Instance-dependent analysis of learning algorithms

by

Ruitong Huang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

Abstract

On the one hand, theoretical analyses of machine learning algorithms are typically performed based on various probabilistic assumptions about the data. While these probabilistic assumptions are important in the analyses, it is debatable whether such assumptions actually hold in practice. Another question is whether these probabilistic assumptions really catch the essence of "learning" as is implicitly assumed since the introduction of PAC models in learning theory. On the other hand, when avoiding making assumptions about the data, typical analyses tend to follow a worst-case minimax approach, e.g. in the adversarial online learning framework. Oftentimes the results obtained will fail to catch and exploit the 'niceness' of the data that may help speeding up the learning. It is also debatable whether the data encountered in typical learning scenarios would ever be truly adversarial.

Motivated by the above issues, this thesis suggests to perform instance-dependent analysis of learning algorithms to improve our understanding of learning. Special emphasis is put on characterizing the 'niceness' of data from the perspective of learning. In this thesis, we demonstrate this approach in three settings:

- In the unsupervised learning setting, we redefine the problem of independent component analysis (ICA) to avoid any kind of stochastic assumptions and develop (for the first time) a provably polynomial-time learning algorithm based on our deterministic analysis.

- In the supervised learning setting, we start with a statistical framework: We analyze the finite-sample performances of the empirical risk minimization algorithm for a generalized partially linear model under the random design setting. We detect a potential deficiency of the ERM algorithm. Further investigation leads to a high probability instance-dependent generalization bound.

- Finally, in the online learning setting, we take a thorough analysis of the follow the leader (FTL) algorithm in the online linear prediction problem, and discover a broad

range of previously unknown favourable conditions of the data under which FTL achieves a fast learning rate.

Our approach leads to various instance-dependent results that are more general, expressive, and meaningful, in the sense that these results are able to catch the important factors of the data on which the performances of the learning algorithms heavily rely.

To my parents.

Acknowledgements

There are many people without whom this work would never have been possible. I worry that my words below cannot adequately express my appreciation for them.

My utmost gratitude goes to my supervisors, Csaba Szepesvári and Dale Schuurmans. Along the long journey of my PhD, I have received invaluable guidance and extensive support from them in research and in life. Csaba has been guiding me through every step from the very beginning of my research to the final thesis revisions. His patience and indulgence has tremendously helped me grow as a researcher. Dale has been generous in sharing with me his wisdom, his insightful critiques and constructive ideas, and for being an inspiration to me. I am also very grateful to my supervisors for their encouragement and the freedom they gave me to explore and pursue various research directions in Machine Learning.

I am deeply grateful to András György for serving in my supervisory committee, for many fruitful discussions we had during the last few years, and for sharing his humor. I also thank the other members of my thesis examining committee, Zachary Friggstad, Russell Greiner, and Nicolò Cesa-Bianchi, for taking time from their tight schedule to gather for my defense, for their great feedback on this work, as well as the lively discussions during my defense.

I would also like to thank Björn Hoffmeister, Sree Hari Krishnan Parthasarathi, Brian King, and Roland Maas for taking me as an intern at Amazon Alexa in 2015. I gained valuable industrial research experience.

I am lucky to know Fan Xie, Bing Xu, Xinhua Zhang, and Yaoliang Yu during my PhD. I am grateful that they always have belief in the value of my research. During my many years at the University of Alberta, I have also received enormous support from all of my friends: Pooria Joulani, Xiaochao Fan, Özlem Aslan, Dan Han, and many others. Their friendship has brought me a lot of joy. Many thanks go to my friends in the Department of Computing Science for making my time fun and memorable.

I also thank my family for their unwavering support. Finally, I thank Longlu Qin for her support and encouragement with love and understanding. She has been my cheerleader, my project manager, my dietitian, my lover, and my friend. For the last five years through all my ups and downs, I have been fortunate to have her faith and trust, and her tolerance of

all the nonsense I may have made. Despite the completion of my PhD, I suspect that part of my nonsense is going to persist. My hope is that her company will persist as well.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Learning theory is studied to better understand the relative strengths and weaknesses of a learning algorithm under various assumptions. As such, it helps to improve existing algorithms and to design new ones.

However, most existing theoretical results in the literature tend to flip between two polarities: On the one hand, learning theory traditionally has been studied in a statistical framework, discussed at length, for example, by Kearns and Vazirani [1994], Vapnik [1998], Shalev-Shwartz and Ben-David [2014]. The data generating mechanism in the statistical framework is assumed to obey probabilistic assumptions. These strong assumptions make the analysis of the learning algorithm easier by only considering 'nice' data for the algorithm. Examples of assumptions imposing niceness include that the individual observations in the data are independent and follow the same distribution (the so called independent and identically distributed (i.i.d.) assumption), various other restrictions in terms of what dependencies exist between the observed data points, and low noise or variance of the instances, etc. One issue with this approach is that the analysis of the algorithm seems to critically depend on whether the data generating mechanism indeed satisfies the stated probabilistic assumptions, which is difficult to argue in practice. In particular, in practice, it is difficult to argue that the data indeed satisfies these probabilistic assumptions. Does it mean that one should avoid using the algorithms whose performance is analyzed in the statistical framework? Perhaps not: Experimentally, it appears that many algorithms achieve good performance when the stated strong probabilistic assumptions are apparently violated. Theory that uses the strong assumptions is not suitable to explain these successes. This observation raises the question of what should be used as the 'niceness' of the data for a particular learning algorithm, going beyond the statistical framework. On the other hand, existing results in the literature that make the least assumptions on the data generating mechanism, are usually centered around a worst-case analysis. In the statistical learning framework, the majority of the results are concerned about the worst-case data under the assumed probabilistic assumptions. The online learning framework tends to make minimal

assumptions on the data, but performs the analysis in an 'adversarial' environment [Cesa-Bianchi and Lugosi, 2006] and thus it is also associated with a strong worst-case flavour. One classical example of what this worst-case approach entails in the online learning framework is as follows: A deterministic strategy can always be exploited by its adversary in the worst-case, hence achieves only trivial performance guarantee. Such observation inspires the design of randomized strategies, like the follow the perturbed leader (FTPL) algorithm [Rakhlin and Sridharan, 2014, van Erven et al., 2014]. But does this result really mean that one should stay away from deterministic strategies? In fact, under appropriate circumstances, the follow the leader algorithm (FTL), which is deterministic, has been proven to achieve fast learning rates (see, e.g., Merhav and Feder 1992, Gaivoronski and Stella 2000, Hazan et al. 2007, Kakade and Shalev-Shwartz 2009).

In this thesis, we propose to analyze learning problems and learning algorithms in a framework that aims to build a bridge between two polarities: the overly conservative worst-case framework and the overly restrictive statistical framework. More specifically, we start with minimal statistical assumptions on the observed data, and perform a deterministic analysis of the learning algorithm. A successful analysis in this way usually leads to a result that depends on some particular quantities of the data. We call such results *instance-dependent*. These particular quantities of the data can usually serve as measures of the 'niceness' of the data, which catch the essential features of data on which the performance of the algorithm heavily relies. Results of this type lead to better understanding of the successes and failures of various learning algorithms, and also hint at design principles for new algorithms. Our 'niceness' notions apply to a wider range of data including the classical stochastic data and the worst-case data, showing the strength and universality of the approach. The idea of the two-step analysis, first developing an instance-dependent result and then applying this to specific settings, has actually appeared in both batch learning setting and online learning setting [Vito et al., 2005, Cesa-Bianchi and Lugosi, 2006, Pires and Szepesvári, 2012, Rakhlin and Sridharan, 2014]. Demonstrations can also be found in the recent NIPS workshops 'learning faster from easy data'.[1] Another interesting question that we won't study is how these quantities can be computed in practice. While in some cases, the niceness measures are computable from the empirical data, there exist other cases when the 'niceness' measure is not available. We also discuss how we could deal with the latter cases like introducing some form of regularization, or developing an adaptive algorithm that can achieve fast rates for the 'nice' data while still maintains the optimal minimax learning rate. In this thesis, we present 3 demonstrations in different learning settings: unsupervised learning, (batch) supervised learning, and online learning. While in all these settings we emphasized the importance of improving our understanding of the

[1] http://wouterkoolen.info/easydata2013/ and http://event.cwi.nl/easydata2015/

behaviour of the learning algorithms, our new analyses also challenge these algorithms, and lead to the development of new algorithms with improved theoretical guarantees.

1.1 Learning Settings

The purpose of this section is to introduce the 3 types of learning settings used in this thesis: unsupervised learning, supervised learning, and online learning. Under the setting of supervised learning, the data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a pair of the feature space $\mathcal{X}$ and the response space $\mathcal{Y}$, and the data (X, Y) is in a input-output pair where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Assuming there exists an underlying relation between X and Y, the goal is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ such that f is close to this underlying relation. For example, each pair (X, Y) can be generated from some distribution over $\mathcal{X}$ and $Y = \|X\|_2$, and ideally the algorithm will seek the length function to be f. Under the setting of unsupervised learning, no variable is treated as the response (thus $\mathcal{Z} = \mathcal{X}$), and the goal of learning is to models general properties of data from the presented data. Lastly under the setting of online learning, data are presented sequentially and the learning goal is to make prediction for the current sample based on the history. Supervised learning and unsupervised learning are instances of batch (off-line) learning: a batch of data forms the basis of learning. We present a brief introduction in this chapter, and leave the details in the latter corresponding chapters.

This thesis is organized as follows: In the rest of this chapter we introduce briefly our results in the above learning settings. In Chapter 2 we present our first demonstration in the unsupervised learning setting where we consider a deterministic framework for the independent component analysis model. Chapter 3 is devoted to our results in the supervised learning setting, where we consider the performance of the empirical risk minimization algorithm (ERM) for the generalized partially linear regression model. The last demonstration in the online learning setting is presented in Chapter 4. Chapter 5 concludes the thesis and discusses some potential future works.

1.1.1 Unsupervised learning

In the setting of unsupervised learning, one usually assumes the samples are generated from a presumed probabilistic model with unknown parameters. Different from the supervised setting, the goal of unsupervised learning is to estimate these parameters and possibly perform some inference with the model. Examples of unsupervised learning include latent variable models and clustering.

Among various latent variable models, independent component analysis (ICA) is analyzed in this thesis as a typical example. The task of ICA is given in a statistical language: Given T i.i.d. observed d-dimensional signals $x \in \mathbb{R}^{d \times T}$ as a data array that is assumed to

be a linear mixture of d independent hidden source signals, ICA attempts to estimate the unknown mixing matrix A and the hidden signals s such that $x = As$ [Hyvärinen et al., 2001]. For simplicity, we assume that A is a non-singular matrix. The key assumption for the ICA problem to be well defined is that all d source signals are mutually independent, and observations are i.i.d. samples from the source distributions. However, very commonly ICA algorithms are applied to data where strong temporal correlations are apparent or even when the input is deterministic. The success of ICA on such data suggests that the usual statistical notions may not capture the very essence of the task. This motivates us to re-define and analyze ICA in a deterministic framework without probabilistic assumptions on the data.

One approach in unsupervised learning, besides the maximum likelihood estimator, is the method of moments which estimates the model's parameters via the estimates of moments of some specifically designed variables. Recently, the method of moments becomes popular in the machine learning literature, mainly because it leads algorithms with appealing theoretical and computational properties [Arora et al., 2012, Anandkumar et al., 2012c,a, Hsu and Kakade, 2013]. The essence of the method of moments is that the exact value of the first or higher moments can be formulated as a tensor whose decomposition can be used to extract a model's parameters. While, in general, tensor decomposition is NP-hard [Hillar and Lim, 2013], it is shown that in a wide range of latent variable models, the resulting tensor is symmetric and orthogonally decomposable [Zhang and Golub, 2001, Anandkumar et al., 2012a].

In Chapter 2 we show that the problem of ICA can indeed be interpreted and analyzed in a non-stochastic manner by assuming the source signals s to be deterministic. Previous ICA results are extended to more general settings. Such analysis leads to a new provable polynomial-time ICA algorithm free of unspecified parameters.[2] We argue that similar ideas may also work for other latent variable models in the work of Anandkumar et al. [2012a]. The results in Chapter 2 have been published in our ICML and NIPS workshop papers [Huang et al., 2015a,b].

1.1.2 Supervised learning

Supervised learning is also referred as predictive learning. Given a training set of n samples $Z_{1:n} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ generated from some underlying distribution $\mathbb{P}$ over $\mathcal{Z}$, a learning algorithm learns a mapping $f : \mathcal{X} \to \mathcal{Y}$ from a set of functions $\mathcal{F}$ such that $f(X)$ is close to Y for a data pair (X, Y) generated from $\mathbb{P}$. $\mathcal{F}$ is called the *hypothesis set*. Given a loss function $\ell(y, y')$, the 'suitability' of a mapping $f \in \mathcal{F}$ is defined by its expected

[2] An algorithm that is free of unspecified parameters is important in the unsupervised learning setting, especially in the deterministic framework. Because we make no assumptions about the data generating mechanism, parameter tuning is almost impossible.

loss $L(f) = \mathbb{E}[\ell(Y, f(X))]$ with respect to $\mathbb{P}$. As the underlying distribution $\mathbb{P}$ of the data (X, Y) is unknown, it is natural to learn the mapping f by minimizing the empirical loss $L_n(f) = \frac{1}{n}\sum_i \ell(f(X_i), Y_i)$ given the samples. Denote the minimizer of L_n by f_n. We also denote $\{X_1, \ldots, X_n\}$ by $X_{1:n}$, and $\{Y_1, \ldots, Y_n\}$ by $Y_{1:n}$.

Note that the loss $L(f_n)$ depends on both the learning algorithm and the nature of the problem, i.e. the underlying distribution $\mathbb{P}$ in $\mathcal{Z}$. To better describe the performance of the learning algorithm, we compare the loss $L(f_n)$ to the best loss possible over the considered set of functions $\mathcal{F}$, i.e., to $L^* = \inf_{f\in\mathcal{F}} L(f)$. A bound on the *excess risk* $L(f_n) - L^*$ is called a generalization (error) bound. Several approaches have been proposed to analyze the excess risk, including the theory of uniform convergence of empirical processes, and stability analysis etc. [Vapnik, 1995, Bousquet and Elisseeff, 2002].

In the statistical literature, instead of bounding the expected excess risk $\mathbb{E}[L(f_n) - L^*]$ as in machine learning, most of the existing results are concerned with the expected excess risk conditioned on the samples $X_{1:n}$ [van de Geer, 1990, 2000]. With our notations, this amounts to bounding $\mathbb{E}[L_n(f_n)|X_{1:n}] - \inf_{f\in\mathcal{F}} \mathbb{E}[L_n(f)|X_{1:n}]$. We call the conditionally expected excess risk $\mathbb{E}[L_n(f_n)|X_{1:n}] - \inf_{f\in\mathcal{F}} \mathbb{E}[L_n(f)|X_{1:n}]$ as "in-sample" error, and $\mathbb{E}[L(f_n) - L^*]$ as "out-of-sample" error. An intuitive connection between these two different settings is that a result in the random design setting can be developed from a result in the fixed design setting that holds uniformly for arbitrary $X_{1:n}$. In Chapter 3 we surprisingly detect a potential deficiency of the ERM algorithm on a linear regression problem which an asymptotic analysis fails to catch. For the first time in the literature, we analyze the finite-sample performances of the ERM algorithm on a generalized partially linear model. Our instance-dependent finite-sample bound catches a 'niceness' measure of the data which helps partially explain the success of ERM on this model. The results of Chapter 3 have appeared in our AISTATS and ISAIM papers [Huang and Szepesvári, 2014a,b].

1.1.3 Online Learning

In the last two decades, a fair amount of interest in statistical learning theory has been devoted to studying online learning. The idea of not making statistical assumptions on sequential data is not new and goes back to at least Cover [1966]. In fact, the adversarial setting is one of the two classic settings in online learning, where the input sequence is assumed to be intentionally generated to maximize the algorithm's regret and thus it is non-stochastic.[3]

A generic online learning framework is as follows: In round $t = 1, \ldots, n$, the algorithm picks a weight $w_t \in \mathcal{W} \subset \mathbb{R}^d$ for some $\mathcal{W}$ (*e.g.* $\mathcal{W} = \{w : \|w\|_2 \leq 1\}$). Then a convex loss function $\ell_t : \mathbb{R}^d \to [0, 1]$ is generated from some unknown system and the algorithm

[3]The other setting is called stochastic setting where the input sequence are i.i.d. samples generated from some fixed distribution.

suffers the loss $\ell_t(w_t)$ and learn the loss function ℓ_t. The goal of the learning algorithm is to minimize the regret in n rounds:

$$R_n = \sum_{t=1}^{n} \ell_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{n} \ell_t(w).$$

In this thesis we only consider the "online linear prediction" when the loss is linear, i.e. ℓ_t is linear in w. In fact, the linear case is not only simple, but is also fundamental since the case of nonlinear loss functions can be reduced to it using a standard linearization trick: Indeed, even if the losses are nonlinear, defining $f_t \in \partial \ell_t(w_t)$ to be a subgradient[4] of ℓ_t at w_t and letting $\tilde{\ell}_t(u) = \langle f_t, u \rangle$, by the definition of subgradients, $\ell_t(w_t) - \ell_t(u) \leq \ell_t(w_t) - (\ell_t(w_t) + \langle f_t, u - w_t \rangle) = \tilde{\ell}_t(w_t) - \tilde{\ell}_t(u)$, hence for any $u \in \mathcal{W}$,

$$\sum_{t=1}^{n} \ell_t(w_t) - \sum_{t=1}^{n} \ell_t(u) \leq \sum_{t=1}^{n} \tilde{\ell}_t(w_t) - \sum_{t=1}^{n} \tilde{\ell}_t(u)\,.$$

In particular, if an algorithm keeps the regret small no matter how the linear losses are selected (even when allowing the environment to pick losses based on the choices of the learner), the algorithm can also be used to keep the regret small in the nonlinear case.

One of the basic online learning algorithm is the 'Following The Leader' algorithm (FTL) [Cesa-Bianchi and Lugosi, 2006]. In each round t, FTL picks w_t such that w_t minimizes the total loss $\sum_{i=1}^{t-1} \ell_i(w)$ accumulated beforehand. It has been proved that FTL achieves optimal constant regret under the stochastic setting. However, it is also shown that FTL may suffer a linear regret in the adversarial setting [Shalev-Shwartz, 2012, De Rooij et al., 2014].[5] This extreme behavior can be avoided by introducing a regularization to the objective function of FTL. The resulting algorithm, named 'Follow The Regularized Leader' (FTRL), achieves the optimal minimax rate ($\sqrt{n}$ in a sequential data of length n) e.g. on the online linear prediction problem [Abernethy et al., 2008, Shalev-Shwartz, 2012].

In Chapter 4, we take a closer look at the performance of the FTL algorithm on the online linear prediction problem. Our result essentially suggests that although FTL may perform extremely bad for the worst-case data, it actually achieves a fast learning rate for some data. Our analysis also motivates us to develop various new adaptive algorithms for the online linear prediction problem. The results in Chapter 4 have been published in our NIPS paper [Huang et al., 2016].

[4] We let $\partial g(x)$ denote the subdifferential of a convex function $g : \mathrm{dom}(g) \to \mathbb{R}$ at x, i.e., $\partial g(x) = \{\theta \in \mathbb{R}^d \,|\, g(x') \geq g(x) + \langle \theta, x' - x \rangle \ \forall x' \in \mathrm{dom}(g)\}$, where $\mathrm{dom}(g) \subset \mathbb{R}^d$ is the domain of g.

[5] A linear regret in online learning indicates that the learning algorithm is not doing better than a random guessing.

Chapter 2

Independent Component Analysis

In Chapter 1 we suggest to develop instance-dependent theoretical guarantees to achieve more expressive results and have a better understanding of the learning algorithm. In this chapter, we present our first demonstration on the task of Independent Component Analysis (ICA) in the unsupervised learning setting. We will not make probabilistic assumptions about the data generating mechanism, and characterize the important features of the data on which the ICA algorithm is guaranteed to have a good performance. We start with analyzing an ICA algorithm due to Hsu and Kakade [2013], named as HKICA. A new ICA algorithm is then proposed to fix a potential deficiency of HKICA. A key feature of our approach is that no probabilistic assumptions are made on the data[1], and the algorithm is free of hyperparameters.

This chapter is organized as follows: We introduce the classic ICA model in Section 2.1. Previous works are discussed in Section 2.1.1. We then redefine the ICA model in a deterministic framework and present our main results in Section 2.2. The polynomial-time algorithms underlying these results are developed through Section 2.3 and 2.4: Section 2.3 is devoted to the analysis of the HKICA algorithm. Our new algorithms are presented and analyzed from Section 2.4.1 to Section 2.4.3. Then Section 2.4.4 proposes a recursive version of our algorithm. Lastly, we present some simulation results in Section 2.5 and conclusions in Section 2.6.

The results in this chapter have appeared in Huang et al. [2015a,b].

2.1 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) attempts to explain an observed $x \in \mathbb{R}^{d\times T}$ array by decomposing it into the product As where $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix and $s \in \mathbb{R}^{d\times T}$ is viewed as T d-dimensional vectors such that the components of these T vectors

[1] In Chapter 3 we still keep the probabilistic assumptions on the noise ϵ.

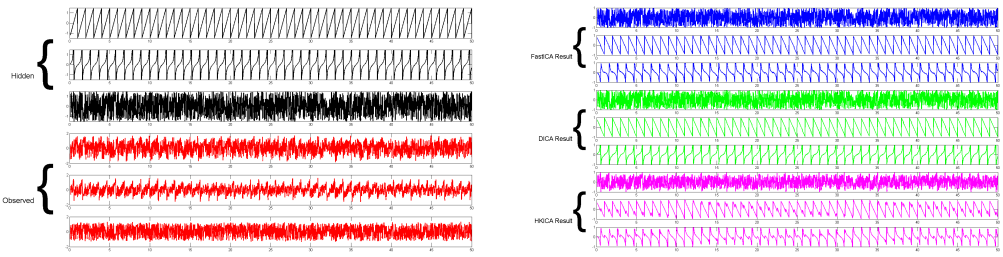

Figure 2.1: Example of ICA: On the left-hand side, the bottom three plots depict the $d = 3$ components of the observed signal x. This observed data x is generated by mixing the sources shown by the top three graphs on the left-hand side. The x axis represents time: The numbers shown are scaled by a factor of 50; thus $T = 2500$. Reconstruction results by three algorithms (FastICA, HKICA, and DICA, see Section 2.1.1 and 2.4.1 for details)

are "statistically independent" [Hyvärinen et al., 2001]. The ICA literature is vast in both practical algorithms and theoretical analyses; we only discuss those closely related to our work in Section 2.1.1, but refer to the book of Comon and Jutten [2010] for a comprehensive survey. Oftentimes, ICA is illustrated by data as shown in Figure 2.1. Three hidden source signals are shown in black, and their linear mixtures as the observations for three different ICA algorithms are in red. The reconstruction of the source signals by the ICA algorithms is shown on the right-hand side. As can be seen, up to scaling and the ordering of the reconstructed components, the reconstruction is quite successful no matter the algorithms. But are these hidden signals "statistically independent"? Note that the all components of the source data are periodic functions of time. This is quite obvious for the first two components, while the third, being generated using a pseudo-random number generator has a long period and thus "looks random". Also, as the reader may recall, any two constant numbers $a, b \in \mathbb{R}$ are independent of each other when viewed as degenerate random variables. Thus, for any single t, the three components of the source signals $s_1(t)$, $s_2(t)$, and $s_3(t)$ are independent! Does the success of the algorithms on this example imply that they will also work for other mixtures of arbitrary deterministic sources? Of course not. For example, if one source is a linear function of other sources, then no algorithm will be able to recover the sources from their mixture. Another question is whether the temporal dependency of the sources may hamper performance. If $s(1) = s(2) = \cdots = s(T)$ then the algorithms effectively need to work with a single vector observation and no algorithm will be able to perform a successful reconstruction.

We attempted to provide some answers to the question in this chapter: to what extent can ICA algorithms separate the mixture of some sources? In particular, can we extend/redefine the problem of ICA in a meaningful way so that we can explain the success of the particular ICA algorithms on the above example? In this chapter, we give a positive answer to this question. The essence of our approach is to define an empirical measure of the "niceness" of

data and then postulating the requirement that good algorithms are those that get better results on 'nice' data. The niceness measure will be so that in the classical ICA setting, the usual statistical results can be recovered from our results. We also propose a *provably polynomial-time* algorithm that has no free (unspecified) parameters for the noisy ICA model and analyze its performance based on the "niceness" of the observed data. The key feature of our approach is that no probabilistic assumptions are made on the data (the algorithm may randomize though), and thus this work can be thought as the natural extension of online learning where learning algorithms are analyzed without making any probabilistic assumptions [Cesa-Bianchi and Lugosi, 2006].

2.1.1 Related Works

We explain the difference between our ICA algorithm and the previous methods in this section. However, as mentioned above, the key feature of our results is the removal of the probabilistic assumptions on the observed data.

A popular approach to the ICA problem is to find a linear transformation W for X by optimizing a *contrast function* that measures dependence or non-gaussianity of the resulting coordinates of WX. The optimal W then can serve as an estimate of A^{-1}, thereby recovering the mixing matrix A. One of the most popular ICA algorithms, FastICA [Hyvarinen, 1999], follows this approach for a specific contrast function. FastICA has been analyzed theoretically from many aspects [Tichavsky et al., 2006, Oja and Yuan, 2006, Ollila, 2010, Dermoune and Wei, 2013, Wei, 2014, Miettinen et al., 2014]. In particular, recently Miettinen et al. [2014] showed that in the *noise-free* case (i.e., when $X = As$), the error of FastICA (when using a particular fourth-moments-based contrast function) vanishes at a rate of $1/\sqrt{T}$ where T is the sample size. In addition, several other methods have been shown to achieve similar error rates in the noise-free setting [e.g., Eriksson and Koivunen, 2003, Samarov and Tsybakov, 2004, Chen and Bickel, 2005, 2006]. However, to our knowledge, no similar finite sample results are available in the noisy case. Algorithms developed based on the *noise-free* assumptions are also observed to be sensitive to the noise in practice [Mollah et al., 2007].

On the other hand, promising algorithms are available in the noisy case that make significant advances towards provably efficient and effective ICA algorithms, albeit fall short of providing a complete solution. Using a quasi-whitening procedure, Arora et al. [2012] reduces the problem to finding all the local optima of a specific function defined using the fourth order cumulant, and propose a polynomial-time algorithm to find them with appealing theoretical guarantees. However, the results depend on an unspecified parameter (β in the original paper) whose proper tuning is essential; note that even an exhaustive search over β could be problematic, since it is unclear how one can use data to infer the

range for β.

The exploitation of the special algebraic structure of the fourth moments induced by independence leads to several other works related to ICA [Hsu and Kakade, 2013, Anandkumar et al., 2012a,b]. A similar idea is also discussed earlier as an intuitive argument to construct a contrast function [Cardoso, 1999]. The first rigorous proofs for this idea are developed using matrix perturbation tools in a general tensor perspective [Anandkumar et al., 2012a,b, Goyal et al., 2014]. A common problem faced by these methods is a minimal gap of the eigenvalues, which may result in an exponential dependence of the sample complexity on the number of source signals d. More precisely, these methods all require an eigen-decomposition of some flattened tensor where the minimal gap between the eigenvalues plays an essential role. Although the exact size of this gap is not yet understood, a naive analysis introduces an exponential dependence on the dimension d. Such dependence is also observed in the literature [Cardoso, 1999, Goyal et al., 2014]. One way to circumvent such dependence is to directly decompose a high-order tensor using the power method, which requires no flattening procedure [Anandkumar et al., 2014]. However, when applied to the ICA problem, this introduces a bias term and so the error does not approach 0 as the sample size approaches infinity. Another issue is the well-known fact that the power method is unstable in practice for high-order tensors. Goyal et al. [2014] proposed another method by exploring the characteristic function rather than the fourth moments. However, their guarantees hold only if a parameter of their algorithm (σ in the original paper) happens to be smaller than some instance-dependent quantity which in general is unknown, making their guarantee weak. Recently, Vempala and Xiao [2014] proposed an ICA algorithm based on an elegant, recursive version of the method of Goyal et al. [2014] that avoids dealing with the aforementioned minimal gap; however, they still need an oracle to set the unspecified parameter of the algorithm of Goyal et al. [2014].

Our ICA algorithm is a refined version of the ICA method proposed by Hsu and Kakade [2013] (HKICA). However, we propose two simpler ways, one inspired by the works of Frieze et al. [1996] and Arora et al. [2012], and another based on Vempala and Xiao [2014], to deal with the spacing problem of the eigenvalues under similar conditions to those of Goyal et al. [2014]. Unlike the method proposed by Goyal et al. [2014], our first method can force the eigenvalues to be well-separated with a gap that is independent of the mixing matrix A, while our second method, based on the recursive decomposition idea of Vempala and Xiao [2014], avoids dealing with the minimum gap (at the price of introducing other complications). We prove that our methods achieve an $O(1/\sqrt{T})$ error in estimating A in the classic setting, with high probability, such that both the convergence rate and the computational complexity scale *polynomially* with the natural parameters of the problem. Our method needs no parameter tuning, making it the first method to provably handle noisy

ICA without relying on a "lucky" parameter choice.

The problem of separating mixture of deterministic signals is also considered in [Kirimoto et al., 2011] and [Forootan and Kusche, 2013]. However their analysis is restricted to certain particular signals, while our result is applicable to general ones.

2.1.2 Notation

All vectors, matrices and tensors are real or complex valued, unless otherwise stated. Symbol K denotes either the set of real or complex numbers. We denote the set of real and natural numbers by $\mathbb{R}$ and $\mathbb{N}$, respectively. A vector $v \in K^d$ is assumed to be a column vector. Let $\|v\|_2$ denote its L_2-norm, and for any matrix Z let $\|Z\|_2 = \max_{v:\|v\|_2=1} \|Zv\|_2$ denote the corresponding induced matrix-norm. Denote the maximal and minimal singular value of Z by $\sigma_{\max}(Z)$ and $\sigma_{\min}(Z)$, respectively. Also, let Z_i and $Z_{i:}$ denote the ith column and, resp., row of Z, and let $Z_{(2,\min)} = \min_i \|Z_i\|_2$, $Z_{(2,\max)} = \max_i \|Z_i\|_2$ and $Z_{\max} = \max_{i,j} |Z_{i,j}|$. Clearly, $\sigma_{\max}(Z) = \|Z\|_2 \geq Z_{(2,\max)} \geq Z_{\max}$, and $\sigma_{\min}(Z) \leq Z_{(2,\min)}$. For a tensor (including vectors and matrices) T, its Frobenious norm (or L_2 norm) $\|T\|_F$ is defined as the square root of the sum of the square of all its entries. For a vector $v = (v_1, \ldots, v_d) \in K^d$, $|v|$ is defined coordinatewise: $|v| = (|v_1|, \ldots, |v_d|)$. Similarly for a matrix $M \in K^{d\times d}$, $|M|$ is also defined coordinatewise. The transpose of a vector/matrix Z is denoted by $Z^\top$, while the inverse of the transpose is denoted by $Z^{-\top}$. The outer product of two vectors $v, u \in K^d$ is denoted by $u \otimes v = uv^\top$. The symbol $v^{\otimes k}$ denotes the k-fold outer product of v with itself, that is, $v \otimes v \otimes v \ldots \otimes v$, which is a k-dimensional tensor. Given a 4-dimensional tensor T, $T(\eta, \eta, \cdot, \cdot)$ denotes a matrix Z that is generated by marginalizing and scaling the first two coordinates of T on the direction η: $Z_{i,j} = \sum_{k_1,k_2=1}^d \eta_{k_1}\eta_{k_2}T_{k_1,k_2,i,j}$. (Similar definitions apply to marginalizing different coordinates of the tensor.) For a real vector v and some real number C, $v \leq C$ means that all the entries of v are at most C. The bold symbol $\mathbf{1}$ denotes a vector with all its entries equal to one (the dimension of this vector will always be clear from the context). Finally, $\mathrm{Poly}(\cdot, \cdots, \cdot)$ denotes a polynomial function of its arguments.

2.2 Deterministic ICA

We consider the following non-stochastic version of the ICA model. Assume that we are given a $d \times T$ matrix x. For $t \in [T]$, let $x(t) \in \mathbb{R}^d$ be the tth column of x. We consider the problem of reconstructing a $d \times d$ non-singular mixing matrix A from x such that

$$x(t) = As(t) + \epsilon(t), \qquad 1 \leq t \leq T\,, \tag{2.1}$$

and (A, s, ϵ) is "nice" in a way to be defined shortly. Intuitively, s is the source whose components are "independent" while ϵ is "noise". We measure how well a matrix $\hat{A}$ constructed

by an algorithm working on data x recovers A by the reconstruction error defined as

$$d(\hat{A}, A) = \inf_{\substack{\pi \in \mathrm{Perm}([d]) \\ c \in \mathbb{R}^d}} \max_k ||c_k \hat{A}_{:\pi(k)} - A_{:k}||_2 \,,$$

where $A_{:i}$ stands for the ith column of A and $\mathrm{Perm}([d])$ is the set of all the permutations on the set $[d]$. This measure compensates for the inherent indeterminacy in reconstructing the scale and ordering of sources. In this chapter, without loss of generality (WLOG), we assume $s(t) \in [-C, C]^d, t \in [T]$ for some constant C (by scaling A if necessary). Since $\sigma_{\min}(A)$ and $\sigma_{\min}(A)$ will appear frequently, when A is clear from the context, we use the shorthands $\sigma_{\min} = \sigma_{\min}(A)$ and $\sigma_{\max} = \sigma_{\max}(A)$.

Let us now develop the "niceness" measure of the tuple (x, A, s, ϵ). We start with defining a family of "distances" between distributions (strictly speaking, these are only pseudo-distances): given two distributions ν_1 and ν_2 over $\mathbb{R}^d$ and an integer $k \geq 1$, let

$$D_k(\nu_1, \nu_2) = \sup_{f \in \mathcal{F}} \left| \int f(s)\nu_1(ds) - \int f(s)\nu_2(ds) \right|,$$

where $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} : f(s) = \prod_{j=1}^k s_{i_j}, 1 \leq i_1, \ldots, i_k \leq d\}$ is the set of all monomials up to degree k. When μ is a product measure, $D_k(\mu, \nu)$ measures how close the components of $X \sim \nu$ are to being independent. When ν is a measure of $p+q$ variables (i.e., $X \in \mathbb{R}^{p+q}$), we also need a measure that quantifies the degree of independence of the vectors $(X_1, \ldots, X_p)$ and $(X_{p+1}, \ldots, X_{p+q})$. We will denote this measure by $D_k^{(p,q)}(\nu)$ and is defined as

$$D_k^{(p,q)}(\nu) = D_k(\mu_1 \otimes \mu_2, \nu),$$

where μ_1 (respectively, μ_2) is the marginal measure of ν on the first p (respectively, last q) variables, i.e. for a Borel set $B \subset \mathbb{R}^p$ (respectively, $\mathbb{R}^q$), $\mu_1(B) = \nu(B \times \mathbb{R}^q)$ (respectively, $\mu_2(B) = \nu(\mathbb{R}^p \times B)$). We will use $D_k^{(p,q)}(\nu)$ to measure the degree of independence between the source (s) and the noise (ϵ).

For a distribution μ over $\mathbb{R}$ we let $\kappa(\mu)$ be the (absolute) 4^{th}-order cumulant of μ:

$$\kappa(\mu) = \left| \int x^4 \mu(dx) - 3\left(\int x^2 \mu(dx)\right)^2 \right|,$$

which, for brevity, we also call "kurtosis" (by slightly abusing terminology). For a product distribution $\mu = \mu_1 \otimes \ldots \otimes \mu_d$ over $\mathbb{R}^d$, we let $\kappa_{\min}(\mu) = \min_{1 \leq i \leq d} \kappa(\mu_i)$ to denote the minimum kurtosis of the components of μ. We also denote $\kappa(\mu_i)$ by κ_i. When μ is a distribution over $\mathbb{R}^d$, we define the d-dimensional absolute 4^{th}-order cumulant of μ by

$$\mathcal{Q}(\mu) = \max_{\|\eta\|_2 \leq 1} \| \left(\mathbb{E}_{Y \sim \mu}[Y^{\otimes 4}] - (\mathbb{E}_{Y \sim \mu}[Y^{\otimes 2}])^{\otimes 2}\right)(\eta, \eta, \cdot, \cdot) - 2(\mathbb{E}_{Y \sim \mu}[Y^{\otimes 2}])^{\otimes 2}(\eta, \cdot, \eta, \cdot)\|_F.$$

We will also use $N(\nu) = \| \int x\nu(dx)\|_F$ to denote the magnitude of the mean of the distribution ν. Now, for any $t, k \geq 1$ and signal $u : [t] \to \mathbb{R}^k$, we introduce the empirical distribution of u, $\nu_t^{(u)}$, which assigns the value

$$\nu_t^{(u)}(B) = \tfrac{1}{t}|\{\tau \in [t] : u(\tau) \in B\}|$$

to a Borel set $B \subset \mathbb{R}^k$. We omit the index t when the time range is T, e.g. denote $\nu_T^{(\epsilon)}$ by $\nu^{(\epsilon)}$. Similarly, for signals $u, v : [t] \to \mathbb{R}^k$ and a Borel set $B \subset \mathbb{R}^{2k}$, we define

$$\nu_t^{(u,v)}(B) = \tfrac{1}{t}|\{\tau \in [t] : (u(\tau), v(\tau)) \in B\}|.$$

Starting from the classical setting, we will define the tuple (x, A, s, ϵ) "nice" (or compliant) if the respective *empirical* distributions approximately satisfy the usual assumptions:

(a) the source (s) components are independent, captured by $D_4(\nu^{(s)}, \mu)$;

(b) the noise and source are independent, captured by $D_4^{(d,d)}(\nu^{(As,\epsilon)})$;

(c) the source (s) have high absolute 4^{th}-order cumulants (kurtosis), captured by combining $D_4(\nu^{(s)}, \mu)$ and the minimal kurtosis of μ, $1/\kappa_{\min}(\mu)$;

(d) the noise (ϵ) has low absolute 4^{th}-order cross cumulants, captured by $\mathcal{Q}(\nu^{(\epsilon)})$;

(e) the source (s) have zero mean, captured by $N(\nu^{(s)})$;

(f) the noise (ϵ) has zero mean, captured by $N(\nu^{(\epsilon)})$;

Finally, we let

$$L = \max\left(\|\textstyle\int [y^{\otimes 2}]\nu^{(\epsilon)}(dy)\|_F, \|\int [y^{\otimes 3}]\nu^{(\epsilon)}(dy)\|_F\right),$$

which captures the magnitude of second and third moments of the noise, and

$$\Pi_0 = \{\mu_1 \otimes \mu_2 \ldots \otimes \mu_d \,|\, \mu_i \text{ is zero mean over } \mathbb{R} \text{ with nonzero kurtosis}, 1 \le i \le d\}$$

to be the set of zero mean product distributions with components of nonzero kurtosis over $\mathbb{R}^d$.

Given all the above notations, the goal of learning is to find an algorithm that, for any tuple (x, A, s, ϵ), produces $\hat{A}$ so that $d(A, \hat{A})$ scales with degree of ICA-compliance ("niceness") of (x, A, s, ϵ).

2.2.1 Main result

Now we are ready to state our main result. The algorithm that achieves this bound will be described later in the next section.

Theorem 2.2.1. *There exists a randomized algorithm such that, for any $A \in \mathbb{R}^{d\times d}$, and $x, s, \epsilon : [T] \to \mathbb{R}^d$ satisfying Equation (2.1), the algorithm returns $\hat{A}$ such that with probability at least $1 - \delta$,*

$$d(\hat{A}, A) \le \inf_{\mu\in\Pi_0} \mathcal{C}(\mu) \min\left(D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)}), \Theta(\mu)\right),$$

where $\mathcal{C}(\mu)$ and $\Theta(\mu)$ are problem dependent constants, polynomial in $(\sigma_{\max}(A), 1/\sigma_{\min}(A), 1/\kappa_{\min}(\mu), 1/\delta, d, L)$. Further, the computational complexity of the algorithm is $O(d^3T)$ when used on any data x of dimensions $T \times d$.

Remark 2.2.1. Our algorithms are randomized algorithms. The high probability event in the first part of Theorem 2.2.1 comes from the random sampling of the algorithm. The observations x are deterministic and thus have no randomness.

Remark 2.2.2. The dependence on δ in our result is Poly $\left(\frac{1}{\delta}\right)$, which is weaker than the usual $\log(\frac{1}{\delta})$.

2.2.2 Applicability of Theorem 2.2.1

Theorem 2.2.1 can be applied to a variety of different ICA settings, including the classical stochastic setting and other various ones.

Proposition 2.2.3. *Let $(s(t))_{t\in[T]}$ be an zero-mean i.i.d. sequence bounded by C in ℓ_∞ norm, independent of the i.i.d. Gaussian noise $\mathcal{N}(0,\Sigma)$ sequence $(\epsilon(t))_{t\in[T]}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, $L = \text{Poly}\left(\|\Sigma\|_2, d, \frac{1}{\delta}\right)$. Moreover, if μ is the product measure of the sources (i.e., $s(t) \sim \mu$), then $D_4(\nu^{(s)}, \mu)$, $\mathcal{Q}(\nu^{(\epsilon)})$, $D_4^{(d,d)}(\nu^{(As,\epsilon)})$, $N(\nu^{(\epsilon)})$, $N(\nu^{(s)})$ are all of orders $O(1/\sqrt{T})$.*

Note that the above result implies that in the standard stochastic setting with independent sources and Gaussian noise, independently generated from the sources, with probability at least $1-\delta$,

$$d(\hat{A}, A) \leq \mathcal{C} \min\left(\frac{1}{\sqrt{T}}, \Theta\right),$$

for some problem-dependent constants $\mathcal{C}$ and Θ.

Our setting can also cover some other examples excluded by the traditional setting, such as the example of Figure 2.1 in Section 2.1.

Example 2.2.4. Assume that the unknown sources s_i $(1 \leq i \leq d)$ are deterministic and periodic. Our observation $x = As+\epsilon$ is a linear mixture of s contaminated by i.i.d. Gaussian noise for each time step, where A is a non-singular matrix and $\epsilon \sim \mathcal{N}(0,\Sigma)$ is Gaussian. Even though ϵ is i.i.d. for every time step, the observations do not satisfy the i.i.d. assumption, since the source s is deterministic. However, it can be proved that if the ratio of the periods of each pair of (s_i, s_j) is irrational, then the reconstruction error would approach 0 for T large enough.

Remark 2.2.5. To have a concrete (noise-free) example, let $s_1(t) = 0.5(-1)^t$, $s_2(t) = \cos(t)$. It is easy to see that the limit distribution of source 1 is a Bernoulli distribution μ_1 with $\mu_1(\{0.5\}) = 1/2$ and $\mu_1(\{-0.5\}) = 1/2$, and the limit distribution of source 2 is a distribution μ_2 with density function $p(x) = \frac{1}{\pi\sqrt{1-x^2}}$ for $-1 \leq x \leq 1$. Pick $\mu = \mu_1 \otimes \mu_2$. Let $T = 2u + b$ as the division with remainder, where u is integer and $0 \leq b < 2$. Moreover, assume $b \leq 1$ (similar analysis will go through for the case of $b > 1$). The induced distribution ν^{s_1} of source 1 is $\nu^{s_1}(\{0.5\}) = \frac{u+b}{T}$ and $\nu^{s_1}(\{-0.5\}) = \frac{u}{T}$. Thus the total variation distance of μ_1

and ν^{s_1} is at most $1/(2T)$. Similarly, it can be verified that the total variation distance of ν^{s_2} and μ_2 and that of ν and μ also decay as $1/T$. Thus, D_4 is $O(1/T)$, since the monomials $f(s)$ in the definition of D_4 are bounded from above by 1, and $N(\nu^{(s)})$ is also $O(1/T)$. Therefore, by Theorem 2.2.1 $d(\hat{A}, A) = O(\frac{1}{T})$.

Our setting also extends the traditional one to a practically important case, Markovian sources.

Example 2.2.6. Assume that s_i is a stationary and ergodic Markovian source, and the sources are independent of each other for $1 \leq i \leq d$. Our observations are similar to the setting in Example 2.2.4. Because of the Markov property, the observations do not satisfy the i.i.d. assumptions.

In Section 2.4, we will present two algorithms that satisfy Theorem 2.2.1. Our algorithm builds on the works of Frieze et al. [1996], Hsu and Kakade [2013], Arora et al. [2012].

2.3 Estimating Moments: the HKICA Algorithm

In this section we will introduce the ICA algorithm of Hsu and Kakade [2013] and analyze its performance. Hsu and Kakade [2013] claimed that HKICA is easy to analyze using matrix perturbation techniques. While one can indeed use matrix perturbation results, our computations reveal some unexpected and unpleasant complications. This will motivate us to refine the algorithm, leading to our algorithm, deterministic ICA (DICA).

2.3.1 The HKICA algorithm

The ICA algorithm of Hsu and Kakade [2013] is based on the well-known excess-kurtosis-like quantity defined as follows: For any $p \geq 1$, $\eta \in \mathbb{R}^d$, and distribution ν over $\mathbb{R}^d$, let

$$m_p^{(\nu)}(\eta) = \mathbb{E}_{X\sim\nu}[(\eta^\top X)^p] \tag{2.2}$$

and let

$$f_\nu(\eta) = \frac{1}{12}\left(m_4^{(\nu)}(\eta) - 3m_2^{(\nu)}(\eta)^2\right). \tag{2.3}$$

For $\mu \in \Pi_0$, let $A\mu$ stand for $\nu^{(As)}$ where s has the product distribution μ. The Hessian of the function $f_{A\mu}(\eta)$ plays an important role in this chapter because of its following property.

Proposition 2.3.1. *Given $\mu \in \Pi_0$ and any vector $\psi \in \mathbb{R}^d$, $\nabla^2 f_{A\mu}(\psi) = AKD_\psi A^\top$, where $K = diag(\kappa_1, \ldots, \kappa_d)$ and $D_\psi = diag((\psi^\top A_1)^2, \ldots, (\psi^\top A_d)^2)$.*

Proof. Note that

$$\nabla^2 \mathbb{E}_{s\sim\mu}[(\psi^\top As)^4] = 12\,\mathbb{E}_{s\sim\mu}[Ass^\top A^\top(\psi^\top As)^2] = 12A\mathbb{E}_{s\sim\mu}[s(\psi^\top As)^2 s^\top]A^\top.$$

Let $h = A^\top \psi$, consider the position (i,i) of the matrix $M_1 := 12\mathbb{E}_{s\sim\mu}[s(h^\top s)^2 s^\top]$,

$$M_{1,(i,i)} = \mathbb{E}_{s\sim\mu}[s_i^2 (h^\top s)^2] = 12h_i^2 \mathbb{E}_{s\sim\mu}[s_i^4] + 12\mathbb{E}_{s\sim\mu}[s_i^2] \sum_{j\neq i} h_j^2 \mathbb{E}_{s\sim\mu}[s_j^2].$$

Similarly, for $i \neq j$,

$$M_{1,(i,j)} = 24 h_i h_j \mathbb{E}_{s\sim\mu}[s_i^2]\mathbb{E}_{s\sim\mu}[s_j^2].$$

Next consider the second derivative of $(\mathbb{E}_{s\sim\mu}[(\psi^\top As)^2])^2$,

$$\begin{aligned}
&\nabla^2 (\mathbb{E}_{s\sim\mu}[(\psi^\top As)^2])^2 \\
&= 4\mathbb{E}_{s\sim\mu}[(h^\top s)^2]\mathbb{E}_{s\sim\mu}[Ass^\top A^\top] + 8\mathbb{E}_{s\sim\mu}[(h^\top s)As]\mathbb{E}_{s\sim\mu}[(h^\top s)s^\top A^\top] \\
&= A\left(4\mathbb{E}_{s\sim\mu}[(h^\top s)^2]\mathbb{E}_{s\sim\mu}[ss^\top] + 8\mathbb{E}_{s\sim\mu}[(h^\top s)s]\mathbb{E}_{s\sim\mu}[(h^\top s)s^\top]\right) A^\top \\
&:= AM_2 A^\top.
\end{aligned}$$

Thus,

$$M_{2,(i,i)} = 12h_i^2(\mathbb{E}_{s\sim\mu}[s_i^2])^2 + 4\mathbb{E}_{s\sim\mu}[s_i^2] \sum_{j\neq i} h_j^2 \mathbb{E}_{s\sim\mu}[s_j^2]; \quad M_{2,(i,j)} = 8h_i h_j \mathbb{E}_{s\sim\mu}[s_i^2]\mathbb{E}_{s\sim\mu}[s_j^2].$$

Hence, we have

$$M_1 - 3M_2 = 12K\,\mathrm{diag}(h_1^2, \ldots, h_d^2),$$

and thus

$$\nabla^2 f_{A\mu}(\psi) = AKD_\psi A^\top.$$

□

Hsu and Kakade [2013]'s algorithm is built around the above algebraic observation concerning $\nabla^2 f_{\nu^{(x)}}(\eta)$, the second derivative of the function $f_{\nu^{(x)}}$. This observation is the subject of the next result:

Theorem 2.3.1 (Hsu and Kakade [2013], Theorem 4)**.** *Assume A is nonsingular. Let $f_{A\mu} : \mathbb{R}^d \to \mathbb{R}$ be defined by Equation (2.3), and $\phi, \psi \in \mathbb{R}^d$ be vectors from the unit sphere of $\mathbb{R}^d$. Then, the matrix*

$$M = (\nabla^2 f_{A\mu}(\phi))(\nabla^2 f_{A\mu}(\psi))^{-1} \tag{2.4}$$

can be written in the diagonal form

$$M = A \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} A^{-1}, \tag{2.5}$$

where $\lambda_i = \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2$.

Consequently, the eigenvectors[2] of the matrix M are the rescaled columns of A if $\frac{\phi^\top A_i}{\psi^\top A_i}$ are distinct for all i. Thus, to obtain an algorithm, one only needs to estimate $\nabla^2 f_{\nu^{(As)}}$ at two appropriate vectors ϕ, ψ. Since $x = As + \epsilon$ is available, the idea is to use $\nu^{(x)}$ in place of $\nu^{(As)}$. The next result quantifies the error induced on $\nabla^2 f_{\nu^{(As)}}$ as a function of the "noise" ϵ

Note that ϵ could indeed have limited effect in the estimation procedure, as shown in Proposition 2.3.2. Similar result is also discussed by Arora et al. [2012] and Belkin et al. [2013].

Proposition 2.3.2. *For any tuple (x, A, s, ϵ) such that $x = As + \epsilon$ and any vector η,*

$$\begin{aligned}
&\|\nabla^2 f_{\nu^{(x)}}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_F \\
&\leq \mathit{Poly}(L, d, \sigma_{\max}, C)\Big(\mathcal{Q}(\nu^{(\epsilon)}) \\
&\quad + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + Q\left(N(\nu^{(s)})N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right)\Big)\|\eta\|_2^2, \qquad (*)
\end{aligned}$$

where $Q(x) = x + x^2$.

In particular, the difference in the estimation of the Hessian matrix caused by the "noise" is at most $\mathrm{Poly}\,(L, d, \sigma_{\max}, C)\left(\mathcal{Q}(\nu^{(\epsilon)}) + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right)\|\eta\|_2$. Note that in the probabilistic setting, this decays at a rate of $1/\sqrt{T}$.

Denote the quantity on the above RHS of Eq. (*) by $P(\|\eta\|_2)$. As to the computation of $\nabla^2 f_{\nu^{(x)}}$, note that for any ν, $\nabla^2 f_\nu(\eta)$ can be written as

$$\nabla^2 f_\nu(\eta) = G_\nu(\eta) := G_1^{(\nu)}(\eta) - G_2^{(\nu)}(\eta) - 2G_3^{(\nu)}(\eta), \qquad (2.6)$$

where

$$\begin{aligned}
G_1^{(\nu)}(\eta) &= \mathbb{E}_{X\sim\nu}[\left(\eta^\top X\right)^2 XX^\top], \\
G_2^{(\nu)}(\eta) &= \mathbb{E}_{X\sim\nu}[\left(\eta^\top X\right)^2]\mathbb{E}_{X\sim\nu}[XX^\top], \\
G_3^{(\nu)}(\eta) &= \mathbb{E}_{X\sim\nu}[\left(\eta^\top X\right)X]\mathbb{E}_{X\sim\nu}[\left(\eta^\top X\right)X^\top].
\end{aligned}$$

These quantities can be computed directly when $\nu = \nu^{(x)}$ using the observed samples. In what follows, we will use the estimate $\nabla^2 \widehat{f} := \nabla^2 f_{\nu^{(x)}}$, $\nabla^2 f := \nabla^2 f_{A\mu}$ and, in general, we will add a "hat" to quantities which are derived from the empirical distribution $\nu^{(x)}$.

Putting everything together, we obtain the algorithm HKICA, named after Hsu and Kakade [2013], which is shown in Algorithm 1,

[2]Throughout the thesis eigenvectors always mean right eigenvectors, unless specified otherwise.

Algorithm 1 The HKICA algorithm.

Input: $x(t)$ for $1 \le t \le T$.
Output: An estimate of the mixing matrix A.

1: Sample ϕ and ψ independently, uniformly from the unit sphere $\mathbb{R}^d$;
2: Evaluate $\nabla^2 \hat{f}(\phi)$ and $\nabla^2 \hat{f}(\psi)$;
3: Compute $\hat{M} = (\nabla^2 \hat{f}(\phi))(\nabla^2 \hat{f}(\psi))^{-1}$;
4: Compute $\{\mu_1, \dots, \mu_d\}$, all the eigenvectors of $\hat{M}$;
5: Return $\hat{A} = (\mu_1, \dots, \mu_d)$.

Remark 2.3.3. Although in theory, HKICA generates a valid, real output almost surely, in practice this may not happen always as, due to numerical errors, $\nabla^2 \hat{f}(\psi)$ may become singular or the ratio $(\phi^\top A_i)/(\psi^\top A_i)$ may become the same for multiple indices i. A simple way to fix this problem is re-sampling ϕ and ψ until a real eigen-decomposition exists.

2.3.2 Analysis of HKICA

In this section we provide a rigorous analysis of the algorithm.

Definition 2.3.4. Given a vector ψ, a matrix A, and constants $0 \le \ell \le 1$, $L_{\text{high}} \ge \sqrt{2d}$, let $L_{\text{low}} = \frac{\sqrt{\pi}}{\sqrt{2}d}\ell$. We use $\mathcal{E}_\psi^A$ to denote the event when $\min_i |\psi^\top A_i| \ge A_{(2,\min)} L_{\text{low}}$ and $\|\psi\|_2 \le L_{\text{high}}$ hold simultaneously.

Let

$$\gamma_A = \min_{i,j:i\neq j} \left| \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2 - \left(\frac{\phi^\top A_j}{\psi^\top A_j}\right)^2 \right|. \tag{2.7}$$

The performance of the HKICA algorithm will essentially depend on this parameter, as shown in the following theorem.

Theorem 2.3.2. *For any ϕ, ψ and a nonsingular A such that $x = As + \epsilon$, on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$, when HKICA is run on x to get $\hat{A}$, it holds that*

$$\begin{aligned} d(\hat{A}, A) \le \mathcal{C}(\mu) \min \Big(& \frac{1}{\gamma_A} (D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) \\ & + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})), \Theta(\mu) \Big), \end{aligned} \tag{2.8}$$

where $\mathcal{C}(\mu)$ and $\Theta(\mu)$ are problem dependent constants, polynomial in $(\sigma_{\max}(A), \frac{1}{\sigma_{\min}(A)}, \frac{1}{\kappa_{\min}(\mu)}, 1/\delta, d, L, L_{high}, \ell)$.

Remark 2.3.5. (i) $\Theta(\mu)$ in the result gives a trivial bound for the problem: The result becomes interesting only when $D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})$ is small enough. (ii) Note that the bound in (2.8) approaches zero at an $O(1/\sqrt{T})$ rate in the stochastic setting. (iii) Since γ_A is the minimum spacing of the eigenvalues of $M = \nabla^2 f_{A\mu}(\phi)(\nabla^2 f_{A\mu}(\psi))^{-1}$, the eigenvalue perturbations imposed by the noise cannot be too large compared to γ_A without potentially ruining the eigenvectors of M. Thus, the dependence on γ_A seems necessary.

Despite the important role that γ_A plays in the efficiency of the HKICA algorithm, it is not clear whether it is well controllable. To the best of our knowledge, even a polynomial (in the dimension d) lower bound of γ_A is not yet available in the literature. The problem of minimal spacings of random variables has been discussed by Hüsler [1987] and Goyal et al. [2014], but results there are unfortunately not applicable to our case.

2.3.3 Proof of Theorem 2.3.2

We go over the key steps of proving Theorem 2.3.2 in this section. The omitted details can be found in Appendix A.3.

Note by Theorem 2.3.1, $\lambda_i = \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2$. WLOG assume that $\phi^\top A_i$ and $\psi^\top A_i$ are all positive for $i \in [d]$, then $\lambda_s = \lambda_t \Leftrightarrow \phi^\top \left(A_s A_t^\top - A_t A_s^\top\right) \psi = 0$. So given ϕ, ψ are sampled independently from the uniform distribution on the unit sphere of $\mathbb{R}^d$, $\lambda_s \neq \lambda_t$ holds almost surely if A_s is not parallel to A_t (i.e. A is nonsingular). Thus the eigenvalues of M are all distinct and the corresponding eigenvectors determine the columns of A up to permutation and scaling.

Let

$$\xi = 6C^2 D_2(\mu, \nu^{(s)}) + D_4(\mu, \nu^{(s)}). \tag{2.9}$$

We can further prove that $\|M - \hat{M}\|_2$ is also bounded.

Lemma 2.3.6. *Given that* $\xi \leq \frac{\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{low}^2}{4 L_{high}^2 d^6 A_{(2,\max)}^2 A_{\max}^2}$ *and* $P(L_{high}) \leq \frac{1}{4}\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{low}^2$, *on the event* $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$,

$$\|M - \hat{M}\|_2 \leq \Phi(\mu)\Big(D_4(\mu, \nu^{(s)}) + \mathcal{Q}(\nu^{(\epsilon)}) + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\Big),$$

where $\Phi(\mu)$ *is a problem-dependent constant that is polynomial in* $L_{high}, d, \sigma_{\max}, 1/\sigma_{\min}$, $\kappa_{\max}$, $1/\kappa_{\min}, \ell, L$ *and* C.

The following perturbation lemma for diagonalizable matrices shows that a small perturbation of M will only result in a small variation of its eigenvectors, at least under some mild regularity conditions. Thus, given a good estimate of M, we can reconstruct A accurately.

Lemma 2.3.7. *Denote* $\hat{M} = M + E$ *where* $M = PDP^{-1}$ *and where* D *is a diagonal matrix* $diag(\sigma_1, \ldots, \sigma_d)$. *Assume* $\hat{M}$ *has distinct eigenvalues. If* $\gamma_D = \min_{i \neq j} |\sigma_i - \sigma_j| > 4\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2$, *and* $\min_{i,j:i\neq j} \|P_i - P_j\|_2 > \frac{8}{\gamma_D}\frac{\sigma_{\max}^2(P)}{\sigma_{\min}(P)}\|E\|_2$, *then there exist constants* $\{c_1, \ldots, c_d\}$ *and a permutation* π, *such that*

$$\max_{1\leq k\leq d} \|c_k \hat{P}_{\pi(k)} - P_k\|_2 \leq 4\frac{\sigma_{\max}^2(P)}{\gamma_D \sigma_{\min}(P)}\|E\|_2\,,$$

and therefore

$$\sum_{k=1}^{d} \|c_k \hat{P}_{\pi(k)} - P_k\|_2 \leq 4d\frac{\sigma_{\max}^2(P)}{\gamma_D \sigma_{\min}(P)}\|E\|_2\,,$$

where $\hat{P}$ *is the matrix of eigenvectors of* $\hat{M}$.

Proof of Theorem 2.3.2. Let

$$\tilde{Q} = \Phi(\mu)\left(D_4(\mu,\nu^{(s)}) + \mathcal{Q}(\nu^{(\epsilon)}) + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right)$$

be the RHS in Lemma 2.3.6. Assume that the following conditions hold:

C1: $\gamma_A > 4\frac{\sigma_{\max}}{\sigma_{\min}}\tilde{Q}$

C2: $\min_{i,j:i\neq j}\|A_i - A_j\|_2 > \frac{8}{\gamma_A}\frac{\sigma_{\max}^2}{\sigma_{\min}}\tilde{Q}$;

C3: $\xi \leq \frac{\kappa_{\min}A_{(2,\min)}^2\sigma_{\min}^2 L_{\text{low}}^2}{4L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2}$;

C4: $P(L_{\text{high}}) \leq \frac{1}{4}\kappa_{\min}A_{(2,\min)}^2\sigma_{\min}^2 L_{\text{low}}^2$.

Note that if the above conditions are not satisfied, simply picking $c_i = 0$ for all the c_i, then

$$\max_{1\leq k\leq d}\|c_1\widehat{A}_{\pi(k)} - A_k\|_2 \leq \sigma_{\max}.$$

Otherwise recall that $M = A\,\text{diag}(\lambda_1,\ldots,\lambda_d)A^{-1}$ where $\lambda_i = (\phi^\top A_i)^2/(\psi^\top A_i)^2$ (Theorem 2.3.1). Assuming $\mathcal{E} = \mathcal{E}_\phi^A \cap \mathcal{E}_\psi^A$ holds. Given conditions C3 and C4 and applying Lemma 2.3.6, we have $\|M - \hat{M}\|_2 \leq \mathcal{Q}$. Now given conditions C1 and C2, if $\hat{M}$ has distinct eigen-values with probability 1 and by Lemma 2.3.7,

$$\max_{1\leq k\leq d}\|c_1\widehat{A}_{\pi(k)} - A_k\|_2 \leq 4\frac{\sigma_{\max}^2}{\gamma_A\sigma_{\min}}\|M - \hat{M}\|_2 \leq 4\frac{\sigma_{\max}^2}{\gamma_A\sigma_{\min}}\tilde{Q}.$$

Combining both upper bounds leads to the final conclusion. □

2.4 A "Deterministic" ICA Algorithm

In this section, we present our ICA algorithm, named as DICA, together with its analysis. A recursive version of the algorithm is also proposed.

2.4.1 A Refined HKICA Algorithm

The problems with γ_A motivate us to refine the HKICA algorithm. The idea is inspired by the works of Arora et al. [2012] and Frieze et al. [1996]. The idea is to use a 'quasi-whitening' procedure:

For a zero mean product distribution μ, denote the kurtosis of its i-th component by κ_i. To derive our procedure, first recall that $\nabla^2 f_{A\mu}(\psi) = AKD_\psi A^\top$, where $K = \text{diag}(\kappa_1,\ldots,\kappa_d)$ and

$$D_\psi = \text{diag}\left((\psi^\top A_1)^2,\cdots,(\psi^\top A_d)^2\right).$$

Hence, the square root of $\nabla^2 f_{A\mu}(\psi)$ is $B = AK^{1/2}D_\psi^{1/2}R^\top$ for some orthonormal matrix R. Now for $i = 1, 2$, defining $T_i = \nabla^2 f_{A\mu}(B^{-\top}\phi_i)$, then $T_i = AKD_{B^{-\top}\phi_i}A^\top$. Consider the jth element on the diagonal of $D_{B^{-\top}\phi_i}$,

$$(\phi_i^\top B^{-1}A_j)^2 = (\phi_i^\top RD_\psi^{-1/2}K^{-1/2}A^{-1}A_j)^2 = (\phi_i^\top RD_\psi^{-1/2}K^{-1/2}e_j)^2 = (\phi_i^\top R_j)^2\kappa_j^{-1}D_{\psi,(j,j)}^{-1},$$

where e_j is the one-hot vector in $\mathbb{R}^d$ with 1 for its jth position and 0 for other positions and R_j denotes the jth column of R. Thus, $T_i = AKK^{-1}D_\psi^{-1}\Lambda_i A^\top = AD_\psi^{-1}\Lambda_i A^\top$, where

$$\Lambda_i = \begin{pmatrix} (\phi_i^\top R_1)^2 & & \\ & \ddots & \\ & & (\phi_i^\top R_d)^2 \end{pmatrix}.$$

Letting $M = T_1 T_2^{-1}$, we see that $M = A\Lambda A^{-1}$ with

$$\Lambda = \Lambda_1\Lambda_2^{-1} = \begin{pmatrix} \left(\frac{\phi_1^\top R_1}{\phi_2^\top R_1}\right)^2 & & \\ & \ddots & \\ & & \left(\frac{\phi_1^\top R_d}{\phi_2^\top R_d}\right)^2 \end{pmatrix}.$$

Thus, A_i are again the eigenvectors of this newly defined matrix M, but now the eigenvalues of M are defined in terms of the orthogonal matrix R instead of A, and so the resulting minimum spacing

$$\gamma_R = \min_{i,j:i\neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i}\right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j}\right)^2 \right| \tag{2.10}$$

depends on A only through R. Since R is orthonormal, γ_R will be shown to be "well-behaved".

The resulting algorithm, called Deterministic ICA (DICA), is shown as Algorithm 2.

Algorithm 2 Deterministic ICA (DICA)

Input: $x(t)$ for $1 \le t \le T$.
Output: An estimate of the mixing matrix A.

1: Sample ψ from a d-dimensional standard Gaussian distribution;
2: Evaluate $\nabla^2 \hat{f}(\psi)$,
3: Compute the SVD of $\nabla^2 \hat{f}(\psi) = U\Sigma V^\top$, and let $\hat{B} = U\Sigma^{1/2}$.
4: Sample ϕ_1 and ϕ_2 independently from the standard Gaussian distribution;
5: Compute $\hat{T}_1 = \nabla^2 \hat{f}(\hat{B}^{-\top}\phi_1)$ and $\hat{T}_2 = \nabla^2 \hat{f}(\hat{B}^{-\top}\phi_2)$;
6: Compute all the eigenvectors $\{\mu_1, \ldots, \mu_d\}$ of $\hat{M} = \hat{T}_1 \left(\hat{T}_2\right)^{-1}$;
7: Return $\hat{A} = (\mu_1, \ldots, \mu_d)$.

Remark 2.4.1. Similarly to HKICA, in theory DICA fails with probability 0 (giving, e.g., complex outputs), but this may be experienced due to numerical errors. The same resampling trick can be applied again, as in Remark 2.3.3.

Similarly to Theorem 2.3.2, one can show that under some technical assumptions, which hold with probability 1,

$$d(\hat{A}, A) \le \mathcal{C}(\mu) \min\Big(\frac{1}{\gamma_R}(D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})), \Theta(\mu)\Big),$$

where $\hat{A}$ is the output of the DICA algorithm, which is the result stated as Theorem 2.2.1.

The proof is very similar to the result of Theorem 2.3.2, with γ_R in place of γ_A, as required. Note that the $\mathcal{C}(\mu)$ here is different from that in Theorem 2.3.2. We will prove that $\frac{1}{\gamma_R}$ is polynomial in d and $\frac{1}{\delta}$ in next section, thus prove Theorem 2.2.1. The detailed proof is postponed to Appendix A.4.

2.4.2 Analysis of γ_R

We first empirically compare the behavior of $1/\gamma_A$ and $1/\gamma_R$. In particular, we construct four kinds of matrices with increasing coherences: $A_1 = P$; $A_2 = v_b\mathbf{1}^\top + 0.3P$; $A_3 = v_b\mathbf{1}^\top + 0.05P$; and $A_4 = v_b\mathbf{1}^\top + 0.005P$. Here, the elements of the vector v_b and the matrix P are both generated from the standard Gaussian distribution (with appropriate dimensions). We also generate an orthonormal mixing matrix R. This matrix is obtained by computing the left-column space of a non-singular random matrix whose components are also drawn from standard normal distribution. For each of the matrices, we generate ϕ and ψ from standard normal distribution 3 times, pick the minimal values of $1/\gamma$, and plot the average value over 200 repetitions. Figure 2.2 below shows the behaviour of $1/\gamma_A$ and $1/\gamma_R$ for mixing matrices, and for some random orthonormal matrix R. As expected, the value of $1/\gamma_A$ increases with the coherence of the matrix. However, it is similar to that of an orthonormal matrix $1/\gamma_R$ unless the coherence is really large. Uncovering the dependence of $1/\gamma_A$ on the properties of A remains an interesting (and challenging) open problem.

Recall that ϕ_1 and ϕ_2 are independently sampled from the standard Gaussian distribution. Thus, $\{\phi_1^\top R_1, \cdots, \phi_1^\top R_d, \phi_2^\top R_1, \cdots, \phi_2^\top R_d\}$ are $2d$ independent standard Gaussian random variables. Let $Z_i = \frac{\phi_1^\top R_i}{\phi_2^\top R_i}$. It follows that Z_i, $1 \leq i \leq d$ are d independent Cauchy$(0, 1)$ random variables. Using this observation, we show in the following lemma that γ_R is large with probability at least $1 - \delta$.

Proposition 2.4.2. *With probability at least* $1 - \delta$,

$$\gamma_R \geq \frac{\pi^2\delta^2}{d^3}.$$

The proof is postponed to Section A.6. For further reference we will denote the event mentioned in the previous proposition by $\mathcal{E}_Z$.

2.4.3 A Modified Version of DICA

In this section, we provide a heuristic modification of DICA (MDICA) that performs better in the experiments. However, proving performance guarantees for this new algorithm has so far defied our efforts. As the cases of the previous analyses, the performance of MDICA depends on the minimal eigenvalue spacings of a specially constructed matrix M. Similar to γ_A, we are not able to bound it polynomially in d.

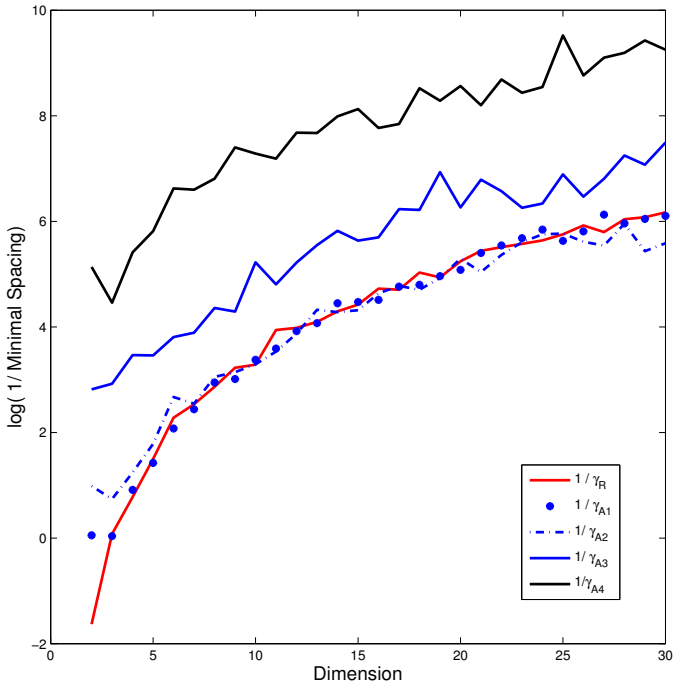

Figure 2.2: The values of $1/\gamma$ for matrices with different coherences

The observation underlying this new algorithm is that errors of the Hessian matrix estimates computed introduce a large estimation error. Thus we want to reduce the number of Hessians used in the procedures, while keeping the minimum spacing under control.

Algorithm 3 DICA Modified (MDICA)

Input: $x(t)$ for $1 \leq t \leq T$.
Output: An estimation of the mixing matrix A.

1: Sample ψ from a d-dimensional standard Gaussian distribution;
2: Evaluate $\nabla^2 \hat{f}(\psi)$,
3: Compute the SVD of $\nabla^2 \hat{f}(\psi) = U\Sigma V^\top$, and let $\hat{B} = U\Sigma^{1/2}$.
4: Sample ϕ from the standard Gaussian distribution;
5: Compute $\hat{T} = \nabla^2 \hat{f}(\hat{B}^{-\top}\phi)$;
6: Compute all the eigenvectors $\{\mu_1, \ldots, \mu_d\}$ of $\hat{M} = \hat{B}^{-1}\hat{T}\hat{B}^{-\top}$. Set $\hat{R} = (\mu_1, \ldots, \mu_d)$;
7: Return $\hat{A} = \hat{B}\hat{R}$.

Remark 2.4.3. Algorithm 3 follows the derivation: Recall that $B = AK^{1/2}D_\psi^{1/2}R^\top$. Let $T = \nabla^2 f_{A\mu}(B^{-\top}\phi)$, then $T = AD_\psi^{-1}\Lambda A^\top$, where

$$\Lambda = \begin{pmatrix} (\phi^\top R_1)^2 & & \\ & \ddots & \\ & & (\phi^\top R_d)^2 \end{pmatrix},$$

and R_j denotes the jth column of R. Thus, $M = B^{-1}TB^{-\top} = RK^{-1}D_\psi^{-2}\Lambda R^\top$, and

its eigen-decomposition recovers the orthonormal matrix R. Finally, $BR = AK^{1/2}D_\psi^{1/2}$ recovers A.

Remark 2.4.4. Note that the eigenvalues of M are $\frac{(\phi^\top R_i)^2}{(\psi^\top A_i)^4\kappa_i}$ for $1 \le i \le d$. When A is highly coherent, we would expect that $\psi^\top A_i$'s are close to each other. Given that $\phi^\top R_i$'s are well separated from each other, we intuitively expect the eigenvalues to be well-separated from each other. However, we do not have a rigorous proof of this. Experimental results show that MDICA consistently outperforms DICA.

The following proposition shows that the minimal spacing of M in MDICA is large when A is highly coherent. Instead of assuming the source signals s are bounded, we assume $\|A_i\|_2 = 1$ for any $i \in [d]$ in this section.

Proposition 2.4.5. *Fix $0 < \delta < 1$ and let $c = \frac{2\sqrt{\pi}\delta}{d(d-1)}$. Under the event of $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^R$, assume that*

- *All κ_i's are equal for $1 \le i \le d$, denoted by κ;*
- *$\langle A_i, A_j\rangle \ge 1 - \epsilon$ for any $1 \le i, j \le d$, such that $\epsilon \le \frac{A_{(2,\min)}^8 L_{low}^8 c^2}{32 L_{high}^{10}}$;*
- *$\|A_i\|_2 = 1$ holds for all $i \in [d]$*

Then with probability at least $1 - \delta$, the minimal spacing of M in the MDICA algorithm is at least $\frac{2\sqrt{\pi}\delta L_{low}}{\kappa L_{high}^4 d(d-1)}$.

Remark 2.4.6. Theoretical guarantee for Algorithm 3 can be developed in a similar way based on Proposition 2.4.5. However, Proposition 2.4.5 is relatively week, as its first two assumptions are not natural.

Proof of Proposition 2.4.5. Note that for any i, $\|A_i\|_2 = 1$. Thus, given $\langle A_i, A_j\rangle \ge 1 - \epsilon$,

$$\|A_i - A_j\|_2^2 \le 2 - 2\langle A_i, A_j\rangle \le 2\epsilon.$$

Also,

$$|(\psi^\top A_i)^2 - (\psi^\top A_j)^2| = |(\psi^\top A_i) - (\psi^\top A_j)| \cdot |(\psi^\top A_i) + (\psi^\top A_j)| \le 2\sqrt{2}\|\psi\|_2^2\sqrt{\epsilon} = 2\sqrt{2}L_u^2\sqrt{\epsilon}.$$

Now WLOG assume $\phi^\top R_i$ and $\phi^\top R_j$ are both nonnegative. Then the minimal eigenvalue spacing

$$\begin{aligned}\left|\frac{(\phi^\top R_i)^2}{(\psi^\top A_i)^4\kappa_i} - \frac{(\phi^\top R_j)^2}{(\psi^\top A_j)^4\kappa_j}\right| &\ge \frac{2}{\kappa}\min_i\left|\frac{\phi^\top R_i}{(\psi^\top A_i)^2}\right| \min_{i\ne j}\left|\frac{\phi^\top R_i}{(\psi^\top A_i)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2}\right| \\ &\ge \frac{2L_{\text{low}}}{\kappa L_{\text{high}}^2}\min_{i\ne j}\left|\frac{\phi^\top R_i}{(\psi^\top A_i)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2}\right|.\end{aligned}$$

It remains to bound

$$\min_{i\ne j}\left|\frac{\phi^\top R_i}{(\psi^\top A_i)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2}\right|.$$

Note that given ϕ uniformly sampled from the standard Gaussian distribution, $\phi^\top R_i - \phi^\top R_j$ is a random variable from a Gaussian distribution with mean 0 and variance 2 for $1 \le i \ne j \le d$. Thus, with probability at least $1-\delta$,

$$\min_{i \ne j} |\phi^\top R_i - \phi^\top R_j| \ge \frac{2\sqrt{\pi}\delta}{d(d-1)}.$$

Thus, for any $1 \le i \ne j \le d$,

$$\begin{aligned}
\left| \frac{\phi^\top R_i}{(\psi^\top A_i)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2} \right| &= \left| \left(\frac{\phi^\top R_i}{(\psi^\top A_j)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2} \right) - \left(\frac{\phi^\top R_i}{(\psi^\top A_j)^2} - \frac{\phi^\top R_i}{(\psi^\top A_i)^2} \right) \right| \\
&\ge \left| \frac{\phi^\top R_i}{(\psi^\top A_j)^2} - \frac{\phi^\top R_j}{(\psi^\top A_j)^2} \right| - \left| \frac{\phi^\top R_i}{(\psi^\top A_j)^2} - \frac{\phi^\top R_i}{(\psi^\top A_i)^2} \right| \\
&\ge \frac{|\phi^\top R_i - \phi^\top R_j|}{L_{\text{high}}^2} - L_{\text{high}} \frac{|(\psi^\top A_i)^2 - (\psi^\top A_j)^2|}{(\psi^\top A_i)^2 (\psi^\top A_i)^2} \\
&\ge \frac{c}{L_{\text{high}}^2} - \frac{2\sqrt{2} L_u^3 \sqrt{\epsilon}}{A_{(2,\min)}^4 L_{\text{low}}^4} \ge \frac{c}{2L_{\text{high}}^2}.
\end{aligned}$$

□

2.4.4 Recursive Versions

Vempala and Xiao [2014] proposed a recursion idea to improve the sample complexity of the Fourier PCA algorithm of Goyal et al. [2014]: Instead of recovering all the columns of A in a single step, their recursive algorithm decomposes the whole space into two subspaces according to the maximal spacing of the eigenvalues, then continues recursively to decompose each of the subspaces obtained until they are all 1-dimensional. The idea underlying this recursive procedure is so that the maximal spacing of the eigenvalues are much larger than the minimal one, so the algorithm may win over a single decomposition even if errors compound through the recursion. As this algorithm assumes that the mixing matrix is orthonormal (so that the projection to its subspaces can always eliminate some component of the source signal), we will need to adapt it to our setting. We will only show this adaptation for HKICA as an example. Our other algorithms can also be modified into a recursive version in a similar way.

To force an orthonormal mixing matrix, we will first compute the square root matrix B from $\nabla^2 f(\psi) = AD_\psi K A^\top$ in the same way as done in DICA. Thus, $B = AD_\psi^{1/2} K^{1/2} R^\top$ for some orthonormal matrix R. Transforming our observations by B^{-1}, we then have the new observation $y(t) = B^{-1}x(t) + B^{-1}\epsilon(t) = RD_\psi^{1/2} K^{1/2} s(t) + B^{-1}\epsilon(t)$. Note that $D_\psi^{1/2} K^{1/2}$ is diagonal, thus $RD_\psi^{1/2} K^{1/2} s(t)$ is an orthonormal mixture of the 'independent' sources $D_\psi^{1/2} K^{1/2} s(t)$ for $t \in [T]$. We then apply the recursive algorithm of Vempala and Xiao [2014] to recover the mixing matrix R. Finally, BR gives an estimation of A up to scaling its columns.

The details of recovering are as follows: We follow the idea of HKICA (and DICA) to compute two Hessian matrices $T_1 = RD_\psi^{-1}\Lambda_1 R^\top$ and $T_2 = RD_\psi^{-1}\Lambda_2 R^\top$. Then, instead of

computing the full eigen-decomposition of $T_0 = T_1 T_2^{-1}$ (as in HKICA), we only decompose its eigenspace into two subspaces, according to the maximal spacing of the eigenvalues of T_0. The *Decompose* helper function takes a projection matrix P of a subspace spanned by some columns of R (WLOG we assume these are the first k columns of R). Then we compute the projection of T_0 as $M = P^\top T_0 P$. Thus the eigenspace of $PMP^\top$ is in the span of P. Lastly, by separating the eigenvectors of M according to its eigenvalues into PP_1 and PP_2, the *Decompose* function recursively decomposes the subspaces into two smaller subspaces.

Algorithm 4 Recursive version of HKICA (HKICA.R)

Input: $x(t)$ for $1 \le t \le T$.
Output: An estimation of the mixing matrix A.

1: Sample ψ from a d-dimensional standard Gaussian distribution;
2: Evaluate $\nabla^2 \hat{f}(\psi) = \hat{G}$;
3: Compute $\hat{B}$ such that $\hat{B}\hat{B}^\top = \hat{G}$;
4: Compute $\hat{y}(t) = \hat{B}^{-1} x(t)$ for $1 \le t \le T$;
5: Let $P = I_d$;
6: Compute $\hat{R} = \text{Decompose}(\hat{y}, P)$;
7: Return $\hat{B}\hat{R}$;

Algorithm 5 The Decompose function

Input: $x(t)$ for $1 \le t \le T$, a projection matrix $P \in \mathbb{R}^{d \times k}$ $(d \ge k)$.
Output: An estimation of the mixing matrix $A \in \mathbb{R}^{d \times k}$.

1: if $k = 1$, return P;
2: Sample ϕ_1 and ϕ_2 independently from a standard Gaussian distribution of dimension d;
3: Evaluate $\nabla^2 \hat{f}(\phi_1)$ and $\nabla^2 \hat{f}(\phi_2)$,
4: Compute $\hat{M} = (\nabla^2 \hat{f}(\phi_1))(\nabla^2 \hat{f}(\phi_2))^{-1}$;
5: Compute $\hat{M}_P = P^\top \hat{T} P$;
6: Compute the eigen-decomposition of $\hat{M}_P$, its eigenvalues $\{\sigma_1, \ldots, \sigma_d\}$ where $\sigma_1 \ge \ldots \ge \sigma_k$ and their corresponding eigenvectors $\{\mu_1, \ldots, \mu_k\}$;
7: Find the index $m = \arg\max_i \sigma_i - \sigma_{i+1}$;
8: Let $P_1 = (\mu_1, \ldots, \mu_m)$, and $P_2 = (\mu_{m+1}, \ldots, \mu_k)$;
9: Compute $W_1 = \text{Decompose}(x, PP_1)$, and $W_2 = \text{Decompose}(x, PP_2)$;
10: Return $[W_1, W_2]$;

Theorem 2.4.1. *With probability at least $1 - \delta$, the recursive version of HKICA returns the mixing matrix with an error bound*

$$d(\hat{A}, A) \le \inf_{\mu \in \Pi_0} \mathcal{C}(\mu) \min\left(K^2(\mu) + K(\mu), \Theta(\mu)\right),$$

where $K(\mu) = D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})$.

Remark 2.4.7. Note that when $K(\mu)$ is small enough, the term $K^2(\mu)$ will be dominated by the error carried over from the quasi-whitening procedure, $K(\mu)$. Compared to the result in Section 2.4.1, the recursive algorithm improves the dependence of $\mathcal{C}(\mu)$ on the dimension d via its eigen-decomposition according to the maximal spacing of the eigen-values.

2.5 Experimental Results

In this section we compare the performance of different ICA algorithms on some synthetic examples, with mixing matrices of different coherences. We test 9 algorithms: HKICA (HKICA), and its recursive version (HKICA.R); DICA (DICA), and its recursive version (DICA.R); the modified version of DICA (MDICA), and its recursive version (MDICA.R); the default FastICA algorithm from the 'ITE' toolbox [Szabó, 2014] (FICA); the recursive Fourier PCA algorithm of Xiao [2014] (FPCA); and random guessing (Random). FPCA is modified so that it can be applied to the case of non-orthogonal mixing matrix. Random guessing is included to provide a scale.

In the simulation, a common mixing matrix A of dimension 6 is generated in the same way as in Section 2.4.2, where A_1 has the lowest coherence and A_4 has the highest coherence. Next, we generate a 6-dimensional "BPSK" signal s as follows. Let $p = (\sqrt{2}, \sqrt{5}, \sqrt{7}, \sqrt{11}, \sqrt{13}, \sqrt{19})$. We generate a $\{+1, -1\}$ valued sequence $q(t)$ uniformly at random for $1 \leq t \leq T$, and set $s_i(t) = q(t)\sin(p_i t)$. Note that in order to have the components of s close to independent, we need the ratios p_i/p_j for all $i \neq j$ to be irrational.

Lastly, the observed signal is generated as $x = As + c\epsilon$ where ϵ is the noise generated from a d-dimensional normal distribution with randomly generated covariance. We take $T = 20000$ instances of the observed signal on time steps $t = 1, \ldots, 20000$. We test the noise ratio c from 0 (noise-free) to 1 (heavily noisy). All the algorithms are evaluated on 150 repetitions. Since the algorithms are randomized, for each repetition we try 3 times and report the best results. A randomly selected example is shown in Figure 2.3.

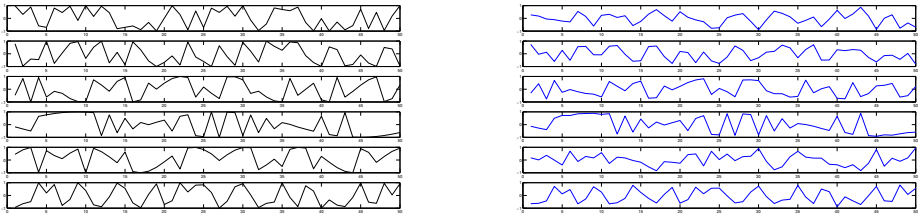

Figure 2.3: An example of the behaviour of the DICA algorithm. Left: the original BPSK signals; Right: the reconstructed signals by DICA with noise ratio $c = 0$ (noise-free case).

We measure the performances of the algorithms by their ability to reconstruct A. In particular, we use $d(\hat{A}, A)$ with the "true" mixing matrix A. The calculation of this measure requires an exhaustive search over all possible permutation.

2.5.1 Results

We investigate the following 4 questions:

(i) How is the performance of moment methods compared to FastICA?

(ii) How does noise affect the performance of the various ICA algorithms?

(iii) How does the coherence of the mixing matrix affect the performance of the various ICA algorithms?

(iv) Do the recursive versions improve performance?

As shown in Figure 2.4, for the low coherence matrices FastICA achieves the best performance. Its performance is specially outstanding in the noise-free case, where it achieves close to 0 reconstruction error. However, its performance degrades quickly as the magnitude of noise, or the coherence of the mixing matrix A increases. On the other hand, the ICA algorithms based on moment methods appear to be more robust to the noise and the coherence of the mixing matrix. One interesting observation is that for the mixing matrix A_4 (high coherent), MDICA seems to be more robust.

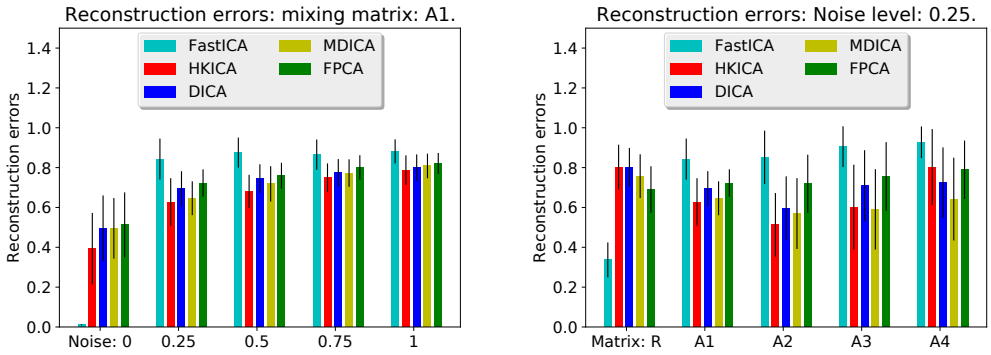

Figure 2.4: Reconstruction errors. Error bars are based on 150 repetitions. Left: Reconstruction errors for different levels of noise; Right: Noisy observations. For comparison, the reconstruction error of "random guessing" for A_1 is 0.9 ± 0.033.)

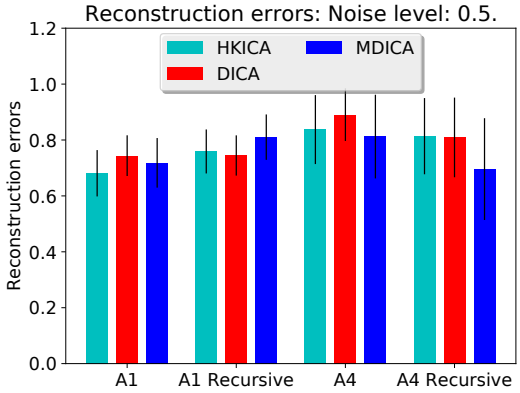

Figure 2.5: Reconstruction errors. Error bars are based on 150 repetitions.

Figure 2.5 shows that the recursive versions of the algorithms tested are not always better than their respective non-recursive versions. In particular, when A has relatively

low coherence, the minimum eigenvalue gap is not too small. For a highly coherent A, the recursive versions outperform their non-recursive counterparts. Note that in this case, A is close to singular (small minimal singular value), and thus it requires more samples.

We expected that DICA will achieve smaller error for an extremely coherent A, since $1/\gamma_A$ will be much larger than $1/\gamma_R$. However, the experimental results indicate the opposite. Note that high coherence implies small minimal singular value. In this case, the estimation error of M in DICA could be much larger than that in HKICA, because of the fourth degree of A^{-1}. This error overwhelms the improvement brought by larger eigenvalue spacings, if the sample size is not large enough. On the other hand, MDICA tries to achieve a small estimation error, meanwhile we expect it to keep the eigenvalue spacing large (intuitively, this eigenvalue spacing is approximately the spacing of the square of d Gaussian random variables), leading to good performance. This is confirmed by the experimental results, in both the non-recursive and recursive versions.

In summary, the results suggest that the moment methods are comparable to each other in practice, while FastICA is better for mixing matrices with low coherence or mild coherence with low noise. If the observations have a large amount of noise and the mixing matrix is not extremely coherent, then HKICA may be the best choice. In the case of an extremely coherent mixing matrix, MDICA performs the best.

2.6 Conclusions

Motivated by the observation that ICA algorithms achieve good performance on separating the mixture of periodic sources, in this chapter we extend and analyze the problem of ICA in a deterministic framework without any probabilistic assumptions. Our analysis leads to provably polynomial-time ICA algorithms that have no unspecified parameters. Our results are instance-dependent, and catches the important features of the data on which the performance of our ICA algorithms relay. These results recover the usual statistical results in the classic ICA setting, and also extend to other sources, e.g. periodic sources.

Appendix A

Omitted Proofs for Chapter 2

This Chapter is devoted to the omitted proofs for Chapter 2.

A.1 Proof of Proposition 2.2.3

Proposition 2.2.3. *Let $(s(t))_{t\in[T]}$ be an zero-mean i.i.d. sequence bounded by C in ℓ_∞ norm, independent of the i.i.d. Gaussian noise $\mathcal{N}(0,\Sigma)$ sequence $(\epsilon(t))_{t\in[T]}$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, $L=\mathrm{Poly}\left(\|\Sigma\|_2,d,\frac{1}{\delta}\right)$. Moreover, if μ is the product measure of the sources (i.e., $s(t)\sim\mu$), then $D_4(\nu^{(s)},\mu)$, $\mathcal{Q}(\nu^{(\epsilon)})$, $D_4^{(d,d)}(\nu^{(As,\epsilon)})$, $N(\nu^{(\epsilon)})$, $N(\nu^{(s)})$ are all of orders $O(1/\sqrt{T})$.*

Proof. Denote the population expectation by $\mathbb{E}$ and the empirical expectation by $\mathbb{E}_t$, i.e. $\mathbb{E}_t[\epsilon]=\mathbb{E}_{X\sim\nu_t^{(\epsilon)}}[X]$. We actually prove a stronger result: there exists a function $g:\mathbb{N}\to\mathbb{R}$ satisfying $g(t)\to 0$ at a rate of $\frac{1}{\sqrt{t}}$ as $t\to\infty$, such that

Claim (i): $D_4(\mu,\nu^{(s)})\le g(T)$; $\|\mathbb{E}_T[s]\|_F\le g(T)$;

Claim (ii): $\|\mathbb{E}_T[\epsilon]\|_F\le g(T)$; $\|\mathbb{E}_T[\epsilon^{\otimes 2}]\|_F=\mathrm{Poly}\left(\|\Sigma\|_2,d,\frac{1}{\delta}\right)$;
$\|\mathbb{E}_T[\epsilon^{\otimes 3}]\|_F=\mathrm{Poly}\left(\|\Sigma\|_2,d,\frac{1}{\delta}\right)/\sqrt{T}$;

Claim (iii): $\|\left(\mathbb{E}_T[\epsilon^{\otimes 4}]-(\mathbb{E}_T[\epsilon^{\otimes 2}])^{\otimes 2}\right)(\eta,\eta,\cdot,\cdot)-2(\mathbb{E}_T[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\cdot,\eta,\cdot)\|_F$
$\le g(T)\|\eta\|_2^2$, for any $\eta\in\mathbb{R}^d$;

Claim (iv): for $i_1,i_2,j_1,j_2\ge 0$ such that $i_1+i_2+j_1+j_2\le 4$, denote $(AS)^{\otimes i_1}\otimes\epsilon^{\otimes j_1}\otimes(AS)^{\otimes i_2}\otimes\epsilon^{\otimes j_2}$ by $\mathcal{T}$, then

$$\|\,\mathbb{E}_{S\sim\nu^{(s)}}[\mathbb{E}_{\epsilon\sim\nu^{(\epsilon)}}[\mathcal{T}]]-\mathbb{E}_{(S,\epsilon)\sim\nu^{(s,\epsilon)}}[\mathcal{T}]\,\|_F\le g(T).$$

Note that claim (i) bounds $D_4(\mu,\nu^{(s)})$ and $N(\nu^{(s)})$, claim (ii) bounds L and $N(\nu^{(\epsilon)})$, claim (iii) bounds $\mathcal{Q}(\nu^{(\epsilon)})$, and claim (iv) bounds $D_4^{(d,d)}(\nu^{(As,\epsilon)})$.

Proof of Claim (i): Recall that the signal s is zero-mean and bounded by C, thus by

Hoeffding's inequality, with probability at least $1-\delta$,

$$\|\mathbb{E}_T[s]\|_F \le C\sqrt{\frac{d\log(d/\delta)}{2T}}. \tag{A.1}$$

Moreover, any monomial with degree ≤ 4 will be bounded by C^4. Similarly, with probability at least $1-\delta$,

$$D_4(\mu,\nu^{(s)}) \le C^4\sqrt{\frac{\log(d^4/\delta)}{2T}}. \tag{A.2}$$

Proof of Claim (ii): Claim (ii) is about the moments of the Gaussian noise. For i.i.d. standard Gaussian random variables, $X_1,\dots,X_t$, note that

- $\mathbb{E}[\sum_j X_j/t]=0$, $\mathrm{Var}(\sum_j X_j/t)=1/t$; thus $\mathbb{P}\left(|\sum_j X_j/t| \le \sqrt{1/(t\delta)}\right) \ge 1-\delta$;
- $\mathbb{E}[\sum_j X_j^2/t]=1$, $\mathrm{Var}(\sum_j X_j^2/t)=2/t$; thus $\mathbb{P}\left(|\sum_j X_j^2/t-1| \le \sqrt{2/(t\delta)}\right) \ge 1-\delta$;
- $\mathbb{E}[\sum_j X_j^3/t]=0$, $\mathrm{Var}(\sum_j X_j^3/t)=15/t$; thus $\mathbb{P}\left(|\sum_j X_j^3/t| \le \sqrt{15/(t\delta)}\right) \ge 1-\delta$;
- $\mathbb{E}[\sum_j X_j^4/t]=3$, $\mathrm{Var}(\sum_j X_j^4/t)=96/t$. thus $\mathbb{P}\left(|\sum_j X_j^4/t-3| \le \sqrt{96/(t\delta)}\right) \ge 1-\delta$;

Here the probability inequalities are due to Chebyshev's inequality.

Given $\epsilon \sim \mathcal{N}(0,\Sigma)$ for some fixed unknown Σ, firstly consider the case when $\Sigma = I$. For the 1-dimensional tensor (vector) ϵ, it is straightforward that with probability at least $1-\delta$,

$$\|\mathbb{E}_t[\epsilon]\|_F \le d\sqrt{1/(t\delta)}. \tag{A.3}$$

Moreover, consider the position (u,v) of the 2-dimensional tensor (matrix) $\epsilon^{\otimes 2}$. If $u=v$ then with probability at least $1-\delta$, $\left|\sum_j \epsilon_u^2(j)/t-1\right| \le \sqrt{2/(t\delta)}$. If $u\neq v$, by Chebysev's inequality with probability at least $1-\delta$, $\left|\sum_j \epsilon_u(j)\epsilon_v(j)/t\right| \le \sqrt{1/(t\delta)}$. Therefore, with probability at least $1-\delta$, all entries are less than $1+\sqrt{2d^2/t\delta}$. Thus

$$\|\mathbb{E}_t[\epsilon^{\otimes 2}]\|_F \le d(1+d\sqrt{2/(t\delta)}). \tag{A.4}$$

Lastly for the 3-dimenstional tensor $\epsilon^{\otimes 3}$, consider the (u,v,w) position where u,v, and w are distinct. The expectation of $\epsilon_u\epsilon_v\epsilon_w$ is 0 and its variance is at most 15. Therefore, with probability at least $1-\delta$, $\left|\sum_j \epsilon_u(j)\epsilon_v(j)\epsilon_w(j)/t\right| \le \sqrt{15/(t\delta)}$. Thus, with probability at least $1-\delta$,

$$\|\mathbb{E}_t[\epsilon^{\otimes 3}]\|_F \le d^3\sqrt{15/(t\delta)}. \tag{A.5}$$

Similar result can be obtained following the same calculations for the cases when u,v and w are not distinct.

Proof of Claim (iii): Consider the (u, v) position of the matrix,

$$\begin{aligned}
&\left|\left(\mathbb{E}_t[\epsilon^{\otimes 4}](\eta,\eta,\cdot,\cdot) - (\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\eta,\cdot,\cdot) - 2(\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\cdot,\eta,\cdot)\right)_{u,v}\right| \\
&\leq \left|\sum_{k_1,k_2} \eta_{k_1}\eta_{k_2}\left(\mathbb{E}_t\left[\epsilon_u\epsilon_v\epsilon_{k_1}\epsilon_{k_2}\right] - \mathbb{E}\left[\epsilon_u\epsilon_v\epsilon_{k_1}\epsilon_{k_2}\right]\right)\right| + \left|\mathbb{E}_t\left[\epsilon_u\epsilon_v\right] - \mathbb{E}\left[\epsilon_u\epsilon_v\right]\right|\left|\sum_{k_1,k_2}\eta_{k_1}\eta_{k_2}\mathbb{E}_t\left[\epsilon_{k_1}\epsilon_{k_2}\right]\right| \\
&\quad + \left|\mathbb{E}\left[\epsilon_u\epsilon_v\right]\right|\left|\sum_{k_1,k_2}\eta_{k_1}\eta_{k_2}\left(\mathbb{E}_t\left[\epsilon_{k_1}\epsilon_{k_2}\right] - \mathbb{E}\left[\epsilon_{k_1}\epsilon_{k_2}\right]\right)\right| \\
&\quad + 2\left|\sum_{k_1,k_2}\eta_{k_1}\eta_{k_2}\left(\mathbb{E}_t\left[\epsilon_u\epsilon_{k_1}\right]\mathbb{E}_t\left[\epsilon_v\epsilon_{k_2}\right] - \mathbb{E}\left[\epsilon_u\epsilon_{k_1}\right]\mathbb{E}\left[\epsilon_v\epsilon_{k_2}\right]\right)\right|. \qquad \text{(A.6)}
\end{aligned}$$

Note that the kurtosis of a Gaussian variable is 0, thus

$$\sum_{k_1,k_2}\eta_{k_1}\eta_{k_2}\mathbb{E}\left[\epsilon_u\epsilon_v\epsilon_{k_1}\epsilon_{k_2}\right] = \sum_{k_1,k_2}\eta_{k_1}\eta_{k_2}\left(\mathbb{E}\left[\epsilon_u\epsilon_v\right]\mathbb{E}\left[\epsilon_{k_1}\epsilon_{k_2}\right] + 2\mathbb{E}\left[\epsilon_u\epsilon_{k_1}\right]\mathbb{E}\left[\epsilon_v\epsilon_{k_2}\right]\right).$$

Note that the RHS of Eq. (A.6) including $5d^2$ deviation terms of moments of Gaussian variables, i.e. differences between the empirical means and their true means. Also, each of the following inequalities holds with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_j \epsilon_u^4(j)/t - 3\right| \leq \sqrt{96/(t\delta)}; \quad \left|\sum_j \epsilon_u^3(j)\epsilon_v(j)/t\right| \leq \sqrt{15/(t\delta)}; \\
&\left|\sum_j \epsilon_u^2(j)\epsilon_v^2(j)/t - 1\right| \leq \sqrt{4/(t\delta)}; \quad \left|\sum_j \epsilon_u^2(j)\epsilon_v(j)\epsilon_w(j)/t\right| \leq \sqrt{2/(t\delta)}; \\
&\left|\sum_j \epsilon_u(j)\epsilon_v(j)\epsilon_w(j)\epsilon_z(j)/t\right| \leq \sqrt{1/(t\delta)};
\end{aligned}$$

Thus, with probability at least $1 - 5d^2\delta$,

$$\begin{aligned}
&\left|\left(\mathbb{E}_t[\epsilon^{\otimes 4}](\eta,\eta,\cdot,\cdot) - (\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\eta,\cdot,\cdot) - 2(\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\cdot,\eta,\cdot)\right)_{u,v}\right| \\
&\quad \leq O(1)d\sqrt{\frac{1}{t\delta}}\|\eta\|_2^2,
\end{aligned}$$

where we used that $\sum_{i,j}\eta_i\eta_j \leq d\|\eta\|_2^2$. Thus,

$$\begin{aligned}
&\|\left(\mathbb{E}_t[\epsilon^{\otimes 4}] - (\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}\right)(\eta,\eta,\cdot,\cdot) - 2(\mathbb{E}_t[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\cdot,\eta,\cdot)\|_F \\
&\quad \leq O(1)d^2\sqrt{\frac{1}{t\delta}}\|\eta\|_2^2. \qquad \text{(A.7)}
\end{aligned}$$

Proof of Claim (iv): For the last claim, by the triangle inequality

$$\begin{aligned}
&\| \mathbb{E}_{S\sim\nu^{(s)}}[\mathbb{E}_{\epsilon\sim\nu^{(\epsilon)}}[\mathcal{T}]] - \mathbb{E}_{(S,\epsilon)\sim\nu^{(s,\epsilon)}}[\mathcal{T}] \|_F \\
\le &\| \mathbb{E}_{S\sim\nu^{(s)}}[\mathbb{E}_{\epsilon\sim\nu^{(\epsilon)}}[\mathcal{T}]] - \mathbb{E}_{S\sim\nu^{(s)}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\mathcal{T}]\right] \|_F \\
&+ \| \mathbb{E}_{S\sim\nu^{(s)}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\mathcal{T}]\right] - \mathbb{E}_{S\sim\mu}\left[\mathbb{E}_{\epsilon\sim\nu^{(\epsilon)}}[\mathcal{T}]\right] \|_F \\
&+ \underbrace{\|\mathbb{E}_{S\sim\mu}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\mathcal{T}]\right] - \mathbb{E}[\mathcal{T}]\|_F}_{=0} + \| \mathbb{E}[\mathcal{T}] - \mathbb{E}_T[\mathcal{T}] \|_F,
\end{aligned}$$

where the cancellation happens because s and ϵ are independent. Note that every term in the RHS is an $i_1 + j_1 + i_2 + j_2$-dimensional tensors. We can bound these differences elementwise as before to get that with probability at least $1 - \delta$,

$$\| \mathbb{E}_{S\sim\nu^{(s)}}[\mathbb{E}_{\epsilon\sim\nu^{(\epsilon)}}[\mathcal{T}]] - \mathbb{E}_T[\mathcal{T}] \|_F \le O(1) A_{\max}^4 C^4 d^4 \sqrt{\frac{1}{t\delta}} \tag{A.8}$$

Combining Eqs. (A.1) to (A.5), (A.7) and (A.8), picking

$$g(t) = O(1) \max\left(C\sqrt{\frac{d\log(d/\delta)}{t}}, C^4\sqrt{\frac{\log(d^4/\delta)}{t}}, d^3\sqrt{\frac{1}{t\delta}}, A_{\max}^4 C^4 d^4 \sqrt{\frac{1}{t\delta}} \right),$$

leads to the claims.

Lastly, for the general case when $\epsilon \sim \mathcal{N}(0, \Sigma)$, the above conclusions will apply when used with $\Sigma^{-1/2}\epsilon$. Thus multiplying $g(t)$ by $\|\Sigma\|_2^4$, all results will still hold. □

A.2 Proof of Proposition 2.3.2

Proposition 2.3.2. *For any tuple (x, A, s, ϵ) such that $x = As + \epsilon$ and any vector η,*

$$\begin{aligned}
&\|\nabla^2 f_{\nu^{(x)}}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_F \\
\le\ & \mathit{Poly}(L, d, \sigma_{\max}, C)\Big(\mathcal{Q}(\nu^{(\epsilon)}) \\
&+ N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + Q\left(N(\nu^{(s)})N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right)\Big)\|\eta\|_2^2, \qquad (*)
\end{aligned}$$

where $Q(x) = x + x^2$.

Remark A.2.1. Note that in Proposition 2.3.2 η can be a function of x, A, and s, in which case $\|\eta\|_2$ on the right-hand side will be replaced by its upper bound.

Proof. Note that $\nabla^2 f_{\nu^{(x)}}(\eta)$ and $\nabla^2 f_{\nu^{(As)}}(\eta)$ are the matrices generated by marginalizing 2 dimensions of the tensors on the direction η:

$$\begin{aligned}
\nabla^2 f_{\nu^{(x)}}(\eta) = &\left(\mathbb{E}_T[(As+\epsilon)^{\otimes 4}] - \left(\mathbb{E}_T[(As+\epsilon)^{\otimes 2}]\right)^{\otimes 2}\right)(\eta, \eta, \cdot, \cdot) \\
&- 2\left(\mathbb{E}_T[(As+\epsilon)^{\otimes 2}]\right)^{\otimes 2}(\eta, \cdot, \eta, \cdot).
\end{aligned}$$

Similarly,

$$\nabla^2 f_{\nu^{(As)}}(\eta) = \left(\mathbb{E}_T[(As)^{\otimes 4}] - \left(\mathbb{E}_T[(As)^{\otimes 2}]\right)^{\otimes 2}\right)(\eta, \eta, \cdot, \cdot) - 2\left(\mathbb{E}_T[(As)^{\otimes 2}]\right)^{\otimes 2}(\eta, \cdot, \eta, \cdot).$$

Hence,

$$\nabla^2 f_{\nu^{(x)}}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta) = (\Delta_1 - \Delta_2)(\eta, \eta, \cdot, \cdot) - 2\Delta_2(\eta, \cdot, \eta, \cdot),$$

where $\Delta_1 = \mathbb{E}_T[(As+\epsilon)^{\otimes 4}] - \mathbb{E}_T[(As)^{\otimes 4}]$ and $\Delta_2 = \left(\mathbb{E}_T[(As+\epsilon)^{\otimes 2}]\right)^{\otimes 2} - \left(\mathbb{E}_T[(As)^{\otimes 2}]\right)^{\otimes 2}$.

Part 1: We start with bounding Δ_1. Recall that

$$\Delta_1 = \mathbb{E}_T[(As+\epsilon)^{\otimes 4}] - \mathbb{E}_T\left[(As)^{\otimes 4}\right] = K_1 + K_2 + K_3 + \mathbb{E}_T\left[\epsilon^{\otimes 4}\right],$$

where

$$\begin{aligned}
K_1 &= \mathbb{E}_T\Big[(As)^{\otimes 3} \otimes \epsilon + (As)^{\otimes 2} \otimes \epsilon \otimes (As) + (As) \otimes \epsilon \otimes (As)^{\otimes 2} + \epsilon \otimes (As)^{\otimes 3}\Big];\\
K_2 &= E_T\Big[(As)^{\otimes 2} \otimes \epsilon^{\otimes 2} + (As) \otimes \epsilon \otimes (As) \otimes \epsilon + (As) \otimes \epsilon^{\otimes 2} \otimes (As)\\
&\quad + \epsilon^{\otimes 2} \otimes (As)^{\otimes 2} + \epsilon \otimes (As) \otimes \epsilon \otimes (As) + \epsilon \otimes (As)^{\otimes 2} \otimes \epsilon\Big];\\
K_3 &= E_T\Big[\epsilon^{\otimes 3} \otimes (As) + \epsilon^{\otimes 2} \otimes (As) \otimes \epsilon + \epsilon \otimes (As) \otimes \epsilon^{\otimes 2} + (As) \otimes \epsilon^{\otimes 3}\Big].
\end{aligned}$$

To bound $\|K_1\|_F$, we show a bound for the term $\|\mathbb{E}_T[(As)^{\otimes 3} \otimes \epsilon]\|_F$ first, noting that the other terms consisting K_1 can be bounded similarly. Note that $\mathbb{E}_T[(As)^{\otimes 3} \otimes \epsilon] - \mathbb{E}_T[(As)^{\otimes 3}] \otimes \mathbb{E}_T[\epsilon]$ is a 4-th order tensor, each element of which is bounded by $D_4^{(d,d)}(\nu^{(As,\epsilon)})$ (by the definition of $D_4^{(d,d)}(\nu^{(As,\epsilon)})$). Thus, by the triangle inequality,

$$\begin{aligned}
\|\mathbb{E}_T[(As)^{\otimes 3} \otimes \epsilon]\|_F &\le \|\mathbb{E}_T\left[(As)^{\otimes 3} \otimes \mathbb{E}_T[\epsilon]\right]\|_F + d^2 D_4^{(d,d)}(\nu^{(As,\epsilon)})\\
&\le d^2 \sigma_{\max}^3 C^3 N(\nu^{(\epsilon)}) + d^2 D_4^{(d,d)}(\nu^{(As,\epsilon)}).
\end{aligned}$$

The bounds for the other terms behave identically, hence,

$$\|K_1\|_F \le 4d^2 \sigma_{\max}^3 C^3 N(\nu^{(\epsilon)}) + 4d^2 D_4^{(d,d)}(\nu^{(As,\epsilon)}).$$

Similarly, one can show that

$$\|K_3\|_F \le 4L\sigma_{\max} N(\nu^{(s)}) + 4d^2 D_4^{(d,d)}(\nu^{(As,\epsilon)}).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_T[(As+\epsilon)^{\otimes 4}] - \mathbb{E}_T[(As)^{\otimes 4}]\\
&= \mathbb{E}_{Y\sim\nu^{(\epsilon)}}\left[Y^{\otimes 4}\right] + K_2 + \mathrm{Poly}(L, d, \sigma_{\max}, C)\left(N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right). \quad \text{(A.9)}
\end{aligned}$$

Part 2: Let us turn to bounding

$$\Delta_2 = \left(\mathbb{E}_T[(As+\epsilon)^{\otimes 2}]\right)^{\otimes 2} - \left(\mathbb{E}_T[(As)^{\otimes 2}]\right)^{\otimes 2}.$$

Note that

$$(As+\epsilon)^{\otimes 2} = (As)^{\otimes 2} + \epsilon^{\otimes 2} + (As) \otimes \epsilon + \epsilon \otimes (As),$$

and

$$\|\mathbb{E}_T[(As)\otimes\epsilon+\epsilon\otimes(As)]\|_F \le 2\sigma_{\max}N(\nu^{(s)})N(\nu^{(\epsilon)})+2dD_4^{(d,d)}(\nu^{(As,\epsilon)}).$$

Thus,

$$\Delta_2 = \left(\mathbb{E}_T[\epsilon^{\otimes 2}]\right)^{\otimes 2}+K_4+\mathrm{Poly}(L,d,\sigma_{\max},C)Q\left(N(\nu^{(s)})N(\nu^{(\epsilon)})+D_4^{(d,d)}(\nu^{(As,\epsilon)})\right), \tag{A.10}$$

where $Q(x)=x+x^2$ and $K_4=\mathbb{E}_T[(As)^{\otimes 2}]\otimes\mathbb{E}_T[\epsilon^{\otimes 2}]+\mathbb{E}_T[\epsilon^{\otimes 2}]\otimes\mathbb{E}_T[(As)^{\otimes 2}]$.

Part 3: Finally, note that by the definition of $\mathcal{Q}(\nu^{(\epsilon)})$,

$$\|\left(\mathbb{E}_T[\epsilon^{\otimes 4}]-(\mathbb{E}_T[\epsilon^{\otimes 2}])^{\otimes 2}\right)(\eta,\eta,\cdot,\cdot)-2(\mathbb{E}_T[\epsilon^{\otimes 2}])^{\otimes 2}(\eta,\cdot,\eta,\cdot)\|_F \le \mathcal{Q}(\nu^{(\epsilon)})\|\eta\|_2^2. \tag{A.11}$$

Combining Eqs. (A.9) to (A.11), we have

$$\begin{aligned}
&\|\nabla^2 f_{\nu^{(x)}}(\eta)-\nabla^2 f_{\nu^{(As)}}(\eta)\|_F\\
&\le \|(K_2-K_4)(\eta,\eta,\cdot,\cdot)-2K_4(\eta,\cdot,\eta,\cdot)\|_F\\
&\quad+\mathrm{Poly}(L,d,\sigma_{\max},C)\Big(\mathcal{Q}(\nu^{(\epsilon)})+N(\nu^{(s)})\\
&\qquad\qquad+N(\nu^{(\epsilon)})+Q\big(N(\nu^{(s)})N(\nu^{(\epsilon)})+D_4^{(d,d)}(\nu^{(As,\epsilon)})\big)\Big)\|\eta\|_2^2.
\end{aligned} \tag{A.12}$$

Part 4: It remains to bound $\|(K_2-K_4)(\eta,\eta,\cdot,\cdot)-2K_4(\eta,\cdot,\eta,\cdot)\|_F$. Note that

$$E_T[(As)\otimes\epsilon\otimes(As)\otimes\epsilon](\eta,\eta,\cdot,\cdot)=E_T[(As)^{\otimes 2}\otimes\epsilon^{\otimes 2}](\eta,\cdot,\eta,\cdot).$$

Then,

$$\begin{aligned}
&\|(K_2-K_4)(\eta,\eta,\cdot,\cdot)-2K_4(\eta,\cdot,\eta,\cdot)\|_F\\
&\le \|\left(E_T[(As)^{\otimes 2}\otimes\epsilon^{\otimes 2}]-\mathbb{E}_T[(As)^{\otimes 2}]\otimes\mathbb{E}_T[\epsilon^{\otimes 2}]\right)(\eta,\eta,\cdot,\cdot)\|_F\\
&\quad+\|\left(E_T[\epsilon^{\otimes 2}\otimes(As)^{\otimes 2}]-\mathbb{E}_T[\epsilon^{\otimes 2}]\otimes\mathbb{E}_T[(As)^{\otimes 2}]\right)(\eta,\eta,\cdot,\cdot)\|_F\\
&\quad+2\|\left(E_T[(As)^{\otimes 2}\otimes\epsilon^{\otimes 2}]-\mathbb{E}_T[(As)^{\otimes 2}]\otimes\mathbb{E}_T[\epsilon^{\otimes 2}]\right)(\eta,\cdot,\eta,\cdot)\|_F\\
&\quad+2\|\left(E_T[\epsilon^{\otimes 2}\otimes(As)^{\otimes 2}]-\mathbb{E}_T[\epsilon^{\otimes 2}]\otimes\mathbb{E}_T[(As)^{\otimes 2}]\right)(\eta,\cdot,\eta,\cdot)\|_F\\
&\le 6d^2D_4^{(d,d)}(\nu^{(As,\epsilon)})\|\eta\|_2^2.
\end{aligned}$$

Combining Eq. (A.12) with the above inequality leads to the conclusion. □

A.3 Proofs for Section 2.3.3

Recall that

$$\xi = 6C^2D_2(\mu,\nu^{(s)})+D_4(\mu,\nu^{(s)}).$$

We start with presenting the proofs for Lemmas 2.3.6 and 2.3.7. The technical lemmas used in the proofs are presented in the latter part of this section.

Lemma 2.3.6. *Given that* $\xi \le \frac{\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{low}^2}{4 L_{high}^2 d^6 A_{(2,\max)}^2 A_{\max}^2}$ *and* $P(L_{high}) \le \frac{1}{4}\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{low}^2$, *on the event* $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$,

$$\|M - \hat{M}\|_2 \le \Phi(\mu)\Big(D_4(\mu, \nu^{(s)}) + \mathcal{Q}(\nu^{(\epsilon)}) + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\Big),$$

where $\Phi(\mu)$ *is a problem-dependent constant that is polynomial in* $L_{high}, d, \sigma_{\max}, 1/\sigma_{\min}$, $\kappa_{\max}$, $1/\kappa_{\min}, \ell, L$ *and* C.

Proof. Recall that $\hat{f} = f_{\nu^{(x)}}$. Let $E_1 = \nabla^2 f_{A\mu}(\phi) - \nabla^2 \hat{f}(\phi)$ and $E_2 = \nabla^2 f_{A\mu}(\psi) - \nabla^2 \hat{f}(\psi)$. Then by Lemma A.3.2, $\|E_1\|_2, \|E_2\|_2 \le L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(L_{\text{high}})$. Note that by Proposition 2.3.1, $\nabla^2 f_{A\mu}(\phi) = AKD_\phi A^\top$. Given that $\xi \le \frac{\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{\text{low}}^2}{4 L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2}$ and $P(L_{\text{high}}) \le \frac{1}{4}\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{\text{low}}^2$, the condition in Lemma A.3.4 holds on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$. Thus we have

$$\begin{aligned}
&\|M - \hat{M}\|_2 \\
&= \|(\nabla^2 f_{A\mu}(\phi))(\nabla^2 f_{A\mu}(\psi))^{-1} - (\nabla^2 \hat{f}(\phi))(\nabla^2 \hat{f}(\psi))^{-1}\|_2 \\
&\le \frac{2\|\nabla^2 f_{A\mu}(\phi)\|_2}{\sigma_{\min}^2(\nabla^2 f_{A\mu}(\psi))}\|E_2\|_2 + \frac{2}{\sigma_{\min}(\nabla^2 f_{A\mu}(\psi))}\|E_1\|_2 \\
&\le 2\left(\frac{L_{\text{high}}^2 A_{(2,\max)}^2 \kappa_{\max}\sigma_{\max}^2}{\kappa_{\min}^2 A_{(2,\min)}^4 \sigma_{\min}^4 L_{\text{low}}^4} + \frac{1}{\kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2 L_{\text{low}}^2}\right)\left(L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(L_{\text{high}})\right) \\
&\le \Phi(\mu)\left(D_4(\mu, \nu^{(s)}) + \mathcal{Q}(\nu^{(\epsilon)}) + N(\nu^{(s)}) + N(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)})\right),
\end{aligned}$$

where the first inequality is due to Lemma A.3.4, and the second inequality is due to Lemma A.3.5. □

Lemma 2.3.7. *Denote* $\hat{M} = M + E$ *where* $M = PDP^{-1}$ *and where* D *is a diagonal matrix* $diag(\sigma_1, \ldots, \sigma_d)$. *Assume* $\hat{M}$ *has distinct eigenvalues. If* $\gamma_D = \min_{i\neq j}|\sigma_i - \sigma_j| > 4\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2$, *and* $\min_{i,j:i\neq j}\|P_i - P_j\|_2 > \frac{8}{\gamma_D}\frac{\sigma_{\max}^2(P)}{\sigma_{\min}(P)}\|E\|_2$, *then there exist constants* $\{c_1, \ldots, c_d\}$ *and a permutation* π, *such that*

$$\max_{1\le k\le d}\|c_k\hat{P}_{\pi(k)} - P_k\|_2 \le 4\frac{\sigma_{\max}^2(P)}{\gamma_D\sigma_{\min}(P)}\|E\|_2,$$

and therefore

$$\sum_{k=1}^{d}\|c_k\hat{P}_{\pi(k)} - P_k\|_2 \le 4d\frac{\sigma_{\max}^2(P)}{\gamma_D\sigma_{\min}(P)}\|E\|_2,$$

where $\hat{P}$ *is the matrix of eigenvectors of* $\hat{M}$.

Proof. For $1 \le k \le d$, assume

$$P_{(k)}^{-1}EP_{(k)} = \begin{pmatrix} F_{1k} & F_{2k} \\ F_{3k} & F_{4k} \end{pmatrix},$$

where $P_{(k)}$ is the matrix $(P_k, P_1, \cdots, P_{k-1}, P_{k+1}, \cdots, P_d)$. Let $\gamma_k = \|F_{3k}\|_2$, $\eta_k = \|F_{2k}\|_2$, and

$$\delta_k = \min_{j:j\neq k}\left|\left(\frac{\phi^\top P_k}{\psi^\top P_k}\right)^2 - \left(\frac{\phi^\top P_j}{\psi^\top P_j}\right)^2\right| - \|F_{1k}\|_2 - \|F_{4k}\|_2.$$

Note that by definition, $\gamma_k = \|F_{3k}\|_2 \le \|P_{(k)}^{-1} E P_k\|_2 \le \frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2$, $\eta_k = \|F_{2k}\|_2 \le \|(P^{-1})_k E P_{(k)}\|_2 \le \frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2$, and $\|F_{1k}\|_2, \|F_{4k}\|_2 \le \|P_{(k)}^{-1} E P_{(k)}\|_2 \le \frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2$. Thus,

$$\begin{aligned}
\delta_k &= \min_{j:j\neq k} \left| \left(\frac{\phi^\top P_k}{\psi^\top P_k}\right)^2 - \left(\frac{\phi^\top P_j}{\psi^\top P_j}\right)^2 \right| - \|F_{1k}\|_2 - \|F_{4k}\|_2 \\
&\ge \min_{j:j\neq k} \left| \left(\frac{\phi^\top P_k}{\psi^\top P_k}\right)^2 - \left(\frac{\phi^\top P_j}{\psi^\top P_j}\right)^2 \right| - 2\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2 \\
&\ge \gamma_D - 2\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2 \\
&> 2\frac{\sigma_{\max}(P)}{\sigma_{\min}(P)}\|E\|_2 > 0,
\end{aligned}$$

and $\delta_k^2 > 4\gamma_k\eta_k$. Therefore, by Theorem 2.8, Chapter V of [Stewart and Sun, 1990], there exist a unique vector v satisfying $\|v\|_2 \le 2\frac{\gamma_k}{\delta_k}$ such that there exists one of a eigenvector $\hat{P}_k$ of $\hat{M}$ satisfying

$$\|\hat{P}_k - P_k\|_2 \le \|P_{ck}\|_2\|v\|_2 \le 2\sigma_{\max}(P)\frac{\gamma_k}{\delta_k} \le \frac{4\sigma_{\max}^2(P)}{\gamma_A\sigma_{\min}(P)}\|E\|_2,$$

where P_{ck} is the $d \times (d-1)$ matrix $(P_1, \ldots, P_{k-1}, P_{k+1}, \ldots, P_d)$. By condition, for $i \neq j$, $\frac{8\sigma_{\max}^2(P)}{\gamma_A\sigma_{\min}(P)}\|E\|_2 < \|P_i - P_j\|_2 \le \|P_i - \hat{P}_i\|_2 + \|P_j - \hat{P}_i\|_2$, thus $\hat{P}_i \neq \hat{P}_j$. Summing up the upper bound gets the result. □

A.3.1 Technical lemmas

Given two matrices A and B, a distribution μ of s, and any vector η, define

$$\begin{aligned}
G_1(A,B,\eta) &= \mathbb{E}[(Bs)^{\otimes 2} \otimes (As)^{\otimes 2}](\eta,\eta,\cdot,\cdot) = \int (\eta^\top Bs)^2 Ass^\top A^\top \, d\mu(s); \\
G_2(A,B,\eta) &= \left(\mathbb{E}[(Bs)^{\otimes 2}] \otimes \mathbb{E}[(As)^{\otimes 2}]\right)(\eta,\eta,\cdot,\cdot) = \int (\eta^\top Bs)^2 \, d\mu(s) \int Ass^\top A^\top \, d\mu(s); \\
G_3(A,B,\eta) &= \mathbb{E}[(Bs)\otimes(As)]^{\otimes 2}(\eta,\cdot,\eta,\cdot) = \Big(\int (\eta^\top Bs) As \, d\mu(s)\Big)\Big(\int (\eta^\top Bs) As \, d\mu(s)\Big)^\top.
\end{aligned}$$

and their empirical estimations

$$\begin{aligned}
\widehat{G_1}(A,B,\eta) &= \frac{1}{n}\sum_{k=1}^n \big(\eta^\top Bs(k)\big)^2 As(k)s(k)^\top A^\top = \int (\eta^\top Bs)^2 Ass^\top A^\top \, d\nu^{(s)}(s); \\
\widehat{G_2}(A,B,\eta) &= \frac{1}{n^2}\sum_{k=1}^n \big(\eta^\top Bs(k)\big)^2 \sum_{k=1}^n As(k)s(k)^\top A^\top \\
&= \int (\eta^\top Bs)^2 \, d\nu^{(s)}(s) \int Ass^\top A^\top \, d\nu^{(s)}(s); \\
\widehat{G_3}(A,B,\eta) &= \frac{1}{n^2}\Big(\sum_{k=1}^n \big(\eta^\top Bs(k)\big) As(k)\Big)\Big(\sum_{k=1}^n \big(\eta^\top Bs(k)\big) As(k)\Big)^\top \\
&= \Big(\int (\eta^\top Bs) As \, d\nu^{(s)}(s)\Big)\Big(\int (\eta^\top Bs) As \, d\nu^{(s)}(s)\Big)^\top.
\end{aligned}$$

We now present the technical lemmas used in the proof of Lemma 2.3.6. The first lemma is to bound the empirical estimation of $G_i(A, B, \eta)$ as defined above, which then can be used to bound $\|\nabla^2 f_{A\mu}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_2$.

Lemma A.3.1. *For any matrices A, B, and any vector η,*

- $\|G_1(A, B, \eta) - \widehat{G_1}(A, B, \eta)\|_2 \le d^5 B_{(2,\max)}^2 A_{\max}^2 D_4(\mu, \nu^{(s)})\|\eta\|_2^2$;
- $\|G_2(A, B, \eta) - \widehat{G_2}(A, B, \eta)\|_2 \le 2d^5 B_{(2,\max)}^2 A_{\max}^2 C^2 D_2(\mu, \nu^{(s)})\|\eta\|_2^2$;
- $\|G_3(A, B, \eta) - \widehat{G_3}(A, B, \eta)\|_2 \le 2d^5 B_{(2,\max)}^2 A_{\max}^2 C^2 D_2(\mu, \nu^{(s)})\|\eta\|_2^2$.

Proof. We use $G_i(\eta)$ to denote $G_i(A, B, \eta)$ in the proof. Without loss of generality assume $\|\eta\|_2 = 1$. Note that all the integral functions of $G_i(\eta)$ or $\widehat{G_i}(\eta)$ are matrices of polynomials in s. Thus, we only need to bound its coefficients.

Part 1: Bounding $\|G_1(\eta) - \widehat{G_1}(\eta)\|_2$. Note that

$$(G_1)_{i,j} = \int (\sum_t \eta^\top B_t s_t)^2 \sum_t A_{i,t} s_t \sum_t A_{j,t} s_t d\mu(s),$$

where the coefficient of the term $s_{t_1} s_{t_2} s_{t_3} s_{t_4}$ is $\eta^\top B_{t_1} \eta^\top B_{t_2} A_{i,t_3} A_{j,t_4}$, which is bounded by $\max_i |\eta^\top B_i|^2 A_{\max}^2 \le B_{(2,\max)}^2 A_{\max}^2$. So,

$$\left|(G_1)_{i,j} - (\hat{G}_1)_{i,j}\right| \le d^4 B_{(2,\max)}^2 A_{\max}^2 D_4(\mu, \nu^{(s)}),$$

and thus

$$\|G_1(\eta) - \widehat{G_1}(\eta)\|_2 \le d^5 B_{(2,\max)}^2 A_{\max}^2 D_4(\mu, \nu^{(s)}).$$

Part 2: Bounding $\|G_2(\eta) - \widehat{G_2}(\eta)\|_2$. Similarly,

$$\left|\int A_{i:} s s^\top A_{j:}^\top \, d\mu(s) - \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \le d^2 A_{\max}^2 D_2(\mu, \nu^{(s)})$$

and

$$\left|\int (\eta^\top B s)^2 \, d\mu(s) - \int (\eta^\top B s)^2 \, d\nu^{(s)}(s)\right| \le d^2 B_{(2,\max)}^2 D_2(\mu, \nu^{(s)}).$$

Also note that $\left|\int (\eta^\top B s)^2 \, d\mu(s)\right| \le d^2 B_{(2,\max)}^2 C^2$, and $\left|\int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \le d^2 A_{\max}^2 C^2$. Now consider the difference between G_2 and $\hat{G}_2$.

$$\begin{aligned}
&\left|(G_2)_{i,j} - (\hat{G}_2)_{i,j}\right| \\
=& \left|\int (\eta^\top B s)^2 \, d\mu(s) \int A_{i:} s s^\top A_{j:}^\top \, d\mu(s) - \int (\eta^\top B s)^2 \, d\nu^{(s)}(s) \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \\
\le& \left|\int (\eta^\top B s)^2 \, d\mu(s) \int A_{i:} s s^\top A_{j:}^\top \, d\mu(s) - \int (\eta^\top B s)^2 \, d\mu(s) \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \\
&\quad + \left|\int (\eta^\top B s)^2 \, d\mu(s) \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s) - \int (\eta^\top B s)^2 \, d\nu^{(s)}(s) \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \\
\le& \left|\int (\eta^\top B s)^2 \, d\mu(s)\right| \left|\int A_{i:} s s^\top A_{j:}^\top \, d\mu(s) - \int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \\
&\quad + \left|\int (\eta^\top B s)^2 \, d\mu(s) - \int (\eta^\top B s)^2 \, d\nu^{(s)}(s)\right| \left|\int A_{i:} s s^\top A_{j:}^\top \, d\nu^{(s)}(s)\right| \\
\le& 2d^4 B_{(2,\max)}^2 A_{\max}^2 C^2 D_2(\mu, \nu^{(s)}).
\end{aligned}$$

Therefore,

$$\|G_2(\eta) - \widehat{G_2}(\eta)\|_2 \le 2d^5 B_{(2,\max)}^2 A_{\max}^2 C^2 D_2(\mu, \nu^{(s)}).$$

Part 3: Bounding $\|G_3(\eta) - \widehat{G_3}(\eta)\|_2$. By a calculation similar to Part 2,

$$\|G_3(\eta) - \widehat{G_3}(\eta)\|_2 \le 2d^5 B_{(2,\max)}^2 A_{\max}^2 C^2 D_2(\mu, \nu^{(s)}).$$

□

Lemma A.3.2. *For any tuple (x, A, s, ϵ) such that $x = As + \epsilon$ and any vector η,*

$$\|\nabla^2 f_{A\mu}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_2 \le \|\nabla^2 f_{A\mu}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_F \le \|\eta\|_2^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi.$$

Thus,

$$\|\nabla^2 f_{A\mu}(\eta) - \nabla^2 \widehat{f}(\eta)\|_2 \le \|\eta\|_2^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(\|\eta\|_2).$$

Proof. Without loss of generality assume $\|\eta\|_2 = 1$. Recall that

$$\nabla^2 f_{A\mu}(\eta) = G_1(A, A, \eta) - G_2(A, A, \eta) - 2G_3(A, A, \eta).$$

Similarly,

$$\nabla^2 f_{\nu^{(As)}}(\eta) = \widehat{G_1}(A, A, \eta) - \widehat{G_2}(A, A, \eta) - 2\widehat{G_3}(A, A, \eta).$$

Applying Lemma A.3.1,

$$\begin{aligned}\|\nabla^2 f_{A\mu}(\eta) - \nabla^2 \hat{f}_{\nu^{(As)}}(\eta)\|_2 &\le \|\nabla^2 f_{A\mu}(\eta) - \nabla^2 \hat{f}_{\nu^{(As)}}(\eta)\|_F \\ &\le d^5 A_{(2,\max)}^2 A_{\max}^2 \left(6C^2 D_2(\mu, \nu^{(s)}) + D_4(\mu, \nu^{(s)})\right).\end{aligned}$$

Lastly, combining with Proposition 2.3.2,

$$\begin{aligned}&\|\nabla^2 f_{A\mu}(\eta) - \nabla^2 \hat{f}(\eta)\|_2 \\ &\quad\le \|\nabla^2 f_{A\mu}(\eta) - \nabla^2 f_{\nu^{(As)}}(\eta)\|_2 + \|\nabla^2 f_{\nu^{(As)}}(\eta) - \nabla^2 \hat{f}(\eta)\|_2 \\ &\quad\le \|\eta\|_2^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(\|\eta\|_2).\end{aligned}$$

□

The next lemma shows that $\hat{X}^{-1}$ is close to X^{-1} with respect to matrix induced 2-norm.

Lemma A.3.3. *Let $\hat{X}$ be a non-singular matrix satisfying that $\hat{X} = X + E$ and $\sigma_{\min}(X) \ge 2\|E\|_2$. Then $\|\hat{X}^{-1}\|_2 \le \frac{2}{\sigma_{\min}(X)}$, and $\|\hat{X}^{-1} - X^{-1}\|_2 \le \frac{2}{\sigma_{\min}^2(X)}\|E\|_2$.*

Proof. Note that $\|\hat{X}^{-1}\|_2$ is the inverse of the minimal singular value of $\hat{X}$. Also,

$$\min_{v:\|v\|_2=1} \|\hat{X}v\|_2 = \min_{v:\|v\|_2=1} \|(X+E)v\|_2 \ge \min_{v:\|v\|_2=1} \|Xv\|_2 - \|Ev\|_2 \ge \sigma_{\min}(X) - \|E\|_2.$$

So $\|\hat{X}^{-1}\|_2 \le \frac{1}{\sigma_{\min}(X)-\|E\|_2} \le \frac{2}{\sigma_{\min}(X)}$. Moreover,

$$\|\hat{X}^{-1} - X^{-1}\|_2 \le \|X^{-1}\|_2\|\hat{X}^{-1}\|_2\|\hat{X} - X\|_2 \le \frac{2}{\sigma^2_{\min}(X)}\|E\|_2.$$

□

Now we can estimate the variance between XY^{-1} and $(X+E_1)(Y+E_2)^{-1}$.

Lemma A.3.4. *Assume that* $\sigma_{\min}(Y) \ge 2\|E_2\|_2$*, then*

$$\|XY^{-1} - (X+E_1)(Y+E_2)^{-1}\|_2 \le \frac{2\|X\|_2}{\sigma^2_{\min}(Y)}\|E_2\|_2 + \frac{2}{\sigma_{\min}(Y)}\|E_1\|_2.$$

Proof. Applying Lemma A.3.3,

$$\begin{aligned}
&\|XY^{-1} - (X+E_1)(Y+E_2)^{-1}\|_2 \\
\le &\|XY^{-1} - X(Y+E_2)^{-1}\|_2 + \|X(Y+E_2)^{-1} - (X+E_1)(Y+E_2)^{-1}\|_2 \\
\le &\|X\|_2\|Y^{-1} - (Y+E_2)^{-1}\|_2 + \|E_1\|_2\|(Y+E_2)^{-1}\|_2 \\
\le &\frac{2\|X\|_2}{\sigma^2_{\min}(Y)}\|E_2\|_2 + \frac{2}{\sigma_{\min}(Y)}\|E_1\|_2.
\end{aligned}$$

□

Lemma A.3.5. *On the event* $\mathcal{E}^A_\psi$*,*

$$\sigma_{\max}(\nabla^2 f_{A\mu}(\psi)) \le L^2_{high}\kappa_{\max}A^2_{(2,\max)}\sigma^2_{\max}; \quad \sigma_{\min}(\nabla^2 f_{A\mu}(\psi)) \ge L^2_{low}\kappa_{\min}A^2_{(2,\min)}\sigma^2_{\min}.$$

Proof. Let $D_\psi = \operatorname{diag}\left((\psi^\top A_1)^2, \cdots, (\psi^\top A_d)^2\right)$. Note that by Proposition 2.3.1, $\nabla^2 f_{A\mu}(\psi) = AKD_\psi A^\top$. Since $\nabla^2 f_{A\mu}(\psi)$ is symmetric, it is sufficient to bound $v^\top \nabla^2 f_{A\mu}(\psi) v$ for any unit vector v, as follows.

$$v^\top A D_\psi K A^\top v \ge L^2_{\text{low}}\kappa_{\min}A^2_{(2,\min)}\|v^\top A\|_2^2 \ge L^2_{\text{low}}\kappa_{\min}A^2_{(2,\min)}\sigma^2_{\min}.$$

Thus, $\sigma_{\min}(\nabla^2 f_{A\mu}(\psi)) \ge L^2_{\text{low}}\kappa_{\min}A^2_{(2,\min)}\sigma^2_{\min}$. A similar calculation shows the bound on the maximum singular value. □

A.4 Analysis of DICA – Proof of Theorem 2.2.1

Instead of Theorem 2.2.1, we prove a stronger result, Theorem A.4.1, which is presented below. As we will see, Theorem 2.2.1 will be an immediate corollary of Theorem A.4.1.

Let

$$\begin{aligned}
\bar{\xi} &= \frac{L^2_{\text{high}}d^5A^2_{(2,\max)}A^2_{\max}\xi + P(L_{\text{high}})}{L^2_{\text{low}}\kappa_{\min}A^2_{(2,\min)}\sigma^2_{\min}}; \\
\hat{\xi} &= \frac{\sqrt{6}L^2_{\text{high}}\sigma^2_{\max}}{A^2_{(2,\min)}L^2_{\text{low}}}\bar{\xi} + \frac{16L^2_{\text{high}}d^5A^2_{\max}}{9\kappa_{\min}A_{(2,\min)}L^2_{\text{low}}}\xi + P\left(\frac{4L_{\text{high}}}{3\kappa^{1/2}_{\min}A_{(2,\min)}\sigma_{\min}L_{\text{low}}}\right) \\
\tilde{Q} &= \frac{4L^6_{\text{high}}\sigma^2_{\max}A^4_{(2,\max)}}{L^6_{\text{low}}\sigma^4_{\min}A^2_{(2,\min)}}\hat{\xi}\,.
\end{aligned}$$

Theorem A.4.1. *Assume that the following conditions hold:*

C1: $\hat{T}$ *has distinct eigenvalues;*

C2: $\gamma_R > 4\frac{\sigma_{\max}^2}{\sigma_{\min}}\tilde{Q}$;

C3: $\min_{i,j:i\neq j} \|A_i - A_j\|_2 > \frac{8}{\gamma_R}\frac{\sigma_{\max}^2}{\sigma_{\min}}\tilde{Q}$;

C4: $\xi \leq \frac{L_{low}^2 \kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2}{6L_{high}^2 d^5 A_{(2,\max)}^2 A_{\max}^2 L_{high}^2}$;

C5: $P(L_{high}) \leq \frac{1}{6}L_{low}^2 \kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2$;

C6: $\hat{\xi} \leq \frac{\sigma_{\min}^2 L_{low}^2}{2L_u^2 A_{(2,\max)}^2}$.

Then on the event $\mathcal{E}_{\psi}^{A} \cap \mathcal{E}_{\phi_1}^{\tilde{R}} \cap \mathcal{E}_{\phi_2}^{\tilde{R}}$,

$$d(\hat{A}, A) \leq \frac{4\sigma_{\max}^2}{\gamma_R \sigma_{\min}}\tilde{Q},$$

where $\hat{A}$ *is the output of the DICA algorithm.*

The proof of Theorem A.4.1 is similar to the proof of Theorem 2.3.2. We will still first bound the difference between M and $\hat{M}$, and then apply Lemma 2.3.7 to get the conclusion. Again we postpone the technical lemmas in the latter part of this section.

Proof of Theorem A.4.1. We prove in Lemma A.4.6 that

$$\|\hat{M} - M\|_2 \leq \tilde{Q}.$$

Then by Lemma 2.3.7,

$$d(\hat{A}, A) \leq \frac{4\sigma_{\max}^2}{\gamma_R \sigma_{\min}}\|\hat{M} - M\|_2.$$

□

To finish the proof of Theorem 2.2.1, it remains to pick L_{low} and L_{high}, such that $\text{Prob}(\mathcal{E}_{\psi}^{A} \cap \mathcal{E}_{\phi_1}^{\tilde{R}} \cap \mathcal{E}_{\phi_2}^{\tilde{R}}) \geq 1 - \delta$. Next lemma provides such result, whose proof is deferred to Section A.6:

Lemma A.4.1. *For any* A *and orthonormal matrix* R*, with probability at least* $1 - \delta$*, the following inequalities holds simultaneously:*

- $\min_i |\psi^\top A_i| \geq \frac{\sqrt{\pi} A_{(2,\min)}}{5\sqrt{2}(d+1)}\delta$;
- $\min_i\{|\phi_2^\top R_i|\} \geq \frac{\sqrt{\pi}}{5\sqrt{2}(d+1)}\delta$;
- $\|\phi_1\|_2, \|\phi_2\|_2 \leq \sqrt{2}\left(\sqrt{\log(\frac{5}{\delta})} + \sqrt{d}\right)$;
- $\gamma_R \geq \frac{\pi^2\delta^2}{25d^3}$.

Remark A.4.2. Note that all the constants in Lemma A.4.1 are polynomial in d (or d^{-1} for the lower bound), thus the result of Theorem A.4.1 is polynomial in d and $\frac{1}{\delta}$ with probability at least $1 - \delta$.

A.4.1 Technical lemmas

This section is to prove that

$$\|\hat{M} - M\|_2 \leq \tilde{Q}.$$

Given a vector ψ, for any distribution ν, let $B(\nu) = \left(\nabla^2 f_\nu(\psi)\right)$, $T(\nu, \phi) = \nabla^2 f_\nu(B^{-1}(\nu)\phi)$, and $M(\nu) = T(\nu, \phi_1)T(\nu, \phi_2)^{-1}$. Then $M = M(A\mu)$ and $\hat{M} = M(\nu^{(x)})$. Thus, to bound $\|\hat{M} - M\|_2$ we study $\nu \mapsto M(\nu)$ in this section with $\nu = A\mu$ for some $\mu \in \Pi_0$. By Proposition 2.3.1, $\nabla^2 f_{A\mu}(\psi) = AKD_\psi A^\top$. Thus $B = AK^{1/2}D_\psi^{1/2}R^\top$ for some orthonormal matrix R. We need to first introduce some lemmas. The next two lemmas show that $\hat{B}$ is a good estimate of B, in the sense that $\hat{B}^{-1}B$ is close to some orthonormal matrix. This result depends on the stability of the map $X \mapsto X^{1/2}$.

Lemma A.4.3. *Given two symmetric positive semi-definite matrices X and $\hat{X} = X + E$, where $X = HH^\top$ and $\hat{X} = \hat{H}\hat{H}^\top$, such that $\|X^{-1}\|_2\|E\|_2 < 1$, then every singular value of $H^{-1}\hat{H}$ is bounded between $\sqrt{1 - \|X^{-1}\|_2\|E\|_2}$ and $\sqrt{1 + \|X^{-1}\|_2\|E\|_2}$. Taking inverses, it also follows that every singular value of $\hat{H}^{-1}H$ is bounded between $\frac{1}{\sqrt{1+\|X^{-1}\|_2\|E\|_2}}$ and $\frac{1}{\sqrt{1-\|X^{-1}\|_2\|E\|_2}}$.*

Proof. For any unit vector x,

$$\begin{aligned} x^\top \left(H^{-1}\hat{H}\hat{H}^\top H^{-\top} - I\right) x &= x^\top H^{-1}\left(\hat{H}\hat{H}^\top - HH^\top\right) H^{-\top} x \\ &\leq \|H^{-\top}x\|_2^2 \|E\|_2 \leq \|X^{-1}\|_2\|E\|_2. \end{aligned}$$

Thus every singular value of $H^{-1}\hat{H}$ is lower bounded by $\sqrt{1 - \|X^{-1}\|_2\|E\|_2}$ and upper bounded by $\sqrt{1 + \|X^{-1}\|_2\|E\|_2}$, and every singular value of $\hat{H}^{-1}H$ is bounded between $\frac{1}{\sqrt{1+\|X^{-1}\|_2\|E\|_2}}$ and $\frac{1}{\sqrt{1-\|X^{-1}\|_2\|E\|_2}}$. □

Applying Lemma A.4.3, we can get the stability of B, as follows.

Lemma A.4.4. *Assuming Condition (4) and (5) , under the event $\mathcal{E}_\psi^A$ there exists an orthonormal matrix R^* such that*

$$\sqrt{1 - \bar{\xi}} \leq \|B^{-1}\hat{B}\|_2 \leq \sqrt{1 + \bar{\xi}}, \text{ and } \|\hat{B}^{-1}B - R^*\|_2 \leq \bar{\xi}.$$

Proof. Note that by Lemma A.3.5, under the event $\mathcal{E}_\psi^A$,

$$\|\nabla^2 f_{A\mu}(\psi)\|_2 \geq l_l^2 \kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2.$$

Moreover, by Lemma A.3.2, on event $\mathcal{E}_\psi^A$,

$$\|\nabla^2 f_{A\mu}(\psi) - \nabla^2 \hat{f}(\psi)\|_2 \leq L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(L_{\text{high}}).$$

Thus,

$$\|\left(\nabla^2 f_{A\mu}(\psi)\right)^{-1}\|_2 \|\nabla^2 f_{A\mu}(\psi) - \nabla^2 \hat{f}(\psi)\|_2 \leq \frac{L_{\text{high}}^2 d^5 A_{(2,\max)}^2 A_{\max}^2 \xi + P(L_{\text{high}})}{l_l^2 \kappa_{\min} A_{(2,\min)}^2 \sigma_{\min}^2} = \bar{\xi}.$$

Note that by Condition (5), $\bar{\xi} \leq 1/3$. Thus we can apply Lemma A.4.3 to get that every singular value of $\hat{B}^{-1}B$ belongs to the interval

$$\left[\frac{1}{\sqrt{1+\|\left(\nabla^2 f_{A\mu}(\psi)\right)^{-1}\|_2\|E\|_2}}, \frac{1}{\sqrt{1-\|\left(\nabla^2 f_{A\mu}(\psi)\right)^{-1}\|_2\|E\|_2}}\right] \subset \left[\frac{1}{\sqrt{1+\bar{\xi}}}, \frac{1}{\sqrt{1-\bar{\xi}}}\right].$$

Therefore, there exist an orthonormal matrix R^* such that

$$\|\hat{B}^{-1}B - R^*\|_2 \leq \max\left\{\left|1 - \frac{1}{\sqrt{1+\bar{\xi}}}\right|, \left|\frac{1}{\sqrt{1-\bar{\xi}}} - 1\right|\right\} \leq \bar{\xi},$$

where the last inequality is by $0 \leq \bar{\xi} \leq 1/3$. □

By renaming ϕ_i to $(R^*)^\top \phi_i$ where R^* is the orthonormal matrix from Lemma A.4.4, recall that $T_i = \nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi_i) = AD_\psi^{-1}\Lambda_i A^\top$ for $i \in \{1,2\}$, where

$$\Lambda_i = \operatorname{diag}\left((\phi_i^\top R^* R_1)^2, \cdots, (\phi_i^\top R^* R_d)^2\right).$$

Then,

$$M = A\Lambda_1\Lambda_2^{-1}A^{-1} = A\Lambda A^{-1}, \tag{A.13}$$

where $\Lambda = \operatorname{diag}\left((\frac{\phi_1^\top R^* R_1}{\phi_2^\top R^* R_1})^2, \cdots, (\frac{\phi_1^\top R^* R_d}{\phi_2^\top R^* R_d})^2\right)$. It still remains to bound the perturbation of M. We let $\tilde{R} = R^* R$ to be the orthonormal matrix that appears in the event $\mathcal{E}_\phi^{\tilde{R}}$.

Lemma A.4.5. *Assuming that Condition (4) and (5) hold. Then on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^{\tilde{R}}$,*

$$\|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi) - \nabla^2 \widehat{f}(\hat{B}^{-\top}\phi)\|_2 \leq \hat{\xi}.$$

Proof. By triangular inequality,

$$\begin{aligned}
&\|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi) - \nabla^2 \hat{f}(\hat{B}^{-\top}\phi)\|_2 \\
&\quad \leq \|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi) - \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi)\|_2 + \|\nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi) - \nabla^2 \widehat{f}(\hat{B}^{-\top}\phi)\|_2 \\
&\quad \leq \|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi) - \nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi)\|_2 + \|\nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi) - \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi)\|_2 \\
&\qquad + \|\nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi) - \nabla^2 \widehat{f}(\hat{B}^{-\top}\phi)\|_2
\end{aligned}$$

Part 1: To bound $\|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi) - \nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi)\|_2$, note that on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^{\tilde{R}}$,

$$\begin{aligned}
&\|\nabla^2 f_{A\mu}(B^{-\top}{R^*}^\top \phi_i) - \nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi_i)\|_2 \\
&\quad = \|AD_\psi^{-1}\Lambda_i A^\top - AD_\psi^{-1}\hat{\Lambda}_i A^\top\|_2 \leq \|A\|_2^2 \|D_\psi^{-1}\|_2 \|\Lambda_i - \hat{\Lambda}_i\|_2,
\end{aligned}$$

where $\hat{\Lambda}_i = \operatorname{diag}\left((\phi_i^\top \hat{B}^{-1}BR_1)^2, \cdots, (\phi_i^\top \hat{B}^{-1}BR_d)^2\right)$. Note that $\Lambda_i - \hat{\Lambda}_i$ is diagonal, and the absolute value of its diagonal element $|(\phi_i \hat{B}^{-1}BR_j)^2 - (\phi_i R^* R_j)^2|$ for $1 \leq j \leq d$ can be bounded as follows.

$$\begin{aligned}
|(\phi_i \hat{B}^{-1}BR_j)^2 - (\phi_i R^* R_j)^2| &\leq |(\phi_i \hat{B}^{-1}BR_j) + (\phi_i R^* R_j)||(\phi_i \hat{B}^{-1}BR_j) - (\phi_i R^* R_j)| \\
&\leq 2\|\phi_i\|_2 \left(\|\hat{B}^{-1}B\|_2 + 1\right)\|\phi_i\|_2 \|\hat{B}^{-1}B - R^*\|_2 \\
&\leq 2L_{\text{high}}^2 \bar{\xi}/\sqrt{1-\bar{\xi}},
\end{aligned}$$

where the last inequality is due to Lemma A.4.4. Hence,

$$\|A\|_2^2\|D_\psi^{-1}\|_2\|\Lambda_i - \hat{\Lambda}_i\|_2 \leq 2\frac{\sigma_{\max}^2}{L_{\text{low}}^2 A_{(2,\min)}^2}\frac{L_{\text{high}}^2\bar{\xi}}{\sqrt{1-\bar{\xi}}}.$$

Thus,

$$\|\nabla^2 f_{A\mu}(B^{-\top}R^{*\top}\phi) - \nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi)\|_2 \leq \frac{\sqrt{6}L_{\text{high}}^2\sigma_{\max}^2}{A_{(2,\min)}^2 L_{\text{low}}^2}\bar{\xi}. \tag{A.14}$$

Part 2: To bound $\|\nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi) - \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi)\|_2$, note that

$$\nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi) = G_1(A,\hat{B}^{-1}A,\phi) - G_2(A,\hat{B}^{-1}A,\phi) - 2G_3(A,\hat{B}^{-1}A,\phi),$$

and

$$\nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi) = \widehat{G_1}(A,\hat{B}^{-1}A,\phi) - \widehat{G_2}(A,\hat{B}^{-1}A,\phi) - 2\widehat{G_3}(A,\hat{B}^{-1}A,\phi).$$

Also,

$$\begin{aligned}\|\hat{B}^{-1}A\|_2 &\leq \|\hat{B}^{-1}B\|_2\|B^{-1}A\|_2 = \|\hat{B}^{-1}B\|_2\|R^\top K^{-1/2}D_\psi^{-1/2}\|_2\\ &\leq \frac{1+\bar{\xi}}{\kappa_{\min}^{1/2}A_{(2,\min)}L_{\text{low}}} \leq \frac{4}{3\kappa_{\min}^{1/2}A_{(2,\min)}L_{\text{low}}}.\end{aligned}$$

Thus, by Lemma A.3.1,

$$\|\nabla^2 f_{A\mu}(\hat{B}^{-\top}\phi) - \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi)\|_2 \leq \frac{16L_{\text{high}}^2 d^5 A_{\max}^2}{9\kappa_{\min}A_{(2,\min)}^2 L_{\text{low}}^2}\xi. \tag{A.15}$$

Part 3: It remains to bound the last term. Again note that on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^{\tilde{R}}$, $\|\hat{B}^{-\top}\phi\|_2 \leq \frac{4L_{\text{high}}}{3\kappa_{\min}^{1/2}A_{(2,\min)}\sigma_{\min}L_{\text{low}}}$. Thus by Proposition 2.3.2,

$$\|\nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi) - \nabla^2\hat{f}(\hat{B}^{-\top}\phi)\|_2 \leq P\left(\frac{4L_{\text{high}}}{3\kappa_{\min}^{1/2}A_{(2,\min)}\sigma_{\min}L_{\text{low}}}\right). \tag{A.16}$$

Therefore, combining Eqs. (A.14) to (A.16),

$$\begin{aligned}&\|\nabla^2 f_\mu(B^{-\top}R^{*\top}\phi) - \nabla^2\hat{f}(\hat{B}^{-\top}\phi)\|_2\\ &\quad\leq \frac{\sqrt{6}L_{\text{high}}^2\sigma_{\max}^2}{A_{(2,\min)}^2 L_{\text{low}}^2}\bar{\xi} + \frac{16L_{\text{high}}^2 d^5 A_{\max}^2}{9\kappa_{\min}A_{(2,\min)}^2 L_{\text{low}}^2}\xi + P\left(\frac{4L_{\text{high}}}{3\kappa_{\min}^{1/2}A_{(2,\min)}\sigma_{\min}L_{\text{low}}}\right) = \hat{\xi}.\end{aligned}$$

□

Lemma A.4.6. *Assuming that Condition (4), (5), and (6) hold, on the event* $\mathcal{E}_\psi^A \cap \mathcal{E}_{\phi_1}^{\tilde{R}} \cap \mathcal{E}_{\phi_2}^{\tilde{R}}$,

$$\|M - \hat{M}\|_2 \leq \frac{4L_{high}^4\kappa_{\max}^{1/2}A_{(2,\max)}^4}{L_{low}^5\kappa_{\min}A_{(2,\min)}^5}\hat{\xi} = \tilde{Q}.$$

Proof. Recall $T_i = AD_\psi^{-1}\Lambda_i A^\top$ for $i \in \{1,2\}$. On the event $\mathcal{E}_\psi^A \cap \mathcal{E}_{\phi_1}^{\tilde{R}} \cap \mathcal{E}_{\phi_2}^{\tilde{R}}$,

$$\sigma_{\min}(T_2) \geq \frac{\sigma_{\min}^2 L_{\text{low}}^2}{L_u^2 A_{(2,\max)}^2}; \quad \sigma_{\max}(T_2) \leq \frac{L_{\text{high}}^2\sigma_{\max}^2}{\ell_l^2 A_{(2,\min)}^2}; \quad \sigma_{\max}(T_1) \leq \frac{L_{\text{high}}^2\sigma_{\max}^2}{\ell_l^2 A_{(2,\min)}^2}.$$

Let $E_i = \nabla^2 f_{A\mu}(B^{-\top} R^{*\top} \phi_i) - \nabla^2 \hat{f}(\hat{B}^{-\top} \phi_i)$ for $i \in \{1, 2\}$. By Condition 6, $\sigma_{\min}(T_2) \geq 2\|E_2\|_2$. Then by Lemma A.4.5 and Lemma A.3.4,

$$\begin{aligned}\|M - \hat{M}\|_2 &\leq \frac{2\|T_1\|_2}{\sigma^2_{\min}(T_2)}\|E_2\|_2 + \frac{2}{\sigma_{\min}(T_2)}\|E_1\|_2 \\ &\leq \frac{4L^6_{\text{high}}\sigma^2_{\max}A^4_{(2,\max)}}{L^6_{\text{low}}\sigma^4_{\min}A^2_{(2,\min)}}\hat{\xi} = \tilde{Q}.\end{aligned}$$

. □

A.5 Proof of Theorem 2.4.1

Let

$$\begin{aligned}\xi_{\text{recur}} &= \frac{4P\left(\frac{\sqrt{4}L_{\text{high}}}{\sqrt{3}\kappa^{1/2}_{\min}A_{(2,\min)}\sigma_{min}L_{\text{low}}}\right)}{3\kappa_{\min}A^2_{(2,\min)}\sigma^2_{min}\ell^2_l} + \frac{L^2_{\text{high}}d^5}{\kappa^2_{\min}A^4_{(2,\min)}\ell^4_l}\xi \\ &\quad + 3\bar{\xi}\left(\frac{L^2_{\text{high}}d^5}{\kappa^2_{\min}A^4_{(2,\min)}\ell^4_l}\xi + \frac{L^2_{\text{high}}}{A^4_{(2,\min)}L^4_{\text{low}}}\right); \\ Q_{\text{recur}} &= \frac{L^{10}_{\text{high}}A^8_{(2,\max)}}{A^4_{(2,\min)}L^8_{\text{low}}}\xi_{\text{recur}}.\end{aligned}$$

In Section A.5.1, we will prove that

$$\|M - \hat{M}\|_2 \leq Q_{\text{recur}}. \tag{*}$$

We now follow the idea of Vempala and Xiao [2014] to analyze the error accumulation of the recursion. Recall that $M = \bar{R}\Lambda\bar{R}^\top$, where $\bar{R} = R^*R$ is an orthonormal matrix. Assume we have computed a m-dimensional subspace in a recursion of depth $k-1$ whose orthonormal projection matrix is $V^{(k-1)} \in \mathbb{R}^{d\times m}$, such that there exists m columns of $\bar{R}$ (WLOG assume these are $1, \ldots, m$.) satisfying

$$\sin\left(\Theta\left(V^{(k-1)}, \bar{R}_{1:m}\right)\right) \leq E_{k-1},$$

where $\bar{R}_{1:m}$ is the first m columns of $\bar{R}$ and E_{k-1} is a error upper bound for depth $k-1$ recursion. Then,

$$\begin{aligned}&{V^{(k-1)}}^\top M V^{(k-1)} = \left({V^{(k-1)}}^\top \bar{R}_{1:m},\ {V^{(k-1)}}^\top \bar{R}_{m+1:d}\right)\Lambda\begin{pmatrix}\bar{R}^\top_{1:m}V^{(k-1)} \\ \bar{R}^\top_{m+1:d}V^{(k-1)}\end{pmatrix} \\ &= {V^{(k-1)}}^\top \bar{R}_{1:m}\Lambda_{1:m}\bar{R}^\top_{1:m}V^{(k-1)} + {V^{(k-1)}}^\top \bar{R}_{m+1:d}\Lambda_{m+1:d}\bar{R}^\top_{m+1:d}V^{(k-1)},\end{aligned}$$

where $\Lambda_{1:m}$ and $\Lambda_{m+1:d}$ are the first $m \times m$ and last $(d-m)\times(d-m)$ submatrices of the diagonal matrix Λ.

Recall that the diagonal elements of Λ are squares of Cauchy random variables. The following proposition lower bounds the maximal spacing of i.i.d. Cauchy random variables, whose proof is deferred to Section A.6:

Proposition A.5.1. *Assuming that $Z_1, \ldots, Z_d$ are i.i.d. Cauchy random variables, then with probability at least $1 - 2\delta$,*

$$\max_i \min_{j \neq i} |Z_i^2 - Z_j^2| \geq \frac{\delta}{2} \left(\frac{\delta}{d} \right)^{1/(d-1)}.$$

Also, with probability at least $1 - \delta$,

$$\max_i |Z_i| \leq \frac{3(d+1)}{\pi \delta}.$$

Denote the event of Proposition A.5.1 as $\mathcal{E}_Z$. Therefore, by Proposition A.5.1, with probability at least $1 - 3\delta$, the error of ${V^{(k-1)}}^\top \bar{R}_{1:m} \Lambda_{1:m} \bar{R}_{1:m}^\top V^{(k-1)}$ is

$$\begin{aligned}
&\|{V^{(k-1)}}^\top \hat{M} V^{(k-1)} - {V^{(k-1)}}^\top \bar{R}_{1:m} \Lambda_{1:m} \bar{R}_{1:m}^\top V^{(k-1)}\|_2 \\
&\leq \|{V^{(k-1)}}^\top \hat{M} V^{(k-1)} - {V^{(k-1)}}^\top M V^{(k-1)}\|_2 \\
&\quad + \|{V^{(k-1)}}^\top M V^{(k-1)} - {V^{(k-1)}}^\top \bar{R}_{1:m} \Lambda_{1:m} \bar{R}_{1:m}^\top V^{(k-1)}\|_2 \\
&= \|{V^{(k-1)}}^\top \hat{M} V^{(k-1)} - {V^{(k-1)}}^\top M V^{(k-1)}\|_2 + \|{V^{(k-1)}}^\top \bar{R}_{m+1:d} \Lambda_{m+1:d} \bar{R}_{m+1:d}^\top V^{(k-1)}\|_2 \\
&\leq Q_{\text{recur}} + E_{k-1}^2 \frac{9(d+1)^2}{\pi^2 \delta^2}.
\end{aligned}$$

Now, by Proposition A.5.1 with probability at least $1 - 2\delta$ the maximal spacing of the diagonal elements of $\Lambda_{1:m}$ satisfy

$$\max_i \min_{j \neq i} |Z_i^2 - Z_j^2| \geq \frac{\delta}{2} \left(\frac{\delta}{d} \right)^{1/(d-1)} \geq \frac{\delta^2}{2d}.$$

Now consider the error after the eigen-decomposition of ${V^{(k-1)}}^\top \hat{M} V^{(k-1)}$. Assume the maximal gap in $\Lambda_{1:m}$ divides $\Lambda_{1:m}$ into $\Lambda_{1:t}$ and $\Lambda_{t+1:m}$. Let $E = {V^{(k-1)}}^\top \hat{M} V^{(k-1)} - {V^{(k-1)}}^\top \bar{R}_{1:m} \Lambda_{1:m} \bar{R}_{1:m}^\top V^{(k-1)}$, and

$$(\bar{R}_{1:t}^\top V^{(k-1)}, \bar{R}_{t+1:m}^\top V^{(k-1)})^\top E ({V^{(k-1)}}^\top \bar{R}_{1:t}, {V^{(k-1)}}^\top \bar{R}_{t+1:m}) = \begin{pmatrix} F_1 & F_2 \\ F_3 & F_4 \end{pmatrix}.$$

It is easy to see that if $\frac{\delta^2}{2d} \geq 4\|E\|_2$, then the conditions of Theorem 2.8 in Chapter V, [Stewart and Sun, 1990] are satisfied. Thus,

$$E_k \leq \frac{4}{\frac{\delta^2}{2d}} \left(Q_{\text{recur}} + E_{k-1}^2 \frac{9(d+1)^2}{\pi^2 \delta^2} \right) = \frac{8d Q_{\text{recur}}}{\delta^2} + \frac{72d(d+1)^2}{\pi^2 \delta^4} E_{k-1}^2.$$

Vempala and Xiao [2014] gave the following proposition.

Proposition A.5.2 ([Vempala and Xiao, 2014], V.1)**.** *Fix $a, b > 0$ where $4a/b^2 \leq 1$, and define the recurrence $y_{i+1} = a + (y_i/b)^2$ and $y_i = 0$. Then $y_i \leq 2a$ for all i.*

Using this proposition on $\{E_k\}$, given that $Q_{\text{recur}} \leq \frac{\pi^2 \delta^6}{576 d^2 (d+1)^2}$, for $0 \leq k \leq d$, we have

$$E_k \leq \frac{16d}{\delta^2} Q_{\text{recur}}.$$

Thus Line 6 in Algorithm 4 returns a matrix $\hat{R}$ s.t.

$$\|\hat{R} - R^* R\|_2 = 2 - 2\cos(\Theta) = 2 - 2\sqrt{1 - \sin^2(\Theta)} \leq 2 - 2\sqrt{1 - \frac{256d^2}{\delta^4} Q_{\text{recur}}^2}.$$

Therefore,

$$\begin{aligned}
\|\hat{B}\hat{R} - AD_\phi^{1/2} K^{1/2}\|_2 &= \|\hat{B}\hat{R} - AD_\phi^{1/2} K^{1/2} R^\top {R^*}^\top R^* R\|_2 \\
&\leq \|\hat{B}\hat{R} - \hat{B}R^* R\|_2 + \|\hat{B}R^* R - AD_\phi^{1/2} K^{1/2} R^\top {R^*}^\top R^* R\|_2 \\
&\leq \|B\|_2 \|B^{-1}\hat{B}\|_2 \|\hat{R} - R^* R\|_2 + \|\hat{B} - AD_\phi^{1/2} K^{1/2} R^\top {R^*}^\top\|_2 \\
&\leq \|B\|_2 \|B^{-1}\hat{B}\|_2 \|\hat{R} - R^* R\|_2 + \|B\|_2 \|B^{-1}\hat{B} - {R^*}^\top\|_2.
\end{aligned}$$

By Lemma A.6.1, $\|B\|_2 \leq \sigma_{\max} L_{\text{high}} A_{(2,\max)} \kappa_{\max}^{1/2}$, and by Lemma A.4.4, $\|B^{-1}\hat{B}\|_2 \leq (1+\bar{\xi})^{1/2}$. Also, given that $\bar{\xi} \leq 1/2$, by Lemma A.3.3, $\|B^{-1}\hat{B} - {R^*}^\top\|_2 \leq 2\bar{\xi}$. Adding all the terms together,

$$\begin{aligned}
&\|\hat{B}\hat{R} - AD_\phi^{1/2} K^{1/2}\|_2 \\
&\quad \leq \sigma_{\max} L_{\text{high}} A_{(2,\max)} \kappa_{\max}^{1/2} \left((1+\bar{\xi})^{1/2} \left(2 - 2\sqrt{1 - \frac{256d^2}{\delta^4} Q_{\text{recur}}^2} \right) + 2\bar{\xi} \right).
\end{aligned}$$

Note that for a small enough ϵ, $\sqrt{1-\epsilon} \geq 1 - \frac{2\epsilon}{3}$. Thus, if Q_{recur} is small enough,

$$\begin{aligned}
\|\hat{B}\hat{R} - AD_\phi^{1/2} K^{1/2}\|_2 \leq& \sigma_{\max} L_{\text{high}} A_{(2,\max)} \kappa_{\max}^{1/2} \left((1+\bar{\xi})^{1/2} \frac{512d^2}{3\delta^4} Q_{\text{recur}}^2 + 2\bar{\xi} \right) \\
\leq& \mathcal{C}(\mu) \left(K^2(\mu) + K(\mu) \right),
\end{aligned}$$

where $K(\mu) = D_4(\nu^{(s)}, \mu) + \mathcal{Q}(\nu^{(\epsilon)}) + D_4^{(d,d)}(\nu^{(As,\epsilon)}) + N(\nu^{(\epsilon)}) + N(\nu^{(s)})$.

It remains to bound the probability that the above result holds. We need event $\mathcal{E}_\psi^A$ to be satisfied once, and $\mathcal{E}_\phi^{\tilde{R}}$ and $\mathcal{E}_Z$ for at most d times (we ignore the $\log d$ factor here). Thus, given the conditions of Theorem 2.2.1, with probability $1 - 7d\delta$, the above error bound holds.

A.5.1 Technical lemmas

This section is to prove that in Algorithm 4,

$$\|M_P - \hat{M}_P\|_2 \leq Q_{\text{recur}},$$

by proving that $\|M - \hat{M}\|_2 \leq Q_{\text{recur}}$, and thus $\|M_P - \hat{M}_P\|_2 \leq \|M - \hat{M}\|_2 \leq Q_{\text{recur}}$.

Let $\tilde{A} = \hat{B}^{-1}A$. Thus, $\hat{y} = \hat{B}^{-1}x = \tilde{A}s + \hat{B}^{-1}\epsilon$. We start with bounding $\|\nabla^2 f_{\nu^{(\hat{y})}}(\phi) - \nabla^2 f_{B^{-1}A\mu}(\phi)\|_2$. Then applying Lemma A.3.4 will lead to a bound on $\|M - \hat{M}\|_2$. Note that

$$\begin{aligned}
\|\nabla^2 f_{\nu^{(\hat{y})}}(\phi) - \nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2 \leq& \|\nabla^2 f_{\nu^{(\hat{x})}}(\phi) - \nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi)\|_2 \\
&+ \|\nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi) - \nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2 + \|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi) - \nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2
\end{aligned}$$

Part 1: Bounding $\|\nabla^2 f_{\nu^{(\tilde{x})}}(\phi) - \nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi)\|_2$.

Note that for any unit vector v, by some calculation, we have

$$v^\top \nabla^2 f_{\nu^{(\tilde{x})}}(\phi) v = (v^\top \hat{B}^{-1}) \nabla^2 f_{\nu^{(x)}}(\hat{B}^{-\top}\phi) \hat{B}^{-\top} v,$$

and

$$v^\top \nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi) v = (v^\top \hat{B}^{-1}) \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi) \hat{B}^{-\top} v.$$

Thus

$$\begin{aligned}
\|\nabla^2 f_{\nu^{(\tilde{x})}}(\phi) - \nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi)\|_2 &\le \|\hat{B}^{-1}\|_2^2 \|\nabla^2 f_{\nu^{(x)}}(\hat{B}^{-\top}\phi) - \nabla^2 f_{\nu^{(As)}}(\hat{B}^{-\top}\phi)\|_2 \\
&\le \|\hat{B}^{-1}\|_2^2 P\left(\|\hat{B}^{-\top}\phi\|_2\right) \le \frac{4}{3\kappa_{\min} A_{(2,\min)}^2 \sigma_{min}^2 \ell_l^2} P\left(\frac{\sqrt{4}\|\phi\|_2}{\sqrt{3}\kappa_{\min}^{1/2} A_{(2,\min)} \sigma_{min} L_{\text{low}}}\right),
\end{aligned} \tag{A.17}$$

where the second inequality is by Proposition 2.3.2, and the last inequality is due to that on the event $\mathcal{E}_\psi^A$, $\|\hat{B}^{-1}\|_2 \le \|\hat{B}^{-1}B\|_2 \|B^{-1}\|_2 \le \sqrt{4}/\left(\sqrt{3}\kappa_{\min}^{1/2} A_{(2,\min)} \sigma_{min} L_{\text{low}}\right)$, and $\|\hat{B}^{-\top}\phi\|_2 \le \|\hat{B}^{-\top}\|_2 \|\phi\|_2$.

Part 2: to bound $\|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi) - \nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2$.

Note that $R^* B^{-1} A = R^* R^\top K^{-1/2} D_\psi^{-1/2}$, thus $\sigma_{\max}(R^* B^{-1} A) \le 1/\left(\kappa_{\min}^{1/2} A_{(2,\min)} L_{\text{low}}\right)$ on the event $\mathcal{E}_\psi^A$. Hence, by Lemma A.3.2,

$$\|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi) - \nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2 \le \frac{\|\phi\|_2^2 d^5}{\kappa_{\min}^2 A_{(2,\min)}^4 \ell_l^4} \xi. \tag{A.18}$$

Part 3: to bound $\|\nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi) - \nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2$.

Again for any unit vector v, $v^\top \nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi) v$ and $v^\top \nabla^2 f_{\nu^{(B^{-1}As)}}(\phi) v$ are both achieved by marginalizing 4th order tensors. Further, note that $\|\hat{B}^{-1}B\|_2 \le \sqrt{1+\bar{\xi}}$. Thus

$$\begin{aligned}
\|\nabla^2 f_{\nu^{(\tilde{A}s)}}(\phi) - \nabla^2 f_{\nu^{(B^{-1}As)}}(\phi)\|_2 &= \|\nabla^2 f_{\nu^{(\hat{B}^{-1}BB^{-1}As)}}(\phi) - \nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2 \\
&\le \left((1+\bar{\xi})^2 - 1\right) \|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2 \\
&\le 3\bar{\xi} \|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2.
\end{aligned}$$

Lastly, note that

$$\nabla^2 f_{R^* B^{-1} A\mu}(\phi) = R^* R^\top D_\psi^{-2} \Lambda_\phi R R^{*\top},$$

where $\Lambda_\phi = \text{diag}\left((\phi^\top R^* R_1^\top)^2, \cdots, (\phi^\top R^* R_d^\top)^2\right)$. Thus on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$

$$\|\nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2 = \|R^* R^\top D_\psi^{-2} \Lambda_\phi R R^{*\top}\|_2 \le \frac{L_{\text{high}}^2}{A_{(2,\min)}^4 L_{\text{low}}^4}.$$

and

$$\begin{aligned}
\|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi)\|_2 &= \|\nabla^2 f_{\nu^{(R^* B^{-1} As)}}(\phi) - \nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2 + \|\nabla^2 f_{R^* B^{-1} A\mu}(\phi)\|_2 \\
&\le \frac{L_{\text{high}}^2 d^5}{\kappa_{\min}^2 A_{(2,\min)}^4 \ell_l^4} \xi + \frac{L_{\text{high}}^2}{A_{(2,\min)}^4 L_{\text{low}}^4}.
\end{aligned}$$

Therefore,

$$\begin{aligned}&\|\nabla^2 f_{\nu(\bar{A}s)}(\phi) - \nabla^2 f_{\nu(B^{-1}As)}(\phi)\|_2 \\ &\quad \le 3\bar{\xi}\|\nabla^2 f_{\nu(R^*B^{-1}As)}(\phi)\|_2 \le 3\bar{\xi}\left(\frac{L_{\text{high}}^2 d^5}{\kappa_{\min}^2 A_{(2,\min)}^4 \ell_l^4}\xi + \frac{L_{\text{high}}^2}{A_{(2,\min)}^4 L_{\text{low}}^4}\right). \end{aligned} \tag{A.19}$$

Combining Eqs. (A.17) to (A.19), we have that on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_\phi^A$,

$$\begin{aligned}&\|\nabla^2 f_{\nu^{(\bar{y})}}(\phi) - \nabla^2 f_{R^*B^{-1}A\mu}(\phi)\|_2 \\ &\quad \le \frac{4}{3\kappa_{\min}A_{(2,\min)}^2\sigma_{min}^2\ell_l^2} P\left(\frac{\sqrt{4}L_{\text{high}}}{\sqrt{3}\kappa_{\min}^{1/2}A_{(2,\min)}\sigma_{min}L_{\text{low}}}\right) + \frac{L_{\text{high}}^2 d^5}{\kappa_{\min}^2 A_{(2,\min)}^4 \ell_l^4}\xi \\ &\qquad + 3\bar{\xi}\left(\frac{L_{\text{high}}^2 d^5}{\kappa_{\min}^2 A_{(2,\min)}^4 \ell_l^4}\xi + \frac{L_{\text{high}}^2}{A_{(2,\min)}^4 L_{\text{low}}^4}\right) = \tilde{\xi}.\end{aligned}$$

Finally, on the event $\mathcal{E}_\psi^A \cap \mathcal{E}_{\phi_1}^{\tilde{R}} \cap \mathcal{E}_{\phi_2}^{\tilde{R}}$, for $i \in \{1,2\}$,

$$\sigma_{\max}\left(\nabla^2 f_{R^*B^{-1}A\mu}(\phi_i)\right) \le \frac{L_{\text{high}}^2}{A_{(2,\min)}^4 L_{\text{low}}^4}; \quad \sigma_{\min}\left(\nabla^2 f_{R^*B^{-1}A\mu}(\phi_i)\right) \ge \frac{L_{\text{low}}^2}{A_{(2,\max)}^4 L_{\text{high}}^4}.$$

Thus, by Lemma A.3.4,

$$\|M - \hat{M}\|_2 \le \frac{4L_{\text{high}}^{10}A_{(2,\max)}^8}{A_{(2,\min)}^4 L_{\text{low}}^8}\xi_{\text{recur}} = Q_{\text{recur}}.$$

A.6 Proofs of Proposition 2.4.2, Lemma A.4.1, and Proposition A.5.1

Proposition 2.4.2. *With probability at least* $1-\delta$,

$$\gamma_R \ge \frac{\pi^2\delta^2}{d^3}.$$

Proof of Proposition 2.4.2. Recall that Z_i's are independent Cauchy random variables. With no loss of generality assume both Z_i and Z_j are positive. Then,

$$\begin{aligned}\gamma_R =& \min_{i\neq j}\left|Z_i^2 - Z_j^2\right| \\ =& \min_{i\neq j}|Z_i - Z_i|\,|Z_i + Z_i| \\ \ge& 2\min_i|Z_i|\min_{i\neq j}|Z_i - Z_j|.\end{aligned}$$

We will first bound $\min_i|Z_i|$. Recall that for $t \in [0,\infty]$,

$$\mathbb{P}(|Z| \ge t) = 1 - 2\int_0^t \frac{1}{\pi(1+x^2)}dx \ge 1 - \frac{2t}{\pi}.$$

Thus,

$$\mathbb{P}\left(\min_i|Z_i| \ge t\right) \ge 1 - \frac{2dt}{\pi}.$$

Picking $t = \frac{\pi\delta}{2d}$,

$$\mathbb{P}\left(\min_i|Z_i| \ge t\right) \ge 1 - \delta.$$

To bound $\min_{i\neq j}|Z_i - Z_j|$, note that for $k \neq j$, $Z_k - Z_j \sim$ Cauchy$(0, 2)$. Thus

$$\mathbb{P}\left(|Z_k - Z_j| \leq \ell\right) = \mathbb{P}\left(|Z_k - Z_j|/2 \leq \ell/2\right) = 2\int_0^{\ell/2} \frac{1}{\pi(1+x^2)} dx \leq \frac{\ell}{\pi}.$$

Therefore,

$$\mathbb{P}\left(\min_{i\neq j}|Z_i - Z_j| \geq \ell\right) \geq 1 - \frac{d(d-1)}{2}\frac{\ell}{\pi}. \tag{A.20}$$

Picking $\ell = \frac{2\pi\delta}{d^2}$, we have

$$\mathbb{P}\left(\min_{i\neq j}|Z_i - Z_j| \geq \ell\right) \geq 1 - \delta.$$

Therefore, with probability at least $1 - 2\delta$,

$$\gamma_R \geq \frac{\pi^2\delta^2}{d^3}.$$

□

Lemma A.6.1. *With probability at least* $\left(1 - \frac{\ell}{d}\right)\exp(-\ell) - \exp(-x)$, $\mathcal{E}_\psi^A$ *holds, where* $L_{high} = \sqrt{2x} + \sqrt{2d}$.

Proof of Lemma A.6.1. For a fixed constant $L_{\text{low}} \leq A_{(2,\max)}$, note that $\{\psi : \min_i\{|\psi^\top A_i|\} \geq A_{(2,\min)}L_{\text{low}}\}$ is equivalent to $\cap_i G_i$, where G_i is the set defined as $\{x : x^\top A_i \geq A_{(2,\min)}L_{\text{low}}\}$. Let $V_i = A_i/\|A_i\|_2$, then $G_i \supset G_i' = \{x : x^\top V_i \geq L_{\text{low}}\}$ holds for any i.

Now we consider $\mathbb{P}\left(\cap_i G_i'\right)$. This probability is minimized when V_is are orthogonal to each other. Thus, for any orthonormal matrix R, define $G_i'' = \{x : x^\top R_i \geq L_{\text{low}}\}$. Then

$$\mathbb{P}\left(\cap_i G_i'\right) \geq \mathbb{P}\left(\cap_i G_i''\right) = \mathbb{P}\left(|\psi^\top R| \geq L_{\text{low}}\right) = \mathbb{P}\left(|\psi| \geq L_{\text{low}}\right).$$

Note that $\mathbb{P}\left(|X| \geq L_{\text{low}}\right) \geq 1 - \frac{\sqrt{2}L_{\text{low}}}{\sqrt{\pi}}$ for $X \sim N(0,1)$. Thus, picking $L_{\text{low}} = \frac{\sqrt{\pi}}{\sqrt{2}d}\ell$ for any $0 \leq \ell \leq 1$,

$$\mathbb{P}\left(|\psi| \geq L_{\text{low}}\right) \geq (1 - \frac{\sqrt{2}L_{\text{low}}}{\sqrt{\pi}})^d = (1 - \frac{\ell}{d})^d = \left(1 - \frac{\ell}{d}\right)\left(1 - \frac{\ell}{d}\right)^{d-1} \geq \left(1 - \frac{\ell}{d}\right)\exp(-\ell). \tag{A.21}$$

On the other hand, note that $\mathbb{P}\left(\|\psi\|_2 \leq L_{\text{high}}\right) = \mathbb{P}\left(X \leq L_{\text{high}}^2\right)$ where $X \sim \chi_d$. Thus, by Lemma 1 of Laurent and Massart [2000], picking

$$x = \left(\frac{\sqrt{2}}{2}L_{\text{high}} - \sqrt{d}\right)^2,$$

then $L_{\text{high}}^2 \geq d + 2\sqrt{dx} + 2x$, and

$$\mathbb{P}\left(\|\psi\|_2 \leq L_{\text{high}}\right) = 1 - \mathbb{P}\left(X \geq L_{\text{high}}^2\right) \geq 1 - \mathbb{P}\left(X - d \geq 2\sqrt{dx} + 2x\right) \geq 1 - \exp(-x).$$

(A.22)

Therefore,

$$\mathbb{P}\left(\mathcal{E}_\psi^A\right) \geq (1 - \exp(-x)) + \left(1 - \frac{\ell}{d}\right)\exp(-\ell) - 1 = \left(1 - \frac{\ell}{d}\right)\exp(-\ell) - \exp(-x). \tag{A.23}$$

□

Lemma A.4.1. *For any A and orthonormal matrix R, with probability at least $1-\delta$, the following inequalities holds simultaneously:*

- $\min_i |\psi^\top A_i| \geq \frac{\sqrt{\pi} A_{(2,\min)}}{5\sqrt{2}(d+1)}\delta$;
- $\min_i\{|\phi_2^\top R_i|\} \geq \frac{\sqrt{\pi}}{5\sqrt{2}(d+1)}\delta$;
- $\|\phi_1\|_2, \|\phi_2\|_2 \leq \sqrt{2}\left(\sqrt{\log(\frac{5}{\delta})} + \sqrt{d}\right)$;
- $\gamma_R \geq \frac{\pi^2\delta^2}{25d^3}$.

Proof of Lemma A.4.1. Denote the event defined by the inequalities by $\mathcal{E}$. Note that for $0 \leq \ell \leq 1$,

$$\left(1 - \frac{\ell}{d}\right)\exp(-\ell) \geq 1 - \frac{d+1}{d}\ell.$$

Combining Proposition 2.4.2 and Lemma A.6.1, given that $0 \leq \ell \leq 1$, we get

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}\left(\mathcal{E}_\psi^A\right) + \mathbb{P}(\mathcal{E}_Z) + \mathbb{P}\left(\mathcal{E}_\phi^R\right) - 2 \geq 1 - \frac{\delta}{5} - 2\frac{d+1}{d}\ell - 2\exp(-x)$$

Now choose $\ell = \frac{d}{d+1}\frac{\delta}{5}$ and $x = \log(\frac{5}{\delta})$. Then, with probability $1-\delta$,

- $\min_i |\psi^\top A_i| \geq \frac{\sqrt{\pi} A_{(2,\min)}}{5\sqrt{2}(d+1)}\delta$;
- $\min_i\{|\phi_2^\top R_i|\} \geq \frac{\sqrt{\pi}}{5\sqrt{2}(d+1)}\delta$;
- $\|\phi_1\|_2, \|\phi_2\|_2 \leq \sqrt{2}\left(\sqrt{\log(\frac{5}{\delta})} + \sqrt{d}\right)$;
- $\gamma_R \geq \frac{\pi^2\delta^2}{25d^3}$.

□

Proposition A.5.1. *Assuming that $Z_1, \ldots, Z_d$ are i.i.d. Cauchy random variables, then with probability at least $1-2\delta$,*

$$\max_i \min_{j\neq i} |Z_i^2 - Z_j^2| \geq \frac{\delta}{2}\left(\frac{\delta}{d}\right)^{1/(d-1)}.$$

Also, with probability at least $1-\delta$,

$$\max_i |Z_i| \leq \frac{3(d+1)}{\pi\delta}.$$

Proof of Proposition A.5.1. Let $Z_{(1)} \leq \ldots \leq Z_{(d)}$ denote the order statistics of Z_i. WLOG, we fold the negative part of the Cauchy distribution to its positive part, leading to a density function $p_Z(z) = \frac{2}{\pi(1+z^2)}$ for $0 \leq z$. Note that

$$\max_i \min_{i\neq j} |Z_i^2 - Z_j^2| \geq |Z_{(d)}^2 - Z_{(d-1)}^2| \geq 2Z_{(d-1)}\left(Z_{(d)} - Z_{(d-1)}\right).$$

We will bound both terms on the RHS.

To bound, $Z_{(d-1)}$, recall that for a folded Cauchy random variable Z,

$$\mathbb{P}(Z \leq L) \leq \frac{2L}{\pi}.$$

Therefore,

$$\mathbb{P}\left(Z_{(d-1)} \geq L\right) \geq 1 - d\mathbb{P}\left(Z \leq L\right)^{d-1} \geq 1 - d\left(\frac{2L}{\pi}\right)^{d-1}.$$

Picking $L = \frac{\pi}{2}\left(\frac{\delta}{d}\right)^{1/(d-1)}$, we have $\mathbb{P}\left(Z_{(d-1)} \geq L\right) \geq 1 - \delta$. Thus, with probability at least $1 - \delta$,

$$Z_{(d-1)} \geq \frac{\pi}{2}\left(\frac{\delta}{d}\right)^{1/(d-1)}.$$

On the other hand, let $U_{(1)}, \ldots, U_{(d)}$ denote the order statistics of d i.i.d random variables from the Uniform(0,1) distribution, and $E_{(1)}, \ldots, E_{(d)}$ denote the order statistics from the exponential distribution with density function $p_E(x) = \pi e^{-\pi x}$. We will bound the probability of $Z_{(d)} - Z_{(d-1)} \geq L$ by $E_{(d)} - E_{(d-1)} \geq L$.

First, by the Quantile Transformation Theorem [DasGupta, 2011], the joint distribution of $(F_{\mathrm{C}}(Z_{(1)}), \ldots, F_{\mathrm{C}}(Z_{(d)}))$ has the distribution of $(U_{(1)}, \ldots, U_{(d)})$. So is the distribution of $(F_{\mathrm{E}}(E_{(1)}), \ldots, F_{\mathrm{E}}(E_{(d)}))$. Here $F_{\mathrm{C}}(\cdot)$ and $F_{\mathrm{E}}(\cdot)$ are the c.d.f. of the folded Cauchy distribution and the exponential distribution with parameter π. Recall that $F_{\mathrm{C}}(t) = \frac{2}{\pi}\arctan(t)$ and $F_{\mathrm{E}}(t) = 1 - e^{-\pi t}$. Therefore,

$$\mathbb{P}\left(Z_{(d)} - Z_{(d-1)} \geq t\right) = \mathbb{P}\left(\tan\left(\frac{\pi}{2}U_{(d)}\right) - \tan\left(\frac{\pi}{2}U_{(d-1)}\right) \geq t\right),$$

and

$$\mathbb{P}\left(E_{(d)} - E_{(d-1)} \geq t\right) = \mathbb{P}\left(\frac{1}{\pi}\left(\log\frac{1}{1-U_{(d)}} - \log\frac{1}{1-U_{(d-1)}}\right) \geq t\right).$$

To bound $\mathbb{P}\left(Z_{(d)} - Z_{(d-1)} \geq t\right)$ by $\mathbb{P}\left(E_{(d)} - E_{(d-1)} \geq t\right)$, it suffices to prove

$$\tan\left(\frac{\pi}{2}U_{(d)}\right) - \tan\left(\frac{\pi}{2}U_{(d-1)}\right) \geq \log\frac{1}{1-U_{(d)}} - \log\frac{1}{1-U_{(d-1)}}.$$

Let $f(x) = \tan\left(\frac{\pi}{2}x\right) - \tan\left(\frac{\pi}{2}U_{(d-1)}\right)$, and $g(x) = \log\frac{1}{1-x} - \log\frac{1}{1-U_{(d-1)}}$ for $1 \geq x \geq U_{(d-1)} \geq 0$. Clearly, $f(U_{(d-1)}) = g(U_{(d-1)})$. Taking the derivative of both functions, by simple algebra, $f'(x) \geq g'(x)$. Therefore,

$$\mathbb{P}\left(Z_{(d)} - Z_{(d-1)} \geq t\right) \geq \mathbb{P}\left(E_{(d)} - E_{(d-1)} \geq t\right) = e^{-\pi t} \geq 1 - \pi t.$$

Thus, with probability at least $1-\delta$, $Z_{(d)} - Z_{(d-1)} \geq \frac{\delta}{\pi}$. Therefore, with probability at least $1 - 2\delta$,

$$\max_i \min_{j \neq i} |Z_i^2 - Z_j^2| \geq \frac{\delta}{2}\left(\frac{\delta}{d}\right)^{1/(d-1)}.$$

Lastly, for a Cauchy random variable Z, $\mathbb{P}\left(|Z| \leq \frac{3L}{\pi}\right) = \frac{2}{\pi}\arctan(\frac{3L}{\pi})$. Note that for $L \geq \pi$,

$$\tan(\frac{\pi}{2} - \frac{\pi}{2L}) \leq \frac{1}{\cos(\frac{\pi}{2} - \frac{\pi}{2L})} \leq \frac{1}{\sin(\frac{\pi}{2L})} \leq \frac{1}{\frac{\pi}{2L} - \left(\frac{\pi}{2L}\right)^3} \leq \frac{2L}{\pi}\frac{1}{1 - \left(\frac{\pi}{2L}\right)^2} \leq \frac{3L}{\pi}.$$

Thus,

$$\mathbb{P}\left(|Z| \leq \frac{3L}{\pi}\right) \geq 1 - \frac{1}{L}.$$

Therefore, for $L \geq d$,

$$\left(1 - \frac{1}{L}\right)^d = \left(1 - \frac{d/L}{d}\right)^d \geq \left(1 - \frac{d/L}{d}\right) \exp(-d/L) \geq 1 - \frac{d+1}{d}\frac{d}{L} = 1 - \frac{d+1}{L}.$$

Picking $L = \frac{d+1}{\delta}$, the probability that $\max_i |Z_i| \leq \frac{3L}{\pi}$ is at least $1 - \delta$. □

Chapter 3

Generalized Partially Linear Model

In Chapter 2, we presented an instance-dependent analysis of the task of independent component analysis in the unsupervised learning setting. In this chapter we will present our first demonstration in the setting of supervised learning. In particular, we consider the performance of arguably the most popular supervised learning algorithm, empirical risk minimization (ERM), on the generalized partially linear model. We first investigate the behaviour of ERM in the purely parametric case. Unexpectedly, we detect a potential deficiency of ERM that in the worst case it suffers infinite expected error even on a simple least-square linear regression problem. We then further investigate the performance of ERM on this model, and develop an instance-dependent high probability finite-sample bound, which helps partially explain the success of ERM in practice. More importantly, our results reveal a potential gap between two different evaluation criteria: an 'in-sample' generalization bound in the fixed design that is extensively studied in the statistics literature may not imply an 'out-of-sample' bound in the random design.

One of the technical challenges in this chapter is that we allow the dependent variable Y to be unbounded, and thus standard concentration inequalities can not be simply applied. Our main tool will thus be a ratio-type concentration inequality due to van de Geer [2000].

This chapter is organized as follows: We introduce the problem setup and notations in Section 3.2. In Section 3.3 we first investigate the behaviour of ERM on a least-square linear regression problem. Section 3.4 is devoted to various assumptions required to develop a high probability finite-sample bound. We also discuss the generality of these assumptions. Before proving our main result, we first prove the boundedness of the predictor in Section 3.5. Our main result is a high probability bound for the generalization error in the random design setting, which is presented in Section 3.6. We discuss our observations in Section 3.7. Lastly, Section 3.8 concludes the chapter.

The results in this chapter have appeared in our AISTATS paper [Huang and Szepesvári,

2014a,b].

3.1 Introduction

We consider finite-time risk bounds for empirical risk-minimization algorithms (ERM) for *partially linear models* of the form

$$Y_i = \phi(X_i)^\top \theta + h(X_i) + \epsilon_i, \quad 1 \le i \le n, \tag{3.1}$$

where $X_i \in \mathbb{R}^d$ is an input, $Y_i \in \mathbb{R}$ is an observed, potentially unbounded response, ϵ_i is random noise, ϕ is the known basis function, θ is an unknown, finite dimensional parameter vector, and h is a nonparametric function component. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$, the learning problem is concerned with the case when the (X_i, Y_i) are sampled independently from some underlying distribution and the goal is to find model "θ" and "h" to use Eq. (3.1) to predict Y at X for arbitrary (X, Y) sampled from the same distribution.

The most well-known example of this type of model in machine learning is the case of Support Vector Machines (SVMs) with offset (in this case $\phi(x) \equiv 1$). The general partially linear stochastic model, which perhaps originates from the econometrics literature [e.g., Engle et al., 1986, Robinson, 1988, Stock, 1989, 1991], is a classic example of semiparametric models that combine parametric (in this case $\phi(\cdot)^\top \theta$) and nonparametric components (here h) into a single model. The appeal of semiparametric models has been widely discussed in statistics, machine learning, control theory and other branches of applied sciences [e.g., Bickel et al., 1998, Smola et al., 1998, Härdle et al., 2004, Gao, 2007, Kosorok, 2008, Greblicki and Pawlak, 2008, Horowitz, 2009]. In a nutshell, whereas a purely parametric model gives rise to the best accuracy if correct, it runs the risk of being misspecified. On the other hand, a purely nonparametric model avoids the risk of model misspecification, therefore achieving greater applicability and robustness, though at the price of the estimates perhaps converging at a slower rate. Semiparametric models, by combining parametric and nonparametric components into a single model, aim at achieving the best of both worlds. Another way of looking at them is that they allow to add prior "structural" knowledge to a nonparametric model, thus potentially significantly boosting the convergence rate when the prior is correct. For a convincing demonstration of the potential advantages of semiparametric models, see, e.g., the paper by Smola et al. [1998].

Despite all the interest in semiparametric modeling, to our surprise we were unable to find any work that would have been concerned with the finite-time *predictive performance* (i.e., risk) of semiparametric methods. Rather, existing theoretical works in semiparametrics are concerned with discovering conditions and algorithms for constructing statistically efficient estimators of the unknown parameters of the parametric part. This problem has been more or less settled in the book of Bickel et al. [1998], where sufficient and necessary

conditions are described along with recipes for constructing statistically efficient procedures. Although statistical efficiency (which roughly means achieving the Cramer-Rao lower bound as the sample size increases indefinitely) is of major interest, statistical efficiency does not give rise to finite-time bounds on the excess risk, the primary quantity of interest in machine learning. Here, we make the first initial steps to provide these missing bounds. Surprisingly a perfectly innocent-looking problem exists in the purely parametric case which makes ordinary least squares fail. This observation then suggests a question, as to what extent ERM would perform well as observed in practice. Our next result is an instance-dependent high probability finite-sample bound for ERM on the generalized partially linear model.

The closest to our work are the papers of Chen et al. [2004] and Steinwart [2005], who both considered the risk of SVMs with offset (a special case of our model). Here, as noted by these authors, the main difficulty is bounding the offset. While Chen et al. [2004] bounded the offset based on a property of the optimal solution for the hinge loss and derived finite-sample risk bounds, Steinwart [2005] considered consistency for a larger class of "convex regular losses". Specific properties of the loss functions were used to show high probability bounds on the offset. For our more general model, similarly to these works the bulk of the work will be to prove that with high probability the parametric model will stay bounded (we assume $\sup_x \|\phi(x)\|_2 < +\infty$). The difficulty is that the model is underdetermined and in the training procedures only the nonparametric component is penalized. This suggests that perhaps one could modify the training procedure to penalize the parametric component, as well. However, it appears that the semiparametric literature largely rejects this approach. The main argument is that a penalty would complicate the tuning of the method (because the strength of the penalty needs to be tuned, too), and that the parametric part is added based on a strong prior belief that the features added will have a significant role and thus rather than penalizing them, the goal is to encourage their inclusion in the model. Furthermore, the number of features in the parametric part are typically small, thus penalizing them is largely unnecessary. We will return to discussing this issue at the end of the chapter.

Finally, let us make some comments on the computational complexity of training partially linear models. When the nonparametric component belongs to an RKHS, an appropriate version of the representer theorem can be used to derive a finite-dimensional optimization problem [Smola et al., 1998], leading to quadratic optimization problem subject to linear constrains. Recent work by Kienzle and Schölkopf [2005] and Lee and Wright [2009] concern specialized solvers to find an approximate optimizer of the arising problem. In particular, in their recent work, Lee and Wright [2009] proposed a decomposition algorithm that is capable to deal with large-scale semiparametric SVMs.

3.2 Problem setting

In this chapter, we will assume that the input space $\mathcal{X}$ is a separable, complete metric space, and $\mathcal{Y}$, the label space, will be a subset of the reals $\mathbb{R}$. The random response $Y \in \mathcal{Y}$ is allowed to be unbounded. We equip $\mathcal{X}$ and $\mathcal{Y}$ with their respective Borel σ-algebra.

We start with the problem setup in the random design setting. Given the independent, identically distributed sample $Z_{1:n} = (Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, the partially constrained empirical risk minimization problem with the partially linear stochastic model (3.1) is to find a minimizer of

$$\min_{\theta \in \mathbb{R}^d, h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i, \phi(X_i)^\top \theta + h(X_i)\right) ,$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is a loss function, $\phi : \mathcal{X} \to \mathbb{R}^d$ is a basis function and $h \in \mathcal{H}$ is a set of real-valued functions over $\mathcal{X}$, holding the "nonparametric" component h. Our main interest is when the loss function is quadratic, i.e., $\ell(y, y') = \frac{1}{2}(y - y')^2$, but for the sake of exploring how much we exploit the structure of this loss, we will present the results in an abstract form.

Introducing $\mathcal{G} = \left\{\phi(\cdot)^\top \theta \, : \, \theta \in \mathbb{R}^d\right\}$, the above problem can be written in the form

$$\min_{g \in \mathcal{G}, h \in \mathcal{H}} L_n(g + h), \tag{3.2}$$

where $L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$. Typically, $\mathcal{H}$ arises as the set

$$\{h : \mathcal{X} \to \mathbb{R} \, : \, J(h) \leq K\}$$

with some $K > 0$ and some functional J that takes larger values for "rougher" functions.[1] The goal of learning is to find a predictor with a small expected loss. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, the expected loss, or *risk*, of f is defined to be $L(f) = \mathbb{E}\left[\ell(Y, f(X))\right]$, where $Z = (X, Y)$ is an independent copy of $Z_i = (X_i, Y_i)$ $(i = 1, \ldots, n)$. Let (g_n, h_n) be a minimizer[2] of (3.2) and let $f_n = g_n + h_n$.

When analyzing a learning procedure returning a function f_n, we compare the risk $L(f_n)$ to the best risk possible over the considered set of functions, i.e., to $L^* = \inf_{g \in \mathcal{G}, h \in \mathcal{H}} L(g+h)$. A bound on the *excess risk* $L(f_n) - L^*$ is called a generalization (error) bound. In this work, we seek bounds on the tail behaviour of the excess risk in terms of the entropy-integral of $\mathcal{H}$. Our main result, Theorem 3.6.1, provides such a bound, essentially generalizing the analogue result of Bartlett and Mendelson [2002]. In particular, our result shows that, in line with existing empirical evidence, the price of including the parametric component in terms of the

[1] The penalized empirical risk-minimization problem, $\min_{g \in \mathcal{G}, h} L_n(h+g) + J(h)$ is closely related to (3.2) as suggested by the identity $\min_{g \in \mathcal{G}, h} L_n(g+h) + \lambda J(h) = \min_{K \geq 0} \lambda K + \min_{g \in \mathcal{G}, h : J(h) \leq K} L_n(g+h)$ explored in a specific context by Blanchard et al. [2008].

[2] For simplicity, we assume that this minimizer and in fact all the others that we will need later exist. This is done for the sake of simplifying the presentation: The analysis is simple to extend to the general case. Further, if there are multiple minimizers, we choose one.

increase of the generalization bound is modest, which, in favourable situations, can be far outweighed by the decrease of L^* that can be attributed to including the parametric part. However, due to the unbounded response, the high probability bound that we derive fails to imply a bound on the expected excess risk. Thus, in the case of unbounded response, it may be unwise to use an unregularized parametric component.

To explain this issue, consider the following standard decomposition of the excess risk:

$$L(f_n) - L(f^*) = (L(f_n) - L_n(f_n)) + \underbrace{(L_n(f_n) - L_n(f^*))}_{\leq 0} + (L_n(f^*) - L(f^*)), \tag{3.3}$$

where $f^* = \arg\min_{f \in \mathcal{G}+\mathcal{H}} L(f)$. Here, the third term can be upper bounded as long as f^* is "reasonable" (e.g., bounded). On the other hand, the first term is more problematic, at least for unbounded loss functions and when Y is unbounded. Indeed, in this case f_n can take on large values and correspondingly $L(f_n)$ could also be rather large. Note that this is due to the fact that the parametric component is unconstrained.

The classical approach to deal with this problem is to introduce clipping or truncation of the predictions (cf. Theorem 11.5 of Györfi et al. [2002]). However, clipping requires additional knowledge perhaps that Y is bounded with a known bound. Furthermore, the clipping level appears in the bounds, making the bounds weak when the level is conservatively estimated. In fact, one suspects that clipping is unnecessary in our setting where we will make strong enough assumptions on the tails of Y (though much weaker than assuming that Y is bounded). In fact, in practice, it is quite rare to see clipping implemented. Hence, in this chapter we will keep our original goal and analyze the procedure with no clipping.

To analyze the excess risk we will proceed by showing that with large probability, $\|g_n\|_\infty$ is controlled. This is, in fact, where the bulk of the work will lie. A high probability upper bound for the generalization error is then developed based on this boundedness. We then discuss a potential problem caused by including the unconstrained parametric part, and explain why standard asymptotic analysis cannot detect this problem.

3.2.1 More notations

Before stating our assumptions and results, we introduce some more notation. We will denote the Minkowski-sum $\mathcal{G} + \mathcal{H}$ of $\mathcal{G}$ and $\mathcal{H}$ by $\mathcal{F}$: $\mathcal{F} = \mathcal{G} + \mathcal{H} \doteq \{g + h : g \in \mathcal{G}, h \in \mathcal{H}\}$. The L^2-norm of a function is defined as $\|f\|_2^2 \doteq \mathbb{E}\left[f^2(X)\right]$, while given the sample $X_{1:n} = \{X_1, \ldots, X_n\}$, the n-norm of a function is defined as the ℓ^2-norm of the restriction of the function to $X_{1:n}$: $\|f\|_n^2 = \frac{1}{n}\sum_i f(X_i)^2$. The vector $(f(X_1), \ldots, f(X_n))^\top$ is denoted by $f(X_{1:n})$. The matrix $(\phi(X_1), \ldots, \phi(X_n))^\top \in \mathbb{R}^{n \times d}$ is denoted by Φ (or $\Phi(X_{1:n})$ if we need to indicate its dependence on $X_{1:n}$). We let $\hat{G} = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d \times d}$ be the empirical Grammian matrix and $G = \mathbb{E}\left[\phi(X)\phi(X)^\top\right]$ be the population Grammian matrix underlying ϕ. Denote the minimal positive eigenvalue of G by $\lambda_{\min}$, while let $\hat{\lambda}_{\min}$ be the same for $\hat{G}$.

The rank of $\hat{G}$ (G, respectively) is denoted by $\hat{\rho} = \mathrm{rank}(\hat{G})$ ($\rho = \mathrm{rank}(G)$, respectively). Lastly, let $L_{h,n}(g) = L_n(h+g)$, $\overline{L}_n(f) = \mathbb{E}[L_n(f) \,|\, X_{1:n}]$, $\overline{L}_{h,n}(g) = \mathbb{E}[L_n(h+g) \,|\, X_{1:n}]$, and $\overline{L}_n^* = \inf_{g\in\mathcal{G}, h\in\mathcal{H}} \overline{L}_n(g+h)$. It will be convenient to introduce the alternate notation $\ell((x,y),f)$ for $\ell(y,f(x))$ (i.e., $\ell((x,y),f) \doteq \ell(y,f(x))$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, $f : \mathcal{X} \to \mathbb{R}$. Given $h \in \mathcal{H}$, let $g_{h,n} = \arg\min_{g\in\mathcal{G}} L_n(h+g) = \arg\min_{g\in\mathcal{G}} L_{h,n}(g)$ and $\overline{g}_{h,n} = \arg\min_{g\in\mathcal{G}} \overline{L}_{h,n}(g)$

We summarized the notations of this section in Table 3.1.

Table 3.1: Notation for Chapter 3

$\mathcal{X}$	Input space: separable complete metric space
$\mathcal{Y}$	Unbounded label space in $\mathbb{R}$
ℓ	Loss function $\ell(y, y')$
$\mathcal{G}$	Hypothesis set $\left\{\phi(\cdot)^\top\theta \,:\, \theta \in \mathbb{R}^d\right\}$ for the parametric part
$\mathcal{H}$	Hypothesis set for the nonparametric part
h_n	(g_n, h_n) is the empirical minimizer of Equation (3.2)
g_n	(g_n, h_n) is the empirical minimizer of Equation (3.2)
f_n	The empirical minimizer $f_n = h_n + g_n$
f^*	Minimizer of the expected loss over $\mathcal{H}+\mathcal{G}$
g^*	(g^*, h^*) is the minimizer of the expected loss, $f^* = h^* + g^*$
h^*	(g^*, h^*) is the minimizer of the expected loss, $f^* = h^* + g^*$
L_n	Empirical loss
$\overline{L}_n$	Expected loss conditioned on $X_{1:n}$
L	Expected loss in the random design setting
L^*	$L(f^*)$, minimum loss over $\mathcal{H}+\mathcal{G}$
$\|\cdot\|_2$	L_2 norm
$\|\cdot\|_n$	Empirical L_2 norm
$g_{h,n}$	Empirical minimizer over $\mathcal{G}$ of (3.2) for a fixed h
G	$G = \mathbb{E}\left[\phi(X)\phi(X)^\top\right]$
$\hat{G}$	$\hat{G} = \frac{1}{n}\sum_i \phi(X_i)\phi(X_i)^\top$
ρ	The rank of G
$\lambda_{\min}$	The minimal positive eigenvalue of G
$\hat{\lambda}_{\min}$	The minimal positive eigenvalue of $\hat{G}$

3.3 An infinite expected excess risk

We start with a simple regression example, based on Problem 10.3 of the book by Györfi et al. [2002], with least square loss in the purely parametric setting. This example shows that already in the purely parametric case, there exist perfectly innocent looking problems

that make ordinary least squares fail.

Example 3.3.1 (Failure of Ordinary Least Squares)**.** *Let* $\mathcal{X} = [0,1]$, $\mathcal{Y} = \mathbb{R}$, $\ell(y,p) = (y-p)^2$, $\phi : \mathcal{X} \to \mathbb{R}^3$, $\phi_1(x) = \mathbb{I}[0,1/2](x)$, $\phi_2(x) = x \cdot \mathbb{I}[0,1/2](x)$, $\phi_3(x) = \mathbb{I}(1/2,1](x)$, *where* $\mathbb{I}(A)$ *denotes the indicator of set* $A \subset \mathcal{X}$. *Let* $f_\theta(x) = \phi(x)^\top\theta$, $\theta \in \mathbb{R}^3$ *as shown in Fig. 3.1. As to the data, let* $(X,Y) \in \mathcal{X} \times \{-1,+1\}$ *be such that* X *and* Y *are independent of each other,* X *is uniform on* $\mathcal{X}$ *and* $\mathbb{P}(Y=+1) = \mathbb{P}(Y=-1) = 1/2$. *Note that* $\mathbb{E}[Y|X] = 0$, *hence the model is well-specified (the true regression function lies in the span of the basis functions). Further,* L *and* $\overline{L}_n$ *have the same optimal regression function* $f^*(x) = 0$ *and* $L(f^*) = \overline{L}_n(f^*) = 1$. *Now, let* $(X_1,Y_1),\ldots,(X_n,Y_n)$ *be* n *independent copies of* (X,Y) *and let* $\hat{\theta}_n = \arg\min_{\theta\in\mathbb{R}^3} L_n(\phi^\top\theta)$. *Denote the empirical Grammian on the data by* $\hat{G}_n = \frac{1}{n}\sum_k \phi(X_k)\phi(X_k)^\top$, $\hat{\lambda}_{\min}(n) = \lambda_{\min}(\hat{G}_n)$, $\lambda_{\min} = \lambda_{\min}(\mathbb{E}\left[\phi(X)\phi(X)^\top\right])$. *The following hold:*

(a) $\mathbb{E}\left[L_n(f_{\hat{\theta}_n})\right] = \infty$ *(infinite risk!);*

(b) $\overline{L}_n(f_{\hat{\theta}_n}) - \overline{L}_n(f^*) \to 0$ *as* $n \to \infty$ *(well-behaved in-sample generalization);*

(c) For some event B_n *with* $\mathbb{P}(B_n) \sim e^{-n}$,

$$c(\sqrt{t} - 2t) \le \mathbb{P}\left(\hat{\lambda}_{\min}(n) \le t\lambda_{\min} | B_n\right) \le c'(\sqrt{t} - 2t)$$

for some $0 < c < c'$;

(d) $\mathbb{E}\left[\hat{\lambda}_{\min}^{-1}(n)\right] = +\infty$.

To understand what happens in this example, consider the event A_n that exactly two of the n X_i's belong to the interval $[0,1/2]$. On this event, which has a probability proportional to e^{-n}, $\hat{\theta}_{n,1} = (Y_1+Y_2)/2$ and $\hat{\theta}_{n,2} = \frac{Y_1-Y_2}{X_1-X_2}$, so that $f_{\hat{\theta}_n}(X_i) = Y_i$, $i = 1,2$. Then, the out-of-sample risk can be lower bounded using $\mathbb{E}\left[\left(f_{\hat{\theta}_n}(X) - Y\right)^2\right] = \mathbb{E}\left[f_{\hat{\theta}_n}(X)^2\right] + 1 \ge \left(\mathbb{E}\left[|f_{\hat{\theta}_n}(X)| \,\middle|\, A_n\right] P(A_n)\right)^2 + 1$. Now,

$$\begin{aligned}\mathbb{E}\left[|f_{\hat{\theta}_n}(X) - Y| \,\middle|\, A_n\right] &= 2\mathbb{E}\left[\frac{X}{|X_1 - X_2|} \,\middle|\, A_n\right] \\ &= \mathbb{E}\left[\frac{1}{|X_1 - X_2|} \,\middle|\, A_n\right] = +\infty.\end{aligned}$$

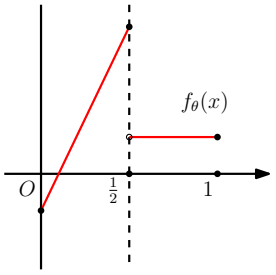

Figure 3.1: The regression function $f_\theta(x)$ for Example 3.3.1 .

A similar calculation shows the rest of the claims. This example leads to multiple conclusions:

(i) Ordinary least squares is guaranteed to have finite expected risk if and only if we have $\mathbb{E}\left[\lambda_{\min}(G_n)^{-1}\right] < +\infty$, a condition that is independent to previous conditions such as "good statistical leverage" [Hsu et al., 2012].

(ii) Neither good in-sample generalization, or in-probability parameter convergence, or that the estimated parameter satisfies the central limit theorem (which all hold in the above example) lead to good expected risk for ordinary least-squares; demonstrating a practical example where out-of-sample generalization error is not implied by any of these "classical" results that are extensively studied in statistics (e.g., [Bickel et al., 1998]).

(iii) Although the "Eigenvalue Chernoff Bound" (Theorem 4.1) of Gittens and Tropp [2011] captures the probability of the smallest positive eigenvalue being significantly underestimated correctly as a function of the sample size, it fails to capture the actual behaviour of the left-tail, and this behaviour can be significantly different for different distributions.

Based on this example, we see that another option to get an expected risk bound without clipping the predictions or imposing an additional restriction on the basis functions and the data generating distribution, is to clip the eigenvalues of the data Grammian before inversion at a level of $O(1/n)$ or to add this amount to all the eigenvalues. One way of implementing the increase of eigenvalues is to employ ridge regression by introducing a penalty of the form $\|\theta\|_2^2$ in the empirical loss minimization criterion. Yet the success of ERM of the partially linear model without regularization in practice still remains vague. In the rest of this chapter, we further investigate the performance of ERM for the general setting. We show that given n is large enough, with high probability, ERM indeed has controllable generalization bound. Unsurprisingly, our analysis also indicates its essential dependence on $\mathbb{E}\left[\hat{\lambda}_{\min}(n)^{-1}\right]$.

3.4 Assumptions

Before presenting our main results, we need some mild assumptions on the setting. In what follows we will assume that the functions in $\mathcal{H}$ are bounded by $r > 0$. If $\mathcal{K}$ is an RKHS space with a continuous reproducing kernel κ and $\mathcal{X}$ is compact (a common assumption in the literature, e.g., Cucker and Zhou [2007], Steinwart and Christmann [2008]), this assumption will be satisfied if $J(h) = \|h\|_{\mathcal{K}}$ and $\mathcal{H} = \{h \in \mathcal{K} : J(h) \leq r\}$, where, without loss of generality (WLOG), we assume that the maximum of κ is below one.

We will also assume that $R = \sup_{x\in\mathcal{X}} \|\phi(x)\|_2$ is finite. If ϕ is continuous and $\mathcal{X}$ is compact, this assumption will be satisfied, too. In fact, by rescaling the basis functions if needed, we will assume WLOG that $R = 1$.

Definition 3.4.1. Let β, Γ be positive numbers. A (noncentered) random variable X is subgaussian with parameters (β, Γ) if

$$\mathbb{E}\left[\exp\left(|\beta X|^2\right)\right] \leq \Gamma < \infty.$$

Let us start with our assumptions that are partly concerned the loss function, ℓ, and are partly concerned with the joint distribution of (X, Y).

Assumption 1 (Loss function)**.**

(i) Convexity: The loss function ℓ is convex with respect to its second argument, i.e., $\ell(y, \cdot)$ is a convex function for all $y \in \mathcal{Y}$.

(ii) There exists a bounded measurable function $\hat{h}$ and a constant $Q < \infty$ such that

$$\mathbb{E}\left[\ell\left(Y, \hat{h}(X)\right) | X\right] \leq Q \quad \text{almost surely.}$$

(iii) Subgaussian Lipschitzness: There exists a function $K_\ell : \mathcal{Y} \times (0, \infty) \to \mathbb{R}$ such that for any constant $c > 0$ and $c_1, c_2 \in [-c, c]$,

$$|\ell(y, c_1) - \ell(y, c_2)| \leq K_\ell(y, c) |c_1 - c_2|,$$

and such that $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2) | X\right] \leq \Gamma_c < \infty$ for some constant Γ_c depending only on c almost surely. WLOG, we assume that $K_\ell(y, \cdot)$ is a monotonically increasing function for any $y \in \mathcal{Y}$.

(iv) Level-Set: For any $X_{1:n} \subset \mathcal{X}$, and any $c \geq 0$, $R_c = \sup_{f \in \mathcal{F}: \overline{L}_n(f) \leq c} \|f\|_n$ is finite and independent of n.

Remark 3.4.2. Assumption 1(ii) requires that Y, even if it is unbounded, still can be approximated by a bounded function in $\mathcal{H}$ at every X with constant expected loss. For quadratic losses, this is satisfied if and only if $\mathbb{E}\left[Y^2 | X\right] < \infty$ almost surely.

Remark 3.4.3. The subgaussian Lipschitzness assumption 1(iii) is a general form of Lipschitzness property that allows the Lipschitzness coefficient to depend on y. If the loss function is the quadratic loss, the subgaussian Lipschitzness assumption is an immediate corollary of the subgaussian property of Y conditioned on X. In particular, $|(Y - c_1)^2 - (Y - c_1)^2| = |2Y - c_1 - c_2||c_1 - c_2|$. Thus we can pick $K_\ell(Y, c) = 2|Y| + 2c$ and $\beta = \frac{1}{2\sqrt{2}}$, then $\mathbb{E}\left[\exp(|\beta K_\ell(Y, c)|^2)\right] = \mathbb{E}\left[\exp(\frac{1}{2}(|Y| + c)^2)\right] \leq \mathbb{E}\left[\exp(|Y|^2)\right] + \exp(c^2)$.

Remark 3.4.4. Unlike the first three assumptions, Assumption 1(iv), which requires that the sublevel sets of $\overline{L}_n(\cdot)$ are bounded in $\|\cdot\|_n$, is nonstandard. This assumption will be crucial for showing the boundedness of the parametric component of the model. We argue that in some sense this assumption, given the method considered, is necessary. The idea is that since f_n minimizes the empirical loss, it should also have a small value of $\overline{L}_n(\cdot)$ (in fact, this is not that simple to show given that it is not known whether f_n is bounded). As such, it will be in some sublevel set of $\overline{L}_n(\cdot)$. Otherwise, nothing prevents the algorithm from choosing a minimizer (even when minimizing $\overline{L}_n(\cdot)$ instead of $L_n(\cdot)$) with an unbounded $\|\cdot\|_n$

norm. One way of weakening Assumption 1(iv) is to assume that there exist a minimizer of $\overline{L}_n(\cdot)$ over $\mathcal{F}$ that has a bounded norm and then modify the procedure to pick the minimizer with the smallest $\|\cdot\|_n$ norm.

Example 3.4.5 (Quadratic Loss)**.** *In the case of quadratic loss, i.e., when* $\ell(y, y') = \frac{1}{2}(y - y')^2$, $R_c^2 \le 4c + 8Q + 4s^2$ *where* $s = \|\hat{h}\|_\infty$. *Indeed, this follows from*

$$\begin{aligned}\|f\|_n^2 &\le \frac{2}{n}\sum_i \mathbb{E}\left[(f(X_i) - Y_i)^2 \,|X_{1:n}\right] + \mathbb{E}\left[Y_i^2 \,|X_{1:n}\right] \\ &\le 4\overline{L}_n(f) + \frac{2}{n}\sum_i \mathbb{E}\left[Y_i^2 \,|X_i\right].\end{aligned}$$

Then $\mathbb{E}\left[Y_i^2 \,|X_i\right] \le 2\mathbb{E}\left[(Y_i - \hat{h}(X_i))^2 \,|X_i\right] + 2\hat{h}^2(X_i) \le 4Q + 2s^2$. *Here, the last inequality is by Assumption 1(ii) and the boundedness of* $\hat{h}$.

Example 3.4.6 (Exponential Loss)**.** *In the case of exponential loss, i.e., when* $\ell(y, y') = \exp(-yy')$ *and if* $\mathcal{Y} = \{+1, -1\}$, *the situation is slightly more complex.* R_c *will be finite as long as the posterior probability of seeing either of the labels is uniformly bounded away from one, as assumed e.g., by Blanchard et al. [2008]. Specifically, if* $\eta(x) \doteq \mathbb{P}(Y = 1|X = x) \in [\epsilon, 1 - \epsilon]$ *for some* $\epsilon > 0$ *then a simple calculation shows that* $R_c^2 \le c/\epsilon$.

The next assumption states that the loss function is locally "non-flat":

Assumption 2 (Non-flat Loss)**.** *Assume that there exists* $\epsilon > 0$ *such that for any* $h \in \mathcal{H}$ *and vector* $a \in [-\epsilon, \epsilon]^n \cap \mathrm{Im}(\Phi)$,

$$\begin{aligned}\frac{\epsilon}{n}\|a\|_2^2 \le \mathbb{E}&\left[\frac{1}{n}\sum_i \ell(Z_i, h + \overline{g}_{h,n} + a_i)\,\middle|\, X_{1:n}\right] \\ &- \mathbb{E}\left[\frac{1}{n}\sum_i \ell(Z_i, h + \overline{g}_{h,n})\,\middle|\, X_{1:n}\right]\end{aligned}$$

holds a.s., where recall that $Z_i = (X_i, Y_i)$.

Note that it is key that the "perturbation" a is in the image space of Φ, and that it is applied at $h + \overline{g}_{h,n}$ and not at an arbitrary function h, as shown by the next example:

Example 3.4.7 (Quadratic loss)**.** *In the case of the quadratic loss, note that* $g(X_{1:n}) = \Phi(X_{1:n})\theta$. *Let* $\overline{\theta}_{h,n}$ *be a minimizer of* $\overline{L}_{h,n}(\cdot)$ *satisfying* $\overline{\theta}_{h,n} = \left(\Phi^\top\Phi\right)^+\Phi^\top(\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n}))$. *Therefore,*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_i \ell((X_i, Y_i), h + \overline{g}_{h,n} + a_i)\,|\, X_{1:n}\right] \\ &\qquad - \mathbb{E}\left[\frac{1}{n}\sum_i \ell((X_i, Y_i), h + \overline{g}_{h,n})\,|\, X_{1:n}\right] \\ &= \frac{1}{n}\sum_i \mathbb{E}\left[a_i\left\{2(\overline{g}_{h,n}(X_i) + h(X_i) - Y_i) + a_i\right\}\,|\, X_{1:n}\right],\end{aligned}$$

which is equal to $\frac{1}{n}\|a\|_2^2 + \frac{2}{n}a^\top \left\{\Phi\left(\Phi^\top\Phi\right)^+\Phi^\top - I\right\}\{\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n})\} = \frac{1}{n}\|a\|_2^2$, *where the last equality follows since* $a \in \mathrm{Im}(\Phi)$.

We will need an assumption that the entropy of $\mathcal{H}$ satisfies an integrability condition. For this, recall the definition of entropy numbers:

Definition 3.4.8. For $\epsilon > 0$, the ϵ-covering number $N(\epsilon, \mathcal{H}, d)$ of a set $\mathcal{H}$ equipped with a pseudo-metric d, is the number of balls with radius ϵ measured with respect to d, necessary to cover $\mathcal{H}$. The ϵ-entropy of $\mathcal{H}$ is $H(\epsilon, \mathcal{H}, d) = \log N(\epsilon, \mathcal{H}, d)$.

The definition is extended in the natural way to the case when $\mathcal{H}$ is equipped with a pseudo-norm. Note that if $d' \leq d$ then the ϵ-balls w.r.t. d' are bigger than the ϵ-balls w.r.t. d. Hence, any ϵ-cover w.r.t. d also gives an ϵ-cover w.r.t. d'. Therefore, $N(\epsilon, \mathcal{H}, d') \leq N(\epsilon, \mathcal{H}, d)$ and also $H(\epsilon, \mathcal{H}, d') \leq H(\epsilon, \mathcal{H}, d)$.

Let $\|\cdot\|_{\infty,n}$ be the infinity empirical norm: For $f : \mathcal{X} \to \mathbb{R}$, $\|f\|_{\infty,n} = \max_{1\leq k\leq n} |f(X_k)|$. Note that trivially $\|f\|_n \leq \|f\|_{\infty,n} \leq \|f\|_\infty$. We use $\|\cdot\|_{\infty,n}$ in our next assumption:

Assumption 3 (Integrable Entropy Numbers of $\mathcal{H}$)**.** *There exists a (non-random) constant* C_H *such that,* $\int_0^1 H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\, dv \leq C_H$ *holds a.s.*

Remark 3.4.9. Assumption 3 is well-known in the literature of empirical processes to guarantee the uniform laws of large numbers [Dudley, 1984, Giné and Zinn, 1984, Tewari and Bartlett, 2013]. The assumption essentially requires that the entropy numbers of $\mathcal{H}$ grow slowly as the scale of ϵ approaches to zero. For example, this assumption holds if for any $0 < u \leq 1$, $H(u, \mathcal{H}, \|\cdot\|_{\infty,n}) \leq cu^{-(2-\epsilon)}$ for some $c > 0$, $\epsilon > 0$. Based on our previous discussion, $H(u, \mathcal{H}, \|\cdot\|_{\infty,n}) \leq H(u, \mathcal{H}, \|\cdot\|_\infty)$. The latter entropy numbers are well-studied for a wide range of function spaces (and enjoy the condition required here). For examples see, e.g., Dudley [1984], Giné and Zinn [1984], Tewari and Bartlett [2013].

For the next assumption let $G_{\lambda_{\min}}$ be the event when $\hat{\lambda}_{\min} \geq \lambda_{\min}/2$.

Assumption 4 (Lipschitzness of the Parametric Solution Path)**.** *Let* P_X *denote the distribution of* X*. There exists a constant* K_h *such that on* $G_{\lambda_{\min}}$ *for* $[P_X]$ *almost all* $x \in \mathcal{X}$*,* $h \mapsto \overline{g}_{h,n}(x)$ *is* K_h*-Lipschitz w.r.t.* $\|\cdot\|_{\infty,n}$ *over* $\mathcal{H}$*.*

Remark 3.4.10. When $\overline{g}_{h,n}$ is uniquely defined, Assumption 4 will be satisfied whenever ℓ is sufficiently smooth w.r.t. its first argument, as follows, e.g., from the Implicit Function Theorem.

Example 3.4.11 (Quadratic loss)**.** *In the case of the quadratic loss, by Example 3.4.7,*

$$\begin{aligned}\overline{g}_{h,n}(x) &= \langle \phi(x), \left(\Phi^\top\Phi\right)^+\Phi^\top(\mathbb{E}\left[Y_{1:n}|X_{1:n}\right] - h(X_{1:n}))\rangle \\ &= \frac{1}{n}\sum_i \langle \phi(x), \hat{G}^+\phi(X_i)\left(\mathbb{E}\left[Y_i|X_{1:n}\right] - h(X_i)\right)\rangle\end{aligned}$$

Thus, for $h, h' \in \mathcal{H}$, on $G_{\lambda_{\min}}$,

$$\begin{aligned}
&|\overline{g}_{h,n}(x) - \overline{g}_{h',n}(x)| \\
&\quad = \left| \langle \phi(x), \left(\Phi^\top \Phi\right)^+ \Phi^\top \left(h'(X_{1:n}) - h(X_{1:n})\right) \rangle \right| \\
&\quad \le \frac{2\,\|\phi(x)\|_2}{\lambda_{\min}} \frac{1}{n} \sum_i |h'(X_i) - h(X_i)|\, \|\phi(X_i)\|_2 \\
&\quad \le \frac{2}{\lambda_{\min}} \|h' - h\|_{\infty,n}
\end{aligned}$$

where we used $\|\phi(x)\|_2 \le 1$ multiple times which holds $[P_X]$ a.e. on $\mathcal{X}$.

3.5 The boundedness of the predictor

Our first main result implies that g_n is bounded with high probability in the random design setting.

Theorem 3.5.1. *Let Assumptions 1 to 4 hold. Then, there exist positive constants c_1, c_2, K such that for any $0 < \delta < 1$ and n such that $n \ge c_1 + c_2 \frac{\log\left(\frac{2\rho}{\delta}\right)}{\lambda_{\min}}$, it holds that*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \|g_{h,n}\|_\infty \ge K\right) \le \delta\,. \tag{3.4}$$

The result essentially states that for some specific value of K, the probability that the event $\sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty > K$ happens is exponentially small as a function of the sample size n. The constant K is inversely proportional to $\lambda_{\min}$ and depends on both R_c from the level-set assumption and r. Here c depends on Q, $\|\hat{h}\|_\infty$ from Assumption 1(ii) and the subgaussian parameters. The actual value of K can be read out from the proof.

The main challenges in the proof of this result are that the bound has to hold uniformly over $\mathcal{H}$ (this allows us to bound $\|g_n\|_\infty$), and also that the response Y is unbounded, as are the functions in $\mathcal{G}$. The main tool is a ratio type tail inequality for empirical processes, allowing us to deal both with the unbounded responses and functions, which is then combined with our assumptions on the loss function, in particular, with the level-set assumption.

Proof of Theorem 3.5.1. Recall that our goal is to derive a bound on

$$\sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty$$

that holds with high probability. Fix $h \in \mathcal{H}$. Then, $g_{h,n}(x) = \langle \theta, \phi(x) \rangle \le \|\theta_{h,n}\|_2 \|\phi(x)\|_2$, where $\theta_{h,n}$ is the parameter vector of $g_{h,n}$. Since $\|\phi(x)\|_2 \le 1$, it suffices to bound $\|\theta_{h,n}\|_2$. On $G_{\lambda_{\min}}$, which is defined as the event $\{\hat{\lambda}_{\min} \ge \lambda_{\min}/2\}$, we have

$$g_{h,n}^2(x) \le \|\theta_{h,n}\|_2^2 \le \frac{\theta_{h,n}^\top \hat{G}\, \theta_{h,n}}{\hat{\lambda}_{\min}} = \frac{2\,\|g_{h,n}\|_n}{\lambda_{\min}}\,. \tag{3.5}$$

Hence, the problem is reduced to proving a uniform (h-independent) upper bound on the empirical norm of $g_{h,n}$ and showing that $G_{\lambda_{\min}}$ happens with "large probability".

For the latter, we use a result of Gittens and Tropp [2011]. This is summarized in the lemma which also includes some observations that will prove to be useful later:

Lemma 3.5.1. *The following hold:*

(i) With probability one, for any $\theta \in \mathbb{R}^d$, $\theta^\top \hat{G}\theta \le \frac{\theta^\top G\theta}{\lambda_{\min}}$.

(ii) Assuming that $n \in \mathbb{N}$ and $\delta \in (0,1)$ are such that

$$n \ge \frac{2}{\lambda_{\min} \log\left(\frac{e}{2}\right)} \log\left(\frac{\rho}{\delta}\right), \tag{3.6}$$

where ρ and $\lambda_{\min}$ are respectively the rank and the smallest positive eigenvalue of G, with probability at least $1-\delta$, it holds that $\hat{\lambda}_{\min} \ge \frac{\lambda_{\min}}{2} > 0$.

(iii) For any n, δ satisfying (3.6), then with probability $1-\delta$ it holds that for any $\theta \in \mathbb{R}^d$ and $[P_X]$ almost every $x \in \mathcal{X}$, $|\langle \theta, \phi(x)\rangle| \le \sqrt{\frac{2\theta^\top \hat{G}\theta}{\lambda_{\min}}}$.

The (easy) proof of the lemma is deferred to Appendix B.2.

To get an upper bound on the empirical norm of $g_{h,n}$, we will use

$$\|g_{h,n}\|_n \le \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n + \left\|\overline{g}_{h,n}\right\|_n \tag{3.7}$$

and develop uniform bound on the two terms on the r.h.s..

Lemma 3.5.2. *It holds almost surely that*

$$\sup_{h \in \mathcal{H}} \left\|\overline{g}_{h,n}\right\|_n \le \bar{R},$$

where $\bar{R} = R_{C_\Gamma} + r$, $C_\Gamma = \frac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1} + Q$, $\hat{r} = \max(r, \|\hat{h}\|_\infty)$ and $\hat{h}$ is the function from Assumption 1(ii).

The constant R_{C_0} that appears in the statement is defined in our "level-set assumption" (cf. Assumption 1(iv)).

Proof. Fix some $h \in \mathcal{H}$. We have $\left\|\overline{g}_{h,n}\right\|_n = \left\|h + \overline{g}_{h,n} + (-h)\right\|_n \le \left\|h + \overline{g}_{h,n}\right\|_n + \|-h\|_n \le \left\|h + \overline{g}_{h,n}\right\|_n + r$ thanks to $\|h\|_\infty \le r$. Hence, it remains to bound $\left\|h + \overline{g}_{h,n}\right\|_n$.

By Assumption 1(iv), for this it suffices if we show a bound on $\overline{L}_n(h + \overline{g}_{h,n})$ since by this assumption if $\overline{L}_n(h + \overline{g}_{h,n}) \le c$ then $\left\|h + \overline{g}_{h,n}\right\|_n \le R_c$. By the optimizing property of $\overline{g}_{h,n}$, we have $\overline{L}_n(h + \overline{g}_{h,n}) = \overline{L}_{n,h}(\overline{g}_{h,n}) \le \overline{L}_{n,h}(0) = \overline{L}_n(h)$. Now, by definition

$$\overline{L}_n(h) = \mathbb{E}\left[\frac{1}{n}\sum_i \ell(Y_i, h(X_i))\middle| X_{1:n}\right],$$

hence, it suffices to bound $\mathbb{E}\left[\ell(Y_i, h(X_i))|X_i\right]$. For this, we have

$$\mathbb{E}\left[\ell(Y_i, h(X_i))|X_i\right] \le \mathbb{E}\left[\left|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))\right| |X_i\right] + \mathbb{E}\left[\ell(Y_i, \hat{h}(X_i))\,|X_i\right],$$

where we used that by assumption the loss is nonnegative. By Assumption 1(ii),

$$\mathbb{E}\left[\ell(Y_i, \hat{h}(X_i))\,|X_i\right] \le Q.$$

Therefore it is sufficient to bound

$$\mathbb{E}\left[|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))|\,|X_i\right].$$

Note that by Assumption 1(iii), almost surely $\mathbb{E}\left[\exp\left(|\beta K_\ell(Y,r)|^2\right)|\,X\right] \le \Gamma_r$. So, by Lemma B.1.1 (i), $\mathbb{E}\left[K_\ell(Y,r)|\,X\right] \le \frac{1}{\beta}\sqrt{\Gamma_r - 1}$ a.s.. Thus, with $\hat{r} = \max(r, \|\hat{h}\|_\infty)$,

$$\mathbb{E}\left[|\ell(Y_i, h(X_i)) - \ell(Y_i, \hat{h}(X_i))|\,|X_i\right] \le \mathbb{E}\left[2\hat{r}K_\ell(Y_i,\hat{r})|\,X_i\right] \le \tfrac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1}.$$

Putting together the inequalities, we obtain that $\overline{L}_n(h + \overline{g}_{h,n}) \le \frac{2\hat{r}}{\beta}\sqrt{\Gamma_{\hat{r}} - 1} + Q =: C_\Gamma$ and thus $\left\|h + \overline{g}_{h,n}\right\|_n \le R_{C_\Gamma}$. □

Let us now consider bounding $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$. In fact, we will only bound this on the event $G_{\lambda_{\min}}$ when $\hat{\lambda}_{\min} \ge \lambda_{\min}/2$. Since we use this event to upper bound $1/\hat{\lambda}_{\min}$ by $2/\lambda_{\min}$, there is no loss in bounding $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$ on this event only. Note that by Lemma 3.5.1 (ii), $G_{\lambda_{\min}}$ holds with probability at least $1 - \delta$.

Lemma 3.5.3. *There exist problem-dependent positive constants C_0 and $L_0 \ge 1$ such that for any $n \ge 16L_0^4$, it holds that*

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \ge 1\,, G_{\lambda_{\min}}\Big) \le \exp\left(-\frac{C_0 n}{4}\right). \tag{3.8}$$

Remark 3.5.4. Although $\overline{g}_{h,n}$ may be bounded, the concentration inequality for $g_{h,n}$ and $\overline{g}_{h,n}$ is still not trivial since $g_{h,n}$ is potentially unbounded. Relaxing the boundedness condition to weaker ones in the concentration inequality has been explored in the literature, e.g. making use of the bounded-difference property [Kutin, 2002], or a Lipschitzness assumption of the loss function w.r.t. the sample space and finite subgaussian diameter [Kontorovich, 2013]. Our result, Lemma 3.5.3, continues to contribute to this line of research.

The proof of this lemma follows the proofs in the paper of van de Geer [1990], who studied the deviations $\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$ for $h = 0$ (see also van de Geer 2000). It turns out the techniques of the mentioned paper are just strong enough to also bound the uniform deviation $\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n$. As the proof is lengthy and technical, it is deferred to Appendix B.3.

Now, combining (3.5), (3.7) and Lemma 3.5.2 we get that on $G_{\lambda_{\min}}$,

$$G_{n,\infty} \doteq \sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \le \frac{2}{\lambda_{\min}} \sup_{h\in\mathcal{H}} \|g_{h,n}\|_n \le \frac{2}{\lambda_{\min}}\left(\overline{R} + \sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n\right). \tag{3.9}$$

Since for any $A > 0$,

$$\mathbb{P}\left(G_{n,\infty} > A\right) \le \mathbb{P}\left(G^c_{\lambda_{\min}}\right) + \mathbb{P}\left(G_{n,\infty} > A, G_{\lambda_{\min}}\right)$$

and by (3.9),

$$\mathbb{P}\left(G_{n,\infty} > A, G_{\lambda_{\min}}\right) \leq \mathbb{P}\left(\frac{2}{\lambda_{\min}}\left(\overline{R} + \sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n\right) > A, G_{\lambda_{\min}}\right),$$

choosing $A = \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)$, we see that

$$\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)\right) \leq \mathbb{P}\left(G^c_{\lambda_{\min}}\right) + \mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq 1, G_{\lambda_{\min}}\right).$$

By Eq. (3.6) and Lemma 3.5.3, provided that $n \geq \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{2\rho}{\delta}\right)$, $n \geq 16L_0^4$ and $n \geq \frac{4\log\left(\frac{2}{\delta}\right)}{C_0}$ we get that

$$\mathbb{P}\left(G_{n,\infty} > \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)\right) \leq \delta,$$

which is the desired statement. In particular, we can choose $K = \frac{2}{\lambda_{\min}}\left(\overline{R}+1\right)$. □

3.6 A high probability bound of the excess risk

Given Theorem 3.5.1, various high-probability risk bounds can be derived using more or less standard techniques. Despite this, when the response is unbounded and clipping is not available, we were not able to identify any result in the literature that would achieve this. In our proof, we use the technique of van de Geer [1990], which allows us to work with unbounded responses without clipping the predictions. Since this technique was developed for the fixed design case, we combine it with a method, which uses Rademacher complexities upper bounded in terms of the entropy integral, so as to get an out-of-sample generalization bound.[3] The bound in our result is of the order $1/\sqrt{n}$, which is expected given our constraints on the nonparametric class $\mathcal{H}$. However, we note in passing, that under stronger conditions, such as $L(f^*) = 0$ [Pollard, 1995, Haussler, 1992], or that $\mathcal{F}$ is convex (which does not hold in our case unless we take the convex hull of $\mathcal{F} = \mathcal{G} + \mathcal{H}$) and the loss is the quadratic loss (or some other loss which is strongly convex), a faster rate of $O(1/n)$ can also be proved [Lee et al., 1998, Györfi et al., 2002, Bartlett et al., 2005, Koltchinskii, 2006, 2011], though the existing works seem to make various assumptions about Y which we would like to avoid. Let $(x)_+ = \max(x, 0)$ denote the positive part of $x \in \mathbb{R}$.

Theorem 3.6.1 (Generalization). *Let Assumptions 1 to 4 hold and let $f^* = g^* + h^*$ be a minimizer of L over $\mathcal{G} + \mathcal{H}$ (i.e., $g^* \in \mathcal{G}$, $h^* \in \mathcal{H}$). There exist positive constants $c, c_1, c_2, c_3, c_4, \alpha$ and $U \geq \max\{1, K, \|g^*\|_\infty\}$ such that for any $0 < \delta < 1$ satisfying $\log\frac{1}{\delta} \geq c$ and $n \geq c_1 + c_2\log\left(\frac{4\rho}{\delta}\right)/\lambda_{\min}$, with probability at least $1 - 3\delta$,*

$$L(f_n) - L(f^*) \leq c_3\frac{C_H + \rho^{1/2}(\log(U))_+}{\sqrt{n}} + 2(r+U)\sqrt{\frac{\log\frac{2}{\delta}}{\alpha n}} + c_4\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{3.10}$$

where $f_n = h_n + g_n$ is a minimizer of $L_n(\cdot)$ over $\mathcal{H} + \mathcal{G}$.

[3] Recall that "In-sample" generalization bounds concern the deviation $\overline{L}_n(f_n) - \overline{L}_n(f^*)$, while "out-of-sample" bounds concern $L(f_n) - L(f^*)$.

Remark 3.6.1. The constants ρ and $\lambda_{\min}$ appear both in U and in the lower bound constraint of n. Defining $\overline{\ell}(x,p) = \mathbb{E}\left[\ell(Y,p)|X=x\right]$, constant c_3 depends on the (essential) Lipschitz coefficient of $\overline{\ell}(X,p)$ when $p \in [-r-U, r+U]$ and constant c_4 depends on the (essential) range of $\ell(X,p)$. Both of them can be shown to be finite based on Assumption 1. The bound has a standard form: The first and the last of the three terms comes from bounding the out-of-sample generalization error, while the term in the middle (containing α) bounds the in-sample generalization error. We use a novel measure-disintegration technique to transfer the results of van de Geer [1990] which are developed for the fixed design setting (i.e., when the covariates $X_{1:n}$ are deterministic) to the random design setting that we consider in this theorem.

Proof of Theorem 3.6.1. Let U be as in Theorem 3.5.1 and let E denote the event when

$$\sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \le K\,.$$

For any $z \ge 0$,

$$\begin{aligned}\mathbb{P}\left(L(f_n) - L(f^*) > z\right) &= \mathbb{P}\left(L(f_n) - L(f^*) > z, E^c\right) + \mathbb{P}\left(L(f_n) - L(f^*) > z, E\right) \\ &\le \mathbb{P}\left(E^c\right) + \mathbb{P}\left(L(f_n) - L(f^*) > z, E\right).\end{aligned} \tag{3.11}$$

Thus, to study the tail probabilities of $L(f_n) - L(f^*)$, it suffices to study $L(f_n) - L(f^*)$ *on the event* E.

Define $\mathcal{G}(U) = \{g \in \mathcal{G} : \|g\|_\infty \le U\}$ and $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$. On E, we claim that $f_n \in \mathcal{C}$. We have $f_n = h_n + g_n$ and since $h_n \in \mathcal{H}$ by definition, it remains to show that $g_n \in \mathcal{G}(U)$. By appropriately selecting $g_{h,n}$, we can arrange for $g_n = g_{h_n,n}$. Hence, $\|g_n\|_\infty \le \sup_{h\in\mathcal{H}} \|g_{h,n}\|_\infty \le K \le U$, showing that $f_n \in \mathcal{C}$ indeed holds.

Note that $f^* = h^* + g^* \in \mathcal{C}$. By Eq. (3.3), on E it holds almost surely that

$$\begin{aligned}L(f_n) - L(f^*) &\le L_n(f^*) - L(f^*) - (L_n(f_n) - L(f_n)) \\ &= (\widetilde{\Delta}_n(f_n) - \widetilde{\Delta}_n(f^*)) + (\overline{\Delta}_n(f^*) - \overline{\Delta}_n(f_n)) \\ &\le \underbrace{\sup_{f\in\mathcal{C}} \widetilde{\Delta}_n(f) - \widetilde{\Delta}_n(f^*)}_{\widetilde{\Delta}_n^*(\mathcal{C})} + \underbrace{\sup_{f\in\mathcal{C}} |\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)|}_{\overline{\Delta}_n^*(\mathcal{C})},\end{aligned} \tag{3.12}$$

where we introduced $\overline{\Delta}_n(f) = L_n(f) - \overline{L}_n(f)$ and $\widetilde{\Delta}_n(f) = L(f) - \overline{L}_n(f)$ with $\overline{L}_n(f) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\ell(Y_k, f(X_k))|X_k\right]$. Note that the first term does not depend on the (unbounded) responses $Y_1, \dots, Y_n$. Furthermore, by our assumptions, $\widetilde{\Delta}_n(f)$ is bounded for f bounded. Hence, we can analyze these terms using tools developed for bounded random variables and empirical processes. Now, while the last term involves $Y_1, \dots, Y_n$, $\overline{\Delta}_n$ compares average losses *over the sample* $X_1, \dots, X_n$, this last term concerns *in-sample* generalization. Hence, as we will show it below, it can be analyzed using tools developed for the "fixed design" setting. In fact, the following result gives tail bounds for this term:

Lemma 3.6.2. *Let Assumptions 1 to 4 hold and WLOG assume that* $U \geq \max(1, K, \|g^*\|_\infty)$. *Then, there exist constants* $c, \alpha > 0$ *such that for any* $0 < \delta < 1$ *satisfying* $\log \frac{1}{\delta} \geq c$ *with probability at least* $1 - \delta$,

$$\overline{\Delta}_n^*(\mathcal{C}) \leq 2(r+U)\sqrt{\frac{\log \frac{2}{\delta}}{\alpha n}}\,. \tag{3.13}$$

The proof, which we defer to Section B.4, is based on Theorem 3.3 of van de Geer [1990]. It remains to bound $\widetilde{\Delta}_n^*(\mathcal{C}) = \sup_{f\in\mathcal{C}} \widetilde{\Delta}_n(f) - \widetilde{\Delta}_n(f^*)$. For this, define

$$\overline{\ell}(x,p) = \mathbb{E}\left[\ell(Y,p) \,|\, X = x\right].$$

With a slight abuse of notation, we also introduce $\overline{\ell}(x,f) = \overline{\ell}(x, f(x))$. Let

$$B(\overline{\ell}, U) = \left\| \sup_{p\in[-r-U, r+U]} \overline{\ell}(X,p) \right\|_{L^\infty},$$

where $\|\cdot\|_{L^\infty}$ denotes the essential supremum of its argument. We also let $\overline{L}$ be the Lipschitz constant of $\overline{\ell}$ when $p \in [-r-U, r+U]$:

$$\mathrm{Lip}(\overline{\ell}, U) = \left\| \sup_{p,p'\in[-r-U,r+U], p\neq p'} \frac{\overline{\ell}(X,p) - \overline{\ell}(X,p')}{|p-p'|} \right\|_{L^\infty}.$$

The next lemma shows that both quantities are finite:

Lemma 3.6.3. *Let* $r' = \max(r+U, \|\hat{h}\|_\infty)$. *Then,* $B(\overline{\ell}, U) \leq Q + \frac{2r'}{\beta}\sqrt{\Gamma_{r'} - 1} < +\infty$ *and* $\mathrm{Lip}(\overline{\ell}, U) < \frac{\sqrt{\Gamma_{r+U}-1}}{\beta} < +\infty$.

Proof. For the second statement, for any $t, s \in [-b, b]$ we have

$$\overline{\ell}(X,t) - \overline{\ell}(X,s) \leq \mathbb{E}\left[|\ell(Y,t) - \ell(Y,s)| \,\middle|\, X\right] \leq \mathbb{E}\left[K_\ell(Y,b)|t-s| \,\middle|\, X\right] \leq \tfrac{\sqrt{\Gamma_b - 1}}{\beta}|t-s|\,,$$

where we used Assumption 1(iii) and Lemma B.1.1(i). Thus, $\mathrm{Lip}(\overline{\ell}, U) \leq \frac{\sqrt{\Gamma_{r+U}-1}}{\beta} < +\infty$.

For the first statement take some $|p| \leq r + U$ and write

$$\begin{aligned}\overline{\ell}(X,p) &\leq \overline{\ell}(X, \hat{h}(X)) + |\overline{\ell}(X,p) - \overline{\ell}(X,\hat{h}(X))| \leq Q + \mathrm{Lip}(\overline{\ell}, r')|p - \hat{h}(X)| \\ &\leq Q + \mathrm{Lip}(\overline{\ell}, r')(|r+U| + \|\hat{h}\|_\infty|) \leq Q + \tfrac{\Gamma_{r'}-1}{\beta}(2r')\,,\end{aligned}$$

where in the second inequality we used Assumption 1(ii), while in the last one we used the bound on the Lipschitz coefficient. □

As it is well known, the Rademacher complexity of $\mathcal{C}$, defined next, captures *exactly* the behavior of $\mathbb{E}\left[\widetilde{\Delta}_n^*(\mathcal{C})\right]$ (e.g., Tewari and Bartlett [2013]).

Definition 3.6.4 (Rademacher Complexity of Subsets of $\mathbb{R}^n$)**.** Let $A \subset \mathbb{R}^n$, $(\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ be independent Rademacher random variables (i.e., $\mathbb{P}(\sigma_k = 1) = 1/2$). The Rademacher complexity of A, $\mathfrak{R}(A)$ is

$$\mathfrak{R}(A) = \frac{1}{n}\,\mathbb{E}\left[\sup_{a\in A} \sum_{i=1}^n \sigma_i a_i\right].$$

Definition 3.6.5 (Rademacher Complexity of Function Sets)**.** Let $\mathcal{F} \subset \{f : f : \mathcal{X} \to \mathbb{R}\}$ and P be a measure on $\mathcal{X}$. Then, the nth Rademacher number of $\mathcal{F}$ induced by P is

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left[\mathfrak{R}(\mathcal{F}(X_{1:n}))\right],$$

where $\mathcal{F}(X_{1:n}) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$ is the projection of $\mathcal{F}$ to an i.i.d. sample $X_{1:n} = (X_1, \ldots, X_n)$ from P. When n and P are uniquely identified from the context, we also call $\mathfrak{R}_n(\mathcal{F})$ the Rademacher-complexity of $\mathcal{F}$.

The Rademacher complexity enjoys a number of useful properties, amongst which we need the following contraction property:

Theorem 3.6.2. *Let $\phi = (\phi_1, \ldots, \phi_n) : \mathbb{R}^n \to \mathbb{R}^n$, denote $\{(\phi_1(a_1), \ldots, \phi_n(a_n)) : a \in A\}$ by $\phi \circ A$ for $A \subset \mathbb{R}^n$. Assume that all the component functions ϕ_i are L-Lipschitz over A. Then, $\mathfrak{R}(\phi \circ A) \le L\mathfrak{R}(A)$.*

Note that this theorem is usually stated for the case when $\phi_1 = \ldots = \phi_n$. The simpler form is sufficient for "margin based losses" (used in classification) that have the form $\ell(y, p) = g(yp)$ with some g. As we will see, here we need this more general form as our losses are less constrained. However, the proof of this more general result still follows the standard reasoning. We defer the proof to Appendix B.5.

Let $\mathcal{L} = \left\{s_f : \mathcal{X} \to \mathbb{R} : s_f(x) = \overline{\ell}(x, f) - \overline{\ell}(x, f^*), f \in \mathcal{C}, x \in \mathcal{X}\right\}$. Note that $\widetilde{\Delta}_n(f) - \widetilde{\Delta}_n(f^*) = (L(f) - \overline{L}_n(f)) - (L(f^*) - \overline{L}_n(f^*)) = \mathbb{E}\left[\overline{\ell}(X, f) - \overline{\ell}(X, f^*)\right] - \frac{1}{n}\sum_{k=1}^n (\overline{\ell}(X_k, f) - \overline{\ell}(X_k, f^*)) = \mathbb{E}\left[s_f(X)\right] - \frac{1}{n}\sum_{k=1}^n s_f(X_k)$. Following the standard argument, since the range of functions in $\mathcal{L}$ is bounded by $B(\overline{\ell}, U)$, by McDiamid's inequality, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\begin{aligned}\widetilde{\Delta}_n^*(\mathcal{C}) &= \sup_{s \in \mathcal{L}} \mathbb{E}\left[s(X)\right] - \tfrac{1}{n}\textstyle\sum_{k=1}^n s(X_k) \\ &\le \mathbb{E}\left[\sup_{s \in \mathcal{L}} \mathbb{E}\left[s(X)\right] - \tfrac{1}{n}\textstyle\sum_{k=1}^n s(X_k)\right] + B(\overline{\ell}, U)\sqrt{\frac{2\log\frac{1}{\delta}}{n}}.\end{aligned}$$

Following the calculation before Theorem 7 in Section 3.2 of Tewari and Bartlett [2013],

$$\mathbb{E}\left[\sup_{s \in \mathcal{L}} \mathbb{E}\left[s(X)\right] - \tfrac{1}{n}\textstyle\sum_{k=1}^n s(X_k)\right] \le 2\mathfrak{R}_n(\mathcal{L}).$$

Let us now bound $\mathfrak{R}_n(\mathcal{L}) = \mathbb{E}\left[\mathfrak{R}(\mathcal{L}(X_{1:n}))\right]$. We can write

$$\mathcal{L}(X_{1:n}) = \{s_f(X_{1:n}) : f \in \mathcal{C}\} = \phi \circ \mathcal{C}(X_{1:n}),$$

where $\phi = (\phi_1, \ldots, \phi_n) : \mathbb{R}^n \to \mathbb{R}^n$ is defined by $\phi_k(t) = \overline{\ell}(X_k, t) - \overline{\ell}(X_k, f^*(X_k))$ (note that ϕ is random). By definition, each component of ϕ is almost surely Lipschitz over any bounded interval $[-b, b]$ with the same Lipschitz constant (depending on b). Indeed, for any

$t, s \in [-b, b]$,

$$\begin{aligned}
\|\|\phi_k(t) - \phi_k(s)\|\|_{L^\infty} &= \|\|\overline{\ell}(X_k, t) - \overline{\ell}(X_k, s)\|\|_{L^\infty} \\
&= \inf\left\{a \in \mathbb{R} \,:\, \mathbb{P}\left(|\overline{\ell}(X_k, t) - \overline{\ell}(X_k, s)| > a\right)\right\} \\
&= \inf\left\{a \in \mathbb{R} \,:\, \mathbb{P}\left(|\overline{\ell}(X, t) - \overline{\ell}(X, s)| > a\right)\right\} \\
&= \|\|\overline{\ell}(X, t) - \overline{\ell}(X, s)\|\|_{L^\infty} \\
&\leq \mathrm{Lip}(\overline{\ell}, b)|t - s|\,,
\end{aligned}$$

where the second and fourth equalities used the definition of $\|\cdot\|_{L^\infty}$ and the third used that X_k and X are identically distributed. Now, since $\mathcal{C}$ contains functions bounded by $r + U$, by Theorem 3.6.2,

$$\mathfrak{R}(\phi \circ \mathcal{C}(X_{1:n})) \leq \mathrm{Lip}(\overline{\ell}, U)\, \mathfrak{R}(\mathcal{C}(X_{1:n})) \quad \text{a.s.}$$

and hence

$$\mathfrak{R}_n(\mathcal{L}) = \mathbb{E}\,\mathfrak{R}(\mathcal{L}(X_{1:n})) = \mathbb{E}\,\mathfrak{R}(\phi \circ \mathcal{C}(X_{1:n})) \leq \mathrm{Lip}(\overline{\ell}, U)\, \mathbb{E}\,\mathfrak{R}(\mathcal{C}(X_{1:n})) = \mathrm{Lip}(\overline{\ell}, U)\, \mathfrak{R}_n(\mathcal{C}).$$

Our next goal is to bound $\mathfrak{R}_n(\mathcal{C})$. By Dudley's entropy integral bound [Dudley, 1967] (e.g., Theorem 10 of Tewari and Bartlett [2013], for a statement with a proof see Theorem 11.4 of Rakhlin and Sridharan [2014]),

$$\mathfrak{R}_n(\mathcal{C}) \leq \frac{12}{\sqrt{n}} \mathbb{E}\left[\int_0^1 H^{1/2}(u, \mathcal{C}, \|\cdot\|_n)\, du\right] \leq \frac{12}{\sqrt{n}}(2C_H + 2C_G(U)),$$

where the second inequality holds thanks to Lemma B.1.2 and we also used that Dudley's bound holds regardless the scale of the range of functions in $\mathcal{C}$ (this is not hard to check by inspecting the proof of the bound). Combining all the inequalities we get that with probability at least $1 - \delta$,

$$\widetilde{\Delta}_n^*(\mathcal{C}) \leq \frac{48(C_H + C_G(U))\ \mathrm{Lip}(\overline{\ell}, U)}{\sqrt{n}} + B(\overline{\ell}, U)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{3.14}$$

Combining Equations (3.11) and (3.12), we have for any $z \geq 0$,

$$\mathbb{P}\left(L(f_n) - L(f^*) > z\right) \leq \mathbb{P}\left(E^c\right) + \mathbb{P}\left(\widetilde{\Delta}_n^*(\mathcal{C}) + \overline{\Delta}_n^*(\mathcal{C}) > z\right). \tag{3.15}$$

Now, by Lemma 3.6.2 and (3.14), for any $0 < \delta < 1$ such that $\log(1/\delta) \geq c$, with probability at least $1 - 2\delta$,

$$\widetilde{\Delta}_n^*(\mathcal{C}) + \overline{\Delta}_n^*(\mathcal{C}) \leq \frac{48(C_H + C_G(U))\ \mathrm{Lip}(\overline{\ell}, U)}{\sqrt{n}} + 2(r + U)\sqrt{\frac{\log \frac{2}{\delta}}{\alpha n}} + B(\overline{\ell}, U)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} =: \pi(\delta)\,.$$

Together with (3.15) and Theorem 3.5.1, we thus get that with probability $1 - 3\delta$, provided that $\log(1/\delta) \geq c$ and $n \geq c_1 + c_2 \frac{\log\left(\frac{2\rho}{\delta}\right)}{\lambda_{\min}}$,

$$L(f_n) - L(f^*) \leq \pi(\delta)\,,$$

thus finishing the proof. □

3.7 Discussions

Notice that the above high probability result holds only if n is large compared to $\log(1/\delta)$, or, equivalently when δ is not too small compared to n, a condition that is inherited from Theorem 3.5.1. Was without this constraint, the tail of $L(f_n) - L(f^*)$ would be of a sub-gaussian type, which we could integrate to get an expected risk bound. However, because of the constraint, this does not work. With no better idea, one can introduce clipping, to limit the magnitude of the prediction errors on an event of probability (say) $1/n$. This still results in an expected risk bound of the order (i.e., $O(1/\sqrt{n})$), as expected, although with an extra logarithmic factor. However, if one needs to introduce clipping, this could be done earlier, reducing the problem to studying the metric entropy of the clipped version of $\mathcal{F}$ (which is almost what is done in Lemma B.1.2 given in the supplementary material). For this, assuming Y is bounded, one can use Theorem 11.5 of Györfi et al. [2002]. Note, however, that in this result, for example, the clipping level, which one would probably select conservatively in practice, appears raised to the 4th power. In comparison, with our technique, the clipping level could actually be made appear only through its logarithm in our bound if we choose $\delta = 1/(Ln)$.

On the other hand, our bound scales with $\lambda_{\min}^{-1}$ through U. This is alarming unless the eigenvalues of the Grammian are well-controlled, in which case $\lambda_{\min}^{-1} = O(\sqrt{\rho})$. Our example in Section 3.3 also shows that the constraint connecting n, $\lambda_{\min}$ and δ in Theorem 3.6.1 is not removable without imposing additional conditions. More importantly, not all high probability bounds are equal. In particular, the type of in Theorem 3.6.1 constraining n to be larger than $\log(1/\delta)$ does not guarantee small expected risk.

Finally based on this example, we see that an expected risk bound can be derived from Theorem 3.6.1, e.g., for the squared loss, by introducing a penalty of the form $\|\theta\|_2^2$ in the empirical loss minimization criterion to add an amount of $O(1/n)$ to all the eigenvalues of the data Grammian and setting $\delta = O(1/n^2)$. Since then outside of an event with probability $O(1/n^2)$, the risk is controlled by the high probability bound of Theorem 3.6.1, while on the remaining "bad event", the prediction error will stay bounded by n^2. Although numerical algebra packages implement pseudo-inverses by cutting the minimum eigenvalue, this may be insufficient since they usually cut at the machine precision level, which translates into sample sizes which may not be available in practice.

3.8 Conclusions and future work

In this chapter we set out to investigate the question whether current practice in semiparametric regression of not penalizing the parametric component is a wise choice from the point of view of finite-time performance. We found that for any error probability level, for sample

sizes $n = \Omega(\log(1/\delta))$, the risk of such a procedure can indeed be bounded with high probability, proving the first finite-sample generalization bound for partially linear stochastic models. The main difficulty of the proof is to guarantee that the parametric component is bounded in the supremum norm. However, we have also found that an additional restriction connecting the data generating distribution and the parametric part is necessary to prove an expected risk bound. This second observation is based on an example where the model is purely parametric. Based on the attained results, unless other additional knowledge is available, we think that it is too risky to follow current practice and recommend introducing some form of regularization for the parametric part and/or clipping the predictions when suitable bounds are available on the range of the Bayes predictor. Our results also help identify that existing bounds in the literature do not capture the behavior of the distribution of the minimum positive eigenvalue of empirical Grammian matrices $\hat{\lambda}_{\min}$, which would be critical for improving our understanding of the basic question of how the expected risk of ordinary least-squares behaves.

Appendix B

Omitted Proofs for Chapter 3

This Chapter is devoted to the omitted proofs for Chapter 3.

B.1 Technical lemmas

We first present some results that we will need later multiple times.

Lemma B.1.1 (Elementary Properties of Subgaussian Random Variables)**.** *Let U be a subgaussian random variable with parameters (β, Γ). Then,*

(i) $\mathbb{E}\left[|U|\right] \leq \frac{1}{\beta}\sqrt{\Gamma - 1}$*;*

(ii) for any constant $c \geq 0$, $U + c$ is subgaussian.

Proof. (i) follows from

$$\Gamma \geq \mathbb{E}\left[\exp\left(|\beta U|^2\right)\right] \geq \exp\left(\mathbb{E}\left[|\beta U|^2\right]\right) \geq \exp\left(\mathbb{E}\left[|\beta U|\right]^2\right) \geq 1 + \mathbb{E}\left[|\beta U|\right]^2 .$$

(ii) follows from

$$\mathbb{E}\left[\exp\left(\tfrac{1}{2}\beta^2(U+c)^2\right)\right] \leq \mathbb{E}\left[\exp\left(\beta^2 U^2 + \beta^2 c^2\right)\right] \leq e^{\beta^2 c^2}\mathbb{E}\left[\exp\left(|\beta U|^2\right)\right] \leq e^{\beta^2 c^2}\Gamma .$$

□

We will also need the following result:

Lemma B.1.2. *Let $U > 0$, $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$. Then, a.s.*

$$\int_0^1 \sqrt{H(u, \mathcal{C}, \|\cdot\|_n)}\, du \leq 2C_H + 2C_G(U) ,$$

where $C_G(U) = \rho^{1/2}\int_0^1 \log^{1/2}\left(\frac{4U+u}{u}\right) du \,(= O(\sqrt{\rho(\log(U))_+}))$.

Proof of Lemma B.1.2. Since $\mathcal{C} = \mathcal{H} + \mathcal{G}(U)$, a standard argument shows that

$$H(u; \sigma) \leq H(u/2; \mathcal{H}; \|\cdot\|_n) + H(u/2, \mathcal{G}(U), \|\cdot\|_n) . \tag{B.1}$$

Now, note that $\|\cdot\|_n \leq \|\cdot\|_{\infty,n}$. Thus,

$$\begin{aligned}
&\int_0^1 H^{1/2}(u/2, \mathcal{H}, \|\cdot\|_n)\, du \\
&\quad = 2\int_0^{1/2} H^{1/2}(u, \mathcal{H}, \|\cdot\|_n)\, du \leq 2\int_0^1 H^{1/2}(u, \mathcal{H}, \|\cdot\|_n)\, du \\
&\quad \leq 2\int_0^1 H^{1/2}(u, \mathcal{H}, \|\cdot\|_{\infty,n})\, du \leq 2C_H,
\end{aligned}$$

where the last inequality is by Assumption 3. Moreover, since $\|g\|_n \leq \|g\|_\infty$, $\mathcal{G}(U)$ is a subset of the ball $B_{\mathcal{G},\|\cdot\|_n}(0, U)$. Thus,

$$\begin{aligned}
&\int_0^1 H^{1/2}(u/2, \mathcal{G}(U), \|\cdot\|_n)\, du \\
&\quad \leq 2\int_0^1 H^{1/2}(u, \mathcal{G}(U), \|\cdot\|_n)\, du \leq 2\int_0^1 H^{1/2}(u, B_{\mathcal{G},\|\cdot\|_n}(0, U), \|\cdot\|_n)\, du \\
&\quad \leq 2\rho^{1/2}\int_0^1 \log^{1/2}\left(\frac{4U+u}{u}\right) du = 2C_G(U)\,,
\end{aligned}$$

where the second inequality is by Corollary 2.6 of [van de Geer, 2000], which states that $H(\epsilon, B_{\mathcal{G},\|\cdot\|_n}(0,\sigma)) \leq \rho\log(\frac{4\sigma+\epsilon}{\epsilon})$. Using (B.1) and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ which holds for $a, b \geq 0$, we conclude that

$$\int_0^1 \sqrt{H(u;\sigma)}\, du \leq 2C_H + 2C_G(U),$$

finishing the proof of the claim. □

B.2 Proof of eigenvalue bound (Lemma 3.5.1)

Lemma 3.5.1. *The following hold:*

(i) With probability one, for any $\theta \in \mathbb{R}^d$, $\theta^\top \hat{G}\theta \leq \frac{\theta^\top G\theta}{\lambda_{\min}}$.

(ii) Assuming that $n \in \mathbb{N}$ and $\delta \in (0,1)$ are such that

$$n \geq \frac{2}{\lambda_{\min}\log\left(\frac{e}{2}\right)}\log\left(\frac{\rho}{\delta}\right), \tag{3.6}$$

where ρ and $\lambda_{\min}$ are respectively the rank and the smallest positive eigenvalue of G, with probability at least $1-\delta$, it holds that $\hat{\lambda}_{\min} \geq \frac{\lambda_{\min}}{2} > 0$.

(iii) For any n, δ satisfying (3.6), then with probability $1-\delta$ it holds that for any $\theta \in \mathbb{R}^d$ and $[P_X]$ almost every $x \in \mathcal{X}$, $|\langle\theta, \phi(x)\rangle| \leq \sqrt{\frac{2\theta^\top \hat{G}\theta}{\lambda_{\min}}}$.

Proof. Part (i): We first show that $\mathrm{Ker}(G) \subseteq \mathrm{Ker}(\hat{G})$ holds almost surely: In particular, this can be seen by proving that $G\theta = 0$ for some $\theta \in \mathbb{R}^d$ then with probability one, $\hat{G}\theta = 0$ also holds. Indeed, if the latter did not hold with probability one, then for some $\epsilon > 0$, $\mathbb{P}(\theta^\top\hat{G}\theta \geq \epsilon) > 0$ would hold. Then, $\theta^\top G\theta = \mathbb{E}[\theta^\top\hat{G}\theta] \geq \epsilon\mathbb{P}(\theta^\top\hat{G}\theta \geq \epsilon) > 0$,

which means that $\theta \notin \mathrm{Ker}(G)$. Now, if we take a set of vectors $\{\theta_1, \ldots, \theta_m\}$ spanning $\mathrm{Ker}(G)$, then on some event E with $\mathbb{P}(E) = 1$, $\hat{G}\theta_i = 0$ holds for *all* $1 \leq i \leq m$. Now, on E, $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$. Indeed, take an arbitrary $\theta \in \mathrm{Ker}(G)$ and expand it using $\{\theta_i\}$: $\theta = \sum_{i=1}^m \lambda_i \theta_i$. Then, $\hat{G}\theta = \sum_i \lambda \hat{G}\theta_i$ and since $\hat{G}\theta_i = 0$ simultaneously for all i, the statement follows.

Now, for proving Part (i), consider the event E where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$. We prove the result on E: Pick any $\theta \in \mathbb{R}^d$ and decompose it into $\theta = \theta_\perp + \theta_{||}$ such that $\theta_\perp \perp \mathrm{Im}(G)$ and $\theta_{||} \in \mathrm{Im}(G)$. Hence, $\theta^\top G \theta = \theta_{||}^\top G \theta_{||}$. Since $\theta_\perp \in \mathrm{Ker}(G)$ and $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$, we have $\hat{G}\theta_\perp = 0$. Hence, $\theta^\top \hat{G} \theta = \theta_{||}^\top \hat{G} \theta_{||}$. Now, since $\|\phi(x)\|_2 \leq 1$ it holds that $\hat{\lambda}_{\max} \leq 1$, where $\hat{\lambda}_{\max}$ denotes the largest eigenvalue of $\hat{G}$. Therefore, on E,

$$\theta^\top \hat{G} \theta = \theta_{||}^\top \hat{G} \theta_{||} \leq \|\theta_{||}\|_2^2 \leq \frac{\theta_{||}^\top G \theta_{||}}{\lambda_{\min}} = \frac{\theta^\top G \theta}{\lambda_{\min}}.$$

Since $\mathbb{P}(E) = 1$, the result follows.

Part (ii): By the "Eigenvalue Chernoff Bound" (Theorem 4.1) of Gittens and Tropp [2011], with probability at least $1 - \rho \exp\Big(- n\lambda_{\min}\big(\epsilon + (1-\epsilon)\log(1-\epsilon) \big)\Big)$, $\hat{\lambda}_{\min} \geq (1-\epsilon)\lambda_{\min}$. Choosing $\epsilon = 1/2$ gives the result.

Part (iii): Fix n, δ as required. Let E be the event where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\hat{G})$ and let F_δ be the event where the inequality of Part (ii) holds. Take the set S of those $x \in \mathrm{supp}(P_X)$ where $\mathrm{Ker}(G) \subset \mathrm{Ker}(\phi(x)\phi(x)^\top)$ holds. It follows from the argument presented in Part (i) that $P_X(\mathcal{X} \setminus S) = 0$.

Since $\mathbb{P}(E \cap F_\delta) \geq 1 - \delta$, it suffices to prove the statement on $E \cap F_\delta$. Hence, in what follows all statements are meant to hold on this event. Pick any $\theta \in \mathbb{R}^d$, $x \in S$ and decompose θ as before. Then, thanks to $x \in S$ it holds that $\theta_\perp \in \mathrm{Ker}(\phi(x)\phi(x)^\top)$. Hence, $\langle \theta, \phi(x) \rangle^2 = \theta^\top \phi(x)\phi(x)^\top \theta = \theta_{||}\phi(x)\phi(x)^\top \theta_{||} = \langle \theta_{||}, \phi(x) \rangle^2$. Now, owning to $\|\phi(x)\|_2 \leq 1$,

$$\langle \theta_{||}, \phi(x) \rangle^2 \leq \big\|\theta_{||}\big\|_2^2 \leq \frac{\theta_{||}^\top \hat{G} \theta_{||}}{\hat{\lambda}_{\min}} \leq \frac{2\,\theta_{||}^\top \hat{G} \theta_{||}}{\lambda_{\min}} = \frac{2\,\theta^\top \hat{G} \theta}{\lambda_{\min}},$$

where the last inequality follows from Part (ii). □

B.3 Proof of Lemma 3.5.3

The proof follows the ideas from the paper of van de Geer [1990]. Lemma 3.5.3 calls for a uniform (in $h \in \mathcal{H}$) bound for $\big\|g_{h,n} - \overline{g}_{h,n}\big\|_n$. Fix $h \in \mathcal{H}$. We consider a self-normalized "version" of the differences $g_{h,n} - \overline{g}_{h,n}$, which are easier to deal with. This is done as follows: For $g \in \mathcal{G}$, define

$$\omega_{g,h} = \frac{g - \overline{g}_{h,n}}{1 + K\big\|g - \overline{g}_{h,n}\big\|_n} \quad \text{and} \quad \Omega_{h,n} = \{\omega_{g,h} \,:\, g \in \mathcal{G}\}\,,$$

where $K > 0$ is to be chosen later. Then, for any $\omega \in \Omega_{h,n}$, $\|\omega\|_n < \frac{1}{K}$ and

$$\begin{aligned}\left\|g - \overline{g}_{h,n}\right\|_n &= \frac{\left\|g - \overline{g}_{h,n}\right\|_n}{1 + K\left\|g - \overline{g}_{h,n}\right\|_n}\left(1 + K\left\|g - \overline{g}_{h,n}\right\|_n\right) = \left\|\omega_{g,h}\right\|_n\left(1 + K\left\|g - \overline{g}_{h,n}\right\|_n\right) \\ &= \frac{\left\|\omega_{g,h}\right\|_n}{1 - K\left\|\omega_{g,h}\right\|_n}.\end{aligned} \tag{B.2}$$

Thus, we see that is enough to control the empirical norm of

$$\hat{\omega}_{h,n} = \omega_{g_{h,n},h} = \frac{g_{h,n} - \overline{g}_{h,n}}{1 + K\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n}.$$

The first step is to bound this norm in terms of the increments of the empirical process

$$\Delta_{h,n}(g) := L_{h,n}(g) - \overline{L}_{h,n}(g).$$

Lemma B.3.1 ("Basic Inequality"). *Let Assumption 2 hold. There exists a constant η, such that on the event $G_{\lambda_{\min}}$, for any $h \in \mathcal{H}$,*

$$\eta\left\|\hat{\omega}_{h,n}\right\|_n^2 \leq \Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}).$$

The proof, which is stated in Section B.3.1, follows standard arguments. Based on this, we can reduce the study of the supremum of the empirical norm of $\hat{\omega}_{h,n}$ to that of the supremum of the increments $\mathcal{V}_{h,n}(\omega) = \sqrt{n}\big(\Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \omega)\big)$ normalized by ω. In particular, it follows from Lemma B.3.1 that for $L, \sigma > 0$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\left\|\hat{\omega}_{h,n}\right\|_n \geq L\sigma, G_{\lambda_{\min}}\Big) \\ &\quad= \mathbb{P}\Big(\exists h \in \mathcal{H} : \left\|\hat{\omega}_{h,n}\right\|_n \geq L\sigma, \frac{\mathcal{V}_{h,n}(\hat{\omega}_{h,n})}{\|\hat{\omega}_{h,n}\|_n^2} \geq \eta\sqrt{n}, G_{\lambda_{\min}}\Big) \\ &\quad\leq \mathbb{P}\left(\sup_{(g,h)\in\mathcal{G}\times\mathcal{H}:\|\omega_{g,h}\|_n\geq L\sigma} \frac{\mathcal{V}_{h,n}(\omega_{g,h})}{\|\omega_{g,h}\|_n^2} \geq \eta\sqrt{n}, G_{\lambda_{\min}}\right).\end{aligned} \tag{B.3}$$

The supremum of normalized increments similar to the one appearing above was studied by van de Geer [1990]. In fact, we will adapt Lemma 3.4 of this paper to our purposes. The lemma requires minimal modifications: In our case, the empirical process is indexed with elements of $\{\omega_{g,h} : g \in \mathcal{G}, h \in \mathcal{H}\}$, the product set $\mathcal{G} \times \mathcal{H}$, whereas van de Geer [1990] considers a similar result for $h = 0$. As a result, whereas van de Geer [1990] reduces the study of this probability to bounding the "size" of balls in the the index space, we will reduce it to bounding the size of "tubes".

To state the generalization of Lemma 3.4 of van de Geer [1990], we introduce the following abstract setting: Let $(V, d_{V,k})$, $(\Lambda, d_{\Lambda,k})$ be pseudo-metric spaces ($k = 1, \ldots, n$), d_k^2 be the pseudo-metric on $V \times \Lambda$, which for $\gamma = (\nu, \lambda)$, $\tilde{\gamma} = (\tilde{\nu}, \tilde{\lambda})$ in $V \times \Lambda$ is defined by

$d_k^2(\gamma,\tilde{\gamma}) = d_{V,k}^2(\nu,\tilde{\nu}) + d_{\Lambda,k}^2(\lambda,\tilde{\lambda})$. Further, let d^2 be the pseudo-metric on $V \times \Lambda$ defined by $d^2 = \frac{1}{n}\sum_{k=1}^n d_k^2$. Consider the real-valued processes $U_1, U_2, \ldots, U_n$ on $V \times \Lambda$ and the process

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n U_k \,.$$

For $\sigma > 0$, denote by $H(\epsilon,\sigma) \doteq H(\epsilon, T(\sigma), d)$, the metric entropy of the σ-"tube"

$$T(\sigma) = \cup_{\nu\in V}\{\nu\} \times \{\lambda \in \Lambda_\nu \,:\, d_\Lambda(\lambda_\nu,\lambda) \le \sigma\} \subset V \times \Lambda \,,$$

where for $\nu \in V$, $\Lambda_\nu \subset \Lambda$ and d_Λ (defining the "tube") is the a pseudo-metric on Λ defined by $d_\Lambda^2(\lambda,\tilde{\lambda}) = \frac{1}{n}\sum_k d_{\Lambda,k}^2(\lambda,\tilde{\lambda})$. For $L > 0$, define

$$\alpha_n(L,\sigma) = \frac{\int_0^1 \sqrt{H(uL\sigma, L\sigma)}du}{\sqrt{n}L\sigma}.$$

With this, we are ready to state our generalization of Lemma 3.4 of van de Geer [1990]:

Lemma B.3.2. *Assume that the following conditions hold:*

(i) $U_1, U_2, \cdots, U_n$ *are independent, centered; for all* $\nu \in V$, $Z_n(\nu,\lambda_\nu) = 0$ *for some* $\lambda_\nu \in \Lambda$, *and*

$$|U_k(\gamma) - U_k(\tilde{\gamma})| \le M_k d_k(\gamma,\tilde{\gamma}) \,, \quad \gamma,\tilde{\gamma} \in V \times \Lambda,$$

where $M_1, M_2, \cdots, M_n$ *are uniformly subgaussian, i.e., for some positive* β *and* Γ,

$$\mathbb{E}[\exp(|\beta M_k|^2)] \le \Gamma < \infty, k = 1, 2, \ldots, n.$$

(ii) Assume that $\sigma > 0$ *is such that* $\sqrt{n}\sigma \ge 1$ *and suppose*

$$\lim_{L\to\infty} \alpha_n(L,\sigma) = 0 \,.$$

Then, there exist constants $L_0 \ge 1$ *and* C_0, *depending only on* (β,Γ) *and the map* $L \mapsto \alpha_n(L,\sigma)$, *such that for all* $L \ge L_0$,

$$\mathbb{P}\Big(\sup_{\nu\in V} \sup_{\substack{\lambda\in\Lambda_\nu:\\ d_\Lambda(\lambda_\nu,\lambda)>L\sigma}} \frac{|Z_n(\nu,\lambda)|}{d_\Lambda^2(\lambda_\nu,\lambda)} \ge \sqrt{n}\Big) \le \exp(-C_0 L^2\sigma^2 n).$$

Remark B.3.3. The proof is obtained by modifying the proof of van de Geer [1990]'s Lemma 3.4 in a straightforward manner and hence it is omitted. A careful investigation of the original proof will find that the result also holds if we find L_0 and C_0 depending on an *upper bound* $\tilde{\alpha}_n(L,\sigma)$ for $\alpha_n(L,\sigma)$ provided that $\lim_{L\to\infty}\tilde{\alpha}_n(L,\sigma) = 0$ still holds. Moreover, if the upper bound is selected such that it does not depend on n and σ but only on L and the "size" of the spaces V, $(\Lambda_\nu)_{\nu\in V}$, then L_0 and C_0 will depend only on (β,Γ) and the mentioned "size".

To apply Lemma B.3.2 to our problem, we choose the spaces to be $V = \mathcal{H}$, $\Lambda = \cup_{h\in\mathcal{H}}\Lambda_h$, where $\Lambda_h = \Omega_{h,n}$. Further, we choose the pseudo-metrics to be $d^2_{V,k}(h,\tilde{h}) = |h(X_k) - \tilde{h}(X_k)|^2 + \|h-\tilde{h}\|^2_{\infty,n}$ $(h,\tilde{h}\in V)$, and $d_{\Lambda,k}(\omega,\tilde{\omega}) = |\omega(X_k) - \tilde{\omega}(X_k)|$ $(\omega,\tilde{\omega}\in\Lambda)$. We also choose $\Lambda_h = \Omega_{h,n} \subset \Lambda$. Since these pseudo-metrics are random (they depend on $X_{1:n}$), for a proper use of Lemma B.3.2 we again need to "condition" on $X_{1:n}$ when using this lemma. Making this argument formal will be discussed in **??**.

For $f \in L^1(\mathcal{X}, P_X)$, $\omega\in\Lambda$, $h\in\mathcal{H}$ set

$$\Delta_k(f) = \frac{1}{\eta}\left(\ell(Z_k,f) - \mathbb{E}_{x_{1:n}}[\ell(Z_k,f)]\right),$$
$$U_k(h,\omega) = \Delta_k(h+\overline{g}_{h,n}) - \Delta_k(h+\overline{g}_{h,n}+\omega).$$

(We remind the reader that, although not shown to minimize clutter, Δ_k and U_k do depend on $x_{1:n}$.)

Now, for $h\in\mathcal{H}$, we set $\lambda_h = 0$. Thus, $U_k(h,\lambda_h) = U_k(h,0) = 0$. Furthermore, for $Z_n(h,\omega) = \frac{1}{\sqrt{n}}\sum_{k=1}^n U_k(h,\omega)$ we have $Z_n(h,\omega) = \frac{1}{\eta}\mathcal{V}_{h,n}(\omega)$ and therefore (using that $\lambda_h = 0$ and $d_\Lambda(\omega,\tilde{\omega}) = \|\omega-\tilde{\omega}\|_n$)

$$\sup_{h\in\mathcal{H}}\sup_{\substack{\omega\in\Lambda_h:\\ d_\Lambda(\lambda_h,\omega)>L\sigma}}\frac{Z_n(h,\omega)}{d^2_\Lambda(\lambda_h,\omega)} = \sup_{h\in\mathcal{H}}\sup_{\substack{\omega\in\Omega_{h,n}:\\ \|\omega\|_n>L\sigma}}\frac{\mathcal{V}_{h,n}(\omega)}{\eta\|\omega\|_n^2} =: Q_n(L\sigma), \tag{B.4}$$

showing that the conclusion of the lemma suffices to bound the quantity of interest appearing in (B.3).

We claim that the condition of Lemma B.3.2 are satisfied for $[P_X]$ almost every $x_{1:n}\in\mathcal{X}^n$ such that $\lambda_{\min}(x_{1:n}) \doteq \lambda_{\min}(\Phi(x_{1:n})^\top\Phi(x_{1:n})) \geq \lambda_{\min}/2$. Let $\mathcal{N}\subset\mathcal{X}^n$ be the $[P_X]$ null-set where the claim is not required to hold (we will construct $\mathcal{N}$ in the proof). That U_k are centered and $Z_n(h,\lambda_h) = 0$ for any $h\in\mathcal{H}$ holds by construction. As far as the remaining conditions are concerned, we have:

Condition (i)*, the independence of* (U_k): This follows from the definition of $\mathbb{P}_{x_{1:n}}$ and the independence of (X_k,Y_k).

Condition (i)*, the Lipschitzness of* U_k: Our goal is to show (for later use) that the Lipschitz coefficients M_k can be chosen independently of n and $x_{1:n}$ as long $\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2$. For this, we will assume that

$$K \geq 1. \tag{B.5}$$

Since U_k is defined as a function of Δ_k, we consider the Lipschitzness of Δ_k first. Using the definition of Δ_k and the Lipschitzness of ℓ (cf. Assumption 1(iii)), for any $f,f'\in L^1(\mathcal{X},P_X)$ we have

$$|\Delta_k(f) - \Delta_k(f')|$$
$$\leq \frac{1}{\eta}\left(\frac{|\ell(Z_k,f)) - \ell(Z_k,f')|}{|f(X_k)-f'(X_k)|} + \frac{\mathbb{E}\left[|\ell(Z_k,f)-\ell(Z_k,f')|\,|X_k\right]}{|f(X_k)-f'(X_k)|}\right)|f(X_k)-f'(X_k)|.$$

Denote $\frac{|\ell(Z_k,f))-\ell(Z_k,f')|}{|f(X_k)-f'(X_k)|}+\frac{\mathbb{E}\left[|\ell(Z_k,f)-\ell(Z_k,f')|\,|X_k\right]}{|f(X_k)-f'(X_k)|}$ by $N_k(f,f')$. Thus, for $h,\tilde{h}\in\mathcal{H}$, $\omega,\tilde{\omega}\in\Lambda$, letting $f=h+\overline{g}_{h,n}$, $\tilde{f}=\tilde{h}+\overline{g}_{\tilde{h},n}$,

$$\begin{aligned}
&|U_k(h,\omega)-U_k(\tilde{h},\tilde{\omega})| \\
&\quad\le \frac{1}{\eta}N_k(f,\tilde{f})\left|f(X_k)-\tilde{f}(X_k)\right| \\
&\qquad+\frac{1}{\eta}N_k(f+\omega,\tilde{f}+\tilde{\omega})\left\{\left|f(X_k)-\tilde{f}(X_k)\right|+|\omega(X_k)-\tilde{\omega}(X_k)|\right\}
\end{aligned}$$

Now, by assumption $|h(x_k)|,|\tilde{h}(x_k)|\le r$. Also from (3.5), Lemma 3.5.2, and $\lambda_{\min}(x_{1:n})\ge\lambda_{\min}/2$, it follows that $|\overline{g}_{h,n}(x_k)|,|\overline{g}_{\tilde{h},n}(x_k)|\le\frac{2\overline{R}}{\lambda_{\min}}$. Also, by the same argument as in Lemma B.3.5, again thanks to $\lambda_{\min}(x_{1:n})\ge\lambda_{\min}/2$, $|\omega(x_k)|,|\tilde{\omega}(x_k)|\le\frac{1}{K(\lambda_{\min}/2)^{1/2}}\le\frac{1}{(\lambda_{\min}/2)^{1/2}}$, where we used (B.5). Hence,

$$N_k(f,\tilde{f})\le K_\ell\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}\right)+\mathbb{E}\left[K_\ell\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}\right)\,|\,X_k\right]$$

and similarly,

$$\begin{aligned}
N_k(f+\omega,\tilde{f}+\tilde{\omega})\le K_\ell&\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}+\frac{1}{(\lambda_{\min}/2)^{1/2}}\right) \\
&+\mathbb{E}\left[K_\ell\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}+\frac{1}{(\lambda_{\min}/2)^{1/2}}\right)\,|\,X_k\right]
\end{aligned}$$

Now,

$$\begin{aligned}
|f(x_k)-\tilde{f}(x_k)|&\le|h(x_k)-\tilde{h}(x_k)|+|\overline{g}_{h,n}(x_k)-\overline{g}_{\tilde{h},n}(x_k)| \\
&\le|h(x_k)-\tilde{h}(x_k)|+K_h\|h-\tilde{h}\|_{\infty,n}\,,
\end{aligned}$$

where the second inequality follows since by Assumption 4, $h\mapsto\overline{g}_{h,n}(x_k)$ is K_h-Lipschitz. Therefore, by the choice of $d_{V,k}$ and $d_{\Lambda,k}$,

$$|U_k(h,\omega)-U_k(\tilde{h},\tilde{\omega})|\le\frac{2M_k}{\eta}\left(d_{V,k}(h,\tilde{h})+d_{\Lambda,k}(\omega,\tilde{\omega})\right)\le M_k'd_k\left((h,\omega),(\tilde{h},\tilde{\omega})\right)$$

where

$$\begin{aligned}
M_k=2K_\ell&\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}+\frac{1}{(\lambda_{\min}/2)^{1/2}}\right) \\
&+2\mathbb{E}\left[K_\ell\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}+\frac{1}{(\lambda_{\min}/2)^{1/2}}\right)\,|\,X_k\right].
\end{aligned}$$

Note that by Lemma B.1.1(i), it holds almost surely that

$$\mathbb{E}\left[K_\ell\left(Y_k,r+\frac{2\overline{R}}{\lambda_{\min}}+\frac{1}{(\lambda_{\min}/2)^{1/2}}\right)\,|\,X_k\right]\le\frac{1}{\beta}\sqrt{\Gamma_{r+2\overline{R}/\lambda_{\min}+1/(\lambda_{\min}/2)^{1/2}}-1}.$$

Then by Lemma B.1.1(ii), M_k is uniformly subgaussian, so is M_k'.

Condition (ii): We want to verify that $\alpha_n(L,\sigma)\to0$ as $L\to\infty$ and show that in fact an upper bound $\tilde{\alpha}(L)$ on $\alpha_n(L,\sigma)$ which is independent of $x_{1:n}$, n, K and σ exists such that

$\tilde{\alpha}(L) \to 0$ still holds. Since $\alpha_n(L,\sigma)$ depends on the entropy numbers $H(\epsilon, T(\sigma), d)$ of the tube w.r.t. $d^2 = \frac{1}{n}\sum_k d_k^2$, first we need to estimate these entropy numbers. For $\gamma = (h,\omega)$, $\tilde{\gamma} = (\tilde{h}, \tilde{\omega})$, we have

$$\begin{aligned} d^2(\gamma,\tilde{\gamma}) &= \frac{1}{n}\sum_k d_{V,k}^2(h,\tilde{h}) + \frac{1}{n}\sum_k d_{\Lambda,k}^2(\omega,\tilde{\omega}) \\ &= \|h-\tilde{h}\|_n^2 + \|h-\tilde{h}\|_{\infty,n}^2 + \|\omega-\tilde{\omega}\|_n^2 \le 2\left(\|h-\tilde{h}\|_{\infty,n}^2 + \|\omega-\tilde{\omega}\|_n^2\right). \end{aligned}$$

Further, $d_\Lambda^2(\omega,\tilde{\omega}) = \|\omega-\tilde{\omega}\|_n^2$ and therefore by the choice $\Lambda_h = \Omega_{h,n}$ and $\lambda_h = 0$,

$$T(\sigma) = \{(h,\omega) \,:\, h \in \mathcal{H}, \omega \in \Omega_{h,n} \text{ s.t. } \|\omega\|_n \le \sigma\}\,.$$

Therefore, it suffices to estimate the metric entropy of $T(\sigma)$ at different scales w.r.t. the pseudo-norm $\|\cdot\|_T$ defined by $\|(h,\omega_{g,h})\|_T = \|h\|_{\infty,n} + \|\omega_{g,h}\|_n$. This is done in the following Proposition B.3.4, which also shows that the integrability assumption is satisfied (the proof is presented in the appendix):

Proposition B.3.4. *Let Assumptions 1 to 4 hold. Take $n \ge 1$, $K > 0$, $\epsilon > 0$, $1 \ge \sigma \ge \epsilon$ such that $K\sigma \le 1/2$. Then on $G_{\lambda_{\min}}$,*

$$H(\epsilon, T(\sigma), \|\cdot\|_T) \le \rho\log(\sigma/\epsilon) + \rho\log(241) + AH(\tfrac{\epsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n})$$

holds a.s. for some positive (non-random) constant A that depends only on K_h.

Furthermore, on $G_{\lambda_{\min}}$,

$$\int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\,du \le A'\sqrt{\rho} + \frac{A''}{\sigma}\,,$$

holds a.s. for some universal constant $A' > 0$ and some non-random constant A'' that depends on C_H and K_h only.

Now, $H(\epsilon,\sigma) = H(\epsilon, T(\sigma), d) \le CH(\epsilon, T(\sigma), \|\cdot\|_T)$ with some universal constant C, hence $H(uL\sigma, L\sigma) \le CH(uL\sigma, T(L\sigma), \|\cdot\|_T)$ and by the previous result,

$$\begin{aligned} \int_0^1 H^{1/2}(uL\sigma, L\sigma)du &\le C^{1/2}\int_0^1 H^{1/2}(uL\sigma, T(L\sigma), \|\cdot\|_T)\,du \\ &\le C'\left(1 + \frac{1}{\sigma}\right) \le \frac{2C'}{\sigma} \end{aligned}$$

where C' is a constant that is independent of L, n, K, σ and we assumed that $\sigma \le 1$. Hence,

$$\alpha_n(L,\sigma) \le \frac{2C'}{\sqrt{n}L\sigma^2} \le \frac{2C'}{L}$$

provided that $\sqrt{n}\sigma^2 \ge 1$. Thus, under this condition, $\alpha_n(L,\sigma) \to 0$ as $L \to \infty$, as required. Furthermore, the upper bound on $\alpha_n(L,\sigma)$ is independent of $x_{1:n}$, K, n and σ. Therefore, L_0 and C_0 can be selected independently of $x_{1:n}$, K, n and σ, finishing the verification of the conditions of Lemma B.3.2.

Therefore, using (B.4) we conclude that for any $L \geq L_0$, K, n, σ such that $\sqrt{n}\sigma^2 \geq 1$ and $K\sigma \leq 1/2$ and $K \geq 1$, for $[P_X]$ almost all $x_{1:n}$ such that $\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2$,

$$\mathbb{P}_{x_{1:n}}\Big(Q_n(L\sigma) \geq \sqrt{n}\Big) \leq \exp(-C_0 L^2\sigma^2 n).$$

Now, by the definition of $\mathbb{P}_{x_{1:n}}$,

$$\begin{aligned}
\mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) &= \int \mathbb{P}_{x_{1:n}}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) P_X(dx_{1:n}) \\
&= \int_{\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2} \mathbb{P}_{x_{1:n}}\left(Q_n(L\sigma) \geq \sqrt{n}\right) P_X(dx_{1:n}) \\
&\leq \int_{\lambda_{\min}(x_{1:n}) \geq \lambda_{\min}/2} \exp(-C_0 L^2\sigma^2 n) P_X(dx_{1:n}) \\
&\leq \exp(-C_0 L^2\sigma^2 n)\,,
\end{aligned}$$

where the second equality follows since $G_{\lambda_{\min}}$ is $X_{1:n}$-measurable.

Hence, by combining (B.2) and (B.3), using the definition of $Q_n(L\sigma)$ in (B.4) and choosing $L = L_0$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq \frac{L_0\sigma}{1 - KL_0\sigma}, G_{\lambda_{\min}}\right) &\leq \mathbb{P}\left(Q_n(L\sigma) \geq \sqrt{n}, G_{\lambda_{\min}}\right) \\
&\leq \exp(-C_0 L_0^2\sigma^2 n)\,.
\end{aligned}$$

Choosing $\sigma = 1/(2L_0)$ and $K = 1$, noting that $n \geq \sigma^{-4}$ then translates into $n \geq 16L_0^4$ gives that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\|g_{h,n} - \overline{g}_{h,n}\right\|_n \geq 1, G_{\lambda_{\min}}\right) \leq \exp(-C_0 n/4),$$

which is the desired result (we also used that $L_0 \geq 1$ by assumption and hence $\sigma \leq 1$ which gives that $\sqrt{n}\sigma \geq \sqrt{n}\sigma^2 \geq 1$).

B.3.1 Proof of the "Basic Inequality" (Lemma B.3.1)

We start with a uniform bound for the infinity norm of elements in $\Omega_{h,n}$. Let

$$K_\infty = \frac{1}{K(\lambda_{\min}/2)^{1/2}}.$$

Recall that $G_{\lambda_{\min}}$ is the event when $\hat{\lambda}_{\min} \geq \lambda_{\min}/2$.

Lemma B.3.5. *On the event $G_{\lambda_{\min}}$,*

$$\sup_{\omega\in\Omega_{h,n}} \|\omega\|_\infty < K_\infty.$$

Proof. Introduce $\|x\|_M^2 = x^\top M x$ for M positive definite. Let $\overline{\theta}_{h,n}$ be the parameter of $\overline{g}_{h,n}$. Thus,

$$|\omega(x)| = \frac{|\langle\phi(x), \theta - \overline{\theta}_{h,n}\rangle|}{1 + K\sqrt{\left\|\theta - \overline{\theta}_{h,n}\right\|_{\hat{G}}^2}} \leq \frac{\|\theta - \overline{\theta}_{h,n}\|_2}{1 + K\hat{\lambda}_{\min}^{1/2}\|\theta - \overline{\theta}_{h,n}\|_2} < \frac{1}{K\hat{\lambda}_{\min}^{1/2}} \leq K_\infty.$$

□

With this, we can state the proof of Lemma B.3.1:

Lemma B.3.1 ("Basic Inequality"). *Let Assumption 2 hold. There exists a constant η, such that on the event $G_{\lambda_{\min}}$, for any $h \in \mathcal{H}$,*

$$\eta \, \|\hat{\omega}_{h,n}\|_n^2 \leq \Delta_{h,n}(\overline{g}_{h,n}) - \Delta_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) \, .$$

Proof. The proof follows the ideas underlying the proof of Lemma 12.2 of the book of van de Geer [2000].

First, we will prove that $L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) \geq 0$. Note that by the definition of $g_{h,n}$, $L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(g_{h,n}) \geq 0$. Thus,

$$0 \leq L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(g_{h,n} - \overline{g}_{h,n} + \overline{g}_{h,n}) \leq \frac{1}{\alpha}(L_{h,n}(\overline{g}_{h,n}) - L_{h,n}((1-\alpha)\overline{g}_{h,n} + \alpha g_{h,n}))$$

for any $0 < \alpha \leq 1$, because of the convexity of $L_{h,n}$. Taking $\alpha = \frac{1}{1+K\left\|g_{h,n} - \overline{g}_{h,n}\right\|_n}$, the previous inequality gives

$$\frac{1}{\alpha}(L_{h,n}(\overline{g}_{h,n}) - L_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n})) \geq 0. \tag{B.6}$$

Now take $\epsilon > 0$ small enough so that it satisfies Assumption 2 and also $\frac{\epsilon}{K_\infty} \leq 1$. Then we have $\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) \geq \frac{\epsilon}{K_\infty}(\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}))$ because $\overline{g}_{h,n}$ is a minimizer of $\overline{L}_{h,n}$ (thus $\overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) > 0$) and thus

$$\begin{aligned} \overline{L}_{h,n}(\overline{g}_{h,n} + \hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) &\geq \overline{L}_{h,n}(\overline{g}_{h,n} + \frac{\epsilon}{K_\infty}\hat{\omega}_{h,n}) - \overline{L}_{h,n}(\overline{g}_{h,n}) \\ &\geq \frac{\epsilon^3}{K_\infty^2}\|\hat{\omega}_{h,n}\|_n^2. \end{aligned} \tag{B.7}$$

Here, the first inequality holds by the convexity of $\overline{L}_{h,n}$. The second inequality follows from Assumption 2 used with $a = \hat{\omega}_{h,n}\Big|_{X_{1:n}}$, once we verify that its conditions. That $a \in [-\epsilon, \epsilon]^n$ follows from Lemma B.3.5, while $a \in \mathrm{Im}(\Phi)$ follows since both $g_{h,n}|_{X_{1:n}}$ and $\overline{g}_{h,n}|_{X_{1:n}}$ satisfy this, by construction. Combining (B.6) and (B.7) gives the desired result. □

B.3.2 Proof of Proposition B.3.4

The result we want to prove is as follows:

Proposition B.3.4. *Let Assumptions 1 to 4 hold. Take $n \geq 1$, $K > 0$, $\epsilon > 0$, $1 \geq \sigma \geq \epsilon$ such that $K\sigma \leq 1/2$. Then on $G_{\lambda_{\min}}$,*

$$H(\epsilon, T(\sigma), \|\cdot\|_T) \leq \rho \log(\sigma/\epsilon) + \rho \log(241) + AH(\tfrac{\epsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n})$$

holds a.s. for some positive (non-random) constant A that depends only on K_h.

Furthermore, on $G_{\lambda_{\min}}$,

$$\int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\, du \leq A'\sqrt{\rho} + \frac{A''}{\sigma}\,,$$

holds a.s. for some universal constant $A' > 0$ and some non-random constant A'' that depends on C_H and K_h only.

We start by showing that the mapping $g, h \mapsto (h, \omega_{g,h})$ is Lipschitz w.r.t $\|\cdot\|_T$ ($g \in \mathcal{G}$, $h \in \mathcal{H}$) as this will allow us to bound the entropy of $T(\sigma)$ in terms of the entropy of $\mathcal{H}$ and the entropy of the union of balls in $\cup_{h\in\mathcal{H}}\Omega_{h,n}$, in particular $\cup_{h\in\mathcal{H}}\Omega_{h,n}(\sigma)$.

Let Assumption 4 hold Then, for any $K, \sigma > 0$ satisfying $K\sigma \leq 1/2$ and any (g_1, h_1), $(g_2, h_2) \in \mathcal{G} \times \mathcal{H}$ s.t. $\|\omega_{g_i,h_i}\|_n \leq \sigma$,

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_n \leq K_g \|g_1 - g_2\|_n + K_g K_h \|h_1 - h_2\|_{\infty,n} \tag{B.8}$$

holds a.s. on the event $G_{\lambda_{\min}}$, where $K_g = 4\sqrt{2}$. The constant K_h appearing in the bound is the Lipschitz constant defined in Assumption 4.

Proof. Take any $(g_1, h_1), (g_2, h_2) \in \mathcal{G} \times \mathcal{H}$ with the required property. By the triangle inequality, we have

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_T \leq \|\omega_{g_1,h_1} - \omega_{g_2,h_1}\|_T + \|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_T \ .$$

Let us consider bounding $\|\omega_{g_1,h_1} - \omega_{g_2,h_1}\|_T$ as the first step. To minimize clutter, introduce $h = h_1$, $\omega_i = \omega_{g_i,h}$, $i = 1, 2$. With this, our goal is to bound $\|\omega_1 - \omega_2\|_T$.

We have

$$\begin{aligned}
|\omega_1(x) - \omega_2(x)| &= \left| \frac{(g_1 - \overline{g}_{h,n})(x)}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{(g_2 - \overline{g}_{h,n})(x)}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right| \\
&= \Bigg| g_1(x) \left(\frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right) \\
&\quad + \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n}(g_1 - g_2)(x) \\
&\quad - \overline{g}_{h,n}(x) \left(\frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right) \Bigg|.
\end{aligned}$$

By the triangle inequality,

$$\left| \frac{1}{1 + K\left\|g_1 - \overline{g}_{h,n}\right\|_n} - \frac{1}{1 + K\left\|g_2 - \overline{g}_{h,n}\right\|_n} \right| \leq K \|g_1 - g_2\|_n \ .$$

Thus,

$$|\omega_1(x) - \omega_2(x)| \leq K|g_1(x) - \overline{g}_{h,n}(x)| \ \|g_1 - g_2\|_n + |(g_1 - g_2)(x)|$$

and therefore,

$$\begin{aligned}
n\|\omega_1 - \omega_2\|_n^2 &\leq \sum_{i=1}^n \left\{ K|g_1(X_i) - \overline{g}_{h,n}(X_i)| \ \|g_1 - g_2\|_n + |(g_1 - g_2)(X_i)| \right\}^2 \\
&\leq 2\sum_{i=1}^n \left\{ K^2|g_1(X_i) - \overline{g}_{h,n}(X_i)|^2 \ \|g_1 - g_2\|_n^2 + |(g_1 - g_2)(X_i)|^2 \right\} \\
&\leq 2n(K^2\left\|g_1 - \overline{g}_{h,n}\right\|_n^2 + 1) \ \|g_1 - g_2\|_n^2 .
\end{aligned}$$

By Equation (B.2),

$$\left\|g_1 - \overline{g}_{h,n}\right\|_n = \frac{\|\omega_1\|_n}{1 - K\|\omega_1\|_n}\,.$$

Since $\omega_1 \in \Omega_{h,n}(\sigma)$, $\|\omega_1\|_n \leq \sigma$ and $K\sigma < 1$ by assumption, $\left\|g_1 - \overline{g}_{h,n}\right\|_n \leq \frac{\sigma}{1-K\sigma}$. Combining this with the bound on $n\|\omega_1 - \omega_2\|_n^2$, after simplification we get

$$\begin{aligned}\|\omega_1 - \omega_2\|_n &\leq \sqrt{2 + 2\left(\frac{K\sigma}{1-K\sigma}\right)^2}\,\|g_1 - g_2\|_n \\ &\leq \sqrt{2}\left(1 + \frac{K\sigma}{1-K\sigma}\right)\|g_1 - g_2\|_n \\ &= \frac{2\sqrt{2}}{1-K\sigma}\,\|g_1 - g_2\|_n \leq K_g\,\|g_1 - g_2\|_n\,,\end{aligned} \tag{B.9}$$

where $K_g = 4\sqrt{2}$ in the last two steps we used that by assumption $K\sigma \leq 1/2$.

Let us now consider bounding

$$\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n\,.$$

Noticing that apart from a sign, $\overline{g}_{h,n}$ and g play a symmetric role in the definition of $\omega_{g,h}$, following the derivation in the first part we get that, similarly to (B.9),

$$\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n \leq \frac{2\sqrt{2}}{1-K\sigma}\,\|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n \leq K_g\,\|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n\,.$$

Since by Assumption 4, $\|\bar{g}_{h_1,n} - \bar{g}_{h_2,n}\|_n \leq K_h\,\|h_1 - h_2\|_{\infty,n}$ holds a.s. on $G_{\lambda_{\min}}$, we get

$$\|\omega_{g_2,h_1} - \omega_{g_2,h_2}\|_n \leq K_g K_h\,\|h_1 - h_2\|_{\infty,n}\,.$$

Putting together the bounds obtained, we get that on the event $G_{\lambda_{\min}}$,

$$\|\omega_{g_1,h_1} - \omega_{g_2,h_2}\|_n \leq K_g\,\|g_1 - g_2\|_n + K_g K_h\,\|h_1 - h_2\|_{\infty,n}$$

as required. □

With this, we can state the proof of Proposition B.3.4.

Proof of Proposition B.3.4. We can write

$$T(\sigma) = \cup_{h\in\mathcal{H}}\{h\} \times \Omega_{h,n}(\sigma)\,,$$

where

$$\Omega_{h,n}(\sigma) = \{\omega \in \Omega_{h,n} \,:\, \|\omega\|_n \leq \sigma\}\,.$$

We first show that

$$H(\epsilon, T(\sigma), \|\cdot\|_T) \leq H(\tfrac{\epsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) + H(\tfrac{\epsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n)\,, \tag{B.10}$$

where $\Omega_n(\sigma) = \cup_{h\in\mathcal{H}}\Omega_{h,n}(\sigma)$. In short, this follows since $T(\sigma) \subset \mathcal{H} \times \Omega_n(\sigma)$ and since, by definition, $\|\cdot\|_T$ is obtained by "summing" $\|\cdot\|_{\infty,n}$ and $\|\cdot\|_n$.

In details, we have: Let C be an integer s.t. $C \geq \exp\left(H(\epsilon/2, \mathcal{H}, \|\cdot\|_{\infty,n})\right)$. Then, there exists $\{h_1, \ldots, h_C\} \subset \mathcal{H}$ such that for any $h \in \mathcal{H}$, $\|h - h_i\|_{\infty,n} \leq \epsilon/2$ for some $i \in \{1, \ldots, C\}$. Similarly, let D be an integer s.t.

$$D \geq \exp(H(\tfrac{\epsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n)) \geq \max_{1 \leq i \leq C} \exp(H(\tfrac{\epsilon}{2}, \Omega_{h_i,n}(\sigma), \|\cdot\|_n))$$

and $\{\omega_1, \ldots, \omega_D\} \subset \Omega_n(\sigma)$ be an $\epsilon/2$-net of $\Omega_n(\sigma)$ w.r.t. $\|\cdot\|_n$. Then,

$$\{(h_i, \omega_j) \,:\, 1 \leq i \leq C, 1 \leq j \leq D\}$$

is an ϵ-net of $T(\sigma)$: To show this pick any $(h, \omega) \in T(\sigma)$. Then, take the index i such that $\|h - h_i\|_{\infty,n} \leq \epsilon/2$ and take the index j such that $\|\omega - \omega_j\|_n \leq \epsilon/2$. Then,

$$\|(h, \omega) - (h_i, \omega_j)\|_T = \|h - h_i\|_{\infty,n} + \|\omega - \omega_j\|_n \leq \epsilon,$$

as required. This shows that (B.10) indeed holds.

Next, we bound $H(\epsilon, \Omega_n(\sigma), \|\cdot\|_n)$. We have

$$\begin{aligned} \Omega_n(\sigma) &= \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}, \|\omega_{g,h}\|_n \leq \sigma\right\} \\ &\subset \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}, \left\|g - \overline{g}_{h,n}\right\|_n \leq \tfrac{\sigma}{1-K\sigma}\right\}, \end{aligned} \tag{B.11}$$

where the containment follows since by Equation (B.2), $\left\|g - \overline{g}_{h,n}\right\|_n = \frac{\|\omega_{g,h}\|_n}{1-K\|\omega_{g,h}\|_n}$. For $s \geq 0$ define

$$\mathcal{G}_h(s) = \left\{g \in \mathcal{G} \,:\, \left\|g - \overline{g}_{h,n}\right\|_n \leq s\right\}.$$

Pick $\hat{\mathcal{H}} \subset \mathcal{H}$ and an arbitrary "discretization" map $N : \mathcal{H} \to \hat{\mathcal{H}}$. We claim that on $G_{\lambda_{\min}}$,

$$\Omega_n(\sigma) \subset \left\{\omega_{g,h} \,:\, h \in \mathcal{H}, g \in \mathcal{G}_{N(h)}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h) - h\|_{\infty,n}\right)\right\} \text{ a.s.} \tag{B.12}$$

By (B.11) it suffices to show that for any $h \in \mathcal{H}$ and $g \in \mathcal{G}_h\left(\frac{\sigma}{1-K\sigma}\right)$,

$$g \in \mathcal{G}_{N(h)}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h) - h\|_{\infty,n}\right) \tag{B.13}$$

also holds true. For brevity introduce $h' = N(h)$. Thanks to the choice g and Assumption 4,

$$\left\|g - \overline{g}_{h',n}\right\|_n \leq \left\|g - \overline{g}_{h,n}\right\|_n + \left\|\overline{g}_{h,n} - \overline{g}_{h',n}\right\|_n \leq \frac{\sigma}{1 - K\sigma} + K_h \|h - h'\|_{\infty,n}$$

holds a.s. on $G_{\lambda_{\min}}$, which shows that (B.13) indeed holds.

The following statements holds a.s. on $G_{\lambda_{\min}}$ – hence we will not mention this condition to minimize clutter. If $\hat{\mathcal{H}}$ is an $\epsilon/(2K_gK_h)$-net of $\mathcal{H}$ w.r.t. $\|\cdot\|_{\infty,n}$ and $N(h) = \arg\min_{h' \in \mathcal{H}} \|h - h'\|_{\infty,n}$ then $K_h \|N(h) - h\|_{\infty,n} \leq \epsilon/(2K_g) \leq \epsilon/2$ and therefore for any $h' \in \hat{\mathcal{H}}$,

$$\mathcal{G}_{h'}\left(\tfrac{\sigma}{1-K\sigma} + K_h \|N(h) - h\|_{\infty,n}\right) \subset \mathcal{G}_{h'}\left(2\sigma + \frac{\epsilon}{2}\right).$$

For each $h \in \hat{\mathcal{H}}$, let $\hat{\mathcal{G}}_{h'}$ be an $\epsilon/2K_g$-net of $\mathcal{G}_{h'}\left(2\sigma + \frac{\epsilon}{2}\right)$. We claim that

$$S = \left\{\omega_{g',h'} \,:\, h' \in \hat{\mathcal{H}},\ g' \in \hat{\mathcal{G}}_{h'}\right\}$$

is an ϵ-net of $\Omega_n(\sigma)$ w.r.t. $\|\cdot\|_n$. Indeed, let $\omega = \omega_{g,h} \in \Omega_n(\sigma)$ arbitrary. Let h' be the nearest neighbor of h in $\hat{\mathcal{H}}$ w.r.t. $\|\cdot\|_{\infty,n}$ and let g' be the nearest neighbor of g in $\hat{\mathcal{G}}_{h'}$ w.r.t. $\|\cdot\|_n$. Note that $g \in \mathcal{G}_{h'}(2\sigma + \epsilon/2)$. Then, by Section B.3.2,

$$\|\omega_{g,h} - \omega_{g',h'}\|_n \leq K_g \|g - g'\|_n + K_g K_h \|h - h'\|_{\infty,n} .$$

Now, because $g \in \mathcal{G}_{h'}(2\sigma+\epsilon/2)$ and $\hat{\mathcal{G}}_{h'}$ is an $\epsilon/(2K_k)$-net of this set, we have $K_g \|g - g'\|_n \leq \epsilon/2$. Similarly, by the choice of $\mathcal{H}$, $\|h - h'\|_{\infty,n} \leq \epsilon/2$, showing that S is indeed an ϵ-net of $\Omega_n(\sigma)$. Note that the cardinality of S can be bounded by

$$|S| \leq |\hat{\mathcal{H}}| \max_{h' \in \hat{\mathcal{H}}} |\hat{\mathcal{G}}_{h'}| .$$

Hence,

$$H(\epsilon, \Omega_n(\sigma), \|\cdot\|_n) \leq H(\tfrac{\epsilon}{2K_g}, \mathcal{G}_{h_0}(2\sigma + \epsilon/2), \|\cdot\|_n) + H(\tfrac{\epsilon}{2K_g K_h}, \mathcal{H}, \|\cdot\|_{\infty,n}) ,$$

for an arbitrary $h_0 \in \mathcal{H}$, where we used that $\|\cdot\|_n$ is translation invariant.

Combining this with (B.10), we get

$$\begin{aligned} H(\epsilon, T(\sigma), \|\cdot\|_T) &\leq H(\tfrac{\epsilon}{2}, \Omega_n(\sigma), \|\cdot\|_n) + H(\tfrac{\epsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) \\ &\leq H(\tfrac{\epsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma + \epsilon/2), \|\cdot\|_n) + H(\tfrac{\epsilon}{4K_g K_h}, \mathcal{H}, \|\cdot\|_{\infty,n}) \\ &\quad + H(\tfrac{\epsilon}{2}, \mathcal{H}, \|\cdot\|_{\infty,n}) \\ &\leq H(\tfrac{\epsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma + \epsilon/2), \|\cdot\|_n) + AH(\tfrac{\epsilon}{A}, \mathcal{H}, \|\cdot\|_{\infty,n}) \end{aligned} \tag{B.14}$$

for A large enough.

By Corollary 2.6 in the book of van de Geer [2000], $H(\epsilon, \mathcal{G}_{h_0}(\sigma), \|\cdot\|_n) \leq \rho \log(\frac{4\sigma+\epsilon}{\epsilon})$ a.s.. Hence,

$$H(\tfrac{\epsilon}{4K_g}, \mathcal{G}_{h_0}(2\sigma + \epsilon/2), \|\cdot\|_n) \leq \rho \log\left(\frac{32\sigma K_g + 8K_g\epsilon + \epsilon}{\epsilon}\right) \leq \rho \log(241) + \rho \log(\sigma/\epsilon), \; a.s..$$

Here the second inequality follows from bounding the ϵ in numerator by σ (since $\sigma \geq \epsilon$), and $K_g < 6$. Combining this with (B.14) finishes the proof of the first statement.

To prove the second part, note that $\int_0^1 (-\log(x))^{1/2}\, dx = \sqrt{\pi}/2$. Thus, with $\mathcal{I}(\sigma, \mathcal{H}) = \int_0^1 H^{1/2}(u\sigma, \mathcal{H}, \|\cdot\|_{\infty,n})\, du$,

$$\begin{aligned} \int_0^1 H^{1/2}(u\sigma, T(\sigma), \|\cdot\|_T)\, du &\leq \rho^{\frac{1}{2}} \int_0^1 \log^{\frac{1}{2}}(\tfrac{1}{u})\, du + \rho^{\frac{1}{2}} \log^{\frac{1}{2}}(241) + A^{\frac{1}{2}} \mathcal{I}(\tfrac{\sigma}{A}, \mathcal{H}) \\ &\leq \rho^{\frac{1}{2}} \sqrt{\pi}/2 + \rho^{\frac{1}{2}} \log^{\frac{1}{2}}(241) + A^{\frac{1}{2}} \mathcal{I}(\tfrac{\sigma}{A}, \mathcal{H}). \end{aligned}$$

Now, note that

$$\begin{aligned} \int_0^1 H^{1/2}(u\sigma, \mathcal{H}, \|\cdot\|_{\infty,n})\, du &= \frac{1}{\sigma} \int_0^\sigma H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\, dv \\ &\leq \frac{1}{\sigma} \int_0^1 H^{1/2}(v, \mathcal{H}, \|\cdot\|_{\infty,n})\, dv \leq C_H/\sigma, \end{aligned}$$

where the first inequality follows since $\sigma \leq 1$, while the last inequality follows by Assumption 3. The desired result follows by choosing $A' = \sqrt{\pi}/2 + \log^{\frac{1}{2}}(241)$ and $A'' = A^{3/2} C_H$. □

B.4 Proof of Lemma 3.6.2

Let (Λ, d) be a pseudo-metric space and for $u > 0$ let $B_{\Lambda,d}(\lambda, u)$ be the d-ball in Λ that has radius u and is centered at λ. We will allow d to be replaced with a pseudo-norm meaning the ball where the pseudo-metric is defined by the chosen pseudo-norm. The theorem of van de Geer bounds the tails of the suprema of centered, Lipschitz empirical processes of Λ over balls of Λ:

Theorem B.4.1 (Theorem 3.3 of van de Geer [1990]). *Let (Λ, d) be a pseudo-metric space with $d^2 = (1/n)\sum_{k=1}^n d_k^2$ where $d_1, \dots, d_n$ pseudo-metrics on Λ. Let $U_1, \dots, U_n$ be real-valued, independent, centered process on Λ such that for $Z_n = \frac{1}{\sqrt{n}}\sum U_k$, $Z_n(\lambda_0) = 0$ for some $\lambda_0 \in \Lambda$. For $u > 0$, denote by $H(u;\sigma) = H(u, B_{\Lambda,d}(\lambda_0, \sigma), d)$, the u-entropy of the ball $B_{\Lambda,d}(\lambda_0, \sigma)$. Assume further that $|U_k(\lambda) - U_k(\lambda')| \le M_k d_k(\lambda, \lambda')$ with $M_k \ge 0$ random such that $\mathbb{E}[\exp(|\beta M_k|^2)] \le \Gamma < \infty$ for some positive constants β and Γ. Then, there exist $\alpha, \eta, C_1, C_2 > 0$ depending only on β and Γ such that*

$$\mathbb{P}\left(\sup_{\lambda \in B_{\Lambda,d}(\lambda_0,\sigma)} |Z_n(\lambda)| \ge t\right) \le 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right)$$

holds for any $t > 0$ and $\sigma > 0$ that satisfies $t/\sigma > C_1$ and $t > C_2 \int_0^{t_0} \sqrt{H(u;\sigma)}\, du$ where $t_0 \ge \inf\{u : H(u;\sigma) \le \eta t^2/\sigma^2\}$.

Let us now turn to the proof of Lemma 3.6.2.

Proof of Lemma 3.6.2. We denote the probability space that holds $(X_1, Y_1), \dots, (X_n, Y_n)$ by $(W, \mathcal{W}, \mathbb{P})$. Note that with no loss of generality, we can assume that $(W, \mathcal{W})$ is a Borel-space (this is because all our random variables leave in complete, separable metric spaces). For $x_1, \dots, x_n \in \mathcal{X}$, let $x_{1:n} = (x_1, \dots, x_n)$. Similarly, let $X_{1:n} = (X_1, \dots, X_n)$. Define $(\mathbb{P}_{x_{1:n}})_{x_{1:n} \in \mathcal{X}^n}$ to be the disintegration of the probability measure $\mathbb{P}$ with respect to $X_{1:n}$, also known as the regular conditional probability measure obtained from $\mathbb{P}$ by conditioning on $X_{1:n}$.[1] The expectation operator corresponding to $\mathbb{P}_{x_{1:n}}$ will be denoted by $\mathbb{E}_{x_{1:n}}$. Note that, by the definition of $\mathbb{P}_{x_{1:n}}$, for any random variable Z on $(W, \mathcal{W}, \mathbb{P})$, $\mathbb{E}_{X_{1:n}}[Z] = \mathbb{E}[Z|X_{1:n}]$ holds everywhere and in particular, for a measurable function $s : \mathcal{Y} \to \mathbb{R}$, $\mathbb{E}_{X_{1:n}}[s(Y_k)] = \mathbb{E}[s(Y_k)|X_{1:n}] = \mathbb{E}[s(Y_k)|X_k]$, where the last equality holds since by assumption (X_t, Y_t) is i.i.d.

Let $U_{k,x_{1:n}}(f) = \Delta_{k,x_{1:n}}(f) - \Delta_{k,x_{1:n}}(f^*)$, where

$$\Delta_{k,x_{1:n}}(f) = \ell(Y_k, f(x_k)) - \mathbb{E}_{x_{1:n}}[\ell(Y_k, f(x_k))]\,, \qquad f \in \Lambda\,.$$

[1] The defining properties of $(\mathbb{P}_{x_{1:n}})$ are that for each $x_{1:n} \in \mathcal{X}^n$, $\mathbb{P}_{x_{1:n}}$ is a probability measure on $(W, \mathcal{W})$ concentrated on $\{X_{1:n} = x_{1:n}\}$, $x_{1:n} \mapsto \mathbb{P}_{x_{1:n}}$ is measurable and for any $f : (W, \mathcal{W}) \to [0, \infty)$ measurable function $\int f(w)\mathbb{P}(dw) = \int(\int f(w)\mathbb{P}_{x_{1:n}}(dw))P_{X_{1:n}}(dx_{1:n})$. The existence of $(\mathbb{P}_{x_{1:n}})$, which is also called a regular conditional probability distribution is ensured thanks to the assumption that $(W, \mathcal{W})$ is Borel. Moreover, $(\mathbb{P}_{x_{1:n}})$ is unique up to an almost sure equivalence in the sense that if $(\hat{\mathbb{P}}_{x_{1:n}})$ is another disintegration of $\mathbb{P}$ w.r.t. $X_{1:n}$ then $P_X(\{x_{1:n} : \mathbb{P}_{ux} \ne \hat{\mathbb{P}}_{x_{1:n}}\}) = 0$. For background on disintegration and conditioning, the reader is referred to Chang and Pollard [1997].

Note that $U_{k,x_{1:n}}$'s are independent, centered processes over $W_{x_{1:n}}$. Moreover, let $Z_{n,x_{1:n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} U_{k,x_{1:n}}$. By construction, $Z_{n,x_{1:n}}(f^*) = 0$. We now show that it is enough to study the deviations of the suprema of $Z_{n,x_{1:n}}(f)$ over the probability spaces $W_{x_{1:n}}$.

We have

$$\overline{\Delta}_n(f) = L_n(f) - \overline{L}_n(f) = \frac{1}{n} \sum_{k=1}^{n} \Delta_{k,X_{1:n}}(f)$$

and so

$$\sqrt{n}(\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)) = Z_{n,X_{1:n}}(f)\,.$$

By the construction of $\mathbb{P}_{x_{1:n}}$, for $z \geq 0$,

$$\mathbb{P}(\overline{\Delta}_n^*(\mathcal{C}) \geq z) = \int \mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) \geq z) P_{X_{1:n}}(dx_{1:n})\,, \tag{B.15}$$

and

$$\mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) \geq z/\sqrt{n}) = \mathbb{P}_{x_{1:n}}(\sqrt{n} \sup_{f \in \mathcal{C}} |\overline{\Delta}_n(f) - \overline{\Delta}_n(f^*)| \geq z) = \mathbb{P}_{x_{1:n}}(\sup_{f \in \mathcal{C}} Z_{n,x_{1:n}}(f) \geq z)\,. \tag{B.16}$$

Let $\Lambda = \mathcal{C}$, $d_{k,x_{1:n}}(f,f') = |f(x_k) - f'(x_k)|$ and $d^2_{x_{1:n}}(f,f') = \frac{1}{n}\sum_{k=1}^{n} d^2_{k,x_{1:n}}(f,f')$. By construction, $d_{x_{1:n}}(f,f') = \|f - f'\|_n$. Since for any $f = h + g \in \mathcal{C}$,

$$\|f - f^*\|_n = \|h - h^*\|_n + \|g - g^*\|_n \leq \|h - h^*\|_\infty + \|g - g^*\|_\infty \leq 2(r + U) =: \sigma\,,$$

thus, $\mathcal{C} \subset B_{\Lambda, d_{x_{1:n}}}(f^*, \sigma) \subset \Lambda = \mathcal{C}$ and

$$\mathbb{P}_{x_{1:n}}(\overline{\Delta}_n^*(\mathcal{C}) > z/\sqrt{n}) = \mathbb{P}_{x_{1:n}}\left(\sup_{f \in B_{\Lambda, d_{x_{1:n}}}(f^*,\sigma)} Z_{n,x_{1:n}}(f) > z \right)\,. \tag{B.17}$$

Thus, it remains to bound this latter probability. Fix $x_{1:n} \in \mathcal{X}^n$ such that

$$\mathbb{E}_{x_{1:n}}[\exp((\beta K_\ell(Y,c)^2)] \leq \Gamma_c, \text{ for all } c > 0\,. \tag{B.18}$$

Let us now apply Theorem B.4.1 to $W_{x_{1:n}} = (W, \mathcal{W}, \mathbb{P}_{x_{1:n}})$ with Λ, $(d_{k,x_{1:n}})$ and $(U_{k,x_{1:n}})$ $(k = 1, \ldots, n)$, as defined above. To verify the uniform subgaussian property of the Lipschitz coefficient of $U_{k,x_{1:n}}$, note that for $f, f' \in \mathcal{C}$, by Assumption 1(iii),

$$\begin{aligned}
|U_{k,x_{1:n}}(f) - U_{k,x_{1:n}}(f')| &= |\Delta_{k,x_{1:n}}(f) - \Delta_{k,x_{1:n}}(f')| \\
&\leq |\ell(Y_k, f(x_k)) - \ell(Y_k, f'(x_k))| + |\mathbb{E}_{x_{1:n}}[\ell(Y_k, f(x_k)) - \ell(Y_k, f'(x_k))]| \\
&\leq K_l(Y_k, r+U)|f(x_k) - f'(x_k)| + \mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)]|f(x_k) - f'(x_k)|\,.
\end{aligned}$$

By Lemma B.1.1(i), $\mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)] \leq \frac{1}{\beta}\sqrt{\Gamma_{r+U} - 1}$ and so by part (ii) of the same lemma, $K_l(Y_k, r+U) + \mathbb{E}_{x_{1:n}}[K_l(Y_k, r+U)]$ is subgaussian, with parameters β' and Γ' only depending on $r + U$.

Therefore, from Theorem B.4.1 we conclude that there exists $C_1, C_2, \eta > 0$ such that for any $t > 0$ satisfying $\eta t^2/\sigma^2 \geq H(1;\sigma)$, $t > C_1\sigma$ and $t > C_2 \int_0^1 \sqrt{H(u;\sigma)}\, du$, it holds that

$$\mathbb{P}_{x_{1:n}}\left(\sup_{f\in\mathcal{C}} |Z_{n,x_{1:n}}(f)| \geq t\right) = \mathbb{P}_{x_{1:n}}\left(\sup_{f\in B_{\Lambda,d_{x_{1:n}}}(f^*,\sigma)} |Z_{n,x_{1:n}}(f)| \geq t\right) \leq 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right). \tag{B.19}$$

It still remains to check that $H(1,\sigma)$ and $\int_0^1 \sqrt{H(u;\sigma)}\, du$ are finite (otherwise the result is vacuous). By definition, $H(u;\sigma) = H(u, B_{\Lambda,d_{x_{1:n}}}(f^*,\sigma), d_{x_{1:n}}) = H(u,\mathcal{C},d_{x_{1:n}}) = H(u,\mathcal{C},\|\cdot\|_n)$. Hence, by Lemma B.1.2, $\int_0^1 H^{1/2}(u;\sigma)\, du \leq 2C_H + 2C_G(U)$. Noting that $H(u;\sigma)$ is monotonically decreasing in u, we calculate $H^{1/2}(1;\sigma) \leq \int_0^1 H^{1/2}(u;\sigma)\, du \leq 2C_H + 2C_G(U)$ and so $H(1;\sigma) \leq (2C_H + 2C_G(U))^2 < \infty$. We conclude that (B.19) holds for any $t \geq t_{\min} := \max\{C_1\sigma, C_2(2C_H + 2C_G(U)), (2C_H + 2C_G(U))\sigma\eta^{-1/2}\}$.

Since by Assumption 1(iii), (B.18) holds $[P_X]$-almost surely, combining (B.15), (B.17) and (B.19), we get

$$\mathbb{P}\left(\overline{\Delta}_n^*(\mathcal{C}) \geq t/\sqrt{n}\right) \leq 2\exp\left(-\frac{\alpha t^2}{\sigma^2}\right). \tag{B.20}$$

Inverting this inequality, we see that for any $0 < \delta < 1$ such that $\log(2/\delta) \geq t_{\min}^2\alpha/\sigma^2$, with probability at least $1-\delta$,

$$\overline{\Delta}_n^*(\mathcal{C}) \leq \sigma\sqrt{\frac{\log\frac{2}{\delta}}{\alpha n}},$$

finishing the proof. □

B.5 Proof of Theorem 3.6.2

Theorem 3.6.2. *Let $\phi = (\phi_1, \ldots, \phi_n) : \mathbb{R}^n \to \mathbb{R}^n$, denote $\{(\phi_1(a_1), \ldots, \phi_n(a_n)) : a \in A\}$ by $\phi \circ A$ for $A \subset \mathbb{R}^n$. Assume that all the component functions ϕ_i are L-Lipschitz over A. Then, $\mathfrak{R}(\phi \circ A) \leq L\mathfrak{R}(A)$.*

Proof. We follow the proof of Theorem 11.9 in [Rakhlin and Sridharan, 2014] and write

$$\begin{aligned}
n\mathfrak{R}(\phi\circ A) &= \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^n \sigma_i\phi_i(a_i)\right] \\
&= \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i) + \phi_n(a_n)\middle|\,\sigma_n = 1\right]\right. \\
&\qquad \left. + \mathbb{E}\left[\sup_{b\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(b_i) - \phi_n(b_n)\middle|\,\sigma_n = -1\right]\right\} \\
&= \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i)+\phi_i(b_i)) + (\phi_n(a_n) - \phi_n(b_n))\right]\right\} \\
&\leq \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i)+\phi_i(b_i)) + L|a_n - b_n|\right]\right\}.
\end{aligned}$$

Now assume that some (a^*, b^*) achieves the supremum (the proof when the supremum is not achieved is easy once we know how to prove the statement for the case when the supremums involved are all achieved). If $a_n^* \geq b_n^*$, the absolute value can be removed. Otherwise, (b^*, a^*) will achieve the same supremum, and again the absolute value can be removed. Thus, the last expression is bounded by

$$\begin{aligned}
&\frac{1}{2}\left\{\mathbb{E}\left[\sup_{a,b\in A}\sum_{i=1}^{n-1}\sigma_i(\phi_i(a_i)+\phi_i(b_i))+L(a_n-b_n)\right]\right\} \\
&\qquad = \frac{1}{2}\left\{\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+La_n|\,\sigma_n=1\right]+\mathbb{E}\left[\sup_{b\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(b_i)-Lb_n|\,\sigma_n=-1\right]\right\} \\
&\qquad = \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+L\sigma_n a_n\right].
\end{aligned}$$

Continuing this way,

$$\begin{aligned}
\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-1}\sigma_i\phi_i(a_i)+L\sigma_n a_n\right] &\leq \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n-2}\sigma_i\phi_i(a_i)+L(\sigma_{n-1}a_{n-1}+\sigma_n a_n)\right] \\
&\leq L\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right],
\end{aligned}$$

thus finishing the proof. □

Chapter 4

Online Linear Prediction

We switch to the domain of online learning in this chapter. Much work has been devoted recently to studying learning algorithms in the online learning framework [Cesa-Bianchi and Lugosi, 2006]. This framework usually makes no probabilistic assumptions about the data generating mechanism.However, by following a minimax approach, results proven, at least initially, were rather conservative and the algorithm designs also failed to take advantage of more regular, "nice" data. Recently, the design of adaptive algorithms that perform better in the presence of "nice" data has received an increasing attention. In this chapter, we study one of the simplest but fundamental setting, online linear prediction, together with the simplest possible online learning method, "Follow the Leader" (FTL) algorithm. Our algorithmcentric study reveals interesting previously unknown features of data that can help to make FTL perform exceptionally well. A lower bound is presented to show that our characterization is tight. One problem with FTL is that it fails to meet the worst-cse guarantee that other algorithms can achieve. To remedy this, we present adaptive algorithms that simultaneously enjoys the worst-case guarantees and the bound available for FTL.

This chapter is organized as follows: We introduce the background and existing results in Section 4.1. Preliminaries and problem setup are given in Section 4.2. A non-stochastic analysis of the FTL algorithm is presented in Section 4.3. In particular our main result is presented in Section 4.3.1 which shows that FTL achieves fast learning rate in the online prediction problem on a curved constraint set. Section 4.3.2 is devoted to a matching asymptotic lower bound, showing that our previous result is tight. Section 4.3.3 is devoted to the case when the constraint set is a polytope (thus non-curved). We then propose some adaptive algorithms in Section 4.4. Lastly, Section 4.5 presents some experimental results. The chapter is concluded in Section 4.6.

The results of this chapter have appeared in our NIPS paper [Huang et al., 2016].

4.1 Introduction

The online learning framework makes minimal assumptions about the data generating mechanism, while still allowing one to replicate results of the statistical framework through online-to-batch conversions [Cesa-Bianchi et al., 2004]. However, the potential downside of such an approach is that if the data-generating mechanism is treated as an adversary, as is done more often than not, the results will have a worst-case flavour. It also encourages the design of algorithms that, while meeting the strong worst-case guarantee, fail to take advantage of "nicer" data. Also, it is hard to argue that data resulting from passive data collection, such as weather data, would ever be adversarially generated (though it is equally hard to defend that such data satisfies the type of stochastic assumptions often used in theoretical studies). Realizing this issue, during recent years, much work has been devoted to understanding what regularities and how can lead to faster learning speed. For example, much work has been devoted to showing that faster learning speed (smaller "regret") can be achieved in the online convex optimization setting with various additional assumptions that can be exploited in an adaptive manner, such as: (i) the loss functions are "curved" such as when the loss functions are strongly convex or exp-concave; (ii) the losses show small variations; (iii) the best prediction in hindsight has a small total loss [Merhav and Feder, 1992, Freund and Schapire, 1997, Gaivoronski and Stella, 2000, Cesa-Bianchi and Lugosi, 2006, Hazan et al., 2007, Bartlett et al., 2007, Kakade and Shalev-Shwartz, 2009, Orabona et al., 2012, Rakhlin and Sridharan, 2013, van Erven et al., 2015, Foster et al., 2015].

In this chapter we contribute to this growing literature by studying online linear prediction and the follow the leader (FTL) algorithm. Online linear prediction is arguably the simplest yet most fundamental of all the learning settings. It lies at the heart of online convex optimization, while it also serves as an abstraction of core learning problems such as prediction with expert advice. FTL, the online analogue of empirical risk minimization of statistical learning, is the simplest learning strategy, one can think of. Although the linear setting of course removes the possibility of exploiting the curvature of losses, as we will see, there are multiple ways online learning problems can present data that allows for small regret, even for FTL. As is it well known, in the worst case, FTL suffers a linear regret (e.g., Example 2.2 of Shalev-Shwartz [2012]), which is why FTL is sometimes discarded. The approach taken here is to acknowledge this, but also to take a more thorough look at FTL to see what FTL can offer in non-worst-case scenarios.

Our motivation comes from the simple observation that, for prediction over the simplex, when the loss vectors are selected independently of each other from a distribution with a bounded support with a nonzero mean, FTL quickly locks onto selecting the loss-minimizing vertex of the simplex, achieving finite expected regret. In this case, FTL is arguably an excellent algorithm. In fact, FTL is shown to be the minimax optimizer for the binary

losses in the stochastic expert setting in the paper of Kotłowski [2016]. Thus, we ask the question of whether there are other regularities that allow FTL to achieve nontrivial performance guarantees. Our main result shows that when the decision set (or constraint set) has a sufficiently "curved" boundary (equivalently, if this set is strongly convex) and the average linear loss is bounded away from 0, FTL is able to achieve logarithmic regret even in the adversarial setting, thus opening up a new way to prove fast rates not based on the curvature of losses, but on that of the boundary of the constraint set and non-singularity of the linear loss. In a matching lower bound we show that our regret bound is essentially unimprovable. We also show an alternate bound for polytope constraint sets, which allows us to prove that (under certain technical conditions) for stochastic problems the expected regret of FTL will be finite. To finish, we use the so-called ($\mathcal{A}$, $\mathcal{B}$)-prod algorithm of Sani et al. [2014] to design an algorithm that adaptively interpolates between the worst-case $O(\sqrt{n \log n})$ regret and the smaller regret bounds, which we prove here for "easy data." We also show that if the constraint set is the unit ball, both the "follow the regularized leader" (FTRL) algorithm and a combination of FTL and shrinkage, which we call "follow the shrunken leader" (FTSL), achieve logarithmic regret for easy data. Simulation results on artificial data complement the theoretical findings.

While we believe that we are the first to point out that the curvature of the constraint set $\mathcal{W}$ can help in speeding up learning, this effect is known in convex optimization since at least the work of Levitin and Polyak [1966], who showed that exponential rates are attainable for strongly convex constraint sets if the norm of the gradients of the objective function admit a uniform lower bound. More recently, Garber and Hazan [2015] proved an $O(1/n^2)$ optimization error bound (with problem-dependent constants) for the Frank-Wolfe algorithm for strongly convex and smooth objectives and strongly convex constraint sets. The effect of the shape of the constraint set was also discussed by Abbasi-Yadkori [2010] who demonstrated $O(\sqrt{n})$ regret in the linear bandit setting. While at a high level these results are similar to ours, our proof technique is rather different than that used there.

4.2 Preliminaries, online learning and the follow the leader algorithm

We consider the standard framework of online convex optimization, where a learner and an environment interact in a sequential manner in n rounds, as shown in Fig. 4.1. Here, $\mathcal{W}$ is a non-empty, compact convex subset of $\mathbb{R}^d$ and $\mathcal{L}$ is a set of convex functions, mapping $\mathcal{W}$ to the reals. The elements of $\mathcal{L}$ are called loss functions. The performance of the learner is measured in terms of its regret,

$$R_n = \sum_{t=1}^{n} \ell_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{n} \ell_t(w) \,.$$

The simplest possible case, which will be the focus of this chapter, is when the losses are linear, i.e., when $\ell_t(w) = \langle f_t, w\rangle$ for some $f_t \in \mathcal{F} \subset \mathbb{R}^d$. Hence, in what follows we let $\ell_t(w) = \langle f_t, w\rangle$. We will study the regret of the so-called "Follow The Leader" (FTL) learner, which for the first round picks $w_1 \in \mathcal{W}$ in an arbitrary manner, while in round $t \geq 2$, FTL picks

1: **for** $t = 1$ to n **do**
2: Learner predicts $w_t \in \mathcal{W}$;
3: Environment picks $\ell_t \in \mathcal{L}$;
4: Learner suffers $\ell_t(w_t)$ and learns ℓ_t.
5: **end for**

Figure 4.1: Online Learning

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \sum_{i=1}^{t-1} \ell_i(w)\,.$$

When $\mathcal{W}$ is compact, the optimal w of $\min_{w\in\mathcal{W}} \sum_{i=1}^{t-1}\langle w, f_t\rangle$ is attainable, which we will assume henceforth. If multiple minimizers exist, we simply fix one of them as w_t. We will also assume that $\mathcal{F}$, which holds the loss vectors $f_t \in \mathbb{R}^d$, is non-empty, compact and convex.

4.2.1 Support functions

Let $\Theta_t = -\frac{1}{t}\sum_{i=1}^{t} f_i$ be the negative average of the first t vectors of $(f_t)_{t=1}^n$, $f_t \in \mathcal{F}$. For convenience, we define $\Theta_0 := 0$. Thus, for $t \geq 2$,

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \sum_{i=1}^{t-1}\langle w, f_i\rangle = \operatorname*{argmin}_{w\in\mathcal{W}}\langle w, -\Theta_{t-1}\rangle = \operatorname*{argmax}_{w\in\mathcal{W}}\langle w, \Theta_{t-1}\rangle\,.$$

Let $\Phi(\Theta) = \max_{w\in\mathcal{W}}\langle w, \Theta\rangle$ denote the so-called *support function* of $\mathcal{W}$. The support function, being the maximum of linear and hence convex functions, is itself convex. Further Φ is positive homogenous: for $a \geq 0$ and $\theta \in \mathbb{R}^d$, $\Phi(a\theta) = a\Phi(\theta)$. Thus the epigraph $\mathrm{epi}(\Phi) = \left\{(\theta, z)\,|\,z \geq \Phi(\theta), z \in \mathbb{R}, \theta \in \mathbb{R}^d\right\}$ of Φ is a cone, since for any $(\theta, z) \in \mathrm{epi}(\Phi)$ and $a \geq 0$, $az \geq a\Phi(\theta) = \Phi(a\theta)$, hence $(a\theta, az) \in \mathrm{epi}(\Phi)$ also holds.

The differentiability of the support function is closely tied to whether in the FTL algorithm the choice of w_t is uniquely determined:

Proposition 4.2.1. *Let $\mathcal{W} \neq \emptyset$ be convex and closed. Fix Θ and let $\mathcal{Z} := \{w \in \mathcal{W}\,|\,\langle w, \Theta\rangle = \Phi(\Theta)\}$. Then, $\partial\Phi(\Theta) = \mathcal{Z}$ and, in particular, $\Phi(\Theta)$ is differentiable at Θ if and only if $\max_{w\in\mathcal{W}}\langle w, \Theta\rangle$ has a unique optimizer. In this case, $\nabla\Phi(\Theta) = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \Theta\rangle$.*

The proposition follows from Danskin's theorem when $\mathcal{W}$ is compact (e.g., Proposition B.25 of Bertsekas 1999), but a simple direct argument can also be used to show that it also remains true even when $\mathcal{W}$ is unbounded. By Proposition 4.2.1, when Φ is differentiable at Θ_{t-1}, $w_t = \nabla\Phi(\Theta_{t-1})$.

Proof of Proposition 4.2.1. We need to show that $\mathcal{Z} = \partial\varphi(\Theta)$ where recall that

$$\begin{aligned}\partial\varphi(\Theta) &= \left\{u \in \mathbb{R}^d \,|\, \varphi(\Theta) + \langle u, \cdot - \Theta\rangle \leq \varphi(\cdot)\right\} \\ &= \left\{u \in \mathbb{R}^d \,|\, \varphi(\Theta) \leq \langle u, \Theta\rangle + \varphi(\cdot) - \langle u, \cdot\rangle\right\}.\end{aligned}$$

Since $\mathcal{Z} \subset \mathcal{W}$, if $w \in \mathcal{Z}$, $\varphi(\Theta') \geq \langle w, \Theta'\rangle$ for any Θ' by the definition of φ. Hence, $\varphi(\Theta) = \langle w, \Theta\rangle \leq \langle w, \Theta\rangle + \varphi(\Theta') - \langle w, \Theta'\rangle$ for any Θ', implying that $w \in \partial\varphi(\Theta)$.

On the other hand, assume $w \in \partial\varphi(\Theta)$. Then $\varphi(\Theta) \leq \langle w, \Theta\rangle$ since $\varphi(0) = \langle w, 0\rangle = 0$. Since $\mathcal{W}$ is closed, $\mathcal{Z}$ is also closed. Therefore, if $w \notin \mathcal{Z}$, the strict separation theorem (applied to $\{w\}$, a convex compact set, and $\mathcal{Z}$, a convex closed set) implies that there exists $\rho \in \mathbb{R}^d$ such that $\langle z, \rho\rangle < \langle w, \rho\rangle$ for all $z \in \mathcal{Z}$. Let $\Theta' = \Theta + \rho$. Then,

$$\varphi(\Theta') = \max_{u\in\mathcal{W}} \langle u, \Theta\rangle + \langle u, \rho\rangle < \varphi(\Theta) + \langle w, \Theta' - \Theta\rangle \leq \langle w, \Theta'\rangle \leq \varphi(\Theta'),$$

a contradiction. Hence, $w \in \mathcal{Z}$. □

4.2.2 Positive principal curvature

Before presenting our main result, we define some basic notions from differential geometry related to the curvature (all differential geometry concept and results that we need can be found in Section 2.5 of Schneider, 2014).

Given a C^2 (twice continuously differentiable) planar curve γ in $\mathbb{R}^2$, there exists a parametrization with respect to the curve length s, such that its derivative $\gamma'(s)$ satisfies $\|\gamma'(s)\| = \|\,(x'(s), y'(s))\,\| = \sqrt{x'(s)^2 + y'(s)^2} = 1$. Under the curve length parametrization, the curvature of γ at $\gamma(s)$ is the length of its second derivative $\|\gamma''(s)\|$. Define the unit normal vector $\mathbf{n}(\mathbf{s})$ as the unit vector that is perpendicular to $\gamma'(s)$.[1] Note that $\mathbf{n}(\mathbf{s})\cdot\gamma'(\mathbf{s}) = \mathbf{0}$. Thus $0 = (\mathbf{n}(\mathbf{s})\cdot\gamma'(\mathbf{s}))' = \mathbf{n}'(\mathbf{s})\cdot\gamma'(\mathbf{s}) + \mathbf{n}(\mathbf{s})\cdot\gamma''(\mathbf{s})$, and $\|\gamma''(s)\| = \|\mathbf{n}(\mathbf{s})\cdot\gamma''(\mathbf{s})\| = \|\mathbf{n}'(\mathbf{s})\cdot\gamma'(\mathbf{s})\| = \|\mathbf{n}'(\mathbf{s})\|$. Therefore, the curvature of γ at point $\gamma(s)$ is the length of the differential of its unit normal vector.

Denote the boundary of $\mathcal{W}$ by $\mathrm{bd}(\mathcal{W})$. We shall assume that $\mathcal{W}$ is C^2, that is, $\mathrm{bd}(\mathcal{W})$ is a twice continuously differentiable submanifold of $\mathbb{R}^d$. We denote the tangent plane of $\mathrm{bd}(\mathcal{W})$ at point w by $T_w\mathcal{W}$. Now there exists a unique unit vector at w that is perpendicular to $T_w\mathcal{W}$ and points outward of $\mathcal{W}$. In fact, one can define a continously differentiable normal unit vector field on $\mathrm{bd}(\mathcal{W})$, $u_{\mathcal{W}} : \mathrm{bd}(\mathcal{W}) \to \mathbb{S}^{d-1}$, the so-called Gauss map, which maps a boundary point $w \in \mathrm{bd}(\mathcal{W})$ to the unique outer normal vector to $\mathcal{W}$ at w, where $\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d \,|\, \|x\|_2 = 1\right\}$ denotes the unit sphere in d-dimensions. The differential of the Gauss map, $\nabla u_{\mathcal{W}}(w)$, defines a linear endomorphism of $T_w\mathcal{W}$. Moreover, $\nabla u_{\mathcal{W}}(w)$ is a self-adjoint operator, with nonnegative eigenvalues. The differential of the Gauss map, $\nabla u_{\mathcal{W}}(w)$, describes the curvature of $\mathrm{bd}(\mathcal{W})$ via the second fundamental form. In particular,

[1] There exist two unit vectors that are perpendicular to $\gamma'(s)$ for each point on γ. Pick the ones that are consistently oriented.

the *principal curvatures* of $\mathrm{bd}(\mathcal{W})$ at $w \in \mathrm{bd}(\mathcal{W})$ is defined as the eigenvalues of $\nabla u_{\mathcal{W}}(w)$. Perhaps a more intuitive, yet equivalent definition, is that the principal curvatures are the eigenvalues of the Hessian of $f = f_w$ in the parameterization $t \mapsto w + t - f_w(t)u_{\mathcal{W}}(w)$ of $\mathrm{bd}(\mathcal{W})$ which is valid in a small open neighborhood of w, where $f_w : T_w\mathcal{W} \to [0, \infty)$ is a suitable convex, nonnegative valued function that also satisfies $f_w(0) = 0$ and where $T_w\mathcal{W}$, a hyperplane of $\mathbb{R}^d$, denotes the tangent space of $\mathcal{W}$ at w, obtained by taking the support plane H of $\mathcal{W}$ at w and shifting it by $-w$. Thus, the principal curvatures at some point $w \in \mathrm{bd}(\mathcal{W})$ describe the local shape of $\mathrm{bd}(\mathcal{W})$ up to the second order. In this chapter, we are interested in the minimum principal curvature at $w \in \mathrm{bd}(\mathcal{W})$, which can be intepreated as the minimum curvature at w over all the planar curves $\gamma \in \mathrm{bd}(\mathcal{W})$ that go through w.

A related concept that has been used in convex optimization to show fast rates is that of a strongly convex constraint set [Levitin and Polyak, 1966, Garber and Hazan, 2015]: $\mathcal{W}$ is *λ-strongly convex* with respect to the norm $\|\cdot\|$ if, for any $x, y \in \mathcal{W}$ and $\gamma \in [0, 1]$, the $\|\cdot\|$-ball with origin $\gamma x + (1-\gamma)y$ and radius $\gamma(1-\gamma)\lambda \|x - y\|^2 / 2$ is included in $\mathcal{W}$. That is, for any $z \in \mathbb{R}^d$ with $\|z\| = 1$, $\gamma x + (1-\gamma)y + \gamma(1-\gamma)\frac{\lambda}{2}\|x - y\|^2 z \in \mathcal{W}$. Let $B_r(x) = \left\{y \in \mathbb{R}^d \,|\, \|x - y\|_2 \le r\right\}$ denote the Euclidean ball of radius r centered at x. The next proposition shows that a convex body $\mathcal{W} \in C^2$ is λ-strongly convex with respect to $\|\cdot\|_2$ if and only if the principal curvatures of the surface $\mathrm{bd}(\mathcal{W})$ are all at least λ.

Proposition 4.2.2. *Let $\mathcal{W} \subset \mathbb{R}^d$ be a C^2 convex body with support function φ, and let λ be an arbitrary positive number. Then the following statements are equivalent:*

(i) The smallest principal curvature of $\mathcal{W}$ is at least λ.

(ii) $\mathcal{W} = \cap_{\theta \in \mathbb{S}^{d-1}} B_{1/\lambda}(w_\theta - \theta/\lambda)$ where $w_\theta \in \partial\varphi(\theta) \subset \mathrm{bd}(\mathcal{W})$.

(iii) $\mathcal{W}$ is λ-strongly convex.

Condition (ii), which is actually the definition of Polovinkin [1996] for strongly convex sets, means that $\mathcal{W}$ can be obtained as the intersection of closed balls of radius $1/\lambda$, such that there is one ball for every boundary point w and tangent hyperplane P where the ball touches P in w. Note that a ball with radius $1/\lambda$ satisfies all conditions: (i) and (ii) by definition, while (iii) holds, e.g., by Example 13 of Journée et al. [2010].

Proof. We show that (i) implies (ii), (ii) implies (iii), and (iii) implies (i).

We start by showing that (i) implies (ii). First note that all principal curvatures of the d-dimensional ball $B = B_{1/\lambda}(0)$ with radius $1/\lambda$ (centered at the origin) are λ. Therefore, (i) and Theorem 3.2.9 of Schneider [2014] implies that there is a convex body $\mathcal{M}$ such that $\mathcal{W} + \mathcal{M} = B$, where for two sets, $S_1, S_2 \subset \mathbb{R}^d$, $S_1 + S_2$ is defined as $\{s_1 + s_2 \,|\, s_1 \in S_1, s_2 \in S_2\}$. For any $\theta \in \mathbb{S}^{d-1}$, let $m_\theta \in \mathrm{argmax}_{m \in \mathcal{M}} \langle m, \theta \rangle$. Then clearly $w_\theta + m_\theta$ maximizes $\langle b, \theta \rangle$ for $b \in \mathcal{W} + \mathcal{M}$. Therefore, $\mathcal{W} + m_\theta$ is a subset of B and touches it at $w_\theta + m_\theta$, or equivalently

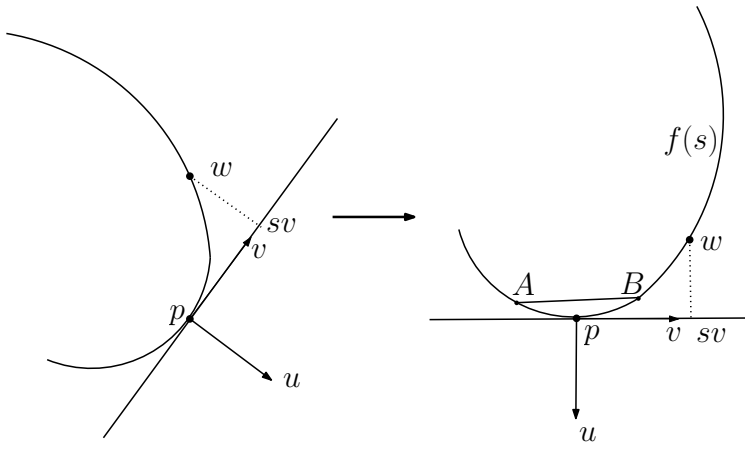

Figure 4.2: The local coordinate system at p.

$\mathcal{W} \subset B - m_\theta$ and they touch each other, and a tangent hyperplane with normal vector θ, in w_θ. This proves that (i) implies (ii).

Next we prove that (ii) implies (iii). Assuming (ii) holds, let $w \in \mathcal{W}$ be any point in the interior of $\mathcal{W}$, and let $p \in \mathrm{bd}(\mathcal{W})$ be the closest boundary point to w, and recall that $T_p\mathcal{W}$ is the tangent space of $\mathcal{W}$ at p. By construction, $B_{\|w-p\|_2}(w)$ touches the boundary of $\mathcal{W}$ at p (in the sense that they do not intersect, but they can have multiple common points), and so $w - p$ is orthogonal to $T_p\mathcal{W}$. Therefore, $B_{\|w-p\|_2}(w)$ also touches the boundary of the ball $B = B_{1/\lambda}(p + \frac{w-p}{\lambda\|w-p\|_2})$, which contains $\mathcal{W}$ by assumption (ii). Now consider any two points $x, y \in \mathcal{W}$ and $\gamma \in [0,1]$ such that $w = \gamma x + (1-\gamma)y$. Then the ball with radius $\lambda\gamma(1-\gamma)\|x-y\|_2^2/2$ centered at w is contained in B, since B is λ-strongly convex. But then its radius is at most $\|p - w\|_2$, and so it is also contained in $\mathcal{W}$. This shows that $\mathcal{W}$ is λ-strongly convex, thus (iii) holds. To finish the proof of the proposition, assume (iii). To prove that (i) holds, we have to show, that for any point p on $\mathrm{bd}(\mathcal{W})$ and for any unit vector $v \in T_p\mathcal{W}$, the curvature of the boundary along v is at least λ. Let P be the hyperplane spanned by v and the outer normal vector u of $\mathcal{W}$ at point p, and consider the planar curve γ defined by $\mathrm{bd}(\mathcal{W}) \cap P$. Using v as the axis of a local coordinate system, a point $w(s)$ on the curve γ in the neighbourhood of p can be expressed as $w(s) = p + sv - f(s)u$ for an appropriate function f, as illustrated in Fig. 4.2.

Note that $f'(0) = 0$, and by Proposition 2.1 of Pressley [2010], the curvature of γ at p can be obtained as

$$\left.\frac{f''(s)}{(1+f'(s)^2)^{3/2}}\right|_{s=0} = f''(0)\ .$$

Now since $w(s), w(-s) \in \mathcal{W}$ for a sufficiently small s, the strong convexity of $\mathcal{W}$ applied to $w(s)$ and $w(-s)$ with $\gamma = 1/2$ implies that $q = \frac{w(s)+w(-s)}{2} + \frac{\lambda}{8}\|w(s) - w(-s)\|_2^2 u \in \mathcal{W}$.

Substituting the definition of $w(s)$ and $w(-s)$, we get

$$q = p - u\left[\frac{f(s)+f(-s)}{2} - \frac{\lambda}{8}\Big(4s^2 + (f(s)-f(-s))^2\Big)\right].$$

Therefore, $q \in \mathcal{W}$ implies $f(s)+f(-s) \geq \lambda s^2$, and so

$$f''(0) = \lim_{s\to 0} \frac{\frac{f(s)-f(0)}{s} - \frac{f(0)-f(-s)}{s}}{s} = \frac{f(s)+f(-s)}{s^2} \geq \lambda.$$

Thus (i) holds, finishing the proof of the proposition. □

4.3 Non-stochastic analysis of FTL

We start by rewriting the regret of FTL in an equivalent form, which shows that we can expect FTL to enjoy a small regret when successive weight vectors move little. A noteworthy feature of the next proposition is that rather than bounding the regret from above, it gives an equivalent expression for it.

Proposition 4.3.1. *The regret R_n of FTL satisfies*

$$R_n = \sum_{t=1}^{n} t\,\langle w_{t+1} - w_t, \Theta_t\rangle\,.$$

The result is a direct corollary of Lemma 9 of McMahan [2010], which holds for any sequence of losses, even in the lack of convexity. It is also a tightening of the well-known inequality $R_n \leq \sum_{t=1}^{n} \ell_t(w_t) - \ell_t(w_{t+1})$, which again holds for arbitrary loss sequences (e.g., Lemma 2.1 of Shalev-Shwartz [2012]). To keep the thesis self-contained, we give an elegant, short direct proof, based on the summation by parts formula:

Proof. The summation by parts formula states that for any $u_1, v_1, \ldots, u_{n+1}$, v_{n+1} reals, $\sum_{t=1}^{n} u_t\,(v_{t+1} - v_t) = (u_{t+1}v_{t+1} - u_1 v_1) - \sum_{t=1}^{n}(u_{t+1} - u_t)\,v_{t+1}$. Applying this to the definition of regret with $u_t := w_{t,\cdot}$ and $v_{t+1} := t\Theta_t$, we get

$$\begin{aligned} R_n &= -\sum_{t=1}^{n} \langle w_t, t\Theta_t - (t-1)\Theta_{t-1}\rangle + \langle w_{n+1}, n\Theta_n\rangle \\ &= -\left\{\cancel{\langle w_{n+1}, n\Theta_n\rangle} - 0 - \sum_{t=1}^{n}\langle w_{t+1} - w_t, t\Theta_t\rangle\right\} + \cancel{\langle w_{n+1}, n\Theta_n\rangle}. \end{aligned}$$

□

Our next proposition gives another formula that is equal to the regret. Although this formula is not directly needed for the rest of the chapter, it provides interesting insights: as opposed to the previous result, it is independent of w_t, and directly connects the sequence $(\Theta_t)_t$ to the geometric properties of $\mathcal{W}$ through the support function Φ. A similar expression for a general "Follow the Regularized Leader" algorithm can also be founded in the work of Abernethy et al. [2014]. For this proposition we will momentarily assume that Φ is differentiable at $(\Theta_t)_{t\geq 1}$; a more general statement will follow later.

Proposition 4.3.2. *If Φ is differentiable at $\Theta_1, \dots, \Theta_n$,*

$$R_n = \sum_{t=1}^{n} t\, D_\Phi(\Theta_t, \Theta_{t-1})\,, \tag{4.1}$$

where $D_\Phi(\theta', \theta) = \Phi(\theta') - \Phi(\theta) - \langle \nabla\Phi(\theta), \theta' - \theta\rangle$ is the Bregman divergence of Φ and we use the convention that $\nabla\Phi(0) = w_1$.

Proof. Let $v = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \theta\rangle$, $v' = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \theta'\rangle$. When Φ is differentiable at θ,

$$\begin{aligned} D_\Phi(\theta', \theta) &= \Phi(\theta') - \Phi(\theta) - \langle \nabla\Phi(\theta), \theta' - \theta\rangle \\ &= \langle v', \theta'\rangle - \langle v, \theta\rangle - \langle v, \theta' - \theta\rangle = \langle v' - v, \theta'\rangle\,. \end{aligned} \tag{4.2}$$

Therefore, by Proposition 4.3.1,

$$R_n = \sum_{t=1}^{n} t\langle w_{t+1} - w_t, \Theta_t\rangle = \sum_{t=1}^{n} t\, D_\Phi(\Theta_t, \Theta_{t-1}).$$

□

When Φ is non-differentiable at some of the points $\Theta_1, \dots, \Theta_n$, the equality in the above proposition can be replaced with inequalities. Defining the upper Bregman divergence $\overline{D}_\Phi(\theta', \theta) = \sup_{w\in\partial\Phi(\theta)} \Phi(\theta') - \Phi(\theta) - \langle w, \theta' - \theta\rangle$ and the lower Bregman divergence $\underline{D}_\Phi(\theta', \theta)$ similarly with inf instead of sup,

$$\sum_{t=1}^{n} t\, \underline{D}_\Phi(\Theta_t, \Theta_{t-1}) \le R_n \le \sum_{t=1}^{n} t\, \overline{D}_\Phi(\Theta_t, \Theta_{t-1})\,. \tag{4.3}$$

While Proposition 4.3.2 and Eq. (4.3) are insightful, we will only use Proposition 4.3.1 in the rest of this chapter.

4.3.1 Constraint sets with positive curvature

The previous results show in an implicit fashion that the curvature of $\mathcal{W}$ controls the regret. Our next result explicitly connects the principal curvatures of $\operatorname{bd}(\mathcal{W})$ to the regret of FTL and shows that FTL enjoys logarithmic regret for highly curved surfaces, as long as $\|\Theta_t\|_2$ is bounded away from zero.

Theorem 4.3.1. *Let $\mathcal{W} \subset \mathbb{R}^d$ be a C^2 convex body[2] with $d \ge 2$. Let $M = \max_{f\in\mathcal{F}} \|f\|_2$ and assume that Φ is differentiable at $(\Theta_t)_t$. Assume that the principal curvatures of the surface $\operatorname{bd}(\mathcal{W})$ are all at least λ_0 for some constant $\lambda_0 > 0$ and $L_n := \min_{1\le t\le n} \|\Theta_t\|_2 > 0$. Choose any $w_1 \in \operatorname{bd}(\mathcal{W})$. Then*

$$R_n \le \frac{2M^2}{\lambda_0 L_n}(1 + \log(n))\,.$$

[2] Following Schneider [2014], a convex body of $\mathbb{R}^d$ is any non-empty, compact, convex subset of $\mathbb{R}^d$.

As we will show later in an essentially matching lower bound, this bound is tight, showing that the forte of FTL is when L_n is bounded away from zero and λ_0 is large. Note that the bound is vacuous as soon as $L_n = O(\log(n)/n)$ and is worse than the minimax bound of $O(\sqrt{n})$ when $L_n = o(\log(n)/\sqrt{n})$. One possibility to reduce the bound's sensitivity to L_n is to use the trivial bound $\langle w_{t+1} - w_t, \Theta_t \rangle \le LW = L \sup_{w,w' \in \mathcal{W}} \|w - w'\|_2$ for indices t when $\|\Theta_t\| \le L$. Then, by optimizing the bound over L, one gets a data-dependent bound of the form $\inf_{L>0} \left(\frac{2M^2}{\lambda_0 L}(1 + \log(n)) + LW \sum_{t=1}^n t \, \mathbb{I}\left(\|\Theta_t\| \le L\right) \right)$, which is more complex, but is free of L_n and thus reflects the nature of FTL better. Note that in the case of stochastic problems, where $f_1, \dots, f_n$ are independent and identically distributed (i.i.d.) with $\mu := -\mathbb{E}\left[\Theta_t\right] \neq 0$, the probability that $\|\Theta_t\|_2 < \|\mu\|_2 / 2$ is exponentially small in t. Thus, selecting $L = \|\mu\|_2 / 2$ in the previous bound, the contribution of the expectation of the second term is $O(\|\mu\|_2 W)$, giving an overall bound of the form $O(\frac{M^2}{\lambda_0 \|\mu\|_2} \log(n) + \|\mu\|_2 W)$. After the proof we will provide some simple examples that should make it more intuitive how the curvature of $\mathcal{W}$ helps keeping the regret of FTL small.

Proof. Fix $\theta_1, \theta_2 \in \mathbb{R}^d$ and let

$$w^{(1)} = \operatorname*{argmax}_{w \in \mathcal{W}} \langle w, \theta_1 \rangle, \quad \text{and} \quad w^{(2)} = \operatorname*{argmax}_{w \in \mathcal{W}} \langle w, \theta_2 \rangle.$$

Note that if $\theta_1, \theta_2 \neq 0$ then $w^{(1)}, w^{(2)} \in \mathrm{bd}(\mathcal{W})$. Below we will show that

$$\langle w^{(1)} - w^{(2)}, \theta_1 \rangle \le \frac{1}{2\lambda_0} \frac{\|\theta_2 - \theta_1\|_2^2}{\|\theta_2\|_2} \,. \tag{4.4}$$

Proposition 4.3.1 suggests that it suffices to bound $\langle w_{t+1} - w_t, \Theta_t \rangle$. By Equation (4.4), we see that it suffices to bound how much Θ_t moves. A straightforward calculation shows that Θ_t cannot move much: for any norm $\|\cdot\|$ on $\mathcal{F}$, we have

$$\begin{aligned}
\|\Theta_t - \Theta_{t-1}\| &= \left\| \frac{1}{t-1} \sum_{i=1}^{t-1} f_i - \frac{1}{t} \sum_{i=1}^{t} f_i \right\| = \left\| \sum_{i=1}^{t-1} \left(\frac{1}{t-1} - \frac{1}{t} \right) f_i - \frac{1}{t} f_t \right\| \\
&\le \left\| \sum_{i=1}^{t-1} \left(\frac{1}{t-1} - \frac{1}{t} \right) f_i \right\| + \left\| \frac{1}{t} f_t \right\| = \left\| \sum_{i=1}^{t-1} \frac{1}{t(t-1)} f_i \right\| + \left\| \frac{1}{t} f_t \right\| \\
&= \frac{1}{t} \left\| \frac{1}{t-1} \sum_{i=1}^{t-1} f_i \right\| + \frac{1}{t} \|f_t\| \le \frac{2}{t} M \,.
\end{aligned} \tag{4.5}$$

where $M = \max_{f \in \mathcal{F}} \|f\|$ is a constant that depends on $\mathcal{F}$ and the norm $\|\cdot\|$.

Combining Equation (4.4) with Proposition 4.3.1 and Equation (4.5), we get

$$\begin{aligned}
R_n &= \sum_{t=1}^n t \langle w_{t+1} - w_t, \Theta_t \rangle \le \sum_{t=1}^n \frac{t}{2\lambda_0} \frac{\|\Theta_t - \Theta_{t-1}\|_2^2}{\|\Theta_{t-1}\|_2} \\
&\le \frac{2M^2}{\lambda_0} \sum_{t=1}^n \frac{1}{t \|\Theta_{t-1}\|_2} \le \frac{2M^2}{\lambda_0 L_n} \sum_{t=1}^n \frac{1}{t} \le \frac{2M^2}{\lambda_0 L_n} (1 + \log(n)) \,.
\end{aligned}$$

To finish the proof, it thus remains to show Equation (4.4).

The following elementary lemma relates the cosine of the angle between two vectors θ_1 and θ_2 to the squared normalized distance between the two vectors, thereby reducing our problem to bounding the cosine of this angle. For brevity, we denote by $\cos(\theta_1, \theta_2)$ the cosine of the angle between θ_1 and θ_2.

Lemma 4.3.3. *For any non-zero vectors $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$1 - \cos(\theta_1, \theta_2) \leq \frac{1}{2} \frac{\|\theta_1 - \theta_2\|_2^2}{\|\theta_1\|_2 \|\theta_2\|_2}. \tag{4.6}$$

Proof. Note that $\|\theta_1\|_2 \|\theta_2\|_2 \cos(\theta_1, \theta_2) = \langle \theta_1, \theta_2 \rangle$. Therefore, Equation (4.6) is equivalent to $2\|\theta_1\|_2 \|\theta_2\|_2 - 2\langle \theta_1, \theta_2 \rangle \leq \|\theta_1 - \theta_2\|_2^2$, which, by algebraic manipulations, is itself equivalent to $0 \leq (\|\theta_1\|_2 - \|\theta_2\|_2)^2$. □

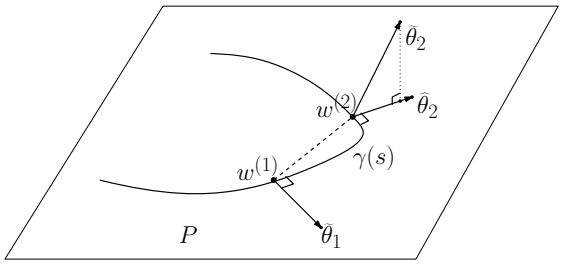

Figure 4.3: Illustration of the construction used in the proof of Equation (4.4).

With this result, we see that it suffices to upper bound $\cos(\theta_1, \theta_2)$ by $1 - \lambda_0 \langle w^{(1)} - w^{(2)}, \frac{\theta_1}{\|\theta_1\|_2} \rangle$. To develop this bound, let $\tilde{\theta}_i = \frac{\theta_i}{\|\theta_i\|_2}$ for $i = 1, 2$. The angle between θ_1 and θ_2 is the same as the angle between the normalized vectors $\tilde{\theta}_1$ and $\tilde{\theta}_2$. To calculate the cosine of the angle between $\tilde{\theta}_1$ and $\tilde{\theta}_2$, let P be a plane spanned by $\tilde{\theta}_1$ and $w^{(1)} - w^{(2)}$ and passing through $w^{(1)}$ (P is uniquely determined if $\tilde{\theta}_1$ is not parallel to $w^{(1)} - w^{(2)}$; if there are multiple planes, just pick any of them). Further, let $\hat{\theta}_2 \in \mathbb{S}^{d-1}$ be the unit vector along the projection of $\tilde{\theta}_2$ onto the plane P, as indicated in Fig. 4.3. Clearly, $\cos(\tilde{\theta}_1, \tilde{\theta}_2) \leq \cos(\tilde{\theta}_1, \hat{\theta}_2)$.

Consider a curve $\gamma(s)$ on $\mathrm{bd}(\mathcal{W})$ connecting $w^{(1)}$ and $w^{(2)}$ that is defined by the intersection of $\mathrm{bd}(\mathcal{W})$ and P and is parametrized by its curve length s so that $\gamma(0) = w^{(1)}$ and $\gamma(l) = w^{(2)}$, where l is the length of the curve γ between $w^{(1)}$ and $w^{(2)}$. Let $u_{\mathcal{W}}(w)$ denote the outer normal vector to $\mathcal{W}$ at w as before, and let $u_\gamma : [0, l] \to \mathbb{S}^{d-1}$ be such that $u_\gamma(s) = \hat{\theta}$ where $\hat{\theta}$ is the unit vector parallel to the projection of $u_{\mathcal{W}}(\gamma(s))$ on the plane P. By definition, $u_\gamma(0) = \tilde{\theta}_1$ and $u_\gamma(l) = \hat{\theta}_2$. Note that in fact γ exists in two versions since $\mathcal{W}$ is a compact convex body, hence the intersection of P and $\mathrm{bd}(\mathcal{W})$ is a closed curve. Of these two versions we choose the one that satisfies that $\langle \gamma'(s), \tilde{\theta}_1 \rangle \leq 0$ for $s \in [0, l]$.[3] Given

[3] γ' and u'_γ denote the derivatives of γ and u, respectively, which exist since $\mathcal{W}$ is C^2.

the above, we have

$$\cos(\tilde{\theta}_1, \hat{\theta}_2) = \langle \hat{\theta}_2, \tilde{\theta}_1 \rangle = 1 + \langle \hat{\theta}_2 - \tilde{\theta}_1, \tilde{\theta}_1 \rangle = 1 + \Big\langle \int_0^l u'_\gamma(s)\,\mathrm{d}s, \tilde{\theta}_1 \Big\rangle$$
$$= 1 + \int_0^l \langle u'_\gamma(s), \tilde{\theta}_1 \rangle\,\mathrm{d}s. \tag{4.7}$$

Note that γ is a planar curve on $\mathrm{bd}(\mathcal{W})$, thus its curvature $\lambda(s)$ satisfies $\lambda(s) \geq \lambda_0$ for $s \in [0, l]$. Also, for any w on the curve γ, $\gamma'(s)$ is a unit vector parallel to P. Moreover, $u'_\gamma(s)$ is parallel to $\gamma'(s)$ and $\lambda(s) = \|u'_\gamma(s)\|_2$. Therefore,

$$\langle u'_\gamma(s), \tilde{\theta}_1 \rangle = \|u'_\gamma(s)\|_2 \langle \gamma'(s), \tilde{\theta}_1 \rangle \leq \lambda_0 \langle \gamma'(s), \tilde{\theta}_1 \rangle,$$

where the last inequality holds because $\langle \gamma'(s), \tilde{\theta}_1 \rangle \leq 0$. Plugging this into Equation (4.7), we get the desired

$$\cos(\tilde{\theta}_1, \hat{\theta}_2) \leq 1 + \lambda_0 \int_0^l \langle \gamma'(s), \tilde{\theta}_1 \rangle\,\mathrm{d}s = 1 + \lambda_0 \Big\langle \int_0^l \gamma'(s)\,\mathrm{d}s, \tilde{\theta}_1 \Big\rangle$$
$$= 1 - \lambda_0 \langle w^{(1)} - w^{(2)}, \tilde{\theta}_1 \rangle\,.$$

Reordering and combining with Equation (4.6) we obtain

$$\langle w^{(1)} - w^{(2)}, \tilde{\theta}_1 \rangle \leq \frac{1}{\lambda_0} \Big(1 - \cos(\tilde{\theta}_1, \hat{\theta}_2)\Big) \leq \frac{1}{\lambda_0} \left(1 - \cos(\theta_1, \theta_2)\right)$$
$$\leq \frac{1}{2\lambda_0} \frac{\|\theta_1 - \theta_2\|_2^2}{\|\theta_1\|_2 \|\theta_2\|_2}\,.$$

Multiplying both sides by $\|\theta_1\|_2$ gives Equation (4.4), thus, finishing the proof. □

The next proposition provides some examples to illustrate how principal curvature behaves.

Proposition 4.3.4. *The smallest principal curvature of some common convex bodies are as follows:*

(i) The smallest principal curvature λ_0 of the Euclidean ball of radius r, $\mathcal{W} = \{w \,|\, \|w\|_2 \leq r\}$, satisfies $\lambda_0 = \frac{1}{r}$.

(ii) Let Q be a positive definite matrix. If $\mathcal{W} = \{w \,|\, w^\top Q w \leq 1\}$ then $\lambda_0 = \lambda_{\min}/\sqrt{\lambda_{\max}}$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal, respectively, maximal eigenvalues of Q.

(iii) More generally, let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a C^2 convex function. Then, for $\mathcal{W} = \{w \,|\, \phi(w) \leq 1\}$, $\lambda_0 = \min_{w \in \mathrm{bd}(\mathcal{W})} \min_{v\,:\,\|v\|_2 = 1, v \perp \phi'(w)} \frac{v^\top \nabla^2 \phi(w) v}{\|\phi'(w)\|_2}$.

We only prove part (iii), since it implies part (ii), which implies (i). Polovinkin [1996] also derived part (ii) for the strong convexity definition (ii) of Proposition 4.2.2.

Proof. Fix $w \in \mathrm{bd}(\mathcal{W})$. Note that $\phi'(w)$ is a normal vector of $\mathrm{bd}(\mathcal{W})$ at w for $\mathrm{bd}(\mathcal{W})$, thus $T_w\mathcal{W} = \{v : v \perp \phi'(w)\}$. Then the Gauss map $u_\mathcal{W}$ of $\mathcal{W}$ satisfies $u_\mathcal{W}(w) = \frac{\phi'(w)}{\|\phi'(w)\|_2}$ for $w \in \mathrm{bd}(\mathcal{W})$.

Next we compute the so-called Weingarten map, $W_w(v) : T_w\mathcal{W} \to T_w\mathcal{W}$, which, by definition, is the differential of $u_{\mathcal{W}}(w)$ restricted to $T_w\mathcal{W}$. Note that the Weingarten map is a linear map. We have that for $v \in T_w\mathcal{W}$,

$$W_w(v) = \frac{\mathrm{d}u_{\mathcal{W}}}{\mathrm{d}w}(v) = \frac{\nabla\phi^2(w)v}{\|\phi'(w)\|_2} - \frac{\phi'(w)^\top\nabla^2\phi(w)\phi'(w)v}{\|\phi'(w)\|_2^3} = \frac{\nabla\phi^2(w)v}{\|\phi'(w)\|_2},$$

where the last equality is due to $\phi'(w)v = 0$.

By Schneider [2014, page 105], the eigenvalues of the Weingarten map $W_w(v)$ are exactly the principal curvature of $\mathcal{W}$ at w. Therefore, the smallest principal curvature at w is $\min_{v\,:\,\|v\|_2=1, v\perp\phi'(w)} \frac{v^\top\nabla^2\phi(w)v}{\|\phi'(w)\|_2}$. Taking minimum over all $w \in \mathrm{bd}(\mathcal{W})$ finishes the proof. □

4.3.2 An asymptotic lower bound

The result in this section is an asymptotic lower bound for the linear game, showing that FTL achieves the optimal rate under the condition that $\min_t \|\Theta_t\|_2 \geq L > 0$.

Theorem 4.3.2. *Let $\lambda, L \in (0,1)$. Assume that $\{(1,-L),(-1,-L)\} \subset \mathcal{F}$ and let*

$$\mathcal{W} = \left\{(x,y) \in \mathbb{R}^2 : x^2 + \frac{y^2}{\lambda^2} \leq 1\right\},$$

as shown in Fig. 4.4, be an ellipsoid with principal curvature h. Then, for any learning strategy, there exists a sequence of losses in $\mathcal{F}$ such that $R_n = \Omega\left(\log(n)/(Lh)\right)$ and $\|\Theta_t\|_2 \geq L$ for all t.

Note that by Proposition 4.3.4 (ii), since that $\lambda_{\max} = \frac{1}{\lambda^2}$ and $\lambda_{\min} = 1$, the principal curvature of $\mathcal{W}$ is $\frac{\lambda_{\min}}{\sqrt{\lambda_{\max}}} = \lambda$. In fact, it is not too hard to extend Theorem 4.3.2 for any set $\mathcal{W}$ such that there is $w \in \mathrm{bd}(\mathcal{W})$ where the curvature is h, and the curvature is a continuous function in a neighbourhood of w over the boundary $\mathrm{bd}(\mathcal{W})$. The constants in the bound would then depend on how fast the curvature changes within this neighbourhood.

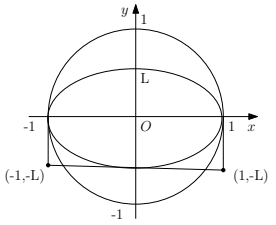

Figure 4.4: The constraint set $\mathcal{W}$ of Theorem 4.3.2.

Proof. We define a random loss sequence, and we will show that no algorithm on this sequence can achieve an $o(\log n/(hL)$ regret. Let P be a random variable with $\mathrm{Beta}(K,K)$ distribution for some $K > 1 + \frac{1}{h^2L^2}$, and, given P, assume that $X_t, t \geq 1$ are i.i.d. Bernoulli random variables with parameter P. Let $f_t = X_t(1,-L) + (1-X_t)(-1,-L) = (2X_t - 1, -L)$. Thus, the second coordinate of f_t is always $-L$, and so $\|\Theta_t\|_2 = \left\|\frac{1}{t}\sum_{i=1}^t f_i\right\|_2 \geq L$. Furthermore, the conditional expectation of the loss vector is $f^p \doteq \mathbb{E}\left[f_t \middle| P = p\right] = (2p-1, -L)$.

Note that $2X_t - 1 = f_{t,1}$ for all t; thus the conditional expectation of P, given $f_1, \ldots, f_{t-1}$, can be determined by the well-known formula $\hat{P}_{t-1} = \mathbb{E}\left[P \middle| f_1 \ldots f_{t-1}\right] = \frac{K + \sum_{i=1}^{t-1} X_i}{2K + t - 1}$. Given

p, denote the optimizer of f^p by w^p, that is, $w^p = \operatorname{argmin}_{w\in\mathcal{W}} \langle w, f^p\rangle$. Then the Bayesian optimal choice in round t is

$$\begin{aligned}\operatorname*{argmin}_{w\in\mathcal{W}} \mathbb{E}\left[\left[\left\langle w, f^P\right\rangle\right| f_1 \dots f_{t-1}\right] &= \operatorname*{argmin}_{w\in\mathcal{W}} \left\langle w, \mathbb{E}\left[f^P\right| f_1 \dots f_{t-1}\right]\right\rangle \\ &= \operatorname*{argmin}_{w\in\mathcal{W}} \left\langle w, f^{\hat{P}_{t-1}}\right\rangle \\ &= w^{\hat{P}_{t-1}}, \end{aligned} \tag{4.8}$$

where the first equality follows by linearity of the inner product, the second since f^p is a linear function of p and the third by the definition of w^p.

Thus, denoting by W_t the prediction of an arbitrary algorithm in round t, the expected regret can be bounded from below as

$$\begin{aligned}\mathbb{E}[R_n] &= \mathbb{E}\left[\max_{w\in\mathcal{W}} \sum_{t=1}^n \langle W_t - w, f_t\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\max_{w\in\mathcal{W}} \sum_{t=1}^n \langle W_t - w, f_t\rangle\right| P\right]\right] \\ &\geq \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^n \langle W_t - w^P, f_t\rangle\right| P\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle W_t - w^P, f_t\right\rangle\right| P, f_1, \dots, f_{t-1}\right]\right]. \end{aligned} \tag{4.9}$$

Further because of the independence of the f_s given P, $\mathbb{E}\left[\left\langle w^P, f_t\right\rangle\right| P, f_1, \dots, f_{t-1}] = \left\langle w^P, \mathbb{E}\left[f_t\right| P, f_1, \dots, f_{t-1}]\right\rangle = \left\langle w^P, \mathbb{E}\left[f_t\right| P]\right\rangle = \left\langle w^P, f^P\right\rangle$. Also since W_t is chosen based on $f_1, \dots, f_{t-1}$ (but not on P),

$$\begin{aligned}\mathbb{E}\left[\langle W_t, f_t\rangle\right| P, f_1, \dots, f_{t-1}] =& \mathbb{E}\left[\mathbb{E}\left[\langle W_t, f_t\rangle\right| P, f_1, \dots, f_{t-1}, W_t]\right| P, f_1, \dots, f_{t-1}] \\ =& \mathbb{E}\left[\langle W_t, \mathbb{E}\left[f_t\right| P, f_1, \dots, f_{t-1}, W_t]\rangle\right| P, f_1, \dots, f_{t-1}] \\ =& \mathbb{E}\left[\left\langle W_t, f^P\right\rangle\right| P, f_1, \dots, f_{t-1}]. \end{aligned}$$

Hence the RHS of Eq. (4.9) can be rewritten as

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle W_t - w^P, f^P\right\rangle\right| f_1, \dots, f_{t-1}]\right] &\geq \mathbb{E}\left[\sum_{t=1}^n \min_{w\in\mathcal{W}} \mathbb{E}\left[\left\langle w - w^P, f^P\right\rangle\right| f_1, \dots, f_{t-1}]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\right| f_1, \dots, f_{t-1}]\right] \\ &= \sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\right], \end{aligned} \tag{4.10}$$

where Equation (4.10) holds by Equation (4.8). We will need several lemmas to continue our calculation. Their proofs are postponed to Section 4.7.

Lemma 4.3.5. *Under the assumptions of Theorem 4.3.2, for any $0 < p_1, p_2 < 1$,*

$$\langle w^{p_2} - w^{p_1}, f^{p_1}\rangle \geq \frac{hL}{2} \frac{\left(\frac{2p_2 - 2p_1}{hL}\right)^2}{\sqrt{1 + \left(\frac{1-2p_1}{hL}\right)^2}\left(1 + \left(\frac{1-2p_2}{hL}\right)^2\right)}.$$

Recall that $\hat{P}_t = \frac{K+\sum_{i=1}^t X_i}{2K+t}$, and so

$$\hat{P}_t - P = \frac{K}{2K+t}(1-2P) + \frac{t}{2K+t}(\frac{S_t}{t} - P), \tag{4.11}$$

where $S_t = \sum_{i=1}^t X_i$.

Lemma 4.3.6. *For any $u > 0$,*

$$\mathbb{P}\left[|\hat{P}_t - P| > \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}u \,\middle|\, P\right] \le 2\exp(-tu^2)\ .$$

Lemma 4.3.7. *For any $t \ge 0$,*

$$\mathbb{E}\left[(P-\hat{P}_t)^2 \,\middle|\, P\right] = \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{tP(1-P)}{(2K+t)^2}.$$

Now to continue our proof, by Lemma 4.3.5 we have

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\right] &\ge \frac{hL}{2}\sum_{t=1}^n \mathbb{E}\left[\frac{\left(\frac{2\hat{P}_{t-1}-2P}{hL}\right)^2}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}\left(1+\left(\frac{1-2\hat{P}_{t-1}}{hL}\right)^2\right)}\right] \\
&= \frac{2}{hL}\sum_{t=1}^n \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_{t-1}-P)^2}{1+\left(\frac{1-2\hat{P}_{t-1}}{hL}\right)^2}\,\middle|\, P\right]\right] \\
&\ge \frac{2}{hL}\sum_{t=1}^n \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_{t-1}-P)^2}{1+2\left(\frac{1-2P}{hL}\right)^2+2\left(\frac{2P-2\hat{P}_{t-1}}{hL}\right)^2}\,\middle|\, P\right]\right],
\end{aligned} \tag{4.12}$$

where in the last step we used $(a+b)^2 \le a^2+b^2$. Let $\mathcal{G}_t$ be the event that $|\hat{P}_t - P| \le \frac{K|1-2P|}{2K+t} + \frac{thL}{2K+t}$. Note that $\mathcal{G}_t$ holds with probability at least $1-2e^{-t(hL)^2}$ by Lemma 4.3.6. Then, lower bounding the first term by 0, the RHS of Equation (4.12) can be lower bounded by

$$\begin{aligned}
&\frac{2}{hL}\sum_{t=0}^{n-1} \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_t-P)^2}{1+2\left(\frac{1-2P}{hL}\right)^2+2\left(\frac{2P-2\hat{P}_t}{hL}\right)^2}\mathbb{I}(\mathcal{G}_t)\,\middle|\, P\right]\right] \\
&\ge \frac{2}{hL}\sum_{t=0}^{n-1} \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\mathbb{I}(\mathcal{G}_t)\,\middle|\, P\right]}{\left(1+2\left(\frac{1-2P}{hL}\right)^2+2\left(\frac{2K}{2K+t}\frac{|1-2P|}{hL}+\frac{2t}{2K+t}\right)^2\right)}\right] \\
&\ge \frac{2}{hL}\sum_{t=0}^{n-1} \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\mathbb{I}(\mathcal{G}_t)\,\middle|\, P\right]}{\left(9+4\left(\frac{1-2P}{hL}\right)^2+8\frac{|1-2P|}{hL}\right)}\right].
\end{aligned}$$

Combining the above, and using $(\hat{P}_t - P)^2 \leq 1$ together with $\mathbb{P}\left(\mathcal{G}_t^c\right) \leq 2e^{-t(hL)^2}$, we get

$$\begin{aligned}
&\mathbb{E}\left[R_n\right] \\
&\geq \frac{2}{hL}\sum_{t=0}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|P\right]-\mathbb{P}\left(\mathcal{G}_t^c\right)}{\left(9+4\left(\frac{1-2P}{hL}\right)^2+8\frac{|1-2P|}{hL}\right)}\right] \\
&\geq \frac{2}{hL}\sum_{t=0}^{n-1}\left(\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|P\right]}{\left(9+4\left(\frac{1-2P}{hL}\right)^2+8\frac{|1-2P|}{hL}\right)}\right]-\frac{2}{9}e^{-th^2L^2}\right) \\
&\geq \frac{2}{hL}\left(-\frac{2}{9(1-e^{-h^2L^2})}+\sum_{t=0}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|P\right]}{\left(9+4\left(\frac{1-2P}{hL}\right)^2+8\frac{|1-2P|}{hL}\right)}\right]\right). \qquad (4.13)
\end{aligned}$$

Now, by Lemma 4.3.7, we have

$$\mathbb{E}\left[(\hat{P}_t-P)^2\middle|P\right]=\frac{K^2(1-2P)^2}{(2K+t)^2}+\frac{tP(1-P)}{(2K+t)^2}.$$

Combining this with Equation (4.13) and introducing the constant

$$C=\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{hL}\right)^2}}\frac{P(1-P)}{\left(9+4\left(\frac{1-2P}{hL}\right)^2+8\frac{|1-2P|}{hL}\right)}\right]$$

we obtain, for any $K>0$,

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{\mathbb{E}\left[R_n\right]}{\log n} \\
&\geq \liminf_{n\to\infty}\frac{2}{hL\log n}\left[-\frac{2}{9(1-e^{-h^2L^2})}+\sum_{t=1}^{n-1}C\left(\frac{t}{(2K+t)^2}-\frac{K^2(1-2P)^2}{P(1-P)(2K+t)^2}\right)\right]=\frac{2C}{hL}\,. \qquad (4.14)
\end{aligned}$$

It remains to calculate a constant lower bound for C that is independent of h and L. Denote $\frac{|1-2P|}{hL}$ by Y. Then $0\leq P(1-P)=\frac{1-Y^2h^2L^2}{4}\leq 1/4$. Define $\widehat{\mathcal{G}}$ to be the event when $|Y|\leq 1$. Since P has Beta(K,K) distribution, $\mathbb{E}\left[P\right]=\frac{1}{2}$ and $\mathrm{Var}(P)=\frac{1}{8K}$. Therefore, by Chebyshev's inequality,

$$\mathbb{P}\left(\widehat{\mathcal{G}}^c\right)=\mathbb{P}\left(\left|P-\frac{1}{2}\right|>\frac{hL}{2}\right)\leq\frac{1}{2Kh^2L^2}\,,$$

and thus,

$$\begin{aligned}
C&=\mathbb{E}\left[\frac{1}{\sqrt{1+Y^2}}\frac{1-Y^2h^2L^2}{4(9+4Y^2+8Y)}\right]\geq\mathbb{E}\left[\frac{1}{\sqrt{1+Y^2}}\frac{1-Y^2h^2L^2}{4(9+4Y^2+8Y)}\mathbb{I}(\widehat{\mathcal{G}})\right] \\
&\geq\frac{1}{84\sqrt{2}}\mathbb{E}\left[(1-Y^2h^2L^2)\mathbb{I}(\widehat{\mathcal{G}})\right]\geq\frac{1}{84\sqrt{2}}\left(\mathbb{E}\left[1-Y^2h^2L^2\right]-\mathbb{P}\left(\widehat{\mathcal{G}}^c\right)\right) \\
&\geq\frac{1}{84\sqrt{2}}\left(1-\mathbb{E}\left[(1-2P)^2\right]-\frac{1}{2Kh^2L^2}\right)>\frac{1}{84\sqrt{2}}\frac{1}{2},
\end{aligned}$$

for $K>1+\frac{1}{h^2L^2}$. Hence,

$$\liminf_{n\to\infty}\frac{\mathbb{E}\left[R_n\right]}{\log n}\geq\frac{1}{84\sqrt{2}}\frac{1}{hL}.$$

The result is completed by noting that the worst-case regret is at least as big as the expected regret, thus, for every n, there exist a P and a sequence of loss vectors $f_1,\ldots,f_n$ such that the regret R_n is at least $\Omega(\frac{\log n}{hL})$. □

4.3.3 Other regularities

So far we have looked at the case when FTL achieves low regret due to the curvature of $\mathrm{bd}(\mathcal{W})$. The next result characterizes the regret of FTL when $\mathcal{W}$ is a polytope, which has a flat, non-smooth boundary and thus Theorem 4.3.1 is not applicable. For this statement recall that given some norm $\|\cdot\|$, its dual norm is defined by $\|w\|_* = \sup_{\|v\|\le 1}\langle v, w\rangle$.

Theorem 4.3.3. *Assume that $\mathcal{W}$ is a polytope and that Φ is differentiable at Θ_i, $i = 1,\dots,n$. Let $w_t = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \Theta_{t-1}\rangle$ and $W = \sup_{w_1,w_2\in\mathcal{W}}\|w_1 - w_2\|_*$. Also, we let $F = \sup_{f_1,f_2\in\mathcal{F}}\|f_1 - f_2\|$. Then the regret of FTL is*

$$R_n \le W\sum_{t=1}^{n} t\,\mathbb{I}(w_{t+1}\neq w_t)\|\Theta_t - \Theta_{t-1}\| \le FW\sum_{t=1}^{n}\mathbb{I}(w_{t+1}\neq w_t)\,.$$

Note that when $\mathcal{W}$ is a polytope, w_t is expected to "snap" to some vertex of $\mathcal{W}$. Hence, we expect the regret bound to be non-vacuous, if, e.g., Θ_t "stabilizes" around some value. Some examples after the proof will illustrate this.

Proof. Let $v = \operatorname{argmax}_{w\in\mathcal{W}}\langle w,\theta\rangle$, $v' = \operatorname{argmax}_{w\in\mathcal{W}}\langle w,\theta'\rangle$. Similarly to the proof of Theorem 4.3.1,

$$\begin{aligned}\langle v'-v,\theta'\rangle &= \langle v',\theta'\rangle - \langle v',\theta\rangle + \langle v',\theta\rangle - \langle v,\theta\rangle + \langle v,\theta\rangle - \langle v,\theta'\rangle\\ &\le \langle v',\theta'\rangle - \langle v',\theta\rangle + \langle v,\theta\rangle - \langle v,\theta'\rangle = \langle v'-v,\theta'-\theta\rangle\\ &\le W\,\mathbb{I}(v'\neq v)\|\theta'-\theta\|,\end{aligned}$$

where the first inequality holds because $\langle v',\theta\rangle \le \langle v,\theta\rangle$. Therefore, by Eq. (4.5),

$$\begin{aligned}R_n &= \sum_{t=1}^{n} t\,\langle w_{t+1}-w_t,\Theta_t\rangle \le W\sum_{t=1}^{n} t\,\mathbb{I}(w_{t+1}\neq w_t)\|\Theta_t-\Theta_{t-1}\|\\ &\le FW\sum_{t=1}^{n}\mathbb{I}(w_{t+1}\neq w_t)\,.\end{aligned}$$

□

Remark 4.3.8. Theorem 4.3.3 bounds the regret of FTL by the number of switches of the maximizers $\sum_{t=1}^{n}\mathbb{I}(w_t\neq w_{t-1})$.

As noted before, since $\mathcal{W}$ is a polytope, w_t is (generally) attained at the vertices. In this case, the epigraph of Φ is a polyhedral cone. Then, the event when $w_{t+1}\neq w_t$, i.e., when the "leader" switches corresponds to when Θ_t and Θ_{t-1} belong to different linear regions corresponding to different linear pieces of the graph of Φ.

We now spell out a corollary for the stochastic setting. In particular, FTL will often enjoy a constant regret in this case:

Corollary 4.3.9 (Stochastic setting)**.** *Assume that $\mathcal{W}$ is a polytope and that $(f_t)_{1\le t\le n}$ is an i.i.d. sequence of random variables such that $\mathbb{E}[f_i] = \mu$ and $\|f_i\|_\infty \le M$. Let $W = \sup_{w_1,w_2\in\mathcal{W}} \|w_1 - w_2\|_1$. Further assume that there exists a constant $r > 0$ such that Φ is differentiable for any ν such that $\|\nu - \mu\|_\infty \le r$. Then,*

$$\mathbb{E}[R_n] \le 2MW\left(1 + 4dM^2/r^2\right).$$

To minimize the upper bound while meeting the conditons, r can be selected to be the radius of the largest ball such that the optimal decisions for expected losses μ and ν (i.e., the maximizers defining $\Phi(-\mu)$ and $\Phi(-\nu)$) belong to the same face of $\mathcal{W}$.

Proof. Let $V = \{\nu \,|\, \|\nu - \mu\|_\infty \le r\}$. Note that the epigraph of the function Φ is a polyhedral cone. Since Φ is differentiable in the interior of V, $\{(\theta, \Phi(\theta)) \,|\, \theta \in V\}$ is a subset of a linear subspace. Therefore, for $-\Theta_t, -\Theta_{t-1} \in V$, $w_{t+1} = w_t$. Hence, by Theorem 4.3.3,

$$\mathbb{E}[R_n] \le 2MW \sum_{t=1}^{n} \mathbb{P}\left(\{-\Theta_t, -\Theta_{t-1}\} \not\subset V\right) \le 4MW\left(1 + \sum_{t=1}^{n} \mathbb{P}\left(-\Theta_t \notin V\right)\right).$$

On the other hand, note that $\|f_i\|_\infty \le M$. Hence,

$$\begin{aligned}\mathbb{P}\left(-\Theta_t \notin V\right) &= \mathbb{P}\left(\left\|\frac{1}{t}\sum_{i=1}^{t} f_i - \mu\right\|_\infty \ge r\right) \\ &\le \sum_{j=1}^{d} \mathbb{P}\left(\left|\frac{1}{t}\sum_{i=1}^{t} f_{i,j} - \mu_j\right| \ge r\right) \le 2de^{-\frac{tr^2}{2M^2}},\end{aligned}$$

where the last inequality is due to Hoeffding's inequality. Now, using that for $\alpha > 0$, $\sum_{t=1}^{n} \exp(-\alpha t) \le \int_0^n \exp(-\alpha t)dt \le \frac{1}{\alpha}$, we get $\mathbb{E}[R_n] \le 2MW\left(1 + 4dM^2/r^2\right)$. □

The condition that Φ is differentiable in a neighbourhood of μ is equivalent to that Φ is differentiable at μ. By Proposition 4.2.1, this condition requires that at μ, $\max_{w\in\mathcal{W}}\langle w, \theta\rangle$ has a unique optimizer. Note that the volume of the set of vectors θ with multiple optimizers is zero.

4.4 Adaptive algorithms

While as shown in Theorem 4.3.1, FTL can exploit the curvature of the surface of the constraint set to achieve $O(\log n)$ regret, it requires the curvature condition and $\min_t \|\Theta_t\|_2$ being bounded away from zero. When these conditions are not met, FTL may even suffer linear regret. On the other hand, many algorithms, such as the "follow the regularized leader" (FTRL) algorithm [see, e.g., Shalev-Shwartz, 2012], are known to achieve a regret guarantee of $O(\sqrt{n})$ even in the worst case (assuming bounded loss vector and bounded $\mathcal{W}$). This raises the question whether one can have an algorithm that can achieve constant or $O(\log n)$ regret in the respective settings of Corollary 4.3.9 or Theorem 4.3.1, while it still

Algorithm 6 Follow The Shrunken Leader (FTSL)

Predict $w_1 = 0$;
for $t = 2, ..., n-1$ **do**
  FTL: Compute $\tilde{w}_t = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, F_{t-1} \rangle$
  Shrinkage: Predict $w_t = \frac{\|F_{t-1}\|_2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2}} \tilde{w}_t$
end for
FTL: Compute $\tilde{w}_n = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, F_{n-1} \rangle$
Shrinkage: Predict $w_n = \frac{\|F_{n-1}\|_2}{\sqrt{\|F_{n-1}\|_2^2 + n}} \tilde{w}_n$

maintains $O(\sqrt{n})$ regret for worst-case data. One way to design an adaptive algorithm is to use the $(\mathcal{A}, \mathcal{B})$-prod algorithm of Sani et al. [2014], leading to the following result:

Proposition 4.4.1. *Consider $(\mathcal{A}, \mathcal{B})$-prod of Sani et al. [2014], where algorithm $\mathcal{A}$ is chosen to be FTRL with an appropriate regularization term, while $\mathcal{B}$ is chosen to be FTL. Then the regret of the resulting hybrid algorithm $\mathcal{H}$ enjoys the following guarantees:*

- *If FTL achieves constant regret as in the setting of Proposition 4.3.9, then the regret of $\mathcal{H}$ is also constant.*
- *If FTL achieves a regret of $O(\log n)$ as in the setting of Theorem 4.3.1, then the regret of $\mathcal{H}$ is also $O(\log n)$.*
- *Otherwise, the regret of $\mathcal{H}$ is at most $O(\sqrt{n \log n})$.*

In the next section we show that if the constraint set is the unit ball, it is possible to design adaptive algorithms directly.

4.4.1 Adaptive Algorithms for the Unit Ball Constraint Set

In this section we provide some interesting results about adaptive algorithms for the case when $\mathcal{W}$ is the unit ball in $\mathbb{R}^d$ (naturally, the results easily generalize to any ball centered at the origin). First, we show that a variant of FTL using shrinkage as regularization has $O(\log(n))$ regret when $\|\Theta_t\|_2 \geq L > 0$ for all t, but it also has $O(\sqrt{n})$ worst case guarantee. Furthermore, we show that the standard FTRL algorithm is adaptive if the constraint set is the unit ball and the loss vectors are stochastic. Throughout the section we will use the notation $F_t = -(t-1)\Theta_t = \sum_{i=1}^{t-1} f_i$.

Follow the Shrunken Leader

In this section we are going to analyze a combination of the FTL algorithm and the idea of shrinkage often used for regularization purposes in statistics. We assume that $\mathcal{W} = \left\{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq 1\right\}$ is the unit ball and, without loss of generality, we further assume that $\|f\|_2 \leq 1$ for all $f \in \mathcal{F}$.

The Follow The Shrunken Leader (FTSL) algorithm is given in Algorithm 6. The main idea of the algorithm is to predict a shrunken version of the FTL prediction, in this way keeping it away from the boundary of $\mathcal{W}$. The next theorem shows that the right amount of shrinkage leads to a robust, adaptive algorithm.

Theorem 4.4.1. *Consider the FTSL algorithm, the following hold:*

- *If there exists L such that $\|\Theta_t\|_2 \geq L > 0$ for any t, then the regret of FTSL is $O(\log(n)/L)$.*
- *Otherwise, the regret of FTSL is at most $O(\sqrt{n})$.*

Proof. By the definition of F_t and $\mathcal{W}$, $\tilde{w}_t = -F_{t-1}/\|F_{t-1}\|_2$. Let $\sigma_n = \frac{\|F_{n-1}\|_2}{\sqrt{\|F_{n-1}\|_2^2+n}}$. Our proof follows the idea of the proof of Theorem 6 by Abernethy et al. [2008]. We compute the upper bound on the value of the game for each round backwards for $t = n, n-1, \ldots, 1$, by solving the optimal strategies for f_t. The worst-case regret of FTSL is by definition

$$\begin{aligned} V_n &= \max_{f_1,\ldots,f_n} \sum_{t=1}^{n} \langle w_t, f_t \rangle - \min_{w \in \mathcal{W}} \langle w, F_n \rangle \\ &= \max_{f_1,\ldots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t \rangle + \underbrace{\max_{f_n} \|F_{n-1} + f_n\|_2 + \langle f_n, w_n \rangle}_{=:U_n} \end{aligned}$$

We first prove that U_n, the second term above, is bounded from above by $\sqrt{\|F_{n-1}\|_2^2 + n}$. To see this, let $f_n = a_n \tilde{F}_{n-1} + b_n \Omega_{n-1}$ where $\tilde{F}_{n-1}$ is the unit vector parallel to F_{n-1} and Ω_{n-1} is a unit vector orthogonal to F_{n-1}. Since $\|f_n\|_2 \leq 1$, we have $a_n^2 + b_n^2 \leq 1$. Thus, thanks to $F_{n-1} + f_n = (\|F_{n-1}\|_2 + a_n)\tilde{F}_{n-1} + b_n \Omega_{n-1}$ and $\tilde{w}_n = -\tilde{F}_{n-1}$,

$$\begin{aligned} U_n &= \max_{f_n} \sqrt{\|F_{n-1}\|_2^2 + 2a_n \|F_{n-1}\|_2 + a_n^2 + b_n^2} - a_n \sigma_n \\ &\leq \max_{a} \sqrt{\|F_{n-1}\|_2^2 + 2a\|F_{n-1}\|_2 + n} - a\sigma_n \\ &= \sqrt{\|F_{n-1}\|_2^2 + n}, \end{aligned}$$

where the last equality follows since the maximum is attained at $a = 0$. A similar statement holds for the other time indices: for any $t \geq 1$,

$$\max_{f_t} \sqrt{\|F_{t-1} + f_t\|_2^2 + t + 1} + \langle f_t, w_t \rangle \leq \sqrt{\|F_{t-1}\|_2^2 + t} + \frac{1}{\sqrt{t}} \,. \tag{4.15}$$

Before proving this inequality, let us see how it implies the second statement of the theorem:

$$\begin{aligned}
V_n &\le \max_{f_1,\dots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t\rangle + \sqrt{\|F_{n-1}\|_2^2 + n} \\
&\le \max_{f_1,\dots,f_{n-2}} \sum_{t=1}^{n-2} \langle w_t, f_t\rangle + \sqrt{\|F_{n-2}\|_2^2 + n - 1} + \frac{1}{\sqrt{n}} \\
&\le \dots \\
&\le \max_{f_1} \langle w_1, f_1\rangle + \sqrt{\|F_1\|_2^2 + 2} + \frac{1}{\sqrt{3}} + \dots + \frac{1}{\sqrt{n}} \\
&\le 1 + \sqrt{3} + \sum_{t=3}^{n} \frac{1}{\sqrt{t}} = O(\sqrt{n}).
\end{aligned}$$

Moreover, if $\|\Theta_t\|_2 \ge L$ for $1 \le t \le n$, a stronger version of Equation (4.15) also holds:

$$\max_{f_t} \sqrt{\|F_{t-1} + f_t\|_2^2 + t + 1} + \langle f_t, w_t\rangle \le \sqrt{\|F_{t-1}\|_2^2 + t} + \frac{1}{(t-1)L}. \tag{4.16}$$

This implies the first statement of the theorem, since

$$\begin{aligned}
V_n &\le \max_{f_1,\dots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t\rangle + \sqrt{\|F_{n-1}\|_2^2 + n} \\
&\le \max_{f_1,\dots,f_{n-2}} \sum_{t=1}^{n-2} \langle w_t, f_t\rangle + \sqrt{\|F_{n-2}\|_2^2 + n - 1} + \frac{1}{(n-1)L} \\
&\le \dots \\
&\le 1 + \sum_{t=1}^{n-1} \frac{1}{tL} = O(\log(n)/L).
\end{aligned}$$

To finish the proof, it remains to show Eqs. (4.15) and (4.16). As before, let $f_t = a_t \tilde{F}_{t-1} + b_t \Omega_{t-1}$ where $\tilde{F}_{t-1}$ is the unit vector parallel to F_{t-1} and Ω_{t-1} is a unit vector orthogonal to F_{t-1}. Again, since $\|f_t\|_2 \le 1$, observe that $a_t^2 + b_t^2 = \|f_t\|_2 \le 1$. Now, let $\sigma_t = \frac{\|F_{t-1}\|_2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2}}$. Then, for any $t \ge 1$,

$$\begin{aligned}
\Delta_t &= \max_{f_t} \sqrt{\|F_t\|_2^2 + t + 1} - a_t\sigma_t - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&= \max_{f_t} \sqrt{\|F_{t-1}\|_2^2 + 2a_t\|F_{t-1}\|_2 + a_t^2 + b_t^2 + t + 1} - a_t\sigma_t - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&\le \max_{a_t} \sqrt{\|F_{t-1}\|_2^2 + 2a_t\|F_{t-1}\|_2 + t + 2} - a_t\sigma_t - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&= \sqrt{\|F_{t-1}\|_2^2 + t + 2} - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&= \frac{2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2} + \sqrt{\|F_{t-1}\|_2^2 + t}} \\
&\le \frac{1}{\sqrt{t}}.
\end{aligned} \tag{4.17}$$

This proves Equation (4.15). Moreover, if $\|F_{t-1}\|_2 = \|(t-1)\Theta_{t-1}\|_2 \ge (t-1)L$, by Equation (4.17) we obtain

$$\Delta_t \le \frac{2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2} + \sqrt{\|F_{t-1}\|_2^2 + t}} \le \frac{1}{\|F_{t-1}\|_2} \le \frac{1}{(t-1)L},$$

proving Equation (4.16). □

FTRL for the case of the unit ball constraint set

This section is to show that in the case when $\mathcal{W}$ is the unit ball in ℓ_2 norm, FTRL regularized with $R(w) = \frac{1}{2}\|w\|^2$ is adaptive in the sense that its regret improves dramatically in the benign stochastic setting. To fix the notation, in round t, FTRL predicts

$$w_t = \operatorname*{argmin}_{w \in \mathcal{W}} \eta_t \langle F_{t-1}, w \rangle + R(w),$$

if $t > 1$, while we let $w_1 = 0$. It is known that FTRL with $\eta_t = 1/\sqrt{t-1}$ is guaranteed to achieve $O(\sqrt{n})$ regret in the adversarial setting, see, e.g., Shalev-Shwartz [2012]. It remains to prove that FTRL indeed achieves a fast rate in the stochastic setting.

Theorem 4.4.2. *Assume that the sequence of loss vectors, $f_1, \dots, f_n \in \mathbb{R}^d$ satisfies $\|f_t\|_2 \leq 1$ almost surely and $\mathbb{E}[f_t] = \mu$ for all t with some $\|\mu\|_2 > 0$. Then the regret of FTRL with $\eta_t = 1/\sqrt{t-1}$ satisfies that*

$$\mathbb{E}[R_n] \leq \frac{8}{\|\mu\|_2^2} + \|\mu\|_2^2 + O(\frac{\log n}{\|\mu\|_2}).$$

Proof. Using $R(w) = \frac{1}{2}\|w\|^2$ as its regularization, in round $t > 1$ FTRL predicts

$$w_t = \operatorname*{argmin}_{w \in \mathcal{W}} \eta_t \langle F_{t-1}, w \rangle + R(w) = \begin{cases} \frac{1}{\sqrt{t-1}} F_{t-1}, & \text{if } \|F_{t-1}\| \leq \sqrt{t-1}; \\ \frac{F_{t-1}}{\|F_{t-1}\|}, & \text{otherwise.} \end{cases} \tag{4.18}$$

For any $1 \leq t \leq n$, denote the event $\|F_t\| \geq \sqrt{t}$ by $\mathcal{E}_t$. Note that if $\|F_{t-1}\| \geq \sqrt{t-1}$, FTRL predicts exactly the same w_t as FTL. Denote the total loss of FTL in n rounds by $\mathcal{L}_n^{FTL}$. Thus, the regret of FTRL is

$$\begin{aligned} \mathbb{E}[R_n] &= \mathbb{E}\left[\sum_{t=1}^n \langle f_t, w_t \rangle - \min_{w \in \mathcal{W}} \sum_{t=1}^n \langle f_t, w \rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \langle f_t, w_t \rangle - \mathcal{L}_n^{FTL}\right] + \mathbb{E}\left[\mathcal{L}_n^{FTL} - \min_{w \in \mathcal{W}} \sum_{t=1}^n \langle f_t, w \rangle\right] \\ &\leq 2\sum_{t=1}^n \mathbb{P}(\mathcal{E}_t^c) + O(\frac{\log n}{\|\mu\|_2}), \end{aligned} \tag{4.19}$$

where, to obtain the last inequality, we applied Equation (4.18) to the first term, while the second term is $O(\log n)$ by the discussion following Theorem 4.3.1. It remains to bound the first term, $2\sum_{t=1}^n \mathbb{P}(\mathcal{E}_t^c)$ in Eq. (4.19). For any $t > \frac{4}{\|\mu\|_2^2}$,

$$\begin{aligned} \mathbb{P}\left(\|F_t\|_2 \leq \sqrt{t}\right) &\leq \mathbb{P}\left(\|F_t\|_2 < \frac{t}{2}\|\mu\|_2\right) \leq \sum_{i=1}^d \mathbb{P}\left(|F_{t,i}| < \frac{t}{2}|\mu_i|\right) \\ &\leq \sum_{i=1}^d \mathbb{P}\left(|F_{t,i} - t\mu_i| > \frac{t}{2}|\mu_i|\right) \leq 2\sum_{i=1}^d e^{-\frac{\mu_i^2}{4}t}, \end{aligned}$$

where the last inequality follows by Hoeffding's inequality. Thus,

$$\begin{aligned}\sum_{t=1}^{n}\mathbb{P}\left(\mathcal{E}_t^c\right) &= \sum_{t=1}^{\lceil 4/\|\mu\|_2^2\rceil-1}\mathbb{P}\left(\mathcal{E}_t^c\right)+\sum_{t=\lceil 4/\|\mu\|_2^2\rceil}^{n}\mathbb{P}\left(\mathcal{E}_t^c\right)\\ &\le \frac{4}{\|\mu\|_2^2}+2\sum_{i=1}^{d}\sum_{t=0}^{n}e^{-\frac{\mu_i^2}{4}t}\\ &\le \frac{4}{\|\mu\|_2^2}+2\sum_{i=1}^{d}\frac{1}{1-e^{-\frac{\mu_i^2}{4}}}\\ &\le \frac{4}{\|\mu\|_2^2}+2\sum_{i=1}^{d}\frac{\mu_i^2}{4}=\frac{4}{\|\mu\|_2^2}+\frac{\|\mu\|_2^2}{2}\ .\end{aligned}$$

where in the last inequality we used $1/(1-e^{-a}) \le a$ which holds for any $a \ge 0$. Therefore, if $\|\mu\| > 0$, the regret of FTRL satisfies

$$\mathbb{E}\left[R_n\right] \le \frac{8}{\|\mu\|_2^2}+\|\mu\|_2^2+O(\frac{\log n}{\|\mu\|_2}).$$

□

4.5 Experimental results

We performed three simulations to illustrate the differences between FTL, FTRL with the regularizer $R(w) = \frac{1}{2}\|w\|_2^2$ and the adaptive algorithm ($\mathcal{A}$, $\mathcal{B}$)-prod (AB) using FTL and FTRL as its candidates. We will call this latter algorithm AB(FTL,FTRL).

For the experiments the constraint set $\mathcal{W}$ was chosen to be a slightly elongated ellipsoid in the 4-dimensional Euclidean space, with volume matching that of the 4-dimensional unit ball. The actual ellipsoid is given by $\mathcal{W} = \left\{w \in \mathbb{R}^4 \,|\, w^\top Q w \le 1\right\}$ where Q is chosen "randomly" to be

$$Q = \begin{pmatrix} 4.3367 & 3.6346 & -2.2250 & 3.5628 \\ 3.6346 & 3.9966 & -2.3613 & 3.2817 \\ -2.2250 & -2.3613 & 2.0589 & -2.1295 \\ 3.5628 & 3.2817 & -2.1295 & 3.4206 \end{pmatrix}.$$

We experimented with 3 types of data to illustrate the behavior of the different algorithms: stochastic, "half-adversarial", and "worst-case" data (worst-case for FTL), as will be explained below. The first two datasets are random, so the experiments were repeated 100 times, and we report the average regret with its standard deviation; the worst case data is deterministic, so there no repetition was needed. For each experiment, we set $n = 2500$. The regularization coefficient for FTRL, and the learning rate for AB were chosen based on their theoretical bounds minimizing the worst-case regret.

Stochastic data. In this setting we used the following model to generate f_t: Let $(\hat{f}_t)_t$ be an i.i.d. sequence drawn from the 4-dimensional standard normal distribution, and let

$\tilde{f}_t = \hat{f}_t / \left\|\hat{f}_t\right\|_2$. Then, f_t is defined as $f_t = \tilde{f}_t + Le_1$ where $e_1 = (1, 0, \dots, 0)^\top$. Therefore, $\mathbb{E}\left[\left\|\frac{1}{t}\sum_{s=1}^{t} f_s\right\|_2\right] \to L$ as $t \to \infty$. In the experiments we picked $L \in \{0, 0.1\}$.

The results are shown in Fig. 4.5. On the left-hand side we plotted the regret against the logarithm of the number of rounds, while on the right-hand side we plotted the regret against the square root of the number of rounds, together with the standard deviation of the results over the 100 independent runs. As can be seen from the figures, when $L = 0.1$, the growth-rate of the regret of FTL is indeed logarithmic, while when $L = 0$, the growth-rate is $\Theta(\sqrt{n})$. In particular, when $L = 0.1$, FTL enjoys a major advantage compared to FTRL, while for $L = 0$, FTL and FTRL perform essentially the same (in this special case, the regret of FTL will indeed be $O(\sqrt{n})$ as w_t will stay bounded but $\|\Theta_t\| = O(1/\sqrt{t})$). As expected, AB(FTL,FTRL), gets the better of the two regrets with little to no extra penalty.

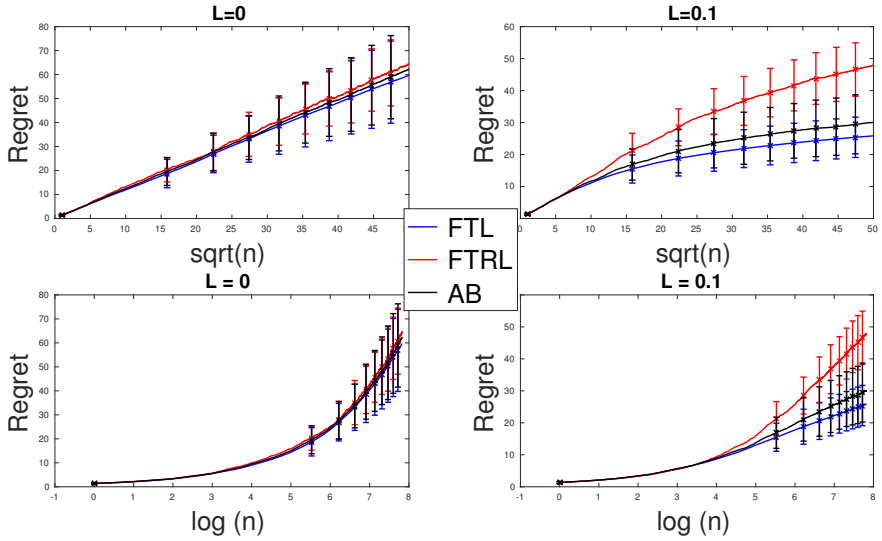

Figure 4.5: Experimental results for stochastic data.

"Half-adversarial" data The half-adversarial data used in this experiment is the optimal solution for the adversary in the *linear game* when $\mathcal{W}$ is the unit ball [Abernethy et al., 2008]. This data is generated as follows: The sequence $\hat{f}_t$ for $t = 1, \dots, n$ is generated randomly in the $(d-1)$-dimensional subspace $S = \text{span}\{e_2, \dots, e_d\}$ (here e_i is the ith unit vector in $\mathbb{R}^d$) as follows: $\hat{f}_1$ is drawn from the uniform distribution on the unit sphere of S (actually $\mathbb{S}_{d-2}$). For $t = 2, \dots, n$, $\hat{f}_t$ is drawn from the uniform distribution on the unit sphere of the intersection of S and the hyperplane perpendicular to $\sum_{i=1}^{t-1} \hat{f}_i$ and going through the origin. Then, $f_t = Le_1 + \sqrt{1 - L^2}\hat{f}_t$ for some $L \geq 0$.

The results are reported in Fig. 4.6. When $L = 0$, the regret of both FTL and FTRL grows as $O(\sqrt{n})$. When $L = 0.1$, FTL achieves $O(\log n)$ regret, while the regret of FTRL

appears to be $O(\sqrt{n})$. AB(FTL,FTRL) closely matches the regret of FTL.

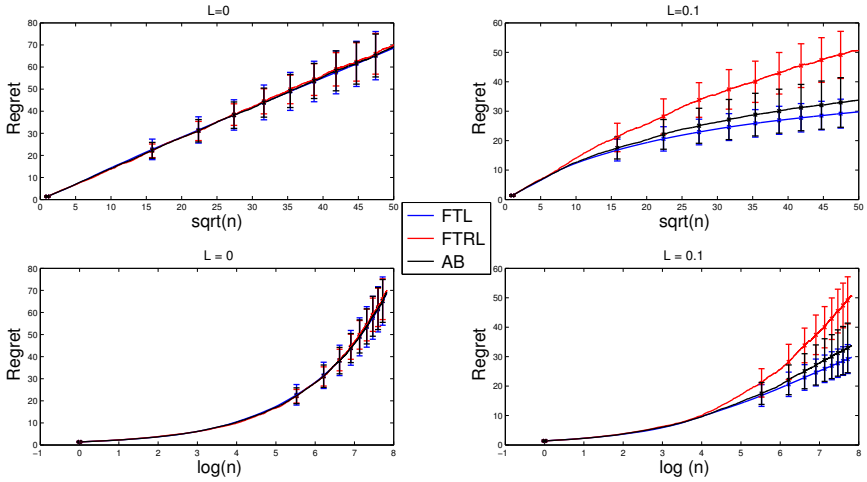

Figure 4.6: Experimental results for "half-adversarial" data.

Worst-case data We also tested the algorithms on data where FTL is known to suffer linear regret, mainly to see how well AB(FTL,FTRL) is able to deal with this setting. In this case, we set $f_{t,i} = 0$ for all t and $i \geq 2$, while for the first coordinate, $f_{1,1} = 0.9$, and $f_{t,1} = 2(t \mod 2) - 1$ for $t \geq 2$.

The results are reported in Fig. 4.7. It can be seen that the regret of FTL is linear (as one can easily verify theoretically), and AB(FTL,FTRL) succeeds to adapt to FTRL, and they both achieve a much smaller $O(\sqrt{n})$ regret.

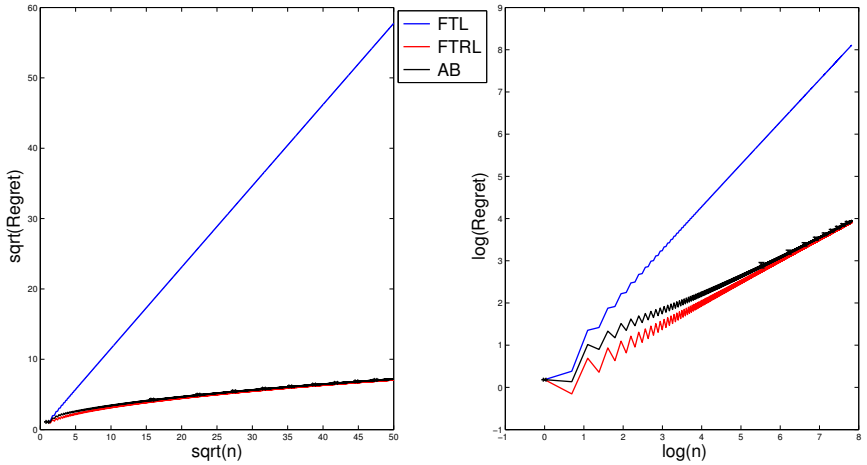

Figure 4.7: Experimental results for the worst-case data.

The unit ball We close this section by comparing the performance of our adaptive algorithms on the unit ball, namely, FTL, FTSL, FTRL, and AB(FTL, FTRL). All these algorithms are parametrized as above. The problem setup is similar to the stochastic data setting and the worst-case data setting. Again, we consider a 4-dimensional setting, that is, $\mathcal{W}$ is the unit ball in $\mathbb{R}^4$ centered at the origin. The worst-case data is generated exactly as above, while the generation process of the stochastic data is slightly modified to increase the difference between FTRL and FTL: we sample the i.i.d. vectors $\hat{f}_t$ from a zero-mean normal distribution with independent components whose variance is $1/16$, and let $\tilde{f}_t = \hat{f}_t$ if $\|\hat{f}_t\|_2 \leq 1$ and $\tilde{f}_t = \hat{f}_t / \left\|\hat{f}_t\right\|_2$ when $\left\|\hat{f}_t\right\|_2 > 1$ (i.e., we only normalize if $\hat{f}_t$ falls outside of the unit ball). The reason of this modification is to encourage the occurrence of the event $\|F_{t-1}\|_2 < \sqrt{t-1}$. Recall that when $\|F_{t-1}\|_2 \geq \sqrt{t-1}$, the prediction of FTRL matches that of FTL, so we are trying to create some data where their behavior is actually different. As a result, we will be able to observe that the predictions of FTL and FTRL are different in the early rounds. Finally, as before, we let $f_t = \tilde{f}_t + Le_1$, and set the time horizon to $n = 20,000$.

The results of the simulation of the stochastic data setting are shown in Figure 4.8. In the case of $L = 0.1$, FTRL suffers more regret at the beginning for some rounds, but then succeeds to match the performance of FTL. The results of the simulation of the worst-case data setting are shown in Figure 4.9, where FTSL has similar performance as FTRL.

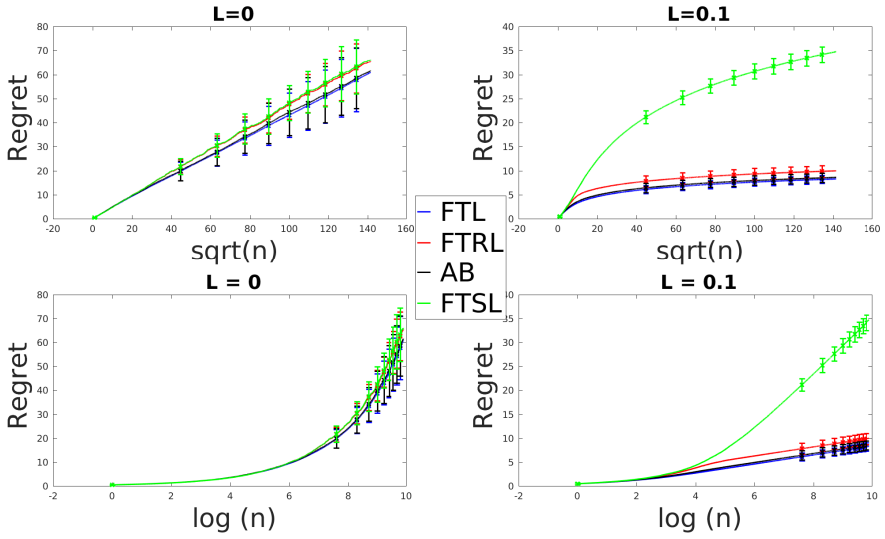

Figure 4.8: Experimental results for stochastic data when $\mathcal{W}$ is the unit ball.

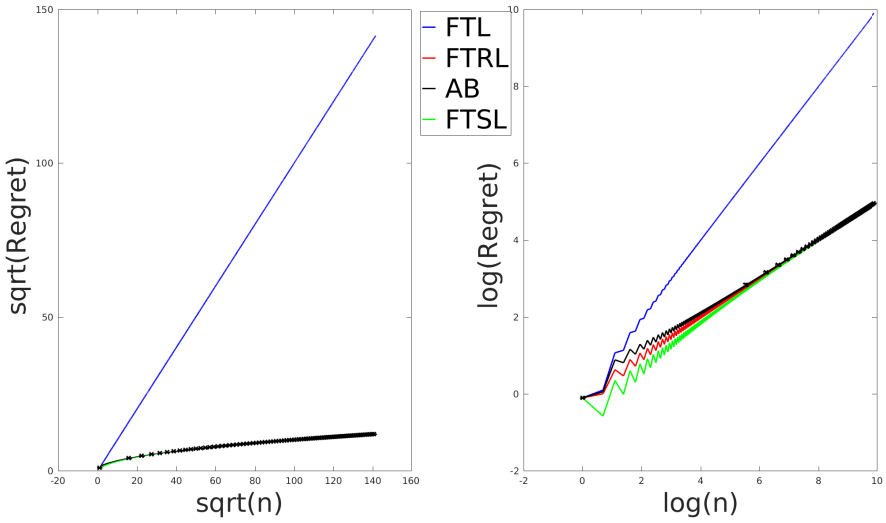

Figure 4.9: Experimental results for the worst-case data when $\mathcal{W}$ is the unit ball.

4.6 Conclusion

FTL is a simple method that is known to perform well in many settings, while existing worst-case results fail to explain its good performance. While taking a thorough look at why and when FTL can be expected to achieve small regret, we discovered that the curvature of the boundary of the constraint and having average loss vectors bounded away from zero help keep the regret of FTL small. These conditions are significantly different from previous conditions on the curvature of the loss functions which have been considered extensively in the literature. It would be interesting to further investigate this phenomenon for other algorithms or in other learning settings.

4.7 Technical lemmas for Theorem 4.3.2

We present the proofs for several technical lemmas used in the proof of Theorem 4.3.2.

Lemma 4.3.5. *Under the assumptions of Theorem 4.3.2, for any* $0 < p_1, p_2 < 1$,

$$\langle w^{p_2} - w^{p_1}, f^{p_1} \rangle \geq \frac{hL}{2} \frac{\left(\frac{2p_2 - 2p_1}{hL}\right)^2}{\sqrt{1 + \left(\frac{1-2p_1}{hL}\right)^2} \left(1 + \left(\frac{1-2p_2}{hL}\right)^2\right)}.$$

Proof. It is easy to see that for any p, w^p is on the boundary of $\mathcal{W}$, that is, $w^p = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, f^p \rangle = (\cos(\varphi^p), h \sin(\varphi^p))$ for some $\varphi^p \in [0, 2\pi]$ that changes smoothly with p. Then $\langle w^p, f^p \rangle = (2p-1)\cos(\varphi^p) - Lh \sin(\varphi^p)$, and so taking the derivative of the loss w.r.t. φ^p, it is easy to verify that $\tan(\varphi^p) = \frac{Lh}{1-2p}$ and $\sin(\varphi^p) = \frac{Lh}{\sqrt{(Lh)^2 + (1-2p)^2}} > 0$. Thus,

$1-2p_1 = \frac{Lh\cos(\varphi^{p_1})}{\sin(\varphi^{p_1})}$. To simplify notation, let $\varphi_1 = \varphi^{P_1}$ and $\varphi_2 = \varphi^{P_2}$. Then,

$$\begin{aligned}
\langle w^{P_2} - w^{P_1}, f^{P_1}\rangle &= \left\langle \begin{pmatrix} \cos\varphi_2 - \cos\varphi_1 \\ h(\sin\varphi_2 - \sin\varphi_1) \end{pmatrix}, \begin{pmatrix} \frac{-hL\cos\varphi_1}{\sin\varphi_1} \\ -L \end{pmatrix} \right\rangle \\
&= -hL\left((\cos(\varphi_2) - \cos(\varphi_1))\frac{\cos(\varphi_1)}{\sin(\varphi_1)} + (\sin(\varphi_2) - \sin(\varphi_1)) \right) \\
&= \frac{-hL}{\sin(\varphi_1)}\Big(\cos(\varphi_2)\cos(\varphi_1) - \cos^2(\varphi_1) \\
&\quad + \sin(\varphi_1)\sin(\varphi_2) - \sin^2(\varphi_1) \Big) \\
&= \frac{hL}{\sin(\varphi_1)}\left(1 - \cos(\varphi_2)\cos(\varphi_1) - \sin(\varphi_1)\sin(\varphi_2)\right) \\
&= \frac{hL}{\sin(\varphi_1)}\left(1 - \cos(\varphi_1 - \varphi_2)\right) \\
&= \frac{hL}{\sin(\varphi_1)}\left(\frac{1}{2}\left(\cos(\varphi_1 - \varphi_2) - 1\right)^2 + \frac{1}{2}\sin^2(\varphi_1 - \varphi_2) \right) \qquad (4.20) \\
&\geq \frac{hL}{2\sin(\varphi_1)}\sin^2(\varphi_1 - \varphi_2) \\
&= \frac{hL}{2}\sin(\varphi_1)\sin^2\varphi_2\left(\cot(\varphi_1) - \cot(\varphi_2)\right)^2 \ . \qquad (4.21)
\end{aligned}$$

The proof is finished by substituting $\cot(\varphi_i) = \frac{1-2P_i}{hL}$, $\sin(\varphi_1) = \frac{1}{\sqrt{1+\left(\frac{1-2P_1}{Lh}\right)^2}}$ and $\sin^2(\varphi_2) = \frac{1}{1+\left(\frac{1-2P_2}{Lh}\right)^2}$. □

Lemma 4.3.6. *For any $u > 0$,*

$$\mathbb{P}\left[|\hat{P}_t - P| > \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}u \,\middle|\, P \right] \leq 2\exp(-tu^2) \ .$$

Proof. From Eq. (4.11) and the triangle inequality, $|\hat{P}_t - P| \leq \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}|\frac{S_t}{t} - P|$, and so

$$\mathbb{P}\left[|\hat{P}_t - P| > \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}u \,\middle|\, P \right] \leq \mathbb{P}\left[\left|\frac{S_t}{t} - P\right| > u \,\middle|\, P \right] \leq 2\exp(-tu^2),$$

where the last inequality is thanks to that conditioned on P, $X_1, \ldots, X_t$ are independent Bernoulli random variables with expectation P, thus it holds by Hoeffding's inequality (see, e.g., [Cesa-Bianchi and Lugosi, 2006, Corollary A.1]). □

Lemma 4.3.7. *For any $t \geq 0$,*

$$\mathbb{E}\left[(P - \hat{P}_t)^2 \,\middle|\, P \right] = \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{tP(1-P)}{(2K+t)^2}.$$

Proof. Starting from Eq. (4.11), we have

$$\begin{aligned}
\mathbb{E}\left[(P - \hat{P}_t)^2 \,\middle|\, P \right] &= \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{1}{(2K+t)^2}\mathbb{E}\left[(S_t - tP)^2 \,\middle|\, P \right] \\
&= \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{tP(1-P)}{(2K+t)^2},
\end{aligned}$$

where the first equality is due to $\mathbb{E}[S_t - Pt|\, P] = 0$, and the last equality is due to that conditioned on P, S_t has a Binomial distribution with parameters t and P. □

Chapter 5

Conclusions and Future Works

In this thesis, we explore the instance-dependent analyses of some learning algorithms in various learning settings. Instead of starting with the stochastic assumptions or overly conservative worst-case examples when analyzing the learning algorithms, we propose to start with minimal statistical assumptions and focus on catching what properties of data essentially affect the performance of the learning algorithms. Such instance-dependent results are usually more expressive and lead to a better understanding of the successes and failures of the learning algorithms.

In the first part of thesis we present an analysis of the task of Independent Component Analysis (ICA) with no stochastic assumptions on the data. We develop the first ICA algorithm in the literature that can recover noisy signals with polynomial computational complexity and provable performance guarantees on the reconstruction error. Several important features of the data are proposed to characterize the "niceness" of the data and the instance-dependent performance guarantee of our algorithm. Originated from a deterministic analysis, our results can recover the usual statistical results in the classic ICA setting, and also extend to deterministic source signals (and potentially other type of source signals) with approximately independent empirical distributions. It remains future work to improve the dependence of the performance on the condition number of the mixing matrix A and the dimension d. Such improvement is necessary for the algorithm to be applied to high dimensional data.

The second contribution of this thesis is an instance-dependent generalization bound for the empirical risk minimization algorithm (ERM) on the partially linear model. Based on a surprising example that ERM achieves infinite expected risk/loss on a perfectly innocent looking least-square linear regression problem, we found that the performance of ERM is highly related to $\mathbb{E}\left[\hat{\lambda}_{\min}^{-1}\right]$, the expectation of the inverse of the minimal positive eigenvalue of the empirical Grammian matrix. We developed a high probability finite-sample bound for this setting, which again showed a dependence on $\mathbb{E}\left[\hat{\lambda}_{\min}^{-1}\right]$. Our result partially explains the success of ERM on the partially linear model. On the other hand, understanding

the behaviour of this quantity for different distributions remains an important problem to study the finite-sample performance of ERM on (partially) linear model. Also, while a high-probability generalization bound has been provided, the question what data is "nice" such that an finite excess risk bound exists remains open.

The last main topic of this thesis is an analysis of the simple "Follow the Leader" (FTL) algorithm in the online learning setting. Although existing worst-case results of FTL imply that FTL may suffer linear regret, we discover several key features of the data that make it possible for FTL to achieve fast rates. One of these is the magnitude of the average loss vector. We prove that the performance of FTL inversely depends on this magnitude, and an asymptotic lower bound shows that this dependence is essential. Lastly we propose various adaptive algorithms that can achieve fast rates when this magnitude is large, while still guaranteed to achieve the standard learning rate in the worst-case. As for future work, it is also interesting to see how our approach can be applied to achieve improvement in other setting, e.g. for strictly convex losses. It is also interesting to convey similar idea in the linear bandit problem to bridge results between the stochastic and the adversarial settings.

Lastly, there are also some general open questions for the proposed "instance-dependent" analysis framework. While the performances of the learning algorithms are shown to essentially depend on some instance-dependent quantities, it remains open how one can compute (or estimate) these quantities in practice, e.g. in the ICA task in Chapter 2 or in the online linear prediction problem in Chapter 4. In the case when these quantities (or their good estimates) are not available, it also remains open how to develop adaptive algorithms in general, preparing for the worst case. Finally, while we succeed in developing instance-dependent upper bound for the learning algorithms and learning problems, the development of instance-and-algorithm-dependent lower bounds seem to be more challenging and as for now remain for future work.

Bibliography

Y. Abbasi-Yadkori. *Forced-exploration based algorithms for playing in bandits with large action sets*. Library and Archives Canada, 2010.

J. Abernethy, P.L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *21st Annual Conference on Learning Theory (COLT)*, 2008.

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. *arXiv preprint arXiv:1405.6076*, 2014.

A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012a.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012b.

A. Anandkumar, Y. Liu, D.J. Hsu, D.P. Foster, and S.M. Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012c.

A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

P.L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 65–72, 2007.

M. Belkin, L. Rademacher, and J. Voss. Blind signal separation in the presence of gaussian noise. In *Conference on Learning Theory*, pages 270–287, 2013.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36:489–531, 2008.

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

J. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Information Theory*, 50(9):2050–2057, 2004.

J.T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Nederlandica*, 51(3): 287–317, 1997.

A. Chen and P.J. Bickel. Consistent independent component analysis and prewhitening. *Signal Processing, IEEE Transactions on*, 53(10):3625–3632, 2005.

A. Chen and P.J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855, 2006.

D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou. Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research*, 5:1143–1175, December 2004.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

T. M. Cover. Behavior of sequential predictors of binary sequences. Technical report, STANFORD UNIV CALIF STANFORD ELECTRONICS LABS, 1966.

F. Cucker and D.-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

A. DasGupta. Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*, pages 221–248. Springer, 2011.

S. De Rooij, T. van Erven, P. Grünwald, and W.M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

A. Dermoune and T. Wei. FastICA algorithm: Five criteria for the optimal choice of the nonlinearity function. *IEEE transactions on signal processing*, 61(5-8):2078–2087, 2013.

R. Dudley. The sizes of compact subsets of Hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

R.M. Dudley. *A course on empirical processes*. Ecole d'Eté de Probabilités de St. Flour, 1982, Lecture Notes in Mathematics. Springer, 1984.

R.F. Engle, C.W.J. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81:310–320, 1986.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003.

E. Forootan and J. Kusche. Separation of deterministic signals using independent component analysis (ica). *Studia Geophysica et Geodaetica*, 57(1):17–26, 2013.

D.J. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3357–3365, 2015.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *37th IEEE Annual Symposium on Foundations of Computer Science*, pages 359–359. IEEE Computer Society, 1996.

A.A. Gaivoronski and F. Stella. Stochastic nonstationary optimization for finding universal portfolios. *Annals of Operations Research*, 100(1–4):165–188, 2000.

J. Gao. *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, volume 108 of *Monographs on Statistics and Applied Probability*. Taylor & Francis, 2007.

D. Garber and E. Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 951, pages 541–549, 2015.

E. Giné and J. Zinn. On the central limit theorem for empirical processes. *Annals of Probability*, 12:929–989, 1984.

A. Gittens and J.A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint arXiv:1104.4513*, 2011.

N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 584–593. ACM, 2014.

W. Greblicki and M. Pawlak. *Nonparametric system identification*. Cambridge University Press, 2008.

L. Györfi, M. Kohler, A. Kryżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

C.J. Hillar and L. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

J.L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer, 2009.

D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

D. Hsu, S.M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *COLT*, pages 9.1–9.24, 2012.

R. Huang and C. Szepesvári. A finite-sample generalization bound for semiparametric regression: Partially linear models. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2014a.

R. Huang and C. Szepesvári. A finite-sample generalization bound for semiparametric regression: Partially linear models. In *AISTATS*, pages 402–410, 2014b.

R. Huang, A. György, and C. Szepesvári. Deterministic independent component analysis. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2521–2530, 2015a.

R. Huang, A. György, and C. Szepesvári. Easy data for independent component analysis. *Workshop "Learning Faster from Easy Data II", Advances in Neural Information Processing Systems (NIPS)*, 2015b.

R. Huang, T. Lattimore, A. György, and C. Szepesvári. Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems*, pages 4970–4978, 2016.

J. Hüsler. Minimal spacings of non-uniform densities. *Stochastic processes and their applications*, 25:73–81, 1987.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.

S. M. Kakade and S. Shalev-Shwartz. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1457–1464, 2009.

M. Kearns and U.V. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

W. Kienzle and B. Schölkopf. Training support vector machines with multiple equality constraints. In *Proceedings of 16th European Conference on Machine Learning*, pages 182–193, 2005.

T. Kirimoto, T. Amishima, and A. Okamura. Separation of mixtures of complex sinusoidal signals with independent component analysis. *IEICE transactions on communications*, 94(1):215–221, 2011.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. *arXiv preprint arXiv:1309.1007*, 2013.

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

W. Kotłowski. Minimax strategy for prediction with expert advice under stochastic assumptions. *Algorithmic Learning Theory (ALT)*, 2016.

S. Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Dept. Comput. Sci., Univ. Chicago, Chicago, IL, Tech. Rep. TR-2002-04*, 2002.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

S. Lee and S.J. Wright. Decomposition algorithms for training large-scale semiparametric support vector machines. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pages 1–14, Berlin, Heidelberg, 2009. Springer-Verlag.

W.S. Lee, P.L. Bartlett, and R.C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

E.S. Levitin and B.T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

H.B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and implicit updates. arXiv, 2010.

N. Merhav and M. Feder. Universal sequential learning and decision from individual data sequences. In *5th Annual ACM Workshop on Computational Learning Theory (COLT)*, pages 413—427. ACM Press, 1992.

J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *arXiv preprint arXiv:1406.4765*, 2014.

M.N.H. Mollah, S. Eguchi, and M. Minami. Robust prewhitening for ica by minimizing β-divergence and its application to fastica. *Neural Processing Letters*, 25(2):91–110, 2007.

E. Oja and Z. Yuan. The FastICA algorithm revisited: Convergence analysis. *Neural Networks, IEEE Transactions on*, 17(6):1370–1381, 2006.

E. Ollila. The deflation-based FastICA estimator: statistical analysis revisited. *Signal Processing, IEEE Transactions on*, 58(3):1527–1541, 2010.

F. Orabona, N. Cesa-Bianchi, and C. Gentile. Beyond logarithmic bounds in online learning. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 823–831, 2012.

B.A. Pires and C. Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. *arXiv preprint arXiv:1206.6444*, 2012.

D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.

E.S. Polovinkin. Strongly convex analysis. *Sbornik: Mathematics*, 187(2):259, 1996.

A.N. Pressley. *Elementary differential geometry*. Springer Science & Business Media, 2010.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *26th Annual Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

A. Rakhlin and K. Sridharan. Stat928: Statistical learning theory and sequential prediction. *Lecture Notes in University of Pennsyvania*, 2014.

P.M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

A. Samarov and A. Tsybakov. Nonparametric independent component analysis. *Bernoulli*, 10(4):565–582, 2004.

A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 810–818, 2014.

R. Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, 2nd edition, 2014.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2012.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

A.J. Smola, T.T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines, 1998.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer-Verlag New York, 2008.

G.W. Stewart and J.-g. Sun. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309.

J.H. Stock. Nonparametric policy analysis. *Journal of the American Statistical Association*, 84(406):567–575, 1989.

J.H. Stock. Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 77–98, 1991.

Z. Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014. (https://bitbucket.org/szzoli/ite/).

A. Tewari and P.L. Bartlett. Learning theory. In R. Chellappa and S. Theodoridis, editors, *Academic Press Library in Signal Processing*, volume 1, chapter 14. Elsevier, 1st edition, 2013. to appear.

P. Tichavsky, Z. Koldovsky, and E. Oja. Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *Signal Processing, IEEE Transactions on*, 54(4):1189–1203, 2006.

S. van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

S. van de Geer. *Empirical processes in M-estimation*, volume 45. Cambridge University Press, 2000.

T. van Erven, W. Kotlowski, and M.K. Warmuth. Follow the leader with dropout perturbations. In *COLT*, pages 949–974, 2014.

T. van Erven, P. Grünwald, N. Mehta, M. Reid, and R. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research (JMLR)*, 16:1793–1861, 2015. Special issue in Memory of Alexey Chervonenkis.

V.N. Vapnik. The nature of statistical learning theory, 1995.

V.N. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

S. Vempala and Y. Xiao. Max vs min: Independent component analysis with nearly linear sample complexity. *CoRR*, abs/1412.2954, 2014.

E.D. Vito, L. Rosasco, A. Caponnetto, U.D. Giovannini, and F. Odone. Learning from examples as an inverse problem. In *Journal of Machine Learning Research*, pages 883–904, 2005.

T. Wei. The convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *arXiv preprint arXiv:1408.0145*, 2014.

Y. Xiao. Fourier pca package. *GitHub*, 2014. URL https://github.com/yingusxiaous/libFPCA.

T. Zhang and G.H. Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550, 2001.