag{6.4.6}$$

satisfies the constraint from (6.4.5) for all $\tilde{v}$. Hence by a union bound and a volumetric argument [see e.g. Wainwright, 2019, Example 5.8] with probability at least $1 - \left(1 + \frac{2}{\varepsilon^2}\right)^d$ (observing that $\varepsilon \leq 1/4$ in (6.4.3)):

$$\begin{aligned}
\left\|\sum_{t=0}^{T-1} X_t X_t^\mathsf{T}\right\|_{\mathsf{op}} &\leq 2 \sup_{\tilde{v}\in\mathbb{S}^{d-1}_{1/4}} \sum_{t=0}^{T-1} \tilde{v}^\mathsf{T} X_t X_t^\mathsf{T} \tilde{v} && \text{(by (6.4.4))} \\
&\leq 2q \sup_{\tilde{v}\in\mathbb{S}^{d-1}_{1/4}} \sum_{t=0}^{T-1} \tilde{v}^\mathsf{T} \mathbf{E} X_t X_t^\mathsf{T} \tilde{v} && \text{(union bound over (6.4.5))} \\
&\leq 2q \left\|\sum_{t=0}^{T-1} \mathbf{E} X_t X_t^\mathsf{T}\right\|_{\mathsf{op}}. && (\mathbb{S}^{d-1}_{1/4} \subset \mathbb{S}^{d-1})
\end{aligned} \tag{6.4.7}$$

Hence with the choice of q from (6.4.6) and another union bound (again observing that $\varepsilon \leq 1/4$):

$$\mathbf{P}(\mathcal{E}_1) + \mathbf{P}(\mathcal{E}_2) \leq 2\left(1 + \frac{2}{\varepsilon^2}\right)^d \exp\left(-\frac{T}{576 K^2 k}\right). \tag{6.4.8}$$

In light of (6.4.3) we may choose with the above choice of q (from (6.4.6)):

$$\begin{aligned}\varepsilon^2 &= \frac{\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}{8q\lambda_{\max}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t\right)} \\ &= \frac{\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}{8\left(\frac{8\sqrt{2}T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{576k\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}+\frac{9\times 8\sqrt{2}}{32}\right)\lambda_{\max}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t\right)}.\end{aligned} \tag{6.4.9}$$

Thus:

$$\begin{aligned}\left(1+\frac{2}{\varepsilon^2}\right)^d &= \left(1+16\frac{\left(\frac{8\sqrt{2}T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{576k\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}+\frac{9\times 8\sqrt{2}}{32}\right)\lambda_{\max}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t\right)}{\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}\right)^d \\ &= \left(1+4\sqrt{2}\frac{\left(\frac{T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{18k\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}+9\right)\lambda_{\max}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t\right)}{\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}\right)^d\end{aligned} \tag{6.4.10}$$

The result has been established. ■

Part II.

Control

7. The Linear Quadratic Regulator

In this lecture we give a brief review of the linear quadratic regulator. Consider a system evolving according to

$$X_{t+1} = AX_t + BU_t + W_{t+1} \qquad X_0 = W_0 \qquad t = 0, 1 \ldots, T \tag{7.0.1}$$

where $A \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$, $B \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{U}}}$ and where $W_{0:T+1}$ is drawn iid with mean zero and covariance matrix $\Sigma_W \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$.

Fix a positive semi-definite matrices $Q, Q_T \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$, and a positive definite matrix $R \in \mathbb{R}^{d_{\mathsf{U}} \times d_{\mathsf{U}}}$. The goal of linear quadratic regulation is to design a policy π, dictating the conditional laws of $U_{0:T-1}$ such that

$$\mathtt{V}_T^\pi = \mathbf{E}^\pi \left[X_T^\mathsf{T} Q_T X_T + \sum_{k=0}^{T-1} X_k^\mathsf{T} Q X_k + U_k^\mathsf{T} R U_k \right] \tag{7.0.2}$$

is rendered minimal. We note that above and in the sequel we write $\mathbf{E}^\pi$ to emphasize that the expectation depends on the particular policy chosen. To be a little more precise, the optimization variable π is a sequence of conditional distributions $\pi_{0:T-1}$ by which $U_t \sim \pi_t(\cdot | X_{0:t}) \mid X_{0:t}$.

7.1. Dynamic Programming Solution to LQR

Dynamic programming—backward induction—is the deep and rather obvious observation that if a sequence of inputs $U_{0:T-1} \sim \pi_\star$ are optimal for (say) the cost in (7.0.2) then the subsequence $U_{t:T-1}, t \in [T-1]$ better be optimal for

$$\mathtt{V}_T^\pi - \mathtt{V}_t^\pi = \mathbf{E}^\pi \left[X_T^\mathsf{T} Q_T X_T + \sum_{k=t+1}^{T-1} X_k^\mathsf{T} Q X_k + U_k^\mathsf{T} R U_k \right] \tag{7.1.1}$$

where for $t < T$ we define:

$$\mathtt{V}_t^\pi = \mathbf{E}^\pi \left[\sum_{k=0}^{t-1} X_k^\mathsf{T} Q X_k + U_k^\mathsf{T} R U_k \right]. \tag{7.1.2}$$

We can use this observation recursively to figure out the optimal control law $\pi_\star$ minimizing (7.0.2).

Indeed, we have that

$$\begin{aligned}
&\mathtt{V}_T^{\pi_\star} - \mathtt{V}_{T-1}^{\pi_\star} \\
&= \mathbf{E}^{\pi_\star}\left[X_T^\mathsf{T} Q_T X_T + X_{T-1}^\mathsf{T} Q X_{T-1} + U_{T-1}^\mathsf{T} R U_{T-1}\right] \\
&= \mathbf{E}^{\pi_\star}\left[(AX_{T-1} + BU_{T-1} + W_T)^\mathsf{T} Q_T (AX_{T-1} + BU_{T-1} + W_T) + X_{T-1}^\mathsf{T} Q X_{T-1} + U_{T-1}^\mathsf{T} R U_{T-1}\right] \\
&= \mathbf{E}^{\pi_\star}\left[\begin{bmatrix} X_{T-1} \\ U_{T-1}\end{bmatrix}^\mathsf{T} \begin{bmatrix} Q + A^\mathsf{T} Q_T A & A^T Q_T B \\ B^\mathsf{T} Q_T^\mathsf{T} A & R + B^\mathsf{T} Q_T B\end{bmatrix} \begin{bmatrix} X_{T-1} \\ U_{T-1}\end{bmatrix} + W_T^\mathsf{T} Q_T W_T\right] \\
&\geq \mathbf{E}^{\pi_\star} \min_u \left[\begin{bmatrix} X_{T-1} \\ u\end{bmatrix}^\mathsf{T} \begin{bmatrix} Q + A^\mathsf{T} Q_T A & A^T Q_T B \\ B^\mathsf{T} Q_T^\mathsf{T} A & R + B^\mathsf{T} Q_T B\end{bmatrix} \begin{bmatrix} X_{T-1} \\ u\end{bmatrix} + W_T^\mathsf{T} Q_T W_T\right].
\end{aligned} \tag{7.1.3}$$

Let us a take a brief detour.

Lemma 7.1.1. *Fix* $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \succeq 0$ *and suppose that* $M_{22} \succ 0$. *Then:*

$$\min_u \begin{bmatrix} x \\ u\end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \begin{bmatrix} x \\ u\end{bmatrix} = x^\mathsf{T}(M_{11} - M_{12}^\mathsf{T} M_{22}^{-1} M_{12})x. \tag{7.1.4}$$

Moreover, the minimum is achieved at $u^\star = -M_{22}^{-1} M_{12} x$.

Proof. The assumption that $M \succeq 0$ is equivalent to convexity for this problem, and so setting the gradient to zero is a sufficient condition:

$$\nabla_u \begin{bmatrix} x \\ u\end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \begin{bmatrix} x \\ u\end{bmatrix} = 2M_{22}u + 2M_{12}x \tag{7.1.5}$$

The optimality of $u^\star$ is thus established by setting the above to zero and and re-arranging. Hence

$$\begin{aligned}
\min_u \begin{bmatrix} x \\ u\end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \begin{bmatrix} x \\ u\end{bmatrix} &= \begin{bmatrix} x \\ u^\star\end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \begin{bmatrix} x \\ u^\star\end{bmatrix} \\
&= x^\mathsf{T} \begin{bmatrix} I \\ -M_{22}^{-1} M_{12}\end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\mathsf{T} & M_{22}\end{bmatrix} \begin{bmatrix} I \\ -M_{22}^{-1} M_{12}\end{bmatrix} x \\
&= x^\mathsf{T}(M_{11} - M_{12}^\mathsf{T} M_{22}^{-1} M_{12})x
\end{aligned} \tag{7.1.6}$$

as was required. ∎

Using Lemma 7.1.1 we see that the minimizer of the right hand side of (7.1.3) is given by $u = -(R + B^\mathsf{T} Q_T B)^{-1} B^\mathsf{T} Q_T A X_{T-1}$. Since this variable is an admissible control action we conclude that $U_{T-1} = -(R + B^\mathsf{T} Q_T B)^{-1} B \mathsf{T} Q_T A X_{T-1}$ is the optimal action so that equality holds throughout (7.1.3).

Moreover, this allows us to compute the residual remaining cost via the Schur complement

$$P_{T-1} \triangleq Q + A^\mathsf{T} Q_T A - A^\mathsf{T} Q_T B (R + B^\mathsf{T} Q_T B)^{-1} B^\mathsf{T} Q_T A \tag{7.1.7}$$

as

$$\mathtt{V}_T^{\pi_\star} - \mathtt{V}_{T-1}^{\pi_\star} = \mathbf{E}^{\pi_\star}[X_{T-1}^\mathsf{T} P_{T-1} X_{T-1} + W_T^\mathsf{T} Q_T W_T]. \tag{7.1.8}$$

Suggestively we select notation $P_T = Q_T$. We may proceed in this fashion:

$$\begin{aligned}
&\mathtt{V}_T^{\pi_\star} - \mathtt{V}_{T-2}^{\pi_\star} \\
&= \mathbf{E}^{\pi_\star}\left[X_T^\mathsf{T} P_T X_T + X_{T-1}^\mathsf{T} Q X_{T-1} + U_{T-1}^\mathsf{T} R U_{T-1} + X_{T-2}^\mathsf{T} Q X_{T-2} + U_{T-2}^\mathsf{T} R U_{T-2}\right] \\
&= \mathbf{E}^{\pi_\star}\left[W_T^\mathsf{T} P_T W_T + X_{T-1}^\mathsf{T} P_{T-1} X_{T-1} + X_{T-2}^\mathsf{T} Q X_{T-2} + U_{T-2}^\mathsf{T} R U_{T-2}\right] \\
&= \mathbf{E}^{\pi_\star}[W_T^\mathsf{T} P_T W_T + W_{T-1}^\mathsf{T} P_{T-1} W_{T-1} + X_{T-2}^\mathsf{T} P_{T-2} X_{T-2}]
\end{aligned} \tag{7.1.9}$$

and the minimizer is again of the same form (repeating the exact same argument as in (7.1.3)) $U_{T-2} = -(R + B^\mathsf{T} P_{T-1} B)^{-1} B^\mathsf{T} P_{T-1} A X_{T-2}$. Similarly the residual is $\mathtt{V}_T^{\pi_\star} - \mathtt{V}_{T-2}^{\pi_\star} = \mathbf{E}^{\pi_\star}[W_T^\mathsf{T} Q_T W_T + W_{T-1}^\mathsf{T} P_{T-1} W_{T-1} + X_{T-2}^\mathsf{T} P_{T-2} X_{T-2}]$ where

$$P_{T-2} \triangleq Q + A^\mathsf{T} P_{T-1} A - A^\mathsf{T} P_{T-1} B (R + B^\mathsf{T} P_{T-1} B)^{-1} B^\mathsf{T} P_{T-1} A. \tag{7.1.10}$$

We may proceed along these lines via induction to establish the following result.

Theorem 7.1.1. *Consider the the problem* $\min_\pi \mathtt{V}_T^\pi$ *where* π *varies over conditional distributions over* $U_{0:T-1}$*. The optimal policy* π *is characterized by*

$$\begin{aligned}
P_{t-1} &\triangleq Q + A^\mathsf{T} P_t A - A^\mathsf{T} P_t B (R + B^\mathsf{T} P_t B)^{-1} B^\mathsf{T} P_t A, \quad P_T = Q_T \\
K_{t-1} &\triangleq -(R + B^\mathsf{T} P_t B)^{-1} B^\mathsf{T} P_t A, \\
U_t &= K_t X_t.
\end{aligned} \tag{7.1.11}$$

Moreover the optimal cost is given

$$\mathtt{V}_T^{\pi_\star} = \sum_{t=0}^{T} \mathrm{tr}(\Sigma_W P_t). \tag{7.1.12}$$

The recursion (7.1.11) for P_t is rather famous and has a name: the Discrete Algebraic Riccati Recursion (DARR). There is also a fixed point analogue, known as the Discrete Algebraic Riccati Equation (DARE):

$$P = Q + A^\mathsf{T} P A - A^\mathsf{T} P B (R + B^\mathsf{T} P B)^{-1} B^\mathsf{T} P A. \tag{7.1.13}$$

The existence of the steady state version (7.1.13) of (7.1.11) is a more subtle question than can be answered by direct calculation and requires some control theory.

7.2. Regret

We will often find ourselves trying to learn the optimal policy $\pi_\star$ given by Theorem 7.1.1. In this case, our learned policy, say π, will typically not achieve the cost $\mathtt{V}_T^{\pi_\star}$ but suffer some degree of sub-optimality. This motivates the definition of the *regret* of a policy, $\mathtt{R}_T^\pi$, as follows

$$\mathtt{R}_T^\pi = \mathtt{V}_T^\pi - \mathtt{V}_T^{\pi_\star}. \tag{7.2.1}$$

Our next result shows that the regret is the expectation of a quadratic form. This perhaps further motivates the common perspective that LQR is the "linear regression of controls".

Theorem 7.2.1. *Let $K_{0:T-1}$ and $P_{1:T}$ be as in (7.1.11). Then for every policy π it holds that:*

$$\mathtt{R}_T^\pi = \sum_{t=0}^{T-1} \mathbf{E}^\pi \left[(U_t - K_t X_t)^\mathsf{T} (R + B^\mathsf{T} P_{t+1} B)(U_t - K_t X_t) \right] \tag{7.2.2}$$

Proof. Let us revisit the proof of Theorem 7.1.1. We have

$$\mathtt{V}_T^\pi = \mathbf{E}^\pi \left[\begin{bmatrix} X_{T-1} \\ U_{T-1} \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q + A^\mathsf{T} P_T A & A^T P_T B \\ B^\mathsf{T} P_T^\mathsf{T} A & R + B^\mathsf{T} P_T B \end{bmatrix} \begin{bmatrix} X_{T-1} \\ U_{T-1} \end{bmatrix} + W_T^\mathsf{T} P_T W_T + \sum_{k=0}^{T-2} X_k^\mathsf{T} Q X_k + U_k^\mathsf{T} R U_k \right] \tag{7.2.3}$$

We know that each step, the optimal policy is found by minimizing

$$f_t(u) \triangleq \begin{bmatrix} X_t \\ u \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q + A^\mathsf{T} P_{t+1} A & A^T P_{t+1} B \\ B^\mathsf{T} P_{t+1}^\mathsf{T} A & R + B^\mathsf{T} P_{t+1} B \end{bmatrix} \begin{bmatrix} X_t \\ u \end{bmatrix}. \tag{7.2.4}$$

The function $f_t(\cdot)$ has hessian $2R + B^\mathsf{T} P_{t+1} B$ and the optimum is attained at $u_\star = K_t X_t$. Moreover, we know by Lemma 7.1.1 that

$$\min f_t(u) = X_t^\mathsf{T} \left(Q + A^\mathsf{T} P_{t+1} - A^T P_{t+1} B (R + B^\mathsf{T} P_{t+1})^{-1} B^\mathsf{T} P_{t+1}^\mathsf{T} A \right) X_t. \tag{7.2.5}$$

Incidentally we recognize that $Q + A^\mathsf{T} P_{t+1} - A^T P_{t+1} B (R + B^\mathsf{T} P_{t+1})^{-1} B^\mathsf{T} P_{t+1}^\mathsf{T} A = P_t$. We now notice that the second order Taylor expansion of f_t is exact and that we have that

$$f_t(u) = (u - K_t x_t)^\mathsf{T} (R + B^\mathsf{T} P_{t+1} B)(u - K_t X_t) + X_t^\mathsf{T} P_t X_t \tag{7.2.6}$$

Hence we have that

$$\begin{aligned} \mathtt{V}_T^\pi = \mathbf{E}^\pi \Bigg[& W_T^\mathsf{T} P_T W_T + (U_{T-1} - K_{T-1} X_{T-1})^\mathsf{T} (R + B^\mathsf{T} P_{T-1} B)(U_{T-1} - K_{T-1} X_{T-1}) \\ & + \begin{bmatrix} X_{T-2} \\ U_{T-2} \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q + A^\mathsf{T} P_{T-1} A & A^T P_{T-1} B \\ B^\mathsf{T} P_{T-1}^\mathsf{T} A & R + B^\mathsf{T} P_{T-1} B \end{bmatrix} \begin{bmatrix} X_{T-2} \\ U_{T-2} \end{bmatrix} + W_{T-1}^\mathsf{T} P_{T-1} W_{T-1} \\ & + \sum_{k=0}^{T-3} X_k^\mathsf{T} Q X_k + U_k^\mathsf{T} R U_k \Bigg] \end{aligned} \tag{7.2.7}$$

The result follows by recusing this on t and subtracting of the constant term $V_T^{\pi_\star} = \sum_{t=0}^T \operatorname{tr} \Sigma_W P_t = \sum_{t=0}^T \mathbf{E} W_t^\mathsf{T} P_t W_t$. ∎

7.3. Elements of Linear Control Theory

One issue with the way we formulated the problem (7.0.2) so far is that their solutions so far are system-theoretically meaningless. Controls is a subject much richer than simply rendering a cost minimal. Rather, we would typically care about trajectories exhibiting *stable* behavior—remaining, for all time forward, within some neighborhood of some fixed point (or more generally some fixed trajectory). For linear systems, the following notion is sufficient to relate the performance of the optimization problem (7.0.2) to stability.

Definition 7.3.1. *We say that a tuple (A, B) is stabilizable if there exists K such that $\rho(A+BK) < 1$. The tuple is (τ, μ)-strongly stabilizable if $\|(A + BK)^k\|_{\mathsf{op}} \leq \tau\mu^k$ for every $k \in \mathbb{N}$.*

Exercise 7.3.1. *Prove that every stabilizable pair (A, B) is strongly stabilizable for some values of τ and μ.*

We will not prove results under this assumption, as they can get quite technical. The following, essentially linear-algebraic, notion, is somewhat easier to work with.

Definition 7.3.2. *The tuple (A, B) is said to be controllable if $\mathsf{C}_k(A, B) \triangleq \begin{bmatrix} B & AB & A^2B & \cdots & A^{k-1}B \end{bmatrix}$ has full rank for some $k \in \mathbb{N}$. Moreover, the tuple (A, B) is (k, ν)-strongly controllable if $\sqrt{\lambda_{\min}(\mathsf{C}_k\mathsf{C}_k^{\mathsf{T}})} > \nu$.*

Exercise 7.3.2. *Prove that (A, B) is controllable if and only $\mathsf{C}_{d_{\mathsf{X}}}(A, B)$ has full rank. Hint: Cayley-Hamilton.*

There is also a dual notion, called observability.

Definition 7.3.3. *The tuple (A, C) is said to be observable if $\mathsf{O}_k(A, B) \triangleq \begin{bmatrix} C & CA & CA^2 & \cdots & CA^{k-1} \end{bmatrix}$ has full rank for some $k \in \mathbb{N}$.*

We will prove the following result.

Theorem 7.3.1. *Let (A, B) be controllable, and suppose there exists C such that 1) $Q = C^{\mathsf{T}}C$ and 2) (A, C) is observable. Then (7.1.13) has a unique positive definite solution P and $A+BK$ is stable, where*

$$K = -(R + B^{\mathsf{T}}PB)^{-1}B^{\mathsf{T}}PA. \tag{7.3.1}$$

The proof passes via deterministic optimal control.

7.4. Deterministic Optimal Control

Our aim is now to study the asymptotics of the backward recursion (7.1.11). The crucial observation here is that the dynamic programming solution above still works, and outputs a particularly simple cost, if we set all the noise variables except the initial condition to be identically zero. This provides us with an excellent opportunity to take a detour to deterministic optimal control, which, strictly speaking, is a simpler problem than we just studied. Namely, the proof of Theorem 7.3.1 works by considering the following family, indexed by $\tau \in \mathbb{N}$, of constrained optimization problems:

$$\begin{aligned} &\min_{u_{0:\tau-1}} x_\tau P_{\text{init}} x_\tau + \sum_{t=0}^{\tau-1} x_t^{\mathsf{T}} Q x_t + u_t^{\mathsf{T}} R u_t \\ &\text{s.t.} \quad x_{t+1} = Ax_t + Bu_t, \qquad x_0 = x \qquad t = 0, 1, \ldots, \tau - 1, \end{aligned} \tag{7.4.1}$$

where $P_{\text{init}} \succeq 0$ takes the role of Q_T before (but we will want to keep this fixed across our family of problems as τ varies and therefore choose notation that is horizon-invariant). We emphasize (7.4.1) as an optimization problem since there is no uncertainty to mitigate; there is no need for feedback. In control parlance, the solution to (7.4.1) is open loop. On a related note, let us add that we have switched to lower case notation for the state and input variables to emphasize the fact that they are *not* random variables.

7.4.1. Proof of Theorem 7.3.1

The proof proceeds by a number of claims. First, we show that a solution to (7.1.13) exists. We achieve this by showing it arises as the limit of the recursion (7.1.11) with zero terminal cost (although this can be relaxed). Our first observation is that the solution to (7.4.1) can still be found by dynamic programming, whence we have the following claim.

Claim 7.4.1. *Consider the problem* (7.4.1)*. Its value is given by*

$$\min_{u_{0:\tau-1}} x_\tau P_{\text{init}} x_\tau + \sum_{t=0}^{\tau-1} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t = x^\mathsf{T} P_0^\tau x. \tag{7.4.2}$$

where

$$P_{t-1}^\tau \triangleq Q + A^\mathsf{T} P_t^\tau A - A^\mathsf{T} P_t^\tau B (R + B^\mathsf{T} P_t B)^{-1} B^\mathsf{T} P_t A, \qquad P_\tau^\tau = P_{\text{init}}. \tag{7.4.3}$$

In principle, we expect the minimal cost to be an increasing function of the horizon τ. This is indeed always the case if $P_{\text{init}} = 0$. Indeed:

$$x^\mathsf{T} P_0^\tau x = \min_{u_{0:\tau-1}} \sum_{t=0}^{\tau-1} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t \le \min_{u_{0:\tau}} \sum_{t=0}^{\tau} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t = x^\mathsf{T} P_0^{\tau+1} x. \tag{7.4.4}$$

Thus, the above claim immediately yields the following corollary by varying over initial conditions $x \in \mathbb{R}^{d_\mathsf{X}}$.

Claim 7.4.2. *Suppose that $P_{\text{init}} = 0$. Then if $\tau' > \tau$ we have that $P_0^{\tau'} \succeq P_0^\tau$.*

Thus, under the additional hypothesis that $P_{\text{init}} = 0$, we have shown that the sequence $\{P_0^\tau\}_{\tau \in \mathbb{N}}$ is monotone is semidefinite order. To establish that a limit point exists, it thus suffices to prove that it further is bounded. For this we use controllability—the idea is that, using controllability, we may reset the system to the origin in $\mathbb{R}^{d_\mathsf{X}}$ in a finite number of steps.

Claim 7.4.3. *Suppose that the tuple (A, B) is controllable. There exists a constant c only depending on A, B, Q and R such that*

$$\sup_\tau \|P_0^\tau\|_{\mathsf{op}} \le c. \tag{7.4.5}$$

Proof. It suffices to prove that $x^\mathsf{T} P_0^\tau x \le c$ for all $x \in \mathbb{S}^{d_\mathsf{X}-1}$. Hence, fix such an x and notice that by controllability, there exists an index $k \in \mathbb{N}$ only depending on (A, B) such that $\begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix}$ has full rank. Notice further that for this index k we may write

$$x_k = x + \begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix} \begin{bmatrix} u_{k-1} \\ u_{k-1} \\ \vdots \\ u_0 \end{bmatrix}. \tag{7.4.6}$$

Controllability precisely reads that $\begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix}$ has a right inverse so that the solution

$$u_{k-1:0} = -\begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix}^\dagger x_0 \tag{7.4.7}$$

renders the right hand side of (7.4.6) zero. Moreover, since k and $\|x_0\|$ are fixed, this means that we can set $u_{k:\tau-1} = 0$ to achieve finite cost depending only on k, A, B, Q, R (and notice that k depends only on (A, B)). ∎

Hence, $\{P_0^\tau\}_{\tau\in\mathbb{N}}$ is monotone and bounded and thus has a limit (convince yourself that you can port this classical analysis fact with monotonicity and boundedness in $\mathbb{R}$ to monotonicity and boundedness in semidefinite order). Hence, we have established the existence part of Theorem 7.3.1.

We next prove that K in (7.3.1) is stabilizing. This is equivalent to proving that

$$x_{t+1} = (A + BK)x_t, \qquad x_0 = x \tag{7.4.8}$$

tends to zero for all x.

Claim 7.4.4 (Lyapunov Equation)**.** *Let P and K satisfy (7.1.13) and (7.3.1). Then*

$$P = Q + (A + BK)^\mathsf{T} P(A + BK) + K^\mathsf{T} RK. \tag{7.4.9}$$

Proof. Observe first that

$$(A + BK)^\mathsf{T} P(A + BK) = A^\mathsf{T} PA + K^\mathsf{T} B^\mathsf{T} PBK + A^\mathsf{T} PBK + K^\mathsf{T} B^\mathsf{T} PA \tag{7.4.10}$$

and that

$$A^\mathsf{T} PBK = -A^\mathsf{T} PB(R + B^\mathsf{T} PB)^{-1} B^\mathsf{T} PA = K^\mathsf{T} B^\mathsf{T} PA. \tag{7.4.11}$$

Using these identities we may write starting from (7.1.13):

$$\begin{aligned} P &= Q + A^\mathsf{T} PA - A^\mathsf{T} PB(R + B^\mathsf{T} PB)^{-1} B^\mathsf{T} PA \\ &= Q + (A + BK)^\mathsf{T} P(A + BK) + A^\mathsf{T} PB(R + B^\mathsf{T} PB)^{-1} B^\mathsf{T} PA - K^\mathsf{T} B^\mathsf{T} PBK \end{aligned} \tag{7.4.12}$$

It thus suffices to rewrite the second last term above as

$$\begin{aligned} A^\mathsf{T} PB(R + B^\mathsf{T} PB)^{-1} B^\mathsf{T} PA & \\ = A^\mathsf{T} PB(R + B^\mathsf{T} PB)^{-1}(R + B^\mathsf{T} PB)(R + B^\mathsf{T} PB)^{-1} B^\mathsf{T} PA & \\ = K^\mathsf{T}(R + B^\mathsf{T} PB)K & \end{aligned} \tag{7.4.13}$$

for us to see that

$$P = Q + (A + BK)^\mathsf{T} P(A + BK) + K^\mathsf{T} RK \tag{7.4.14}$$

as was required. ∎

Using this Lyapunov equation, we can now establish a form of descent lemma, namely:

$$\begin{aligned} x_{t+1}^\mathsf{T} Px_{t+1} - x_t^\mathsf{T} Px_t &= x_t^\mathsf{T} \left[(A + BK)^\mathsf{T} P(A + BK) - P\right] x_t \\ &= -x_t^\mathsf{T} \left[Q + K^\mathsf{T} RK\right] x_t. \end{aligned} \tag{7.4.15}$$

Consequently

$$x_{t+1}^\mathsf{T} Px_{t+1} = x^\mathsf{T} Px - \sum_{k=1}^{t} x_k^\mathsf{T} \left[Q + K^\mathsf{T} RK\right] x_k \tag{7.4.16}$$

by which it becomes clear that $\lim_{k\to\infty} x_k^\mathsf{T} \left[Q + K^\mathsf{T} RK\right] x_k = 0$. In particular, have both that $\lim_{k\to\infty} x_k^\mathsf{T} Qx_k = 0$ and that $\lim_{k\to\infty} x_k^\mathsf{T} K^\mathsf{T} RKx_k = \lim_{k\to\infty} u_k^\mathsf{T} Ru_k = 0$ of which the latter implies that $u_k \to 0$ since $R \succ 0$.

Let us now use our observability condition to establish that also $x_k \to 0$. Indeed, we have assumed that there exists C such that $Q = C^\mathsf{T} C$ with (A, C) observable and we have in this terminology already shown that $Cx_k \to 0$. Indeed, notice that we may write, using the fact that $x_{t+1} = Ax_t + Bu_t = Ax_t + BKx_t$:

$$\begin{bmatrix} CA^{d_\mathsf{X}-1} \\ CA^{d_\mathsf{X}-2} \\ \vdots \\ CA \\ C \end{bmatrix} x_k = \begin{bmatrix} C\left(x_{k+d_\mathsf{X}-1} - \sum_{i=1}^{d_\mathsf{X}-1} A^{i-1} B u_{k+d_\mathsf{X}-i-1}\right) \\ C\left(x_{k+d_\mathsf{X}-2} - \sum_{i=1}^{d_\mathsf{X}-2} A^{i-1} B u_{k+d_\mathsf{X}-i-2}\right) \\ \vdots \\ C(x_{k+1} - Bu_k) \\ Cx_k \end{bmatrix}. \tag{7.4.17}$$

The right hand side above tends to zero by the previously established limits, and hence the left hand side also tends to zero. Moreover, observability precisely implies that the matrix on the left has full rank, meaning that also $x_k \to 0$ as was required. The proof has been concluded. ■

7.5. Notes

The development in this section is mostly standard [see e.g. Bertsekas, 1995, Chapter 4]. The main result of this chapter, Theorem 7.3.1 can be stregthened in a number of ways. It not necessary that the terminal cost is zero for the limiting solution to be P. Moreover, controllability is a more stringent assumption than necessary and in fact stabilizability combined with observability of the costs is sufficient [Kučera, 1972].

Part III.

Experiment Design and Statistical Optimality

Part IV.

Further Topics

Bibliography

Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena scientific Belmont, MA, 1995.

Ben-Zion Bobrovsky, E Mayer-Wolf, and M Zakai. Some Classes of Global Cramér-Rao Bounds. *The Annals of Statistics*, pages 1421–1438, 1987.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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. PMLR, 2018.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of linear systems: Fundamental limits and optimal algorithms. *IEEE Transactions on Automatic Control*, 2022.

Vladimír Kučera. The discrete riccati equation of optimal control. *Kybernetika*, 8(5):430–447, 1972.

Bruce D Lee, Ingvar Ziemann, George J Pappas, and Nikolai Matni. Active learning for control-oriented identification of nonlinear systems. *arXiv preprint arXiv:2404.09030*, 2024.

Lennart Ljung. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, 28, 1999.

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.

Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From Self-Tuning Regulators to Reinforcement Learning and Back Again. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3724–3740. IEEE, 2019.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

Benjamin Recht. A Tour of Reinforcement Learning: The View From Continuous Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

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, 2019.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

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.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652. PMLR, 2021.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Tingnan Zhang, Jie Tan, and Vikas Sindhwani. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning*, pages 1–10. PMLR, 2020.

Ingvar Ziemann. A note on the smallest eigenvalue of the empirical covariance of causal gaussian processes. *IEEE Transactions on Automatic Control*, 2023.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. In *Advances in Neural Information Processing Systems*, volume 35, pages 4626–4637, 2022.

Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J Pappas. A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8921–8939. IEEE, 2023.

Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In *Forty-first International Conference on Machine Learning*, 2024.