# Applications of Empirical Processes in Learning Theory: Algorithmic Stability and Generalization Bounds

by

Alexander Rakhlin

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006



Author .....
Department of Brain and Cognitive Sciences
May 15, 2006

Certified by .....
Tomaso Poggio
Eugene McDermott Professor of Brain Sciences
Thesis Supervisor

Accepted by .....
Matt Wilson
Head, Department Graduate Committee

# Applications of Empirical Processes in Learning Theory: Algorithmic Stability and Generalization Bounds

by

Alexander Rakhlin

Submitted to the Department of Brain and Cognitive Sciences
on May 15, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

This thesis studies two key properties of learning algorithms: their generalization ability and their stability with respect to perturbations. To analyze these properties, we focus on concentration inequalities and tools from empirical process theory. We obtain theoretical results and demonstrate their applications to machine learning.

First, we show how various notions of stability upper- and lower-bound the bias and variance of several estimators of the expected performance for general learning algorithms. A weak stability condition is shown to be equivalent to consistency of empirical risk minimization.

The second part of the thesis derives tight performance guarantees for greedy error minimization methods – a family of computationally tractable algorithms. In particular, we derive risk bounds for a greedy mixture density estimation procedure. We prove that, unlike what is suggested in the literature, the number of terms in the mixture is not a bias-variance trade-off for the performance.

The third part of this thesis provides a solution to an open problem regarding the stability of Empirical Risk Minimization (ERM). This algorithm is of central importance in Learning Theory. By studying the suprema of the empirical process, we prove that ERM over Donsker classes of functions is stable in the $L_1$ norm. Hence, as the number of samples grows, it becomes less and less likely that a perturbation of $o(\sqrt{n})$ samples will result in a very different empirical minimizer. Asymptotic rates of this stability are proved under metric entropy assumptions on the function class. Through the use of a ratio limit inequality, we also prove stability of expected errors of empirical minimizers. Next, we investigate applications of the stability result. In particular, we focus on procedures that optimize an objective function, such as $k$-means and other clustering methods. We demonstrate that stability of clustering, just like stability of ERM, is closely related to the geometry of the class and the underlying measure. Furthermore, our result on stability of ERM delineates a phase transition between stability and instability of clustering methods.

In the last chapter, we prove a generalization of the bounded-difference concentration inequality for almost-everywhere smooth functions. This result can be utilized to

analyze algorithms which are almost always stable. Next, we prove a phase transition in the concentration of almost-everywhere smooth functions. Finally, a tight concentration of empirical errors of empirical minimizers is shown under an assumption on the underlying space.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor of Brain Sciences

*TO MY PARENTS*

# Acknowledgments

I would like to start by thanking Tomaso Poggio for advising me throughout my years at MIT. Unlike applied projects, where progress is observed continuously, theoretical research requires a certain time until the known results are understood and the new results start to appear. I thank Tommy for believing in my abilities and allowing me to work on interesting open-ended theoretical problems. Additionally, I am grateful to Tommy for introducing me to the multi-disciplinary approach to learning.

Many thanks go to Dmitry Panchenko. It is after his class that I became very interested in Statistical Learning Theory. Numerous discussions with Dmitry shaped the direction of my research. I very much value his encouragement and support all these years.

I thank Andrea Caponnetto for being a great teacher, colleague, and a friend. Thanks for the discussions about everything – from philosophy to Hilbert spaces.

I owe many thanks to Sayan Mukherjee, who supported me since my arrival at MIT. He always found problems for me to work on and a pint of beer when I felt down.

Very special thanks to Shahar Mendelson, who invited me to Australia. Shahar taught me the geometric style of thinking, as well as a great deal of math. He also taught me to publish only valuable results instead of seeking to lengthen my CV.

I express my deepest gratitude to Petra Philips.

Thanks to Gadi for the wonderful coffee and interesting conversations; thanks to Mary Pat for her help on the administrative front; thanks to all the past and present CBCL members, especially Gene.

I thank the Student Art Association for providing the opportunity to make pottery and release stress; thanks to the CSAIL hockey team for keeping me in shape.

I owe everything to my friends. Without you, my life in Boston would have been dull. Lots of thanks to Dima, Essie, Max, Sashka, Yanka & Dimka, Marina, Shok, and many others. Special thanks to Dima, Lena, and Sasha for spending many hours fixing my grammar. Thanks to Lena for her support.

Finally, I would like to express my deepest appreciation to my parents for everything they have done for me.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Theory of Learning: Introduction

*Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings –"catching on," "making sense" of things, or "figuring out" what to do.* [30]

The quest for building *intelligent* computer systems started in the 1950's, when the term "artificial intelligence" (AI) was first coined by John McCarthy. Since then, major achievements have been made, ranging from medical diagnosis systems to the `Deep Blue` chess playing program that beat the world champion Gary Kasparov in 1997. However, when measured against the definition above, the advances in Artificial Intelligence are still distant from their goal. It can be argued that, although the current systems can reason, plan, and solve problems in particular constrained domains, it is the "learning" part that stands out as an obstacle to overcome.

Machine learning has been an extremely active area of research in the past fifteen years. Since the pioneering work of Vapnik and Chervonenkis, theoretical foundations of learning have been laid out and numerous successful algorithms developed. This thesis aims to add to our understanding of the theory behind learning processes.

The problem of learning is often formalized within a probabilistic setting. Once such a mathematical framework is set, the following questions can be attacked:

*How many examples are needed to accurately learn a concept? Will a given system be likely to give a correct answer on an unseen example? What is easier to learn and what is harder? How should one proceed in order to build a system that can learn? What are the key properties of predictive systems, and what does this knowledge tell us about biological learning?*

Valuable tools and concepts for answering these questions within a probabilistic framework have been developed in Statistical Learning Theory. The beauty of the results lies in the inter-disciplinary approach to the study of learning. Indeed, a conference on machine learning would likely present ideas in the realms of Computer Science, Statistics, Mathematics, Economics, and Neuroscience.

Learning from examples can be viewed as a high-dimensional mathematical problem, and results from convex geometry and probability in Banach spaces have played an important role in the recent advances. This thesis employs tools from the theory of empirical processes to address some of the questions posed above. Without getting into technical definitions, we will now describe the learning problem and the questions studied by this thesis.

## 1.1 The Learning Problem

The problem of learning can be viewed as a problem of estimating some unknown *phenomenon* from the observed data. The vague word "phenomenon" serves as a common umbrella for diverse settings of the problem. Some interesting settings considered in this thesis are *classification*, *regression*, and *density estimation*. The observed data is often referred to as the *training data* and the learning process as *training*.

Recall that "intelligence ... is not merely book learning." Hence, simply memorizing the observed data does not qualify as learning the phenomenon. Finding the right way to *extrapolate* or *generalize* from the observed data is the key problem of learning.

Let us call the precise method of learning (extrapolating from examples) *an algorithm* or *a procedure*. How does one gauge the success of a learning algorithm? In

other words, how well does the algorithm estimate the unknown phenomenon? A natural answer is to check if a new sample generated by the phenomenon fits the estimate. Finding quantitative bounds on this measure of success is one of the main problems of Statistical Learning Theory.

Since the exposition so far has been somewhat imprecise, let us now describe a few concrete learning scenarios.

One classical learning problem is recognition of hand-written digits (e.g. [24]). Such a system can be used for automatically determining the zip-code written on an envelope. The training data is given as a collection of images of hand-written digits, with the additional information, *label*, denoting the actual digit depicted in each image. Such a labeling is often performed by a human – a process which from the start introduces some inaccuracies. The aim is to constrict a decision rule to predict the label of a new image, one which is not in our collection. Since the new image of a hand-written digit is likely to differ from the previous ones, the system must perform clever extrapolation, ignoring some potential errors introduced in the labeling process.

Prescribing different treatments for a disease can be viewed as a complex learning problem. Assume there is a collection of therapies that could be prescribed to an ill person. The observed data consists of a number of patients' histories, with particular treatment decisions made by doctors at various stages. The number of treatments could be large, and their order might make a profound difference. Taking into account variability of responses of patients and variability of their symptoms turns this into a very complex problem. But the question is simple: what should be the best therapy strategy for a new patient? In other words, is it possible to extrapolate a new patient's treatment from what happened in the observed cases?

Spam filtering, web search, automatic camera surveillance, face recognition, fingerprint recognition, stock market predictions, disease classification – this is only a small number of applications that benefited from the recent advances in machine learning. Theoretical foundations of learning provide performance guarantees for learning algorithms, delineate important properties of successful approaches, and offer suggestions

for improvements. In this thesis, we study two key properties of learning algorithms: their predictive ability (generalization bounds), and their robustness with respect to noise (stability). In the next two sections, we motivate the study of these properties.

## 1.2 Generalization Bounds

Recall that the goal of learning is to estimate the unknown phenomenon from the observed data; that is, the estimate has to be correct on unseen samples. Hence, it is natural to bound the probability of making a mistake on an unseen sample. At first, it seems magical that any such guarantee is possible. After all, we have no idea what the unseen sample looks like. Indeed, if the observed data and the new sample were generated differently, there would be little hope of extrapolating from the data. The key assumption in Statistical Learning Theory is that all the data are independently drawn from the same distribution. Hence, even though we do not know what the next sample will be, we have some idea which samples are more likely.

Once we agree upon the measure of the quality of the estimate (i.e. the error on an unseen example), the goal is to provide probabilistic bounds for it. These bounds are called *performance guarantees* or *generalization bounds.*

Following Vapnik [73], we state key topics of learning theory related to proving performance guarantees:

- *the asymptotic theory of consistency of learning processes;*
- *the non-asymptotic theory of the rate of convergence of learning processes.*

The first topic addresses the limiting performance of the procedures as the number of observed samples increases to infinity. Vaguely speaking, consistency ensures that the learning procedure estimates the unknown phenomenon perfectly with infinite amount of data.

The second topic studies the rates of convergence (as the number of samples increases) of the procedure to the unknown phenomenon which generated the data. Results are given as confidence intervals for the performance on a given number of

samples. These confidence intervals can be viewed as *sample bounds* – number of examples needed to achieve a desired accuracy.

The pioneering work of Vapnik and Chervonenkis [74, 75, 76, 72], addressed the above topics for the simplest learning algorithm, Empirical Risk Minimization (ERM). Vapnik-Chervonenkis (VC) dimension, a combinatorial notion of complexity of a binary function class, turned out to be the key to demonstrating uniform convergence of empirical errors to the expected performance; the result has been extended to the real-valued function classes through the notion of fat-shattering dimension by Alon et al [1]. While the theory of performance of ERM is well understood, the algorithm is impractical. It can be shown (e.g. Ben-David et al [9]) that minimizing mistakes even over a simple class of hypotheses is NP-hard. In recent years, tractable algorithms, such as Support Vector Machines [72] and Boosting [65, 27], became very popular off-the-shelf methods in machine learning. However, their performance guarantees are not as well-understood. In this thesis, we obtain generalization bounds for a family of greedy error minimization methods, which subsume regularized boosting, greedy mixture density estimation, and other algorithms.

The theory of uniform convergence, developed by Vapnik and Chervonenkis, provides a bound on the generalization performance in terms of the empirical performance *for any algorithm* working on a "small" function class. This generality is also a weakness of this approach. In the next section, we discuss an algorithm-based approach to obtaining generalization bounds.

## 1.3 Algorithmic Stability

The motivation for studying stability of learning algorithms is many-fold. Let us start from the perspective of human learning. Suppose a child is trying to learn the distinction between Asian and African elephants. A successful strategy in this case is to realize that the African elephant has large ears matching the shape of Africa, while the Asian elephant has smaller ears which resemble the shape of India. After observing $N$ pictures of each type of elephant, the child has formed some hypothesis

about what makes up the difference. Now, a new example is shown, and the child somewhat changes his mind (forms a new hypothesis). If the new example is an 'outlier' (i.e. not representative of the populations), then the child should ignore it and keep the old hypothesis. If the new example is similar to what has been seen before, the hypothesis should not change much. It can therefore be argued that a successful learning procedure should become more and more stable as the number of observations $N$ increases. Of course, this is a very vague statement, which will be made precise in the following chapters.

Another motivation for studying stability of learning processes is to get a handle on the variability of hypotheses formed from different draws of samples. Roughly speaking, if the learning process is stable, it is easier to predict its performance than if it is unstable. Indeed, if the learning algorithm always outputs the same hypothesis, The Central Limit Theorem provides exponential bounds on the convergence of the empirical performance to the expected performance. This "dumb" learning algorithm is completely stable – the hypothesis does not depend on the observed data. Once this assumption is relaxed, obtaining bounds on the convergence of empirical errors to their expectations becomes difficult. The worst-case approach of Vapnik and Chervonenkis [74, 75] provides loose bounds for this purpose. By studying stability of the specific algorithm, tighter confidence intervals can sometimes be obtained. In fact, Rogers, Devroye, and Wagner [63, 21, 23] showed that bounds on the expected performance can be obtained for $k$-Nearest Neighbors and other local rules even when the VC-based approach fails completely.

If stability of a learning algorithm is a desirable property, why not try to enforce it? Based on this intuition, Breiman [17] advocated averaging classifiers to increase stability and reduce the variance. While averaging helps increase stability, its effect on the bias of the procedure is less clear. We will provide some answers to this question in Chapter 3.

Which learning algorithms are stable? The recent work by Bousquet and Elisseeff [16] surprised the learning community by proving very strong stability of Tikhonov regularization-based methods and by deducing exponential bounds on the difference

of empirical and expected performance solely from these stability considerations. Intuitively, the regularization term in these learning algorithms enforces stability, in agreement with the original motivation of the work of Tikhonov and Arsenin [68] on restoring well-posedness of ill-posed inverse problems.

Kutin and Nyiogi [44, 45] introduced a number of various notions of stability, showing various implications between them. Poggio et al [58, 55] made an important connection between consistency and stability of ERM. This thesis builds upon these results, proving in a systematic manner how algorithmic stability upper- and lower-bounds the performance of learning methods.

In past literature, algorithmic stability has been used as a tool for obtaining bounds on the expected performance. In this thesis, we advocate the study of stability of learning methods also for other purposes. In particular, in Chapter 6 we prove hypothesis (or $L_1$) stability of empirical risk minimization algorithms over Donsker function classes. This result reveals the behavior of the algorithm with respect to perturbations of the observed data, and is interesting on its own. With the help of this result, we are able to analyze sensitivity of various optimization procedures to noise and perturbations of the training data.

## 1.4 Overview

Let us now outline the organization of this thesis. In Chapter 2 we introduce notation and definitions to be used throughout the thesis, as well as provide some background results. We discuss a measure of performance of learning methods and ways to estimate it (Section 2.2). In Section 2.3, we discuss specific algorithms, and in Sections 2.4 and 2.5 we introduce concentration inequalities and the tools from the Theory of Empirical Processes which will be used in the thesis.

In Chapter 3, we show how stability of a learning algorithm can upper- and lower-bound the bias and variance of estimators of the performance, thus obtaining performance guarantees from stability conditions.

Chapter 4 investigates performance of a certain class of greedy error minimization

methods. We start by proving general estimates in Section 4.1. The methods are then applied in the classification setting in Section 4.3 and in the density estimation setting in Section 4.2.

In Chapter 5, we prove a surprising stability result on the behavior of the empirical risk minimization algorithm over Donsker function classes. This result is applied to several optimization methods in Section 5.7.

Connections are made between concentration of functions and stability in Chapter 6. In Section 6.1 we study concentration of almost-everywhere smooth functions.

## 1.5 Contributions

We now briefly outline the contributions of this thesis:

- A systematic approach to upper- and lower-bounding the bias and variance of estimators of the expected performance from stability conditions (Chapter 3). Most of these results have been published in Rakhlin et al [60].

- A performance guarantee for a class of greedy error minimization procedures (Chapter 4) with application to mixture density estimation (Section 4.2). Most of these results appear in Rakhlin et al [61].

- A solution to an open problem regarding $L_1$ stability of empirical risk minimization. These results, obtained in collaboration with A. Caponnetto, are under review for publication [18].

- Applications of the stability result of Chapter 5 for optimization procedures (Section 5.7), such as finding most/least dense regions and clustering. These results are under preparation for publication.

- An extension of McDiarmid's inequality for almost-everywhere Lipschitz functions (Section 6.1). This result appears in Rakhlin et al [60].

- A proof of a phase transition for concentration of real-valued functions on a binary hypercube (Section 6.2). These results are in preparation for publication.

- A tight concentration of empirical errors around the mean for empirical risk minimization under a condition on the underlying space (Section 6.3). These results are in preparation for publication.

# Chapter 2

# Preliminaries

## 2.1 Notation and Definitions

The notion of a "phenomenon", discussed in the previous chapter, is defined formally as the probability space $(\mathcal{Z}, \mathcal{G}, P)$. The measurable space $(\mathcal{Z}, \mathcal{G})$ is usually assumed to be known, while $P$ is not. The only information available about $P$ is through the finite sample $S = \{Z_1, \ldots, Z_n\}$ of $n \in \mathbb{Z}^+$ independent and identically distributed (according to $P$) random variables. Note that we use upper-case letters $X, Y, Z$ to denote random variables, while $x, y, z$ are their realizations.

"Learning" is formally defined as finding a hypothesis $h$ based on the observed samples $Z_1, \ldots, Z_n$. To evaluate the quality of $h$, a bounded real-valued loss (cost) function $\ell$ is introduced, such that $\ell(h; z)$ indicates how well $h$ explains (or fits) $z$. Unless specified otherwise, we assume throughout the thesis that $-M \leq \ell \leq M$ for some $M > 0$.

- *Classification*:

  $\mathcal{Z}$ is defined as the product $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is a discrete output space denoting the labels of inputs. In the case of binary classification, $\mathcal{Y} = \{-1, 1\}$, corresponding to the labels of the two classes. The loss function $\ell$ takes the form $\ell(h; z) = \ell(yh(x))$, and $h$ is called a *binary classifier*. The basic

example of $\ell$ is the indicator loss:

$$\ell(yh'(x)) = I(yh'(x) < 0) = I(y \neq \text{sign}(h'(x))).$$

- *Regression:*

  $\mathcal{Z}$ is defined as the product $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is a real output space denoting the real-valued labels of inputs. The loss function $\ell$ often takes the form $\ell(h; z) = \ell(y - h(x))$, and the basic example is the square loss:

  $$\ell(y - h(x)) = (y - h(x))^2.$$

- *Density Estimation:*

  The functions $h$ are probability densities over $\mathcal{Z}$, and the loss function takes the form $\ell(h; z) = \ell(h(z))$. For instance,

  $$\ell(h(z)) = -\log h(z)$$

  is the likelihood of a point $z$ being generated by $h$.

A learning algorithm is defined as the mapping $\mathcal{A}$ from samples $z_1, \ldots, z_n$ to functions $h$. With this notation, the quality of extrapolation from $z_1, \ldots, z_n$ to a new sample $z$ is measured by $\ell(\mathcal{A}(z_1, \ldots, z_n); z)$.

Whenever $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, it is the function $h : \mathcal{X} \mapsto \mathcal{Y}$ that we seek. In this case, $\mathcal{A}(Z_1, \ldots, Z_n) : \mathcal{X} \mapsto \mathcal{Y}$. Let us denote by $\mathcal{A}(Z_1, \ldots, Z_n; X)$ the evaluation of the function, learned on $Z_1, \ldots, Z_n$, at the point $X$.

Unless indicated, we will assume that the algorithm ignores the ordering of $S$, i.e. $\mathcal{A}(z_1, \ldots, z_n) = \mathcal{A}(\pi(z_1, \ldots, z_n))$ for any permutation $\pi \in S_n$, the symmetric group. If the learning algorithm $\mathcal{A}$ is clear from the context, we will write $\ell(Z_1, \ldots, Z_n; \cdot)$ instead of $\ell(\mathcal{A}(Z_1, \ldots, Z_n); \cdot)$.

The functions $\ell(h; \cdot)$ are called the *loss functions*. If we have a class $\mathcal{H}$ of hypothe-

ses available, the class

$$\mathcal{L}(\mathcal{H}) = \{\ell(h;\cdot) : h \in \mathcal{H}\}$$

is called the *loss class.*

To ascertain the overall quality of a function $h$, we need to evaluate the loss $\ell(h;\cdot)$ on an unseen sample $z$. Since some $z$'s are more likely than others, we integrate over $\mathcal{Z}$ with respect to the measure $P$. Hence, the quality of $h$ is measured by

$$\mathcal{R}(h) := \mathbb{E}\ell(h;Z),$$

called the *expected error* or *expected risk* of $h$. For an algorithm $\mathcal{A}$, its performance is the random variable

$$\mathcal{R}(\mathcal{A}(Z_1,\ldots,Z_n)) = \mathbb{E}\left[\ell(\mathcal{A}(Z_1,\ldots,Z_n);Z)|Z_1,\ldots,Z_n\right].$$

If the algorithm is clear from the context, we will simply write $\mathcal{R}(Z_1,\ldots,Z_n)$.

Since $P$ is unknown, the expected error is impossible to compute. A major part of Statistical Learning Theory is concerned with *bounding* it in probability, i.e. proving bounds of the type

$$\mathbb{P}\left(\mathcal{R}(Z_1,\ldots,Z_n) \geq \varepsilon\right) \leq \delta(\epsilon,n),$$

where the probability is with respect to an i.i.d. draw of samples $Z_1,\ldots,Z_n$.

Such bounds are called *generalization bounds* or *performance guarantees.* In the above expression, $\varepsilon$ sometimes depends on a quantity computable from the data. In the next chapter, we will consider bounds of the form

$$\mathbb{P}\left(\left|\mathcal{R}(Z_1,\ldots,Z_n) - \hat{\mathcal{R}}(Z_1,\ldots,Z_n)\right| \geq \varepsilon\right) \leq \delta(\varepsilon,n), \tag{2.1}$$

where $\hat{\mathcal{R}}(Z_1,\ldots,Z_n)$ is an estimate of the unknown $\mathcal{R}(Z_1,\ldots,Z_n)$ from the data $Z_1,\ldots,Z_n$. The next section discusses such estimates (proxies) for $\mathcal{R}$. The reader is referred to the excellent book of Devroye et al [20] for more information.

Table 2.1: Table of notation

| | |
|---|---|
| $\mathcal{Z}$ | Space of samples |
| $\mathcal{X}$, $\mathcal{Y}$ | Input and output spaces, whenever $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ |
| $P$ | Unknown distribution on $\mathcal{Z}$ |
| $Z_1, \ldots, Z_n$ | I.i.d. sample from $P$ |
| $n$ | Number of samples |
| $S$ | The sample $\{Z_1, \ldots, Z_n\}$ |
| $\ell$ | Loss (cost) function |
| $\mathcal{H}$ | Class of hypotheses |
| $\mathcal{A}$ | Learning algorithm |
| $\mathcal{R}$ | Expected error (exp. loss, exp. risk) |
| $\mathcal{R}_{\text{emp}}$ | Empirical error (resubstitution estimate) |
| $\mathcal{R}_{\text{loo}}$ | Leave-one-out error (deleted estimate) |
| $\widetilde{\mathcal{R}}_{\text{emp}}$ | Defect of the resubstitution estimate: $\widetilde{\mathcal{R}}_{\text{emp}} = \mathcal{R} - \mathcal{R}_{\text{emp}}$ |
| $\widetilde{\mathcal{R}}_{\text{loo}}$ | Defect of the deleted estimate: $\widetilde{\mathcal{R}}_{\text{loo}} = \mathcal{R} - \mathcal{R}_{\text{loo}}$ |
| conv $(\mathcal{H})$ | Convex hull of $\mathcal{H}$ |
| $\text{conv}_k$ $(\mathcal{H})$ | $k$-term convex hull of $\mathcal{H}$ |
| $T_n(Z_1, \ldots, Z_n)$ | A generic function of $n$ random variables |
| $\nu_n$ | Empirical process |

## 2.2 Estimates of the Performance

Several important estimates of the expected error $\mathcal{R}(h)$ can be computed from the sample. The first one is the *empirical error* (or *resubstitution estimate*),

$$\mathcal{R}_{\text{emp}}(Z_1, \ldots, Z_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(Z_1, \ldots, Z_n; Z_i).$$

The second one is the *leave-one-out error* (or *deleted estimate*)[1],

$$\mathcal{R}_{\text{loo}}(Z_1, \ldots, Z_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n; Z_i).$$

These quantities are employed to estimate the expected error, and Statistical Learning Theory is concerned with providing bounds on the deviations of these esti-

[1]It is understood that the first term in the sum is $\ell(Z_2, \ldots, Z_n; Z_1)$ and the last term is $\ell(Z_1, \ldots, Z_{n-1}; Z_n)$.

mates from the expected error. For convenience, denote these deviations

$$\begin{aligned}\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n) &:= \mathcal{R}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n),\\ \widetilde{\mathcal{R}}_{\text{loo}}(Z_1,\ldots,Z_n) &:= \mathcal{R}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{loo}}(Z_1,\ldots,Z_n).\end{aligned}$$

With this notation, Equation 2.1 becomes

$$\mathbb{P}\left(\left|\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n)\right| \geq \varepsilon\right) < \delta(\varepsilon,n) \tag{2.2}$$

or

$$\mathbb{P}\left(\left|\widetilde{\mathcal{R}}_{\text{loo}}(Z_1,\ldots,Z_n)\right| \geq \varepsilon\right) < \delta(\varepsilon,n) \tag{2.3}$$

If one can show that $\widetilde{\mathcal{R}}_{\text{emp}}$ (or $\widetilde{\mathcal{R}}_{\text{loo}}$) is "small", then the empirical error (resp. leave-one-out error) is a good proxy for the expected error. Hence, a small empirical (or leave-one-out error) implies a small expected error, with a certain confidence. In particular, we are often interested in the rate of the convergence of $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ to zero as $n$ increases.

The goal is to derive bounds such that $\lim_{n\to\infty}\delta(\varepsilon,n) = 0$ for any fixed $\varepsilon > 0$. If the rate of decrease of $\delta(\varepsilon,n)$ is not important, we will write $|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$ and $|\widetilde{\mathcal{R}}_{\text{loo}}| \xrightarrow{P} 0$.

Let us focus on the random variable $\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n)$. Recall that the Central Limit Theorem (CLT) guarantees that the average of $n$ i.i.d. random variables converges to their mean (under the assumption of finiteness of second moment) quite fast. Unfortunately, the random variables

$$\ell(Z_1,\ldots,Z_n;Z_1),\ldots,\ell(Z_1,\ldots,Z_n;Z_n)$$

are dependent, and the CLT is not applicable. In fact, the interdependence of these random variables makes the resubstitution estimate *positively biased*, as the next example shows.

**Example 1.** *Let* $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, *and*

$$P(X) = U[0,1],\ P(Y|X) = \delta_{Y=1}.$$

*Suppose* $\ell(h(x),y) = I(h(x) \neq y)$, *and* $\mathcal{A}$ *is defined as* $\mathcal{A}(Z_1,\ldots,Z_n;X) = 1$ *if* $X \in \{X_1,...,X_n\}$ *and* 0 *otherwise. In other words, the algorithm observes* $n$ *data points* $(X_i,1)$, *where* $X_i$ *is distributed uniformly on* $[0,1]$, *and generates a hypothesis which fits exactly the observed data, but outputs* 0 *for unseen points* $X$. *This situation is depicted in Figure 2-1. The empirical error of* $\mathcal{A}$ *is* 0, *while the expected error is* 1, *i.e.* $\widetilde{\mathcal{R}}_{emp}(Z_1,\ldots,Z_n) = 1$ *for any* $Z_1,\ldots,Z_n$.

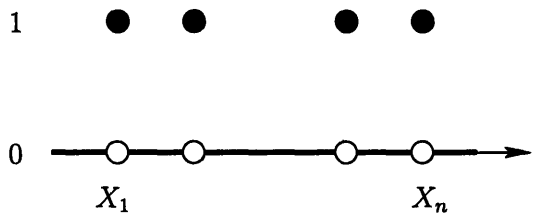



Figure 2-1: Fitting the data.

No guarantee on smallness of $\widetilde{\mathcal{R}}_{\text{emp}}$ can be made in Example 1. Intuitively, this is due to the fact that the algorithm can fit *any* data, i.e. the space of functions $\mathcal{L}(\mathcal{H})$ is too large.

Assume that we have no idea what the learning algorithm is except that it picks its hypotheses from $\mathcal{H}$. To bound $\widetilde{\mathcal{R}}_{\text{emp}}$, we would need to resort to the worst-case approach of bounding the deviations between empirical and expected errors for all functions simultaneously. The ability to make such a statement is completely characterized by the "size" of $\mathcal{L}(\mathcal{H})$, as discussed next.

### 2.2.1 Uniform Convergence of Means to Expectations

The class $\mathcal{L}(\mathcal{H})$ is called *uniform Glivenko-Cantelli* if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left( \sup_{\ell\in\mathcal{L}(\mathcal{H})} \left| \mathbb{E}\ell - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i) \right| \geq \varepsilon \right) = 0,$$

where $Z_1, \ldots, Z_n$ are i.i.d random variables distributed according to $\mu$.

Non-asymptotic results of the form

$$\mathbb{P}\left( \sup_{\ell\in\mathcal{L}(\mathcal{H})} \left| \mathbb{E}\ell - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i) \right| \geq \varepsilon \right) \leq \delta(\varepsilon, n, \mathcal{L}(\mathcal{H}))$$

give *uniform* (over the class $\mathcal{L}(\mathcal{H})$) rates of convergence of empirical means to expectations. Since the guarantee is given for *all* functions in the class, we immediately obtain

$$\mathbb{P}\left( \left| \widetilde{\mathcal{R}}_{\text{emp}}(Z_1, \ldots, Z_n) \right| \geq \varepsilon \right) \leq \delta(\varepsilon, n, \mathcal{L}(\mathcal{H})).$$

We postpone further discussion of Glivenko-Cantelli classes of functions to Section 2.5.

### 2.2.2 Algorithmic Stability

The uniform-convergence approach above ignores the algorithm, except for the fact that it picks its hypotheses from $\mathcal{H}$. Hence, this approach might provide only loose bounds on $\widetilde{\mathcal{R}}_{\text{emp}}$. Indeed, suppose that the algorithm would in fact only pick one function from $\mathcal{H}$. The bound on $\widetilde{\mathcal{R}}_{\text{emp}}$ would then follow immediately from The Central Limit Theorem. It turns out that analogous bounds can be proved even if the algorithm picks diverse functions, as long is it is done in a "smooth" way. In Chapter 3, we will derive bounds on both $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ in terms of various stability conditions on the algorithm. Such algorithm-dependent conditions provide guarantees for $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ even when the uniform-convergence approach of Section 2.2.1 fails.

## 2.3 Some Algorithms

### 2.3.1 Empirical Risk Minimization

The simplest method of learning from observed data is the *Empirical Risk Minimization (ERM)* algorithm

$$\mathcal{A}(Z_1, \ldots, Z_n) = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; Z_i).$$

Note that the ERM algorithm is defined with respect to a class $\mathcal{H}$. Although an exact minimizer of empirical risk in this class might not exist, an almost-minimizer always exists. This situation will be discussed in much greater detail in Chapter 5.

The algorithm in Example 1 is an example of ERM over the function class

$$\mathcal{H} = \bigcup_{n \geq 1} \{h_{\mathbf{x}} : \mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n\},$$

where $h_{\mathbf{x}}(x) = 1$ if $x = x_i$ for some $1 \leq i \leq n$ and $h_{\mathbf{x}}(x) = 0$ otherwise.

ERM over uniform Glivenko-Cantelli classes is a consistent procedure in the sense that the expected performance converges to the best possible within the class of hypotheses.

There are a number of drawbacks of ERM: ill-posedness for general classes $\mathcal{H}$, as well as computational intractability (e.g. for classification with the indicator loss). The following two families of algorithms, *regularization algorithms* and *boosting algorithms*, aim to overcome these difficulties.

### 2.3.2 Regularization Algorithms

One of the drawbacks of ERM is ill-posedness of the solution. Indeed, learning can be viewed as reconstruction of the function from the observed data (inverse problem), and the information contained in the data is not sufficient for the solution to be unique. For instance, there could be an infinite number of hypotheses with zero empirical risk, as shown in Figure 2-2. Moreover, the inverse mapping tends to be

unstable.

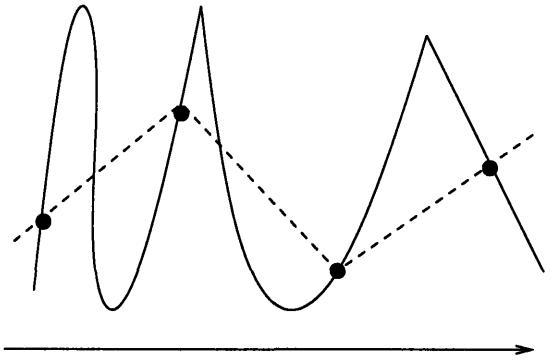


Figure 2-2: Multiple minima of the empirical risk: two dissimilar functions fit the data.

The regularization method described next is widely used in machine learning [57, 77], and arises from the theory of solving ill-posed problems. Discovered by J. Hadamard, ill-posed inverse problems turned out to be important in physics and statistics (see Chapter 7 of Vapnik [72]).

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, i.e. we consider regression or classification. The *Tikhonov regularization method* [68], applied to the learning setting, proposes to solve the following minimization problem

$$\mathcal{A}(Z_1, \ldots, Z_n) = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) + \lambda \|h\|_K^2,$$

where $K$ is a positive definite kernel and $\|\cdot\|$ is the norm in the associated Reproducing Kernel Hilbert Space $\mathcal{H}$. The parameter $\lambda \geq 0$ controls the balance of the fit to the data (the first term) and the "smoothness" of the solution (the second term), and is usually set by a cross-validation method. It is exactly this balance between smoothness and fit to the data that restores the uniqueness of the solution. It also restores stability.

This particular minimization problem owes its success to the following surprising (although simple to prove) fact: even though the minimization is performed over a possibly infinite-dimensional Hilbert Space of functions, the solution always has the

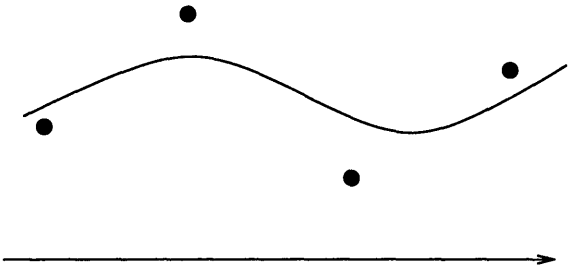


Figure 2-3: Unique minimum of the regularized fit to the data.

form

$$\mathcal{A}(Z_1, \ldots, Z_n; x) = \sum_{i=1}^{n} c_i K(X_i, x),$$

assuming that $\ell$ depends on $h$ only through $h(X_i)$.

### 2.3.3 Boosting Algorithms

We now describe boosting methods, which have become very popular in machine learning [66, 27]. Consider the classification setting, i.e. $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{-1, 1\}$. The idea is to iteratively build a complex classifier $f$ by adding weighted functions $h \in \mathcal{H}$, where $\mathcal{H}$ is typically a set of simple functions. "Boosting" stands for the increase in the performance of the ensemble, as compared to the relatively weak performance of the simple classifiers. The ensemble is built in a greedy stage-wise manner, and can be viewed as an example of additive models in statistics [32].

Given a class $\mathcal{H}$ of "base" functions and the observed data $Z_1, \ldots, Z_n$, a greedy boosting procedure builds the ensemble $f_k$ in the following way. Start with some $f_0 = h_0$. At the $k$-th step, choose $\alpha_k$ and $h_k \in \mathcal{H}$ to approximately minimize the empirical error on the sample. After $T$ steps, output the resulting classifier as

$$\mathcal{A}(Z_1, \ldots, Z_n) = \operatorname{sign}\left(\sum_{i=1}^{T} \alpha_i h_i\right).$$

There exist a number of variations of boosting algorithms. The most popular one, AdaBoost, is an unregularized procedure with a potential to overfit if left running

for enough time. The regularization is performed as a constraint on the norm of the coefficients or via early stopping. The precise details of a boosting procedure with the constraint on the $\ell_1$ norm of the coefficients are given in Chapter 4, where a bound on the generalization performance is proved.

## 2.4 Concentration Inequalities

In the context of learning theory, concentration inequalities serve as tools for obtaining generalization bounds. While deviation inequalities are probabilistic statements about the deviation of a random variable from its expectation, the term "concentration" often refers to the exponential bounds on the deviation of a function of many random variables from its mean. The reader is referred to the excellent book of Ledoux [47] for more information on the concentration of measure phenomenon. The well-known probabilistic statements mentioned below can be found, for instance, in the survey by Boucheron et al [13].

Let us start with some basic probability inequalities. For a non-negative random variable $X$,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)\, dt.$$

The integral above can be lower-bounded by the product $t\mathbb{P}(X \geq t)$ for any fixed $t > 0$. Hence, we obtain *Markov's inequality*:

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}X}{t}$$

for a non-negative random variable $X$ and any $t > 0$.

For a non-negative strictly monotonically increasing function $\phi$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\phi(X) \geq \phi(t)),$$

resulting in

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}$$

for an arbitrary random variable $X$.

Setting $\phi(x) = x^q$ for any $q > 0$ leads to the method of moments

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\mathbb{E}|X - \mathbb{E}X|^q}{t^q}$$

for any random variable $X$. Since the inequality holds for any $q > 0$, one can optimize the bound to get the smallest one. This idea will be used in Chapter 6. Setting $q = 2$ we obtain Chebyshev's inequality which uses the second-moment information to bound the deviation of $X$ from its expectation:

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\mathrm{Var}X}{t^2}.$$

Other choices of $\phi$ lead to useful probability inequalities. For instance, $\phi(x) = e^{sx}$ for $s > 0$ leads to the Chernoff's bounding method. Since

$$\mathbb{P}(X \geq t) = \mathbb{P}\left(e^{sX} \geq e^{st}\right),$$

we obtain

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}e^{sX}}{e^{st}}.$$

Once some information about $X$ is available, one can minimize the above bound over $s > 0$.

So far, we have discussed generic probability inequalities which do not exploit any "structure" of $X$. Suppose $X$ is in fact a function of $n$ random variables. Instead of the letter $X$, let us denote the function of $n$ random variables by $T_n(Z_1, \ldots, Z_n)$.

**Theorem 2.4.1** (Hoeffding [34]). *Suppose $T_n(Z_1, \ldots, Z_n) = \sum_{i=1}^n Z_i$, where $Z_i$'s are independent and $a_i \leq Z_i \leq b_i$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}(|T_n - \mathbb{E}T_n| > \varepsilon) \leq 2e^{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Hoeffding's inequality does not use any information about the variances of $Z_i$'s. A tighter bound can be obtained whenever these variances are small.

**Theorem 2.4.2** (Bennett [10]). *Suppose* $T_n(Z_1, \ldots, Z_n) = \sum_{i=1}^n Z_i$, *where* $Z_i$*'s are independent, and for any* $i$, $\mathbb{E}Z_i = 0$ *and* $|Z_i| \leq M$. *Let* $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\{Z_i\}$. *Then for any* $\varepsilon > 0$,

$$\mathbb{P}\left(|T_n| > \varepsilon\right) \leq 2\exp\left(-\frac{n\sigma^2}{M^2}\phi\left(\frac{\varepsilon M}{n\sigma^2}\right)\right),$$

*where* $\phi(x) = (1+x)\log(1+x) - x$.

Somewhat surprisingly, exponential deviation inequalities hold not only for sums, but for general "smooth" functions of $n$ variables.

**Theorem 2.4.3** (McDiarmid [54]). *Let* $T_n : \mathcal{Z}^n \to \mathbb{R}$ *such that*

$$\forall z_1, \ldots, z_n, z_1', \ldots, z_n' \in \mathcal{Z} \quad |T_n(z_1, \ldots, z_n) - T_n(z_1, \ldots, z_i', \ldots, z_n)| \leq c_i.$$

*Let* $Z_1, \ldots, Z_n$ *be independent random variables. Then*

$$\mathbb{P}\left(T_n(Z_1, \ldots, Z_n) - \mathbb{E}T_n(Z_1, \ldots, Z_n) > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^n c_i^2}\right)$$

*and*

$$\mathbb{P}\left(T_n(Z_1, \ldots, Z_n) - \mathbb{E}T_n(Z_1, \ldots, Z_n) < -\varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^n c_i^2}\right).$$

The following Efron-Stein's inequality can be used to directly upper-bound the variance of functions of $n$ random variables.

**Theorem 2.4.4** (Efron-Stein [26]). *Let* $T_n : \mathcal{Z}^n \mapsto \mathbb{R}$ *be a measurable function of* $n$ *variables and define* $\Gamma = T_n(Z_1, \ldots, Z_n)$ *and* $\Gamma_i' = T_n(Z_1, \ldots, Z_i', \ldots, Z_n)$, *where*

$$Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$$

*are i.i.d. random variables. Then*

$$\mathrm{Var}(T_n) \leq \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[(\Gamma - \Gamma_i')^2\right]. \tag{2.4}$$

A "removal" version of the above is the following:

**Theorem 2.4.5** (Efron-Stein). *Let $T_n : \mathcal{Z}^n \mapsto \mathbb{R}$ be a measurable function of $n$ variables and $T'_n : \mathcal{Z}^{n-1} \mapsto \mathbb{R}$ of $n-1$ variables. Define $\Gamma = T_n(Z_1, \ldots, Z_n)$ and $\Gamma_i = T'_n(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$, where $Z_1, \ldots, Z_n$ are i.i.d. random variables. Then*

$$\mathrm{Var}(T_n) \leq \sum_{i=1}^{n} \mathbb{E}\left[(\Gamma - \Gamma_i)^2\right]. \tag{2.5}$$

A collection of random functions $T_n$ for $n = 1, 2, \ldots$ can be viewed as a sequence of random variables $\{T_n\}$. There are several important notions of convergence of sequences of random variables: *in probability* and *almost surely.*

**Definition 2.4.1.** *A sequence $\{T_n\}$, $n = 1, 2, \ldots$, of random variables converges to $T$ in probability*

$$T_n \xrightarrow{P} T$$

*if for each $\varepsilon > 0$*

$$\lim_{n \to \infty} \mathbb{P}\left(|T_n - T| \geq \varepsilon\right) = 0.$$

*This can also be written as*

$$\mathbb{P}\left(|T_n - T| \geq \varepsilon\right) \to 0.$$

**Definition 2.4.2.** *A sequence $\{T_n\}$, $n = 1, 2, \ldots$, of random variables converges to $T$ almost surely if*

$$\mathbb{P}\left(\lim_{n \to \infty} T_n = T\right) = 0.$$

Deviation inequalities provide specific upper bounds on the convergence of

$$\mathbb{P}\left(|T_n - T| \geq \varepsilon\right) \to 0.$$

Assume for simplicity that $T_n$ is a non-negative function and the limit $T$ is 0. When inspecting inequalities of the type

$$\mathbb{P}\left(T_n \geq \varepsilon\right) \leq \delta(\varepsilon, n)$$

it is often helpful to keep in mind the two-dimensional surface depicted in Figure 2-4. For a fixed $n$, decreasing $\varepsilon$ increases $\mathbb{P}(T_n \geq \varepsilon)$. For a fixed $\varepsilon$, $\mathbb{P}(T_n \geq \varepsilon) \to 0$ as $n \to \infty$. Now, suppose that we would like $\varepsilon$ to decrease with $n$. One can often find the fastest possible decay $\varepsilon(n)$, such that $\mathbb{P}(T_n \geq \varepsilon(n)) \to 0$ as $n \to \infty$. This defines the rate of convergence of $T_n$ to 0 in probability. In Chapter 5 we study rates of decay of certain quantities in great detail.

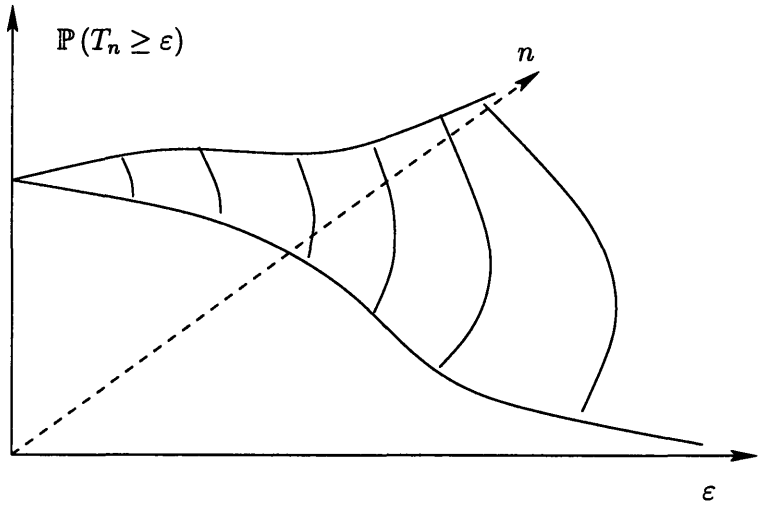



Figure 2-4: Probability surface

Let us conclude this Section by reminding the reader about the order notation. Let $f(n)$ and $g(n)$ be two functions.

**Definition 2.4.3** (Asymptotic upper bound).

$$f(n) \in O(g(n)) \text{ if } \lim_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| < \infty.$$

**Definition 2.4.4** (Asymptotically negligible).

$$f(n) \in o(g(n)) \text{ if } \lim_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| = 0.$$

**Definition 2.4.5** (Asymptotic lower bound).

$$f(n) \in \Omega(g(n)) \text{ if } \lim_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| > 0.$$

In the above definitions, "$\in$" will often be replaced by "$=$".

## 2.5 Empirical Process Theory

In this section, we briefly mention a few results from the Theory of Empirical Processes, relevant to this thesis. The reader is referred to the excellent book by A. W. van der Waart and J. A. Wellner [71].

### 2.5.1 Covering and Packing Numbers

Fix a distance function $d(f,g)$ for $f,g \in \mathcal{H}$.

**Definition 2.5.1.** *Given $\varepsilon > 0$ and $h_1,\ldots,h_N \in \mathcal{H}$, we say that $h_1,\ldots,h_N$ are $\varepsilon$-separated if $d(h_i,h_j) > \varepsilon$ for any $i \neq j$.*

*The $\varepsilon$-packing number, $\mathcal{D}(\mathcal{H},\varepsilon,d)$, is the maximal cardinality of an $\varepsilon$-separated set.*

**Definition 2.5.2.** *Given $\varepsilon > 0$ and $h_1,\ldots,h_N \in \mathcal{H}$, we say that the set $h_1,\ldots,h_N$ is an $\varepsilon$-cover of $\mathcal{H}$ if for any $h \in \mathcal{H}$, there exists $1 \leq i \leq N$ such that $d(h,h_i) \leq \varepsilon$.*

*The $\varepsilon$-covering number, $\mathcal{N}(\mathcal{H},\varepsilon,d)$, is the minimal cardinality of an $\varepsilon$-cover of $\mathcal{H}$. Furthermore, $\log\mathcal{N}(\mathcal{H},\varepsilon,d)$ is called metric entropy.*

It can be shown that

$$\mathcal{D}(\mathcal{H},2\varepsilon,d) \leq \mathcal{N}(\mathcal{H},\varepsilon,d) \leq \mathcal{D}(\mathcal{H},\varepsilon,d).$$

**Definition 2.5.3.** *Entropy with bracketing $\mathcal{N}_{[]}(\mathcal{H},\varepsilon,d)$ is defined as the smallest number $N$ for which there exists pairs $\{h_i,h_i'\}_{i=1}^N$ such that $d(h_i,h_i') \leq \varepsilon$ for all $i$ and for any $h \in \mathcal{H}$ there exists a pair $\{h_j,h_j'\}$ such that $h_j \leq h \leq h_j'$.*

### 2.5.2 Donsker and Glivenko-Cantelli Classes

Let $P_n$ stand for the discrete measure supported on $Z_1, \ldots, Z_n$. More precisely,

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i},$$

the sum of Dirac measures at the samples. Throughout this thesis, we will denote

$$Pf = \mathbb{E}_Z f(Z) \quad \text{and} \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(Z_i).$$

Define the sup norm as

$$\|Qf\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |Qf|$$

for any measure $Q$.

We now introduce the notion of *empirical process.*

**Definition 2.5.4.** *The empirical process $\nu_n$ indexed by a function class $\mathcal{F}$ is defined as the map*

$$f \mapsto \nu_n(f) = \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Z_i) - Pf).$$

The Law of Large Numbers (LLN) guarantees that $P_n f$ converges to $Pf$ for a fixed $f$, if the latter exists. Moreover, the Central Limit Theorem (CLT) guarantees that the empirical process $\nu_n(f)$ converges to $N(0, P(f - Pf)^2)$ if $Pf^2$ is finite. Similar statements can be made for a finite number of functions simultaneously. The analogous statements that hold uniformly over infinite function classes are the core topic of Empirical Process Theory.

**Definition 2.5.5.** *A class $\mathcal{F}$ is called P-Glivenko-Cantelli if*

$$\|P_n - P\|_{\mathcal{F}} \xrightarrow{P^*} 0,$$

*where the convergence is in (outer) probability or (outer) almost surely [71].*

**Definition 2.5.6.** *A class $\mathcal{F}$ is called P-Donsker if*

$$\nu_n \rightsquigarrow \nu$$

*in $\ell^\infty(\mathcal{F})$, where the limit $\nu$ is a tight Borel measurable element in $\ell^\infty(\mathcal{F})$ and " $\rightsquigarrow$" denotes weak convergence, as defined on p. 17 of [71].*

In fact, it follows that the limit process $\nu$ must be a zero-mean Gaussian process with covariance function $\mathbb{E}\nu(f)\nu(f') = \langle f, f' \rangle$ (i.e. a Brownian bridge).

While Glivenko-Cantelli classes of functions have been used in Learning Theory to a great extent, the important properties of Donsker classes have not been utilized. In Chapter 5, $P$-Donsker classes of functions play an important role because of a specific covariance structure they possess. We hypothesize that many more results can be discovered for learning with Donsker classes.

Various Donsker theorems provide sufficient conditions for a class being $P$-Donsker. Here we mention a few known results (see [71], Eqn. 2.1.7 and [70], Thm. 6.3) in terms of entropy $\log \mathcal{N}$ and entropy with bracketing $\log \mathcal{N}_{[]}$.

**Proposition 2.5.1.** *If the envelope $F$ of $\mathcal{F}$ is square integrable and*

$$\int_0^\infty \sup_Q \sqrt{\log \mathcal{N}(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty,$$

*then $\mathcal{F}$ is P-Donsker for every $P$, i.e. $\mathcal{F}$ is a universal Donsker class. Here the supremum is taken over all finitely discrete probability measures.*

**Proposition 2.5.2.** *If $\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon < \infty$, then $\mathcal{F}$ is P-Donsker.*

From the learning theory perspective, however, the most interesting theorems are probably those relating the Donsker property to the VC-dimension. For example, if $\mathcal{F}$ is a $\{0,1\}$-valued class, then $\mathcal{F}$ is *universal* Donsker if and only if its VC dimension is finite (Thm. 10.1.4 of [25] provides a more general result involving Pollard's entropy condition). As a corollary of their Proposition 3.1, [29] show that under the Pollard's entropy condition, the $\{0,1\}$-valued class $\mathcal{F}$ is in fact *uniform* Donsker. Finally,

Rudelson and Vershynin [64] extended these results to the real-valued case: a class $\mathcal{F}$ is *uniform* Donsker if the square root of its VC dimension is integrable.

### 2.5.3 Symmetrization and Concentration

The celebrated result of Talagrand [67] states that the supremum of an empirical process is tightly concentrated around its mean. The version stated below is taken from [6].

**Theorem 2.5.1** (Talagrand). *Let $\mathcal{F}$ be a class of functions such that $\|f\|_\infty \leq b$ for every $f \in \mathcal{F}$. Suppose, for simplicity, that $Pf = 0$ for all $f \in \mathcal{F}$. Let $\sigma^2 = \sqrt{n} \sup_{f \in \mathcal{F}} \mathrm{Var}(f)$. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}\left(|\|\nu_n\|_{\mathcal{F}} - \mathbb{E}\|\nu_n\|_{\mathcal{F}}| \geq \varepsilon\right) \leq C \exp\left(-\frac{\sqrt{n}\varepsilon}{Kb} \log\left(1 + \frac{b\varepsilon}{\sigma^2 + b\mathbb{E}\|\nu_n\|_{\mathcal{F}}}\right)\right),$$

*where $C$ and $K$ are absolute constants.*

The above Theorem is key to obtaining fast rates of convergence for empirical risk minimization. The reader is referred to Bartlett and Mendelson [6].

It can be shown that the empirical process

$$f \mapsto \nu_n(f) = \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Z_i) - Pf)$$

is very closely related to the symmetrized (Rademacher) process

$$f \mapsto \eta_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(Z_i),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables, independent of $Z_1, \ldots, Z_n$, such that $\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = +1) = 1/2$. In fact, the LLN or the CLT for one process holds if and only if it holds for the other [71].

Since we are interested in statements which are uniform over a function class, the object of study becomes the supremum of the empirical process and the supremum of

the Rademacher process. The Symmetrization technique [28] is key to relating these two.

**Lemma 2.5.1** ([71] Symmetrization). *Consider the following suprema:*

$$Z_n(X_1,\ldots,X_n) = \sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right| = \frac{1}{\sqrt{n}}\|\nu_n\|_{\mathcal{F}}$$

*and*

$$R_n(X_1,\ldots,X_n) = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\right| = \frac{1}{\sqrt{n}}\|\eta_n\|_{\mathcal{F}}$$

*Then*

$$\mathbb{E}Z_n \leq 2\mathbb{E}R_n.$$

*The quantity $\mathbb{E}R_n$ is called the Rademacher average of $\mathcal{F}$.*

Consider functions with bounded Lipschitz constant. It turns out that such functions can be "erased" from the Rademacher sum, as stated in Lemma 2.5.2.

**Definition 2.5.7.** *A function $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a contraction if $\phi(0) = 0$ and*

$$|\phi(s) - \phi(t)| \leq |s - t|\,.$$

We will denote $f_i = f(x_i)$. The following inequality can be found in [46], Theorem 4.12.

**Lemma 2.5.2** ([46] Comparison inequality for Rademacher processes). *If $\phi_i : \mathbb{R} \to \mathbb{R}$ $(i = 1, .., n)$ are contractions, then*

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} \varepsilon_i\phi_i(f_i)\right| \leq 2\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} \varepsilon_i f_i\right|.$$

In Chapter 4, this lemma will allow us to greatly simplify Rademacher complexities over convex combinations of functions by "erasing" the loss function.

# Chapter 3

# Generalization Bounds via Stability

The results of this Chapter appear in [60].

## 3.1 Introduction

Albeit interesting from the theoretical point of view, the uniform bounds, discussed in Section 2.2.1, are, in general, loose, as they are worst-case over all functions in the class. As an extreme example, consider the algorithm that always ouputs the same function (the constant algorithm)

$$\mathcal{A}(Z_1, \ldots, Z_n) = f_0, \quad \forall (Z_1, \ldots, Z_n) \in \mathcal{Z}^n.$$

The bound on $\widetilde{\mathcal{R}}_{\text{emp}}(Z_1, \ldots, Z_n)$ follows from the CLT and an analysis based upon the complexity of a class $\mathcal{H}$ does not make sense.

Recall from the previous Chapter that we would like, for a given algorithm, to obtain the following generalization bounds

$$\mathbb{P}\left(\left|\widetilde{\mathcal{R}}_{\text{emp}}(Z_1, \ldots, Z_n)\right| > \varepsilon\right) < \delta(\varepsilon, n) \quad \text{or} \quad \mathbb{P}\left(\left|\widetilde{\mathcal{R}}_{\text{loo}}(Z_1, \ldots, Z_n)\right| > \varepsilon\right) < \delta(\varepsilon, n)$$

with $\delta(\varepsilon, n) \to 0$ as $n \to \infty$.

Throughout this Chapter, we assume that the loss function $\ell$ is bounded and non-negative, i.e. $0 \leq \ell \leq M$. Notice that $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ are bounded random variables. By Markov's inequality,

$$\forall \varepsilon \geq 0,\ \mathbb{P}\left(|\widetilde{\mathcal{R}}_{\text{emp}}| \geq \varepsilon\right) \leq \frac{\mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}|}{\varepsilon}$$

and also

$$\forall \varepsilon' \geq 0,\ \mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}| \leq M\mathbb{P}\left(|\widetilde{\mathcal{R}}_{\text{emp}}| \geq \varepsilon'\right) + \varepsilon'.$$

Therefore, showing

$$|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$$

is *equivalent* to showing

$$\mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}| \to 0.$$

The latter is equivalent to

$$\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2 \to 0$$

since $|\widetilde{\mathcal{R}}_{\text{emp}}| \leq M$. Further notice that

$$\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2 = \text{Var}(\widetilde{\mathcal{R}}_{\text{emp}}) + (\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}})^2.$$

We will call $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}}$ the *bias*, $\text{Var}(\widetilde{\mathcal{R}}_{\text{emp}})$ the *variance*, and $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ the *second moment* of $\widetilde{\mathcal{R}}_{\text{emp}}$. The same derivations and terminology hold for $\widetilde{\mathcal{R}}_{\text{loo}}$.

Hence, studying conditions for convergence in probability of the estimators to zero is equivalent to studying their mean and variance (or the second moment alone).

In this Chapter we consider various *stability* conditions which allow one to bound bias and variance or the second moment, and thus imply convergence of $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ to zero in probability. Though the reader should expect a number of definitions of stability, the common flavor of these notions is the comparison of the "behavior" of the algorithm $\mathcal{A}$ on similar samples. We hope that the present work sheds light on the important stability aspects of algorithms, suggesting principles for designing

predictive learning systems.

We now sketch the organization of this Chapter. In Section 3.2 we motivate the use of stability and give some historical background. In Section 3.3, we show how bias (Section 3.3.1) and variance (Section 3.3.2) can be bounded by various stability quantities. Sometimes it is mathematically more convenient to bound the second moment instead of bias and variance, and this is done in Section 3.4. In particular, Section 3.4.1 deals with the second moment $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{loo}})^2$ in the spirit of [22], while in Sections 3.4.3 and 3.4.2 we bound $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ in the spirit of [55] and [16], respectively. The goal of Sections 3.4.1 and 3.4.2 is to re-derive some known results in a simple manner that allows one to compare the proofs side-by-side. The results of these sections hold for general algorithms. Furthermore, for specific algorithms the results can be improved, i.e. simpler quantities might govern the convergence of the estimators to zero. To illustrate this, in Section 3.4.4 we prove that for the empirical risk minimization algorithm, a bound on the bias $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}}$ implies a bound on the second moment $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$. We therefore provide a simple necessary and sufficient condition for consistency of ERM. If rates of convergence are of importance, rather than using Markov's inequality, one can make use of more sophisticated concentration inequalities with a cost of requiring more stringent stability conditions. In Section 3.6, we discuss the most rigid stability, Uniform Stability, and provide exponential bounds in the spirit of [16]. In Section 3.6.2 we consider less rigid notions of stability and prove exponential inequalities based on powerful moment inequalities of [15]. Finally, Section 3.7 summarizes the Chapter and discusses further directions and open questions.

## 3.2 Historical Remarks and Motivation

Devroye, Rogers, and Wagner (see e.g. [22]) were the first, to our knowledge, to observe that the sensitivity of the algorithms with regard to small changes in the sample is related to the behavior of the leave-one-out estimate. The authors were able to obtain results for the k-Nearest-Neighbor algorithm, where VC theory fails

because of large class of potential hypotheses. These results were further extended for k-local algorithms and for potential learning rules. Kearns and Ron [36] later discovered a connection between finite VC-dimension and stability. Bousquet and Elisseeff [16] showed that a large class of learning algorithms, based on *Tikhonov Regularization*, is stable in a very strong sense, which allowed the authors to obtain exponential generalization bounds. Kutin and Niyogi [44] introduced a number of notions of stability and showed implications between them. The authors emphasized the importance of "almost-everywhere" stability and proved valuable extensions of McDiarmid's exponential inequality [43]. Mukherjee et al [55] proved that a combination of three stability notions is sufficient to bound the difference between the empirical estimate and the expected error, while for empirical risk minimization these notions are necessary and sufficient. The latter result showed an alternative to VC theory condition for consistency of empirical risk minimization. In this Chapter we prove, in a unified framework, some of the important results mentioned above, as well as show new ways of incorporating stability notions in Learning Theory.

We now give some intuition for using algorithmic stability. First, note that without any assumptions on the algorithm, nothing can be said about the mean and the variance of $\widetilde{\mathcal{R}}_{\text{emp}}$. One can easily come up with settings when the mean is converging to zero, but not the variance, or vice versa (e.g. Example 1), or both quantities diverge from zero.

The assumptions of this Chapter that allow us to bound the mean and the variance of $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ are loosely termed as *stability assumptions*. Recall that if the algorithm is a constant algorithm, $\widetilde{\mathcal{R}}_{\text{emp}}$ is bounded by the Central Limit Theorem. Of course, this is an extreme and the most "stable" case. It turns out that the "constancy" assumption on the algorithm can be relaxed while still achieving tight bounds. A central notion here is that of *Uniform Stability* [16]:

**Definition 3.2.1.** *Uniform Stability $\beta_\infty(n)$ of an algorithm $\mathcal{A}$ is*

$$\beta_\infty(n) := \sup_{Z_1,\ldots,Z_n,z\in\mathcal{Z},x\in\mathcal{X}} |\mathcal{A}(Z_1,\ldots,Z_n;X) - \mathcal{A}(z,Z_2,\ldots,Z_n;X)|.$$

Intuitively, if $\beta_\infty(n) \to 0$, the algorithm resembles more and more the constant algorithm when considered on similar samples (although it can produce distant functions on different samples). It can be shown that some well-known algorithms possess Uniform Stability with a certain rate on $\beta_\infty(n)$ (see [16] and section 3.6.1).

In the following sections, we will show how the bias and variance (or second moment) can be upper-bounded or decomposed in terms of quantities over "similar" samples. The advantage of this approach is that it allows one to check stability for a specific algorithm and derive generalization bounds without much further work. For instance, it is easy to show that k-Nearest Neighbors algorithm is $L_1$-stable and a generalization bound follows immediately (see section 3.4.1).

## 3.3 Bounding Bias and Variance

### 3.3.1 Decomposing the Bias

The bias of the resubstitution estimate and the deleted estimate can be written as quantities over similar samples:

$$\begin{aligned}
\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_z\ell(Z_1,\ldots,Z_n;Z) - \ell(Z_1,\ldots,Z_n;Z_i)\right]\right] \\
&= \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z) - \ell(Z_1,\ldots,Z_n;Z_1)\right] \\
&= \mathbb{E}\left[\ell(Z,Z_2,\ldots,Z_n;Z_1) - \ell(Z_1,\ldots,Z_n;Z_1)\right].
\end{aligned}$$

The first equality above follows because

$$\mathbb{E}\ell(Z_1,\ldots,Z_n;Z_k) = \mathbb{E}\ell(Z_1,\ldots,Z_n;Z_m)$$

for any $k, m$. The second equality holds by noticing that

$$\mathbb{E}\ell(Z_1,\ldots,Z_n;Z) = \mathbb{E}\ell(Z,Z_2,\ldots,Z_n;Z_1)$$

because the roles of $Z$ and $Z_1$ can be switched (see [63]). We will employ this trick many times in the later proofs, and for convenience we shall denote this renaming process by $Z \leftrightarrow Z_1$.

Let us inspect the quantity

$$\mathbb{E}\left[\ell(Z, Z_2, \ldots, Z_n; Z_1) - \ell(Z_1, \ldots, Z_n; Z_1)\right].$$

It is the average difference between the loss at a point $Z_1$ when it is not present in the learning sample (out-of-sample) and the loss at $Z_1$ when it is present in the $n$-tuple (in-sample). Hence, the bias $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}}$ will decrease if and only if the average behavior on in-sample and out-of-sample points is becoming more and more similar. This is a stability property and we will give a name to it:

**Definition 3.3.1.** *Average Stability $\beta_{bias}(n)$ of an algorithm $\mathcal{A}$ is*

$$\beta_{bias}(n) := \mathbb{E}\left[\ell(Z, Z_2, \ldots, Z_n; Z_1) - \ell(Z_1, \ldots, Z_n; Z_1)\right].$$

We now turn to the deleted estimate. The bias $\mathbb{E}\widetilde{\mathcal{R}}_{\text{loo}}$ can be written as

$$\begin{aligned}
\mathbb{E}\widetilde{\mathcal{R}}_{\text{loo}} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_Z\ell(Z_1, \ldots, Z_n; Z) - \ell(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n; Z_i))\right] \\
&= \mathbb{E}\left[\ell(Z_1, \ldots, Z_n; Z) - \ell(Z_2, \ldots, Z_n; Z_1)\right] \\
&= \mathbb{E}\left[\mathcal{R}(Z_1, \ldots, Z_n) - \mathcal{R}(Z_2, \ldots, Z_n)\right]
\end{aligned}$$

We will not give a name to this quantity, as it will not be used explicitly later. One can see that the bias of the deleted estimate should be small for reasonable algorithms. Unfortunately, the variance of the deleted estimate is large in general (see e.g. page 415 of [20]). The opposite is believed to be true for the resubstitution estimate. We refer the reader to Chap. 23, 24, and 31 of [20] for more information. Surprisingly, we will show in section 3.4.4 that for empirical risk minimization algorithms, if one shows that the bias of the resubstitution estimate decreases, one also obtains that the variance decreases.

### 3.3.2 Bounding the Variance

Having shown a decomposition of the bias of $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ in terms of stability conditions, we now show a simple way to bound the variance in terms of quantities over "similar" samples. In order to upper-bound the variance, we will use the Efron-Stein's bounds (Theorems 2.4 and 2.5).

The proofs of the Efron-Stein bounds are based on the fact that

$$\text{Var}(\Gamma) \leq \mathbb{E}(\Gamma - c)^2$$

for any constant $c$, and so

$$\text{Var}_i(\Gamma) = \mathbb{E}_{z_i}(\Gamma - \mathbb{E}_{z_i}\Gamma)^2 \leq \mathbb{E}_{z_i}(\Gamma - \Gamma_i)^2.$$

Thus, we artificially introduce a quantity over a "similar" sample to upper-bound the variance. If the increments $\Gamma - \Gamma_i$ and $\Gamma - \Gamma_i'$ are small, the variance is small. When applied to the function $T_n = \widetilde{\mathcal{R}}_{\text{emp}}(Z_1, \ldots, Z_n)$, this translates exactly into controlling the behavior of the algorithm $\mathcal{A}$ on similar samples:

$$\begin{aligned} \text{Var}(\widetilde{\mathcal{R}}_{\text{emp}}) &\leq n\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}}(Z_1, \ldots, Z_n) - \widetilde{\mathcal{R}}_{\text{emp}}(Z_2, \ldots, Z_n))^2 \\ &\leq 2n\mathbb{E}\left(\mathcal{R}(Z_1, \ldots, Z_n) - \mathcal{R}(Z_2, \ldots, Z_n)\right)^2 \\ &+ 2n\mathbb{E}\left(\mathcal{R}_{\text{emp}}(Z_2, \ldots, Z_n) - \mathcal{R}_{\text{emp}}(Z_1, \ldots, Z_n)\right)^2 \end{aligned} \tag{3.1}$$

Here we used the fact that the algorithm is invariant under permutation of coordinates, and therefore all the terms in the sum of Equation (2.5) are equal. This symmetry will be exploited to a great extent in the later sections. Note that similar results can be obtained using the replacement version of Efron-Stein's bound.

The meaning of the above bound is that if the mean *square* of the difference between expected errors of functions, learned from samples differing in one point, is decreasing faster than $n^{-1}$, and if the same holds for the empirical errors, then the variance of the resubstitution estimate is decreasing. Let us give names to the above

quantities.

**Definition 3.3.2.** *Empirical-Error (Removal) Stability of an algorithm $\mathcal{A}$ is*

$$\beta_{emp}^2(n) := \mathbb{E}\left|\mathcal{R}_{emp}(Z_1,\ldots,Z_n) - \mathcal{R}_{emp}(Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)\right|^2.$$

**Definition 3.3.3.** *Expected-Error (Removal) Stability of an algorithm $\mathcal{A}$ is*

$$\beta_{exp}^2(n) := \mathbb{E}\left|\mathcal{R}(Z_1,\ldots,Z_n) - \mathcal{R}(Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)\right|^2.$$

With the above definitions, the following Theorem follows:

**Theorem 3.3.1.**

$$\mathrm{Var}(\widetilde{\mathcal{R}}_{emp}) \leq 2n(\beta_{exp}^2(n) + \beta_{emp}^2(n)).$$

The following example shows that the ERM algorithm is always Empirical-Error Stable with $\beta_{\mathrm{emp}}(n) \leq M(n-1)^{-1}$. We deduce that $\widetilde{\mathcal{R}}_{\mathrm{emp}} \xrightarrow{P} 0$ for ERM whenever $\beta_{\mathrm{exp}} = o(n^{-1/2})$. As we will show in Section 3.4.4, the decay of Average Stability, $\beta_{\mathrm{bias}}(n) = o(1)$, is both necessary and sufficient for $\widetilde{\mathcal{R}}_{\mathrm{emp}} \xrightarrow{P} 0$ for ERM.

**Example 2.** *For an empirical risk minimization algorithm, $\beta_{emp}(n) \leq \frac{M}{n-1}$:*

$$\begin{aligned}
\mathcal{R}_{emp}(Z_2,\ldots,Z_n) - \mathcal{R}_{emp}(Z_1,\ldots,Z_n) &\leq \\
&\leq \frac{1}{n-1}\sum_{i=2}^{n}\ell(Z_2,\ldots,Z_n;Z_i) - \frac{1}{n-1}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) + \frac{M}{n-1} \\
&\leq \frac{1}{n-1}\sum_{i=2}^{n}\ell(Z_2,\ldots,Z_n;Z_i) - \frac{1}{n-1}\sum_{i=2}^{n}\ell(Z_1,\ldots,Z_n;Z_i) \\
&\qquad + \frac{M}{n-1} - \frac{1}{n-1}\ell(Z_1,\ldots,Z_n;Z_1) \leq \frac{M}{n-1}
\end{aligned}$$

*and the other direction is proved similarly.*

We will show in the following sections that a direct study of the second moment leads to better bounds. For the bound on the variance in Theorem 3.3.1 to decrease, $\beta_{\mathrm{exp}}$ and $\beta_{\mathrm{emp}}$ have to be $o(n^{-1/2})$. With an additional assumption, we will be able to

remove the factor $n$ from the bound 3.1 by upper-bounding the second moment and by exploiting the structure of the random variables $\widetilde{\mathcal{R}}_{\text{loo}}$ and $\widetilde{\mathcal{R}}_{\text{emp}}$.

## 3.4 Bounding the 2nd Moment

Instead of bounding the mean and variance of the estimators, we can bound the second moment. The reason for doing so is mathematical convenience and is due to the following straight-forward bounds on the second moment:

$$\begin{aligned}
\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2 &= \mathbb{E}\left[\mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\right]^2 - \mathbb{E}\left[\mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^n \ell(Z_1,\ldots,Z_n;Z_i)\right] \\
&+ \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell(Z_1,\ldots,Z_n;Z_i)\right]^2 - \mathbb{E}\left[\mathbb{E}_z \ell(Z_1,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^n \ell(Z_1,\ldots,Z_n;Z_i)\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_z \ell(Z_1,\ldots,Z_n;Z)\mathbb{E}_{Z'}\ell(Z_1,\ldots,Z_n;Z') - \mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_1)\right] \\
&+ \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z_1)\ell(Z_1,\ldots,Z_n;Z_2) - \mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_1)\right] \\
&+ \frac{1}{n}\mathbb{E}\ell(Z_1,\ldots,Z_n;Z_1)^2,
\end{aligned}$$

and the last term is bounded by $\frac{M^2}{n}$. Similarly,

$$\begin{aligned}
\mathbb{E}(\widetilde{\mathcal{R}}_{\text{loo}})^2 &\leq \mathbb{E}\left[\mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\mathbb{E}_{Z'}\ell(Z_1,\ldots,Z_n;Z') - \mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z_1)\right] \\
&+ \mathbb{E}\left[\ell(Z_2,\ldots,Z_n;Z_1)\ell(Z_1,Z_3,\cdots,Z_n;Z_2) - \mathbb{E}_z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z_1)\right] \\
&+ \frac{1}{n}\mathbb{E}\ell(Z_2,\ldots,Z_n;Z_1)^2,
\end{aligned}$$

and the last term is bounded by $\frac{M^2}{n}$.

In the proofs we will use the following inequality for random variables $X$, $X'$ and $Y$:

$$\mathbb{E}\left[XY - X'Y\right] \leq M\mathbb{E}|X - X'| \tag{3.2}$$

if $-M \leq Y \leq M$. The bounds on the second moments are already sums of terms of the type "$\mathbb{E}\left[XY - WZ\right]$", and we will find a way to use symmetry to change these

terms into the type "$\mathbb{E}[XY - X'Y]$", where $X$ and $X'$ will be quantities over similar samples, and so $\mathbb{E}|X - X'|$ will be bounded by a certain stability of the algorithm.

### 3.4.1 Leave-one-out (Deleted) Estimate

We have seen that

$$\mathbb{E}\widetilde{\mathcal{R}}_{\text{loo}} = \mathbb{E}[\mathcal{R}(Z_1, \ldots, Z_n) - \mathcal{R}(Z_2, \ldots, Z_n)]$$

and thus the bias decreases if and only if the expected errors are similar when learning on similar (one additional point) samples. Moreover, intuitively, these errors have to occur at the same places because otherwise evaluation of leave-one-out functions $\ell(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n; Z)$ will not tell us about $\ell(Z_1, \ldots, Z_n; Z)$. This implies that the $L_1$ distance between the functions on similar (one additional point) samples should be small. This connection between $L_1$ stability and the leave-one-out estimate has been observed by Devroye and Wagner [22] and further studied in [36]. We now define this stability notion:

**Definition 3.4.1.** *$L_1$-Stability of an algorithm $\mathcal{A}$ is*

$$\begin{aligned}\beta_1(n) :&= \|\ell(Z_1, \ldots, Z_n; \cdot) - \ell(Z_2, \ldots, Z_n; \cdot)\|_{L_1(\mu)} \\ &= \mathbb{E}_Z |\ell(Z_1, \ldots, Z_n; Z) - \ell(Z_2, \ldots, Z_n; Z)|.\end{aligned}$$

The following Theorem is proved in [22, 20] for classification algorithms. We give a similar proof for general learning algorithms. The result shows that the second moment (and therefore both bias and variance) of the leave-one-out error estimate is bounded by the $L_1$ distance between loss functions on similar samples.

**Theorem 3.4.1.**

$$\mathbb{E}(\widetilde{\mathcal{R}}_{loo})^2 \leq M(2\beta_1(n-1) + 4\beta_1(n)) + \frac{M^2}{n}.$$

*Proof.* The first term in the decomposition of the second moment of $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{loo}})^2$ can be

bounded as follows:

$$\begin{aligned}
&\mathbb{E}\ \ [\ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z') - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z_1)] \\
&= \mathbb{E}\,[\ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z')] \\
&= \mathbb{E}\,[\ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z') - \ell(Z_2,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z')] \\
&+ \mathbb{E}\,[\ell(Z_2,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z')] \\
&+ \mathbb{E}\,[\ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z') - \ell(z',Z_2,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z')] \\
&\leq 3M\beta_1(n).
\end{aligned}$$

The first equality holds by renaming $Z' \leftrightarrow Z_1$. In doing this, we are using the fact that all the variables $Z_1,\ldots,Z_n,Z,Z'$ are identically distributed and independent. To obtain the inequality above, note that each of the three terms is bounded (using Ineq. 3.2) by $M\beta_1(n)$ .

The second term in the decomposition is bounded similarly:

$$\begin{aligned}
&\mathbb{E}\ \ [\ell(Z_2,\ldots,Z_n;Z_1)\ell(Z_1,Z_3,\ldots,Z_n;Z_2) - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z_1)] \\
&= \mathbb{E}\,[\ell(Z',Z_3,\ldots,Z_n;Z)\ell(Z,Z_3,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z')] \\
&= \mathbb{E}\,[\ell(Z',Z_3,\ldots,Z_n;Z)\ell(Z,Z_3,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z,Z_3,\ldots,Z_n;Z')] \\
&+ \mathbb{E}\,[\ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z,Z_3,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_3,\ldots,Z_n;Z')] \\
&+ \mathbb{E}\,[\ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_3,\ldots,Z_n;Z') - \ell(Z',Z_2,\ldots,Z_n;Z)\ell(Z_2,\ldots,Z_n;Z')] \\
&\leq M\beta_1(n) + 2M\beta_1(n-1)
\end{aligned}$$

The first equality follows by renaming $Z_2 \leftrightarrow Z'$ as well as $Z_1 \leftrightarrow Z$ in the first term, and $Z_1 \leftrightarrow Z'$ in the second term. Finally, we bound the last term by $M^2/n$ to obtain the result. □

### 3.4.2 Empirical Error (Resubstitution) Estimate: Replacement Case

Recall that the the bias of the resubstitution estimate is the Average Stability, $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} = \beta_{\text{bias}}$. However this is not enough to bound the second moment $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ for general algorithms. Nevertheless, $\beta_{\text{bias}}$ measures the average performance of in-sample and out-of-sample errors and this is inherently linked to the closeness of the resubstitution (in-sample) estimate and the expected error (out-of-sample performance). It turns out that it is possible to derive bounds on $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ by using a stronger version of the Average Stability. The natural strengthening is requiring that not only the first, but also the second moment of $\ell(Z_1, \ldots, Z_n; Z_i) - \ell(Z_1, \ldots, Z_i', \ldots, Z_n; Z_i)$ is decaying to 0. We follow [44] in calling this type of stability *Cross-Validation* (CV) stability:

**Definition 3.4.2.** *CV (Replacement) Stability of an algorithm $\mathcal{A}$ is*

$$\beta_{cvr} := \mathbb{E}|\ell(Z_1, \ldots, Z_n; Z_1) - \ell(Z, Z_2, \ldots, Z_n; Z_1)|,$$

*where the expectation is over a draw of $n+1$ points.*

The following Theorem was proved in [16]. Here we give a version of the proof.

**Theorem 3.4.2.**

$$\mathbb{E}(\widetilde{\mathcal{R}}_{emp})^2 \leq 6M\beta_{cvr}(n) + \frac{M^2}{n}.$$

*Proof.* The first term in the decomposition of $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\mathbb{E}_{Z'}\ell(Z_1,\ldots,Z_n;Z') - \mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_2)\right] \\
&\quad = \mathbb{E}\left[\ell(Z_1,Z',Z_3,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right. \\
&\qquad \left. - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_2)\right] \\
&\quad = \mathbb{E}\left[\ell(Z_1,Z',Z_3,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right. \\
&\qquad \left. - \ell(Z_1,Z,Z_3,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right] \\
&\quad + \mathbb{E}\left[\ell(Z_1,Z,Z_3,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right. \\
&\qquad \left. - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right] \\
&\quad + \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,Z',Z_3,\ldots,Z_n;Z_2)\right. \\
&\qquad \left. - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_2)\right] \\
&\quad \leq 3M\beta_{\text{cvr}}(n).
\end{aligned}
$$

The first equality follows from renaming $Z_2 \leftrightarrow Z'$ in the first term. Each of the three terms in the sum above is bounded by $M\beta_{\text{cvr}}(n)$.

The second term in the decomposition of $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}\quad \left[\ell(Z_1,\ldots,Z_n;Z_1)\ell(Z_1,\ldots,Z_n;Z_2) - \mathbb{E}_Z \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_1)\right] \\
&= \quad \mathbb{E}\left[\ell(Z,Z_2,\ldots,Z_n;Z)\ell(Z,Z_2,\ldots,Z_n;Z_2) - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_2)\right] \\
&= \quad \mathbb{E}\left[\ell(Z,Z_2,\ldots,Z_n;Z)\ell(Z,Z_2,\ldots,Z_n;Z_2) - \ell(Z_1,\ldots,Z_n;Z)\ell(Z,Z_2,\ldots,Z_n;Z_2)\right] \\
&+ \quad \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z)\ell(Z,Z_2,\ldots,Z_n;Z_2) - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,Z,Z_3,\ldots,Z_n;Z_2)\right] \\
&+ \quad \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,Z,Z_3,\ldots,Z_n;Z_2) - \ell(Z_1,\ldots,Z_n;Z)\ell(Z_1,\ldots,Z_n;Z_2)\right] \\
&\leq \quad 3M\beta_{\text{cvr}}(n).
\end{aligned}
$$

The first equality follows by renaming $Z_1 \leftrightarrow Z$ in the first term. Again, each of the three terms in the sum above can be bounded by $M\beta_{\text{cvr}}(n)$.

□

### 3.4.3 Empirical Error (Resubstitution) Estimate

Mukherjee et al [55] considered the removal version of the CV stability defined in Section 3.4.3, the motivation being that addition of a new point $Z'$ complicates the cross-validation nature of the stability. Another motivation is the fact that $\ell(Z_1, \ldots, Z_n; Z_1) - \ell(Z_2, \ldots, Z_n; Z_1)$ is non-negative for Empirical Risk Minimization. It turns out that this removal version of the CV stability together with Expected and Empirical Stabilities upper-bound $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}}$. Following [55], we have the following definition:

**Definition 3.4.3.** *CV (Removal) Stability of an algorithm $\mathcal{A}$ is*

$$\beta_{cv}(n) := \mathbb{E}|\ell(Z_1, \ldots, Z_n; Z_1) - \ell(Z_2, \ldots, Z_n; Z_1)|.$$

The following theorem was proved in [55]. Here we give a version of the proof.

**Theorem 3.4.3.**

$$\mathbb{E}(\widetilde{\mathcal{R}}_{emp})^2 \leq M(\beta_{cv}(n) + 4\beta_{exp}(n) + 2\beta_{emp}(n)) + \frac{M^2}{n}.$$

*Proof.* The first term in the decomposition of the second moment of $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\ [\ell(Z_1, \ldots, Z_n; Z)\ell(Z_1, \ldots, Z_n; Z') - \ell(Z_1, \ldots, Z_n; Z)\ell(Z_1, \ldots, Z_n; Z_1)] \\
&= \mathbb{E}\,[\ell(Z', Z_2, \ldots, Z_n; Z)\ell(Z', Z_2, \ldots, Z_n; Z_1) - \ell(Z_1, \ldots, Z_n; Z)\ell(Z_1, \ldots, Z_n; Z_1)] \\
&= \mathbb{E}\,[\ell(Z', Z_2, \ldots, Z_n; Z)\mathbb{E}_{Z_1}\ell(Z', Z_2, \ldots, Z_n; Z_1) - \ell(Z', Z_2, \ldots, Z_n; Z)\mathbb{E}_{Z_1}\ell(Z_2, \ldots, Z_n; Z_1)] \\
&+ \mathbb{E}\,[\mathbb{E}_Z\ell(Z', Z_2, \ldots, Z_n; Z)\ell(Z_2, \ldots, Z_n; Z_1) - \mathbb{E}_Z\ell(Z_2, \ldots, Z_n; Z)\ell(Z_2, \ldots, Z_n; Z_1)] \\
&+ \mathbb{E}\,[\mathbb{E}_Z\ell(Z_2, \ldots, Z_n; Z)\ell(Z_2, \ldots, Z_n; Z_1) - \mathbb{E}_Z\ell(Z_1, \ldots, Z_n; Z)\ell(Z_2, \ldots, Z_n; Z_1)] \\
&+ \mathbb{E}\,[\ell(Z_1, \ldots, Z_n; Z)\ell(Z_2, \ldots, Z_n; Z_1) - \ell(Z_1, \ldots, Z_n; Z)\ell(Z_1, \ldots, Z_n; Z_1)] \\
&\leq M(3\beta_{\text{exp}}(n) + \beta_{\text{cv}}(n)).
\end{aligned}$$

The first equality follows by renaming $Z_1 \leftrightarrow Z$ in the first term. In the sum above,

the first three terms are each bounded by $M\beta_{\text{exp}}(n)$, while the last one is bounded by $M\beta_{\text{cv}}(n)$. Since the Expected (and Empirical) Error stability has been defined in Section 3.3.2 as expectation of a square, we used the fact that $\mathbb{E}|X| \leq (\mathbb{E}X^2)^{1/2}$.

The second term in the decomposition of $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ is bounded as follows:

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right)^2 - \mathbb{E}_z\ell(Z_1,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right]$$
$$= \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z_1)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) - \ell(Z_1,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right]$$

$$= \mathbb{E}\left[\ell(Z_1,\ldots,Z_n;Z_1)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) - \ell(Z_2,\ldots,Z_n;Z_1)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right]$$
$$+ \mathbb{E}\left[\ell(Z_2,\ldots,Z_n;Z_1)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) - \ell(Z_2,\ldots,Z_n;Z)\frac{1}{n-1}\sum_{i=2}^{n}\ell(Z_2,\ldots,Z_n;Z_i)\right]$$
$$+ \mathbb{E}\left[\ell(Z_2,\ldots,Z_n;Z)\frac{1}{n-1}\sum_{i=2}^{n}\ell(Z_2,\ldots,Z_n;Z_i) - \ell(Z_2,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right]$$
$$+ \mathbb{E}\left[\mathbb{E}_Z\ell(Z_2,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) - \mathbb{E}_Z\ell(Z_1,\ldots,Z_n;Z)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)\right]$$
$$\leq M(\beta_{\text{cv}}(n) + 2\beta_{\text{emp}}(n) + \beta_{\text{exp}}(n)).$$

The first equality follows by symmetry:

$$\ell(Z_1,\ldots,Z_n;Z_k)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i) = \ell(Z_1,\ldots,Z_n;Z_m)\frac{1}{n}\sum_{i=1}^{n}\ell(Z_1,\ldots,Z_n;Z_i)$$

for all $k, m$. First term in the sum above is bounded by $M\beta_{\text{cv}}(n)$. The second term is bounded by $M\beta_{\text{emp}}(n)$ (and $Z_1 \leftrightarrow Z$). The third term is also bounded by $M\beta_{\text{emp}}(n)$, and the last term by $M\beta_{\text{exp}}(n)$. □

### 3.4.4 Resubstitution Estimate for the Empirical Risk Minimization Algorithm

It turns out that for the ERM algorithm, $\widetilde{\mathcal{R}}_{\text{emp}}$ is "almost positive". Intuitively, if one minimizes the empirical error, then the expected error is likely to be larger than the empirical estimate. Since $\widetilde{\mathcal{R}}_{\text{emp}}$ is "almost positive", $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} \to 0$ implies $|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$. We now give a formal proof of this reasoning.

Recall that an ERM algorithm searches in the function space $\mathcal{H}$. Let

$$f^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_Z \ell(f; Z),$$

the minimizer of the expected error[1]. Consider the shifted loss class

$$\mathcal{L}'(\mathcal{H}) = \{\ell'(f;\cdot) = \ell(f;\cdot) - \ell(f^*;\cdot) | f \in \mathcal{H}\}$$

and note that $\mathbb{E}_Z \ell'(f; Z) \geq 0$ for any $f \in \mathcal{H}$. Trivially, if $\ell(Z_1, \ldots, Z_n; \cdot)$ is an empirical minimizer over the loss class $\mathcal{L}(\mathcal{H})$, then $\ell'(f;\cdot) = \ell(Z_1, \ldots, Z_n; \cdot) - \ell(f^*;\cdot)$ is an empirical minimizer over the shifted loss class $\mathcal{L}'(\mathcal{H})$

$$\begin{aligned}\mathbb{E}_z \ell'(Z_1, \ldots, Z_n; Z) - \frac{1}{n}\sum_{i=1}^{n} \ell'(Z_1, \ldots, Z_n; Z_i) = \\ \mathbb{E}_z \ell(Z_1, \ldots, Z_n; Z) - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_1, \ldots, Z_n; Z_i) \\ - \left(\mathbb{E}_Z \ell(f^*; Z) - \frac{1}{n}\sum_{i=1}^{n} \ell(f^*; Z_i)\right).\end{aligned}$$

Note that $\frac{1}{n}\sum_{i=1}^{n} \ell'(Z_1, \ldots, Z_n; Z_i) \leq 0$ because $\mathcal{L}'(\mathcal{H})$ contains the zero function. Therefore, the left-hand side is non-negative and the second term on the right-hand

[1] If the minimizer does not exist, we consider $\varepsilon$-minimizer

side is small with high probability because $f^*$ is non-random. We have

$$\mathbb{P}\left(\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n) < -\varepsilon\right) \leq \mathbb{P}\left(\mathbb{E}_Z \ell(f^*;Z) - \frac{1}{n}\sum_{i=1}^{n} \ell(f^*;Z_i) < -\varepsilon\right)$$
$$\leq e^{-2n\varepsilon^2/M^2}.$$

Therefore,

$$\mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}| \leq \mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} + 2\varepsilon + 2Me^{-2n\varepsilon^2/M^2}.$$

If $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} \to 0$, the right-hand side can be made arbitrarily small for large enough $n$, thus proving $\mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}| \to 0$. Clearly, $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} \to 0$ whenever $\mathbb{E}|\widetilde{\mathcal{R}}_{\text{emp}}| \to 0$. Hence, we have the following Theorem:

**Theorem 3.4.4.** *For empirical risk minimization, $\beta_{bias}(n) \to 0$ is equivalent to $|\widetilde{\mathcal{R}}_{emp}| \xrightarrow{P} 0$.*

**Remark 3.4.1.** *With this approach the rate of convergence of $\mathcal{R}_{emp}(Z_1,\ldots,Z_n)$ to $\mathcal{R}(Z_1,\ldots,Z_n)$ is limited by the rate of convergence of $\frac{1}{n}\sum_{i=1}^{n} \ell(f^*;Z_i)$ to $\mathbb{E}_Z\ell(f^*;Z)$, which is $O(n^{-1/2})$ without further assumptions.*

For ERM, one can show that

$$|\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z_2,\ldots,Z_n)| \leq \frac{M}{n}.$$

Hence, a "removal" version of Average Stability is closely related to Average Stability:

$$\mathbb{E}\left(\ell(Z_1,\ldots,Z_n;Z_1) - \ell(Z_2,\ldots,Z_n;Z_1)\right) = \mathbb{E}\left(\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) - \mathcal{R}(Z_2,\ldots,Z_n)\right)$$
$$= \beta_{\text{bias}}(n-1) + \mathbb{E}\left(\mathcal{R}_{\text{emp}}(Z_2,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n)\right).$$

Thus,

$$\mathbb{E}\left(\ell(Z_1,\ldots,Z_n;Z_1) - \ell(Z_2,\ldots,Z_n;Z_1)\right) \to 0$$

is also equivalent to $|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$.

Furthermore, one can show that

$$\ell(Z_1, \ldots, Z_n; Z_1) - \ell(Z_2, \ldots, Z_n; Z_1) \geq 0$$

for ERM (see [55]), and so CV (Removal) Stability, defined in Section 3.4.3, is equal to the above removal version of Average Stability. Hence, $\beta_{\text{cv}}(n) \to 0$ is equivalent to $|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$.

Since empirical risk minimization over a uniform Glivenko-Cantelli class implies that $|\widetilde{\mathcal{R}}_{\text{emp}}| \xrightarrow{P} 0$, it also implies that $\beta_{\text{bias}}(n) \to 0$ and $\beta_{\text{cv}}(n) \to 0$. Thus, ERM over a UGC class is stable in these regards. By using techniques from the Empirical Process Theory, it can be shown (see [19]) that for ERM over a smaller family of classes, called Donsker classes, a much stronger stability in $L_1$ norm (see Definition 3.4.1) holds: $\beta_1(n) \xrightarrow{P} 0$. Donsker classes are classes of functions satisfying the Central Limit Theorem, and for binary classes of function this is equivalent to finiteness of the VC dimension.

## 3.5 Lower Bounds

We lower-bound the second moment $\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2$ as follows.

$$\begin{aligned}
\mathbb{E}(\widetilde{\mathcal{R}}_{\text{emp}})^2 &\geq (\mathbb{E}\ell(Z_1, \ldots, Z_n; Z) - \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} \ell(Z_1, \ldots, Z_n; Z_i))^2 \\
&= (\mathbb{E}\ell(Z_1, \ldots, Z_n; Z) - \mathbb{E}\ell(Z_1, \ldots, Z_n; Z_1))^2 \\
&= (\mathbb{E}\left[\ell(Z, Z_2, \ldots, Z_n; Z_1) - \ell(Z_1, \ldots, Z_n; Z_1)\right])^2 \\
&= \beta_{\text{bias}}^2(n)
\end{aligned}$$

Therefore, convergence of the Average Stability $\beta_{\text{bias}}(n) \to 0$ is a necessary condition for the convergence of the empirical error to the expected error. For ERM, this condition is also sufficient, as shown in the previous Section.

Now, rewrite $\widetilde{\mathcal{R}}_{\text{emp}}$ as

$$\begin{aligned}
\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n) &= \mathcal{R}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) \\
&= \mathcal{R}(Z_1,\ldots,Z_n) - \mathbb{E}\mathcal{R}(Z_1,\ldots,Z_n) \\
&+ \mathbb{E}\mathcal{R}(Z_1,\ldots,Z_n) - \mathbb{E}\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) \\
&+ \mathbb{E}\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) \\
&= (\mathcal{R}(Z_1,\ldots,Z_n) - \mathbb{E}\mathcal{R}(Z_1,\ldots,Z_n)) \\
&+ (\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) - \mathbb{E}\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n)) \\
&+ \beta_{\text{bias}}(n)
\end{aligned}$$

If for an algorithm one shows that $|\mathcal{R} - \mathbb{E}\mathcal{R}| \xrightarrow{P} 0$, then $|\mathcal{R}_{\text{emp}} - \mathbb{E}\mathcal{R}_{\text{emp}}| \xrightarrow{P} 0$ if and only if $\widetilde{\mathcal{R}}_{\text{emp}} \xrightarrow{P} 0$. Same holds in the other direction: if empirical errors converge to their expectations in probability, then $\widetilde{\mathcal{R}}_{\text{emp}} \xrightarrow{P} 0$ if and only if the expected errors also converge.

## 3.6 Rates of Convergence

Previous sections focused on finding rather weak conditions for proving $\widetilde{\mathcal{R}}_{\text{emp}} \xrightarrow{P} 0$ and $\widetilde{\mathcal{R}}_{\text{loo}} \xrightarrow{P} 0$ via Markov's inequality. With stronger notions of stability, it is possible to use more sophisticated inequalities, which is the focus of this section.

### 3.6.1 Uniform Stability

Uniform Stability (see Definition 3.2.1), is a very strong notion, and we would not expect, in general, that $\beta_\infty(n) \to 0$. Surprisingly, for *Tikhonov Regularization* algorithms

$$\mathcal{A}(Z_1,\ldots,Z_n) = \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h;Z_i) + \lambda\|h\|_K^2$$

it can be shown [16] that

$$\beta_\infty(n) \leq \frac{L^2\kappa^2}{2\lambda n},$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel $K$, $K(x,x) \leq \kappa^2 < \infty, \forall x \in \mathcal{X}$, and $L$ is a Lipschitz constant relating distances between functions $f \in \mathcal{H}$ to distances between losses $\ell(f) \in \mathcal{L}(\mathcal{H})$.

Clearly, $\beta_\infty$ dominates all stabilities discussed in the previous sections, and so can be used to bound the mean and variance of the estimators. For this strong stability a more powerful concentration inequality can be used instead of Markov's inequality. McDiarmid's bounded difference inequality (Chapter 2, Theorem 2.4.3) states that if a function of many random variables does not change much when one variable is changed, then the function is almost a constant. This is exactly what we need to bound $\widetilde{\mathcal{R}}_{\text{emp}}$ or $\widetilde{\mathcal{R}}_{\text{loo}}$.

Bousquet and Elisseeff [16] applied McDiarmid's inequality to $T_n = \widetilde{\mathcal{R}}_{\text{emp}}$:

$$\begin{aligned}
&|\widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z_n) - \widetilde{\mathcal{R}}_{\text{emp}}(Z_1,\ldots,Z,\ldots,Z_n)| \\
&\qquad \leq |\mathcal{R}_{\text{emp}}(Z_1,\ldots,Z_n) - \mathcal{R}_{\text{emp}}(Z,Z_2,\ldots,Z_n)| \\
&\qquad + |\mathcal{R}(Z_1,\ldots,Z_n) - \mathcal{R}(Z,Z_2,\ldots,Z_n)| \\
&\qquad \leq \frac{1}{n}|\ell(Z_1,\ldots,Z_n;Z_1) - \ell(Z,Z_2,\ldots,Z_n;Z)| \\
&\qquad + \frac{1}{n}\sum_{j=2}^{n} |\ell(Z_1,\ldots,Z_n;Z_j) - \ell(Z,Z_2,\ldots,Z_n;Z_j)| \\
&\qquad + \mathbb{E}'_Z|\ell(Z_1,\ldots,Z_n;Z') - \ell(Z,Z_2,\ldots,Z_n;Z')| \\
&\qquad \leq 2\beta_\infty(n) + \frac{M}{n} =: \beta_n.
\end{aligned}$$

If $\beta_\infty(n) = o(n^{-1/2})$, McDiarmid's inequality shows that $\widetilde{\mathcal{R}}_{\text{emp}}$ is exponentially concentrated around $\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}}$, which is also small:

$$\mathbb{E}\widetilde{\mathcal{R}}_{\text{emp}} = \beta_{\text{bias}}(n) \leq \beta_\infty(n).$$

Therefore,

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\widetilde{\mathcal{R}}_{\text{emp}} \geq \beta_\infty(n) + \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{(2n\beta_\infty(n) + M)^2}\right).$$

Notice that for ERM,

$$|\mathcal{R}_{\text{emp}}(Z_1, \ldots, Z_n) - \mathcal{R}_{\text{emp}}(Z, Z_2, \ldots, Z_n)| \leq \frac{M}{n}$$

and so it is enough to require

$$\beta_{\text{bias}} \to 0$$

and

$$|\mathcal{R}(Z_1, \ldots, Z_n) - \mathcal{R}(Z, Z_2, \ldots, Z_n)| = o(n^{-1/2})$$

to get exponential bounds. The last requirement is strong, as it requires expected errors on similar samples to be close for *every* sample. The next section deals with "almost-everywhere" stabilities (see [44]), i.e. when a stability quantity is small for most samples.

### 3.6.2 Extending McDiarmid's Inequality

As one extreme, if we know that $\beta_\infty(n) = o(n^{-1/2})$, we can use exponential McDiarmid's inequality. As the other extreme, if we only have information about averages $\beta_{\text{emp}}$ and $\beta_{\text{exp}}$, we are forced to use the second moment and Chebyshev's or Markov's inequality. What happens in-between these extremes? What if we know more about the random variables $\mathcal{R}_{\text{emp}}(Z_1, \ldots, Z_n) - \mathcal{R}_{\text{emp}}(Z, Z_2, \ldots, Z_n)$? One example is the case when we know that these random variables are almost always small. Unfortunately, assumptions of McDiarmid's inequality are no longer satisfied, so other ways of deriving exponential bounds are needed. This section discusses this situation. The proofs of the results are deferred to Chapter 6.

Assume that for a given $\beta_n$, a measurable function $T_n : \mathcal{Z}^n \mapsto [-M, M]$ satisfies the bounded difference condition

$$\forall i, \ \sup_{z_i' \in \mathcal{Z}} |T_n(Z_1, \ldots, Z_n) - T_n(Z_1, \ldots, z_i', \ldots, Z_n)| \leq \beta_n \tag{3.3}$$

on a subset $G \subseteq \mathcal{Z}^n$ of measure $1 - \delta_n$, while

$$\forall (Z_1, \ldots, Z_n) \in \bar{G}, \exists z_i' \in \mathcal{Z} \text{ s.t.}$$

$$\beta_n < |T_n(Z_1, \ldots, Z_n) - T_n(Z_1, \ldots, z_i', \ldots, Z_n)| \leq 2M.$$

Here $\bar{G}$ denotes the complement of the subset $G$. Again, denote $\Gamma = T_n(Z_1, \ldots, Z_n)$, $\Gamma_i' = T_n(Z_1, \ldots, Z_i', \ldots, Z_n)$. A simple application of Efron-Stein inequality (Theorem 2.4) shows that

$$\begin{aligned} \mathrm{Var}(T_n) &\leq \frac{1}{2} n \mathbb{E} \left(T_n(Z_1, \ldots, Z_n) - T_n(Z, Z_2, \ldots, Z_n)\right)^2 \\ &\leq \frac{1}{2} n \mathbb{E} \left[I_{(Z_1, \ldots, Z_n) \in G} \left(T_n(Z_1, \ldots, Z_n) - T_n(Z, Z_2, \ldots, Z_n)\right)^2\right] \\ &+ \frac{1}{2} n \mathbb{E} \left[I_{(Z_1, \ldots, Z_n) \in \bar{G}} \left(T_n(Z_1, \ldots, Z_n) - T_n(Z, Z_2, \ldots, Z_n)\right)^2\right] \\ &\leq \frac{1}{2} n (\beta_n^2 + 4M^2 \delta_n). \end{aligned} \tag{3.4}$$

This leads to a polynomial bound on $\mathbb{P}\left(|T_n - \mathbb{E}T_n| \geq \varepsilon\right)$. Kutin and Niyogi [44, 43] proved an inequality which is exponential when $\delta_n$ decays exponentially with $n$, thus extending McDiarmid's inequality to incorporate a small possibility of a large jump of $T_n$. A more general version of their bound is the following:

**Theorem 3.6.1** (Kutin and Niyogi [44]). *Assume $T_n : \mathcal{Z}^n \mapsto [-M, M]$ satisfies the bounded difference condition (3.3) on a set of measure $1 - \delta_n$ and denote $\Gamma = T_n(Z_1, \ldots, Z_n)$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(|T_n - \mathbb{E}T_n| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-\varepsilon^2}{8n\beta_n^2}\right) + \frac{2Mn\delta_n}{\beta_n}. \tag{3.5}$$

Note that the bound tightens only if $\beta_n = o(n^{-1/2})$ and $\delta_n / \beta_n = o(n^{-1})$. Furthermore, the bound is exponential only if $\delta_n$ decays exponentially[2].

In Chapter 6 we prove the following extension of McDiarmid's inequality.

**Theorem 3.6.2** (Chapter 6, Theorem 6.1.4). *Assume $T_n : \mathcal{Z}^n \mapsto \mathbb{R}$ satisfies the*

[2]By exponential rate we mean decay $o(exp(-n^r))$ for a fixed $r > 0$.

*bounded difference condition (3.3) on a set of measure* $1 - \delta_n$, *and denote* $\Gamma = T_n(Z_1, \ldots, Z_n)$. *Then for any* $q \geq 2$ *and* $\varepsilon > 0$,

$$\mathbb{P}\left(T_n - \mathbb{E}T_n > \varepsilon\right) \leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q},$$

*where* $\kappa \approx 1.271$.

Having proved extension to McDiarmid's inequality, we can use it in a straightforward way to derive bounds on $\mathbb{P}\left(|\widetilde{\mathcal{R}}_{\text{emp}}| > \varepsilon\right)$ and $\mathbb{P}\left(|\widetilde{\mathcal{R}}_{\text{loo}}| > \varepsilon\right)$ when expected and empirical quantities do not change "most of the time", when compared on similar samples (see [44] for examples).

## 3.7 Summary and Open Problems

We have shown how stability of algorithms provides an alternative to classical Statistical Learning Theory approach for controlling the behavior of empirical and leave-one-out estimates. The results presented are by no means a complete picture: one can come up with other notions of algorithmic stability, suited for the problem. Our goal was to present some results in a common framework and delineate important techniques for proving bounds.

One important (and largely unexplored) area of further research is looking at existing algorithms and proving bounds on their stabilities. For instance, work of Caponnetto and Rakhlin [19] showed that empirical risk minimization (over certain classes) is $L_1$ stable. It might turn out that other algorithms are stable in this (or even stronger) sense when considered over restricted function classes, which are nevertheless used in practice. Can these results lead to faster learning rates for algorithms?

Adding a regularization term for ERM leads to an extremely stable Tikhonov Regularization algorithm. How can regularization be used to stabilize other algorithms, and how does this affect the bias-variance tradeoff of fitting the data versus having a simple solution?

Though the results presented in this Chapter are theoretical, there is a potential for estimating stability in practice. Can a useful quantity be computed by running the algorithm many times to determine its stability? Can this quantity serve as a measure of the performace of the algorithm?

# Chapter 4

# Performance of Greedy Error Minimization Procedures

The results of this Chapter appear partially in [61].

## 4.1 General Results

Recall that the empirical risk minimization principle states that one should search for a function minimizing the empirical risk within a class $\mathcal{H}$:

$$\mathcal{A}(Z_1, \ldots, Z_n) = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; Z_i).$$

In practice, such a procedure is not tractable for two reasons:

- the loss function might not be convex;
- the class $\mathcal{H}$ is too large.

Let us discuss the convexity issue. It can be shown (see Arora et al [2], Ben-David et al [9]) that minimizing the empirical error in the classification setting with the indicator loss

$$\ell(yh(x)) = I(yh(x) \leq 0)$$

is computationally intractable even for simple classes of functions. In recent years, several papers addressed this difficulty. Bartlett et al [5], Lugosi and Vayatis [51], and Zhang [79] have studied the statistical consequence of replacing the indicator loss by a convex upper bound $\phi$. For instance, Bartlett et all [5] showed that for any $f$,

$$\psi(\mathcal{R}(f) - \mathcal{R}^*) \leq \mathcal{R}_\phi(f) - \mathcal{R}_\phi^*$$

for a nondecreasing function $\psi : [0,1] \to [0,\infty)$. Here $\mathcal{R}_\phi(f) = \mathbb{E}\phi(Yf(X))$, $\mathcal{R}_\phi^* = \inf_f \mathcal{R}_\phi(f)$, and $\mathcal{R}^* = \inf_f \mathbb{E}I(Yf(X) \leq 0)$. The latter quantity is called the *Bayes risk*. In fact, $\phi$ does not need to be convex for this result, but rather "classification-calibrated" (see [5]). Hence, minimization of the surrogate $\phi$ risk might not only alleviate the computational difficulty of minimizing the indicator loss, but also result in a consistent procedure with respect to the original loss. This type of result provides one of the first connections between computational and statistical issues in learning algorithms.

We now turn to the second computational issue related to the size of $\mathcal{H}$. If the class of functions is large, the search for an empirical minimizer is intractable. A small class $\mathcal{H}$ means we have little hope of capturing the unknown phenomenon, i.e. the approximation error is large. In this chapter we present an *approximate* greedy minimization method, which is computationally tractable. This method allows us to search in a greedy way over a large class, which is a convex hull of a small class.

The greedy procedure described next goes back at least to the paper of Jones [35]. Variants of it have been used by Barron [3, 49] and Mannor et al [52, 53]. The version described here is the most general one, appearing in the paper of Zhang [80].

Suppose $\mathcal{H}$ is a subset of a linear space. Denote the convex hull of $k > 0$ terms as

$$\text{conv}_k\ (\mathcal{H}) = \left\{ \sum_{i=1}^{k} \alpha_i h_i : \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1,\ h_i \in \mathcal{H} \right\}$$

and the convex hull of $\mathcal{H}$ as

$$\text{conv } (\mathcal{H}) = \cup_{k \geq 1} \text{conv}_k \ (\mathcal{H}).$$

We would like to minimize a convex functional $\Upsilon$ over the convex hull conv $(\mathcal{H})$. Define the closeness of a solution $g \in$ conv $(\mathcal{H})$ to the optimal by

$$\Delta\Upsilon(g) = \Upsilon(g) - \inf_{v \in \text{conv } (\mathcal{H})} \Upsilon(v).$$

The objective is to find a sequence $\{g_k\}$ of functions such that $\Delta\Upsilon(g_k) \to 0$ and $g_k \in \text{conv}_k \ (\mathcal{H})$.

**Algorithm 1** Greedy Minimization Algorithm

1: Start with $g_1 \in \mathcal{H}$
2: **for** $k = 2$ to $N$ **do**
3: Find $h \in \mathcal{H}$ and $0 \leq \alpha_k \leq 1$ such that
4: $(h, \alpha_k) = \text{argmin } \Upsilon((1 - \alpha_k)g_{k-1} + \alpha_k h)$
5: Let $g_k = (1 - \alpha_k)g_{k-1} + \alpha_k h$
6: **end for**

The minimization step can be performed approximately, to within some $\varepsilon_k \geq 0$ converging to zero:

$$\Upsilon((1 - \alpha_k)g_{k-1} + \alpha_k h) \leq \inf_{\bar{\alpha}, \bar{h}} \Upsilon((1 - \bar{\alpha})g_{k-1} + \bar{\alpha}_k \bar{h}) + \varepsilon_k.$$

The following Theorem of Zhang [80] states that under very general assumptions on $\Upsilon$, the sequence $g_k$ converges to the optimum at the rate $O(1/k)$.

**Theorem 4.1.1** (Zhang [80]). *Assume $\Upsilon$ is differentiable and*

$$\sup_{g', g'' \in conv \ (\mathcal{H}), \theta \in (0,1)} \frac{d^2}{d\theta^2} \Upsilon((1 - \theta)g' + \theta g'') \leq c < +\infty.$$

*Assume that the optimization at step (3) of Algorithm 1 can be performed exactly for all $k \geq 1$. Then*

$$\Delta\Upsilon(g_k) \leq \frac{2c}{k + 2}.$$

Now apply the Algorithm 1 to the empirical error $\Upsilon(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(g; Z_i)$ for a convex (in the first argument) loss $\ell$. Hence, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \ell(g_k; Z_i) \leq \inf_{g \in \text{conv}\,(\mathcal{H})} \frac{1}{n}\sum_{i=1}^{n} \ell(g; Z_i) + \frac{2c}{k+2}$$

under the assumptions of Theorem 4.1.1. Denoting

$$g^* = \operatorname*{argmin}_{g \in \text{conv}\,(\mathcal{H})} \mathcal{R}(g),$$

we obtain

$$\begin{aligned}
\mathcal{R}(g_k) - \mathcal{R}(g^*) &= \mathcal{R}(g_k) - \mathcal{R}_{\text{emp}}(g_k) \\
&\quad + \mathcal{R}_{\text{emp}}(g_k) - \mathcal{R}_{\text{emp}}(g^*) \\
&\quad + \mathcal{R}_{\text{emp}}(g^*) - \mathcal{R}(g^*) \\
&\leq 2 \sup_{g \in \text{conv}\,(\mathcal{H})} |\mathcal{R}(g) - \mathcal{R}_{\text{emp}}(g)| + \frac{2c}{k+2}. \qquad (4.1)
\end{aligned}$$

The supremum of the deviation between the empirical and expected averages is usually bounded, via a Symmetrization and Concentration steps, by Rademacher averages of the function class. Rademacher averages serve as a complexity measure, which can be upper-bounded by the metric entropy of the function class. Unfortunately, conv $(\mathcal{H})$ can be large even for small $\mathcal{H}$, as the Example 3 shows. This would imply a loose upper bound on the convergence of $\mathcal{R}(g_k) - \mathcal{R}(g^*)$ to zero. Luckily, under an assumption on the loss function $\ell$, we can employ a comparison inequality for Rademacher Averages [46], obtaining a bound on $\mathcal{R}(g_k) - \mathcal{R}(g^*)$ in terms of the metric entropy of a small class $\mathcal{H}$. The main contribution of this Chapter is the statistical analysis of the greedy error minimization method utilizing the Contraction Principle. The idea of employing this method goes back to Koltchinskii and Panchenko [39, 41, 40].

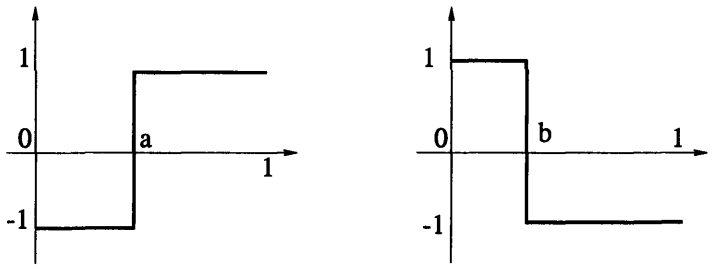



Figure 4-1: Step-up and step-down functions on the $[0,1]$ interval

**Example 3.** *Let $\mathcal{H}$ be the class of simple step-up and step-down functions on the $[0,1]$ interval, parametrized by $a$ and $b$, as shown in Figure 4-1. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is two. Let $\mathcal{F} = \text{conv}\,\mathcal{H}$. First, rescale the functions:*

$$f = \sum_{i=1}^{T} \lambda_i h_i = 2 \sum_{i=1}^{T} \lambda_i \left( \frac{h_i + 1}{2} \right) - 1 = 2f' - 1$$

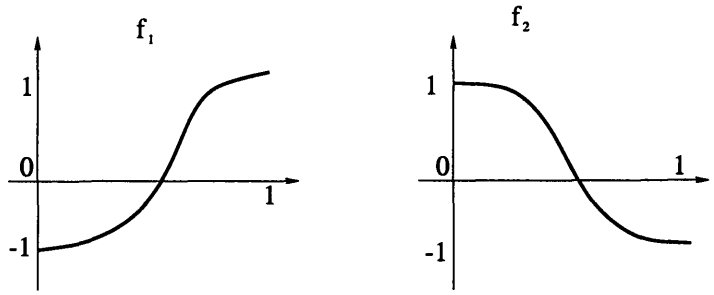
*where*

$$f' = \sum_{i=1}^{T} \lambda_i h_i'$$

*and*

$$h_i' = \frac{h_i + 1}{2}.$$

*We can generate any non-decreasing function $f'$ such that $f'(0) = 0$ and $f'(1) = 1$. Similarly, we can generate any non-increasing $f'$ such that $f'(0) = 1$ and $f'(1) = 0$. Rescaling back to $f$, we can get any non-increasing and non-decreasing functions of the form*



## 4.2 Density Estimation

In the density estimation setting, we are given i.i.d. sample $S = \{Z_1, \ldots, Z_n\}$ drawn from an unknown density $f$ (for convenience of notation, we will denote this unknown density by $f$ instead of $P$). The goal is to estimate $f$ from the given data. We set the loss function to be $\ell(h; Z) = -\log h(Z)$ for a density $h$. The Maximum Likelihood Estimation (MLE) principle is an empirical risk minimization procedure with the "$-\log$" loss. Indeed, minimizing $\frac{1}{n}\sum_{i=1}^{n} -\log h(Z_i)$ is equivalent to maximizing $\sum_{i=1}^{n} \log h(Z_i)$, which is equivalent to maximizing $\prod_{i=1}^{n} h(Z_i)$. Based on the method described in the previous Section, we will perform a greedy stage-wise density estimation procedure. This procedure has been used by Li and Barron [49, 50], where the authors obtained certain estimation and approximation bounds on its performance. By employing the Contraction Principle for Rademacher averages, we obtain tighter and somewhat more general results.

Rates of convergence for density estimation were studied in [11, 69, 70, 78]. For neural networks and projection pursuit, approximation and estimation bounds can be found in [3, 4, 35, 56].

Let $(\mathcal{Z}, \mathcal{G})$ be a measurable space and let $\lambda$ be a $\sigma$-finite measure on $\mathcal{G}$. Whenever we mention below that a probability measure on $\mathcal{G}$ has a density we will understand that it has a Radon-Nikodym derivative with respect to $\lambda$.

The choice of negative logarithm as the loss function leads to the Kullback-Leibler notion of distance. Kullback-Leibler (KL) divergence and Hellinger distance are the most commonly used distances for densities (although KL-divergence is not, strictly speaking, a distance). KL-divergence is defined for two distributions $f$ and $g$ as

$$D(f\|g) = \int f(Z) \log \frac{f(Z)}{g(Z)} d\lambda(Z) = \mathbb{E}_Z \log \frac{f(Z)}{g(Z)}.$$

Here $Z$ has distribution with density $f$.

Consider a parametric family of probability density functions

$$\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$$

over $\mathcal{Z}$. The class of $k$-component mixtures $g_k$ is defined as

$$\mathcal{C}_k = \text{conv}_k\ (\mathcal{H}) = \left\{ g : g(z) = \sum_{i=1}^{k} \alpha_i \phi_{\theta_i}(z), \sum_{i=1}^{k} \alpha_i = 1, \alpha_i \geq 0, \theta_i \in \Theta \right\}.$$

Let us define the class of continuous convex combinations

$$\mathcal{C} = \text{conv}\ (\mathcal{H}) = \left\{ g : g(z) = \int_{\Theta} \phi_\theta(z) P(d\theta),\ P \text{ is a probability measure on } \Theta \right\}.$$

The class $\mathcal{C}$ can be viewed as a closure of the union of all $\text{conv}_k\ (\mathcal{H})$.

Li and Barron prove that a $k$-mixture approximation to $f$ can be constructed by the following greedy procedure: Initialize $g_1 = \phi_\theta$ to minimize $D(f\|g_1)$, and at step $k$ construct $g_k$ from $g_{k-1}$ by finding $\alpha$ and $\theta$ such that

$$D(f\|g_k) \leq \min_{\alpha,\theta} D(f\|(1-\alpha)g_{k-1} + \alpha\phi_\theta).$$

Note that this method is equivalent to the Algorithm 1 with $\Upsilon(g) = -\mathbb{E}\log g$. Indeed,

$$\operatorname*{argmin}_g D(f\|g) = \operatorname*{argmin}_g \mathbb{E}\log\frac{f}{g} = \operatorname*{argmin}_g -\mathbb{E}\log g.$$

The approximation bound of Li and Barron [49, 50] states that for any $f$, there exists a $g_k \in \mathcal{C}_k$, such that

$$D(f\|g_k) \leq D(f\|\mathcal{C}) + \frac{c_{f,P}^2 \gamma}{k}, \tag{4.2}$$

where $c_{f,P}$ and $\gamma$ are constants and $D(f\|\mathcal{C}) = \inf_{g\in\mathcal{C}} D(f\|g)$. Furthermore, $\gamma$ is an upper bound on the log-ratio of any two functions $\phi_\theta(z), \phi_{\theta'}(z)$ for all $\theta, \theta', z$ and therefore

$$\sup_{\theta,\theta',z} \log\frac{\phi_\theta(z)}{\phi_{\theta'}(z)} < \infty \tag{4.3}$$

is a condition on the class $\mathcal{H}$.

A bound similar to the above result follows directly from Theorem 4.1.1 once the condition on the second derivative of $\Upsilon((1-\theta)g' + \theta g'')$ is verified (see [80]).

Of course, greedy minimization of $-\mathbb{E}\log g$ is not possible since $f$ is unknown. As motivated in Chapter 2, we aim to minimize the empirical counterpart. A connection between KL-divergence and Maximum Likelihood suggests the following method to compute the *estimate* $\hat{g}_k$ *from the data* by greedily choosing $\phi_\theta$ at step $k$ so that

$$\sum_{i=1}^{n} \log \hat{g}_k(Z_i) \geq \max_{\alpha,\theta} \sum_{i=1}^{n} \log[(1-\alpha)\hat{g}_{k-1}(Z_i) + \alpha\phi_\theta(Z_i)]. \tag{4.4}$$

This procedure corresponds to Algorithm 1 with $\Upsilon(g) = -\sum_{i=1}^{n} \log g(Z_i)$.

Li and Barron proved the following theorem:

**Theorem 4.2.1.** *Let $\hat{g}_k(x)$ be either the maximizer of the likelihood over $k$-component mixtures or, more generally, any sequence of density estimates satisfying (4.4). Assume additionally that $\Theta$ is a $d$-dimensional cube with side-length $A$, and that*

$$\sup_{z\in\mathcal{Z}} |\log \phi_\theta(z) - \log \phi_{\theta'}(z)| \leq B \sum_{j}^{d} |\theta_j - \theta'_j| \tag{4.5}$$

*for any $\theta, \theta' \in \Theta$. Then*

$$\mathbb{E}\,[D(f\|\hat{g}_k)] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2 k}{n}\log(nc_3), \tag{4.6}$$

*where $c_1, c_2, c_3$ are constants (dependent on $A, B, d$).*

The above bound combines the *approximation* and *estimation* results. Note that the first term decreases with the number of components $k$, while the second term increases. The rate of convergence for the optimal $k$ is therefore $O(\sqrt{\frac{\log n}{n}})$.

### 4.2.1 Main Results

We assume that the densities in $\mathcal{H}$ are bounded above and below by some constants $a$ and $b$, respectively. This boundedness naturally extends to the convex combinations as well. We prove the following results:

**Theorem 4.2.2.** *Suppose $a \leq \phi_\theta \leq b$ for all $\phi_\theta \in \mathcal{H}$. For $\hat{g}_k(z)$ being either the*

*maximizer of the likelihood over k-component mixtures or more generally any sequence of density estimates satisfying (4.4),*

$$\mathbb{E}\left[D(f\|\hat{g}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}\left[\frac{c_2}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_n)d\epsilon\right],$$

*where $c_1, c_2$ are constants (dependent on $a, b$) and $\mathcal{D}(\mathcal{H},\epsilon,d_n)$ is the $\epsilon$-covering number of $\mathcal{H}$ with respect to empirical distance $d_n$ ($d_n^2(\phi_1,\phi_2) = \frac{1}{n}\sum_{i=1}^n(\phi_1(Z_i)-\phi_2(Z_i))^2$).*

**Corollary 4.2.1.** *Under the conditions of Theorem 4.2.1 (i.e. $\mathcal{H}$ satisfying condition (4.5) and $\Theta$ being a cube with side-length $A$) and assuming boundedness of the densities, the bound of Theorem 4.2.2 becomes*

$$\mathbb{E}\left[D(f\|\hat{g}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}},$$

*where $c_1$ and $c_2$ are constants (dependent on $a, b, A, B, d$).*

**Corollary 4.2.2.** *The bound of Corollary 4.2.1 holds for the class of (truncated) Gaussian densities $\mathcal{H} = \{f_{\mu,\sigma} : g_{\mu,\sigma}(z) = \frac{1}{Z_{\mu,\sigma}}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right), |\mu| \leq M, \sigma_{min} \leq \sigma \leq \sigma_{max}\}$ over a compact domain $\mathcal{Z}$ ( $Z_{\mu,\sigma}$ is needed for normalization).*

**Remark 4.2.1.** *Theorem 4.2.2 hides the dependence of constants $c_1, c_2$ on $a$ and $b$ for the sake of easy comparison to Theorem 4.2.1. We now state the result with explicit dependence on $a$ and $b$:*

$$\begin{aligned} D(f\|\hat{g}_k) - D(f\|\mathcal{C}) \leq \frac{1}{\sqrt{n}}\left(\frac{1}{a}\mathbb{E}\left[c_1\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_n)d\epsilon\right] + \frac{8}{a}\right) \\ + \sqrt{\frac{t}{n}}\left(2\sqrt{2}\log\frac{b}{a}\right) + \frac{1}{k}\frac{8b^2}{a^2}\left(2+\log\frac{b}{a}\right) \end{aligned}$$

*with probability at least $1-e^{-t}$, or, by integrating,*

$$\mathbb{E}\left[D(f\|\hat{g}_k)\right] - D(f\|\mathcal{C}) \leq \frac{1}{\sqrt{n}}\left(\frac{1}{a}\mathbb{E}\left[c_1\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_n)d\epsilon\right] + \frac{8}{a} + 2\sqrt{2}\log\frac{b}{a}\right) + \frac{1}{k}\frac{8b^2}{a^2}\left(2+\log\frac{b}{a}\right),$$

*where $c_1$ is an absolute constant.*

**Remark 4.2.2.** *Upper and lower bounds $a$ and $b$ are determined by the class $\mathcal{H}$. Assume there exists a sequence of truncations $\{f_i\}$ of $f$, such that $a_i \leq f_i(z) \leq b_i$ for all $z \in \mathcal{Z}$, and $\{a_i\}$ is decreasing and $\{b_i\}$ increasing. Further assume that each class $\mathcal{H}_i$ contains functions bounded by $a_i$ and $b_i$. As the number of samples $n$ grows, one can choose more and more complex models $\mathcal{H}_i$. If $a_i$ is a decreasing function of $n$ and $b_i$ is an increasing function of $n$, Remark 4.2.1 provides the rate for learning $f_i$, the truncated version of $f$. This could be applied, for instance, to a sequence of classes $\mathcal{H}_i$ of Gaussian densities over an increasing domain and an increasing range of variances.*

### 4.2.2 Discussion of the Results

The result of Theorem 4.2.2 is two-fold. The first implication concerns dependence of the bound on $k$, the number of components. Our results show that there is an estimation bound of the order $O(\frac{1}{\sqrt{n}})$ that does not depend on $k$. Therefore, the number of components is not a trade-off that has to be made with the approximation part (which decreases with $k$). The bound also suggests that the number of components $k$ should be chosen to be $O(\sqrt{n})$.

The second implication concerns the rate of convergence in terms of $n$, the number of samples. The rate of convergence (in the sense of KL-divergence) of the estimated mixture to the true density is of the order $O(1/\sqrt{n})$. As Corollary 4.2.1 shows, for the specific class $\mathcal{H}$ considered by Li and Barron, the Dudley integral converges and does not depend on $n$. We therefore improve the results of Li and Barron by removing the $\log n$ factor. Furthermore, the result of this paper holds for general base classes $\mathcal{H}$

with a converging entropy integral, extending the result of Li and Barron. Note that the bound of Theorem 4.2.2 is in terms of the metric entropy of $\mathcal{H}$, as opposed to the metric entropy of $\mathcal{C}$. This is a strong result because the convex class $\mathcal{C}$ can be very large even for small $\mathcal{H}$ (see Example 3).

Rates of convergence for the MLE in mixture models were studied by Sara van de Geer [69]. As the author notes, the optimality of the rates depends primarily on the optimality of the entropy calculations. Unfortunately, in the results of [69], the entropy of the convex class appears in the bounds, which is undesirable. Moreover, only finite combinations are considered.

Wong and Shen [78] also considered density estimation, deriving rates of convergence in Hellinger distance for a class of bounded Lipschitz densities. In their work, a bound on the metric entropy of the whole class appears.

An advantage of the approach of [69] is the use of Hellinger distance to avoid problems near zero. Li and Barron address this problem by requiring (4.3), which is boundedness of the log of the ratio of two densities. Birgé and Massart ([11], page 122) cite a counterexample of Bahadur (1958) which shows that *even with a compact parameter space, M.L.E. can diverge when likelihood ratios are unbounded.* Unfortunately, boundedness of the ratios of densities is not enough for the proofs of this paper. We assume boundedness of the densities themselves. This is critical in one step of the proof, when the contraction principle is used (for the second time). Although the boundedness condition seems a somewhat strict requirement, note that a class of densities that satisfies (4.3), but not boundedness of the densities, has to contain functions which *all* go to zero (or infinity) in exactly the same manner. Also note that on a non-compact domain $\mathbb{R}$ even a simple class of Gaussian densities does not satisfy (4.3). Indeed, the log-ratio of the tails of two Gaussians with the same variance but different means becomes infinite. If one considers a compact domain $\mathcal{X}$, the boundedness of densities assumption does not seem very restrictive.

The proof technique of this paper seems to be a powerful general method for bounding uniform deviations of empirical and expected quantities. The main ingredients of the proof are the Comparison inequality for Rademacher processes and the

fact that Rademacher averages (as defined in Lemma 2.5.1) of the convex hull are equivalent to those of the base class.

### 4.2.3 Proofs

Assume

$$0 < a \leq \phi_\theta(z) \leq b \ \ \forall z \in \mathcal{Z}, \ \forall \phi_\theta \in \mathcal{H}.$$

Constants which depend only on $a$ and $b$ will be denoted by $c$ with various subscripts. The values of the constants might change from line to line.

**Theorem 4.2.3.** *For any fixed density $f$ and $S = (Z_1, \ldots, Z_n)$ drawn i.i.d from $f$, with probability at least $1 - e^{-t}$,*

$$\sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log g(Z_i) - \mathbb{E} \log g \right| \leq \mathbb{E} \left[ \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon \right] + c_2 \sqrt{\frac{t}{n}},$$

*where $c_1$ and $c_2$ are constants that depend on $a$ and $b$.*

*Proof.* First, we apply Lemma 2.4.3 to the random variable

$$T_n(Z_1, \ldots, Z_n) = \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log g(Z_i) - \mathbb{E} \log g \right|.$$

Let $t_i = \log g(z_i)$ and $t_i' = \log g(z_i')$. The bound on the martingale difference follows:

$$\begin{aligned}
&|T_n(z_1, \ldots, z_i', \ldots, z_n) - Z(Z_1, \ldots, Z_n)| \\
&= \left| \sup_{g \in \mathcal{C}} \left| \mathbb{E} \log g - \frac{1}{n} (t_1 + \ldots + t_i + \ldots + t_n) \right| \right. \\
&\left. - \sup_{g \in \mathcal{C}} \left| \mathbb{E} \log g - \frac{1}{n} (t_1 + \ldots + t_i' + \ldots + t_n) \right| \right| \\
&\leq \sup_{g \in \mathcal{C}} \frac{1}{n} |\log g(z_i') - \log g(z_i)| \leq \frac{1}{n} \log \frac{b}{a} = c_i.
\end{aligned}$$

The above chain of inequalities holds because of the triangle inequality and the prop-

erties of sup. Applying McDiarmid's inequality (see Lemma 2.4.3),

$$\mathbb{P}\left(T_n - \mathbb{E}T_n > u\right) \leq \exp\left(-\frac{u^2}{2\sum c_i^2}\right) = \exp\left(-\frac{nu^2}{2\left(\log\frac{b}{a}\right)^2}\right).$$

Therefore,

$$\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| \leq \mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| + \sqrt{2}\log\frac{b}{a}\sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$ and by Lemma 2.5.1,

$$\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| \leq 2\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\log g(Z_i)\right|.$$

Combining,

$$\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| \leq 2\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\log g(Z_i)\right| + \sqrt{2}\log\frac{b}{a}\sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$.

Therefore, instead of bounding the difference between the "empirical" and the "expectation", it is enough to bound the above expectation of the Rademacher average. This is a simpler task, but first we have to deal with the logarithm in the Rademacher sum. To eliminate this difficulty, we apply Lemma 2.5.2. Once we reduce our problem to bounding the Rademacher sum of the basis functions $\sup_{\phi\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(Z_i)\right|$, we will be able to use the entropy of class $\mathcal{H}$.

Let $p_i = g(x_i) - 1$ and note that $a - 1 \leq p_i \leq b - 1$. Consider $\phi(p) = \log(1+p)$. The largest derivative of $\log(1+p)$ on the interval $p \in [a-1, b-1]$ is at $p = a - 1$ and is equal to $1/a$. So, $a\log(p+1)$ is 1-Lipschitz. Also, $\phi(0) = 0$. By Lemma 2.5.2

applied to $\phi(p)$,

$$\begin{aligned}
2\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \log g(Z_i)\right| &= 2\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{1}^{n}\varepsilon_i\phi(p_i)\right| \\
&\leq 4\frac{1}{a}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i) - \frac{1}{n}\sum_{1}^{n}\epsilon_i\right| \\
&\leq 4\frac{1}{a}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right| + 4\frac{1}{a}\mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\leq 4\frac{1}{a}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right| + \frac{4}{a}\frac{1}{\sqrt{n}}.
\end{aligned}$$

The last inequality holds because

$$\mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \leq \left(\mathbb{E}_\epsilon\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right)^{1/2} = \frac{1}{\sqrt{n}}.$$

Combining the inequalities, with probability at least $1 - e^{-t}$

$$\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| \leq \frac{4}{a}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right| + \sqrt{2}\log\frac{b}{a}\sqrt{\frac{t}{n}} + \frac{4}{a}\frac{1}{\sqrt{n}}.$$

The power of using Rademacher averages to estimate complexity comes from the fact that the Rademacher averages of a class are equal to those of the convex hull. Indeed, consider

$$\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right|$$

with $g(z) = \int_\theta \phi_\theta(z)P(d\theta)$. Since a linear functional over convex combinations achieves its maximum value at the vertices, the above supremum is equal to

$$\sup_{\theta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi_\theta(Z_i)\right|,$$

the corresponding supremum on the basis functions $\phi \in \mathcal{H}$. Therefore,

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(Z_i) \right| = \mathbb{E}_\epsilon \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(Z_i) \right|.$$

Next, we use the following classical result (see [71]),

$$\mathbb{E}_\epsilon \sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(Z_i) \right| \leq \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon,$$

where $d_n$ is the empirical distance with respect to the set $Z_1, \ldots, Z_n$.

Combining the results, the following holds with probability at least $1 - e^{-t}$:

$$\sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log g(Z_i) - \mathbb{E} \log g \right| \leq \left[ \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon \right] + c_2 \sqrt{\frac{t}{n}}.$$

□

**Remark 4.2.3.** *If $\mathcal{H}$ is a VC-subgraph with VC dimension $V$, the Dudley integral above is bounded by $c\sqrt{V}$ and we obtain $O(1/\sqrt{n})$ convergence. One example of such a class is the class of (truncated) Gaussian densities over a compact domain and with bounded variance (see Corollary 4.2.2). Another example is the class considered in [49], and its cover is computed in the proof of Corollary 4.2.1. More information on the classes with converging Dudley integral and examples of VC-subgraph classes can be found in [25, 71].*

We are now ready to prove Theorem 4.2.2:

*Proof.*

$$
\begin{aligned}
D(f\|\hat{g}_k) - D(f\|g_k) &= \mathbb{E}\log g_k - \mathbb{E}\log \hat{g}_k \\
&= \left(\mathbb{E}\log g_k - \frac{1}{n}\sum_{i=1}^{n}\log g_k(Z_i)\right) \\
&+ \left(\frac{1}{n}\sum_{i=1}^{n}\log g_k(Z_i) - \frac{1}{n}\sum_{i=1}^{n}\log \hat{g}_k(Z_i)\right) \\
&+ \left(\frac{1}{n}\sum_{i=1}^{n}\log \hat{g}_k(Z_i) - \mathbb{E}\log \hat{g}_k\right) \\
&\leq 2\sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i) - \mathbb{E}\log g\right| \\
&+ \left(\frac{1}{n}\sum_{i=1}^{n}\log g_k(Z_i) - \frac{1}{n}\sum_{i=1}^{n}\log \hat{g}_k(Z_i)\right) \\
&\leq \mathbb{E}\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_n)d\epsilon\right] + c_2\sqrt{\frac{t}{n}} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{g_k(Z_i)}{\hat{g}_k(Z_i)}
\end{aligned}
$$

with probability at least $1-e^{-t}$ (by Theorem 4.2.3). Note that $\frac{1}{n}\sum_{i=1}^{n}\log\frac{g_k(Z_i)}{\hat{g}_k(Z_i)} \leq 0$ if $\hat{f}_k$ is constructed by maximizing the likelihood over $k$-component mixtures. If it is constructed by the Algorithm 1, Theorem 4.1.1 shows that $\hat{g}_k$ achieves "almost maximum likelihood". Another proof of this result is given on page 27 of [50], or section 3 of [49]:

$$
\forall g \in \mathcal{C},\quad \frac{1}{n}\sum_{i=1}^{n}\log(\hat{g}_k(Z_i))(n_i) \geq \frac{1}{n}\sum_{i=1}^{n}\log(g(Z_i))(n_i) - \gamma\frac{c_{F_n,P}^2}{k}.
$$

Here $c_{F_n,P}^2 = (1/n)\sum_{i=1}^{n}\frac{\int\phi_\theta^2(z_i)P(d\theta)}{\left(\int\phi_\theta(z_i)P(d\theta)\right)^2} \leq \frac{b^2}{a^2}$ and $\gamma = 4\log(3\sqrt{e}) + 4\log\frac{b}{a}$. Hence, with probability at least $1-e^{-t}$,

$$
D(f\|\hat{g}_k) - D(f\|g_k) \leq \mathbb{E}\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_n)d\epsilon\right] + c_2\sqrt{\frac{t}{n}} + \frac{c_3}{k}.
$$

We now write the overall error of estimating an unknown density $f$ as the sum of approximation and estimation errors. The former is bounded by (4.2) and the latter is bounded as above. Note again that $c_{f,P}^2$ and $\gamma$ in the approximation bound

(4.2) are bounded above by constants which depend only on $a$ and $b$. Therefore, with probability at least $1 - e^{-t}$,

$$\begin{aligned} D(f\|\hat{g}_k) - D(f\|\mathcal{C}) &= (D(f\|g_k) - D(f\|\mathcal{C})) + (D(f\|\hat{g}_k) - D(f\|g_k)) \\ &\leq \frac{c}{k} + \left[\frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon\right] + c_2\sqrt{\frac{t}{n}}. \end{aligned} \tag{4.7}$$

**Remark 4.2.4.** *Note that we could have obtained the above bound from the Equation 4.1 without the decomposition into the approximation and estimation errors.*

Finally, we rewrite the above probabilistic statement as a statement in terms of expectations. Let $\zeta = \frac{c}{k} + \mathbb{E}\left[\frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon\right]$ and $\xi = D(f\|\hat{g}_k) - D(f\|\mathcal{C})$. We have shown that $\mathbb{P}\left(\xi \geq \zeta + c_2\sqrt{\frac{t}{n}}\right) \leq e^{-t}$. Since $\xi \geq 0$,

$$\mathbb{E}[\xi] = \int_0^\zeta \mathbb{P}(\xi > u)\, du + \int_\zeta^\infty \mathbb{P}(\xi > u)\, du \leq \zeta + \int_0^\infty \mathbb{P}(\xi > u + \zeta)\, du.$$

Now set $u = c_2\sqrt{\frac{t}{n}}$. Then $t = c_3 n u^2$ and $E[\xi] \leq \zeta + \int_0^\infty e^{-c_3 n u^2} du \leq \zeta + \frac{c}{\sqrt{n}}$. Hence,

$$E[D(f\|\hat{g}_k)] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}\left[\frac{c_2}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon\right].$$

□

**Remark 4.2.5.** *Inequality (4.7) is much stronger than the result of Theorem 4.2.2 because it reveals the tail behavior of $D(f\|\hat{g}_k) - D(f\|\mathcal{C})$. Nevertheless, to be able to compare our results to those of Li and Barron, we present our results in terms of expectations.*

**Remark 4.2.6.** *In the actual proof of the bounds, Li and Barron [50, 49] use a specific sequence of $\alpha_i$ for the finite combinations. The authors take $\alpha_1 = 1$, $\alpha_2 = \frac{1}{2}$,*

*and* $\alpha_k = \frac{2}{k}$ *for* $k \geq 2$. *It can be shown that in this case*

$$g_k = \frac{2}{k(k-1)}\left(\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2 + \sum_{m=3}^{k}(m-1)\phi_m\right),$$

*so the later choices have more weight.*

We now prove Corollary 4.2.1:

*Proof.* Since we consider bounded densities $a \leq \phi_\theta \leq b$, condition (4.5) implies that

$$\forall z,\ \log\left(\frac{\phi_\theta(z) - \phi_{\theta'}(z)}{b} + 1\right) \leq B|\theta - \theta'|_{L_1}.$$

This allows us to bound $L_\infty$ distances between functions in $\mathcal{H}$ in terms of the $L_1$ distances between the corresponding parameters. Since $\Theta$ is a $d$-dimensional cube of side-length $A$, we can cover $\Theta$ by $\left(\frac{A}{\delta}\right)^d$ balls of $L_1$-radius $d\frac{\delta}{2}$. This cover induces a cover of $\mathcal{H}$. For any $\phi_\theta$ there exists an element of the cover $\phi_{\theta'}$, so that

$$d_n(\phi_\theta, \phi_{\theta'}) \leq |\phi_\theta - \phi_{\theta'}|_\infty \leq be^{B\frac{d\delta}{2}} - b = \epsilon.$$

Therefore, $\delta = \frac{2\log\left(\frac{\epsilon}{b}+1\right)}{Bd}$ and the cardinality of the cover is $(\frac{A}{\delta})^d = \left(\frac{ABd}{2\log\left(\frac{\epsilon}{b}+1\right)}\right)^d$. Hence,

$$\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H}, \epsilon, d_n)d\epsilon = \int_0^b \sqrt{d\log\frac{ABd}{2\log\left(\frac{\epsilon}{b}+1\right)}}d\epsilon.$$

A straightforward calculation shows that the integral above converges. □

By creating a simple net over the class $\mathcal{H}$ in Corollary 4.2.2, one can easily show that $\mathcal{H}$ has a finite cover $\mathcal{D}(\mathcal{H}, \epsilon, d_n) = \frac{K}{\epsilon^2}$, for some constant $K$. Corollary 4.2.2 follows.

## 4.3 Classification

The analysis of the greedy algorithm (Algorithm 1) in the classification setting follows along the same lines as that of density estimation. Once the loss function in

Equation 4.1 is specified, the Concentration and Symmetrization steps can be performed similarly to the proofs of the previous Section. Similar proofs have been done independently by Mannor et al [53].

Suppose $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} = \{-1, 1\}$. Let $\ell(g; Z) = \ell(yg(x))$ such that

$$\ell(yg(x)) \geq I(yg(x) \leq 0)$$

and $\ell(\cdot)$ is a convex function with a Lipschitz constant $L$. For natural convex loss functions, the Lipschitz constant is finite if the functions $g$ are bounded.

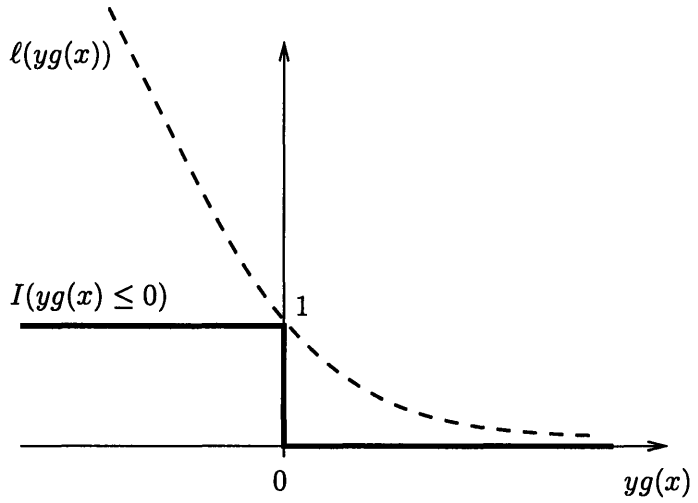



Figure 4-2: Convex loss $\ell$ upper-bounds the indicator loss.

By performing the greedy minimization procedure with $\Upsilon(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i g(X_i))$ over the convex hull conv $(\mathcal{H})$, we obtain $g_k$ such that, according to the Equation 4.1,

$$\mathcal{R}(g_k) - \mathcal{R}(g^*) \leq 2 \sup_{g \in \text{conv}\,(\mathcal{H})} |\mathcal{R}(g) - \mathcal{R}_{\text{emp}}(g)| + \frac{2c}{k+2}.$$

Zhang [80] provides the specific constants $c$ for common loss functions $\ell$.

Using the Symmetrization and Concentration steps as well as employing the comparison inequality for Rademacher processes, we obtain a result similar to Theorem 4.2.3:

**Theorem 4.3.1.** *Suppose* $0 < \ell(g;\cdot) < M$ *for any* $g \in conv\,(\mathcal{H})$. *Then with probability at least* $1 - e^{-t}$,

$$\sup_{g \in conv\,(\mathcal{H})} |\mathcal{R}(g) - \mathcal{R}_{emp}(g)| \leq \mathbb{E}\left[\frac{c_1}{\sqrt{n}} \int_0^M \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_n) d\epsilon\right] + c_2 \sqrt{\frac{t}{n}},$$

*where* $c_1$ *and* $c_2$ *are constants which depend on* $M$ *and* $L$.

The upper bound on the convergence of $\mathcal{R}(g_k)$ to $\mathcal{R}(g^*)$ as ($k \to \infty$ and $n \to \infty$) follows immediately.

# Chapter 5

# Stability of Empirical Risk Minimization over Donsker Classes

The results of this Chapter appear partially in [18, 19].

## 5.1 Introduction

Empirical risk minimization (ERM) algorithm has been studied in learning theory to a great extent. Vapnik and Chervonenkis [74, 76] showed necessary and sufficient conditions for its consistency. In recent developments, [6, 8, 7, 38] proved sharp bounds on the performance of ERM. Tools from empirical process theory have been successfully applied, and, in particular, it has been shown that the *localized Rademacher averages* play an important role in studying the behavior of the ERM algorithm.

In this Chapter we are not directly concerned with rates of performance of ERM. Rather, we prove some properties of ERM algorithms, which, to our knowledge, do not appear in the literature. The analysis of this Chapter has been motivated by the study of *algorithmic stability*: the behavior of a learning algorithm with respect to perturbations of the training set. Algorithmic stability has been studied in the recent years as an alternative to the classical (complexity-oriented) approach to deriving generalization bounds [16, 45, 55, 58, 60]. Motivation for studying algorithmic stability comes, in part, from the work of [22]. Their results indicate that for any al-

gorithm, the performance of the leave-one-out estimator of expected error is bounded by $L_1$-stability of the algorithm, i.e. by the average $L_1$ distance between hypotheses on similar samples. This result can be used to derive bounds on the performance of the leave-one-out estimate for algorithms such as $k$-Nearest Neighbors. It is important to note that no class of finite complexity is searched by algorithms like $k$-NN, and so the classical approach of using complexity of the hypothesis space fails.

Further important results were proved by Bousquet and Elisseeff [16], where a large family of algorithms (*Tikhonov regularization* based methods) has been shown to possess a strong $L_\infty$ stability with respect to changes of single samples of the training set, and exponential bounds have been proved for the generalization error in terms of empirical error. Tikhonov regularization based algorithms minimize the empirical error plus a stabilizer, and are closely related to ERM. Though ERM is not, in general, $L_\infty$-stable, it *is* $L_1$-stable over certain classes of functions, as one of the results below shows. To the best of our knowledge, the outcomes of the present Chapter do not follow directly from results available in the machine learning literature. In fact we had to turn to the empirical process theory (see Section 2.5) for the mathematical tools necessary for studying stability of ERM.

Various assumptions on the function class, over which ERM is performed, have been considered recently to obtain fast rates on the performance of ERM. The importance of having a unique best function in the class has been shown by [48]: the difficult learning problems seem to be the ones where two minimizers of the expected error exist and are far apart. Although we do not address the question of performance rates here, our results does shed some light on the behavior of ERM when two (or more) minimizers of expected error exist. Our results imply that, under a certain weak condition on the class, as the expected performance of empirical minimizers approaches the best in the class, a jump to a different part of the function class becomes less and less likely.

Some algorithmic implications of our results are straight-forward. For example, in the context of on-line learning, when a point is added to the training set, with high probability one has to search for empirical minimizers in a small $L_1$-ball around the

current hypothesis, which can be a tractable problem. Moreover, it seems plausible that $L_1$-stability can have consequences for computational complexity of ERM. While it has been shown that ERM is NP-hard even for simple function classes (see e.g. [9]), our results could allow more optimistic average-case analysis.

Since ERM minimizes empirical error instead of expected error, it is reasonable to require that the two quantities become close uniformly over the class, as the number of examples grows. Hence, ERM is a sound strategy only if the function class is uniform Glivenko-Cantelli, that is, it satisfies the uniform law of large numbers. In this Chapter we focus our attention on more restricted family of function classes: Donsker classes (see Section 2.5). These are classes satisfying not only the law of large numbers, but also a version of the central limit theorem. Though a more restricted family of classes, Donsker classes are still quite general. In particular, uniform Donsker and uniform Glivenko-Cantelli properties are equivalent in the case of binary-valued functions (and also equivalent to finiteness of VC dimension). The central limit theorem for Donsker classes states a form of convergence of the empirical process to a Gaussian process with a specific covariance structure (e.g. [25, 71]). This structure is used in the proof of the main result of this Chapter to control the correlation of the empirical errors of ERM minimizers on similar samples.

This Chapter is organized as follows. In Section 5.2 we introduce the notation and background results. Section 5.3 presents the main result, which is proved in the appendix using tools from empirical process theory. In Section 5.4, we show $L_1$-stability of ERM over Donsker classes as an application of the main result of Section 5.3. In Section 5.5 we show an improvement (in terms of the rates) of the main result under a suitable Komlos-Major-Tusnady condition and an assumption on entropy growth. Section 5.6 combines the results of Sections 5.4 and 5.5 and uses a uniform ratio limit theorem to obtain fast rates of decay on the deviations of expected errors of almost-ERM solutions, thus establishing *strong expected error stability* of ERM (see Chapter 3). Several further applications of the results are considered in Section 5.7. Most of the proofs are postponed to the Appendix. Section 5.8 is a final summary of the results.

## 5.2 Notation

Let $(\mathcal{Z}, \mathcal{G})$ be a measurable space. Let $P$ be a probability measure on $(\mathcal{Z}, \mathcal{G})$ and $Z_1, \ldots, Z_n$ be independent copies of $Z$ with distribution $P$. Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. In the setting of learning theory, samples $Z$ are input-output pairs $(X, Y)$ and for $f \in \mathcal{F}$, $f(Z)$ measures how well the relationship between $X$ and $Y$ is captured by $f$. Hence, $\mathcal{F}$ is usually the loss class of some other function class $\mathcal{H}$, i.e. $\mathcal{F} = \mathcal{L}(\mathcal{H})$ (see Section 2.1). The goal is to minimize $Pf = \mathbb{E}f(Z)$ where information about the unknown $P$ is given only through the finite sample $S = (Z_1, \ldots, Z_n)$. Define the empirical measure as $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$.

**Definition 5.2.1.** *Given a sample $S$,*

$$f_S := \operatorname*{argmin}_{f \in \mathcal{F}} P_n f = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i)$$

*is a minimizer of the empirical risk (empirical error), if the minimum exists.*

Since an exact minimizer of the empirical risk might not exist, as well as for algorithmic reasons, we consider the set of almost-minimizers of empirical risk.

**Definition 5.2.2.** *Given $\xi \geq 0$ and $S$, define the set of almost empirical minimizers*

$$\mathcal{M}_S^{\xi} = \{f \in \mathcal{F} : P_n f - \inf_{g \in \mathcal{F}} P_n g \leq \xi\}$$

*and define its diameter as*

$$\operatorname{diam}\mathcal{M}_S^{\xi} = \sup_{f,g \in \mathcal{M}_S^{\xi}} \|f - g\|.$$

Note that $\mathcal{M}_S^{\xi} \subset \mathcal{M}_S^{\xi'}$ whenever $\xi \leq \xi'$. Moreover, if $f, g \in \mathcal{M}_S^{\xi}$ and $h = (1-\lambda)f + \lambda g \in \mathcal{F}$, then $h \in \mathcal{M}_S^{\xi}$ by linearity of the average. Hence, if $\mathcal{F}$ is convex, so is $\mathcal{M}_S^{\xi}$ for any $\xi > 0$.

The $\|\cdot\|$ in the above definition is the seminorm on $\mathcal{F}$ induced by symmetric

bilinear product

$$\langle f, f' \rangle = P(f - Pf)(f' - Pf').$$

This is a natural measure of distance between functions, as will become apparent later, because of the central role of the covariance structure of Brownian bridges in our proofs. The results obtained for the seminorm $\|\cdot\|$ will be easily extended to the $L_2(P)$ norm, thanks to the close relation of these two notions of distance.

## 5.3 Main Result

We now state the main result of this Chapter.

**Theorem 5.3.1.** *Let $\mathcal{F}$ be a P-Donsker class. For any sequence $\xi(n) = o(n^{-1/2})$,*

$$\text{diam}\mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0.$$

The outer probability $P^*$ above is due to measurability issues. Definitions and results on various types of convergence, as well as ways to deal with measurability issues arising in the proofs, are based on the rigorous book of [71].

Before turning to the proof of Theorem 5.3.1, let us discuss the geometry of the function class $\mathcal{F}$ and its relation to the results. Recall that $P$-Donsker class $\mathcal{F}$ is also a $P$-Glivenko-Cantelli class, and so empirical risk minimization on $\mathcal{F}$ is consistent: as $n \to \infty$,

$$Pf_S \xrightarrow{P^*} \inf_{g \in \mathcal{F}} Pg.$$

For simplicity, assume for a second that functions $f \in \mathcal{F}$ are the square loss functions over some class $\mathcal{H}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$:

$$\mathcal{F} = \{f(z) = \ell(h, z) = \{(h(x) - y)^2 : h \in \mathcal{H}\} = \mathcal{L}(\mathcal{H}).$$

Figures 5-1 through 5-4 depict four important possibilities regarding the geometry of the function class.

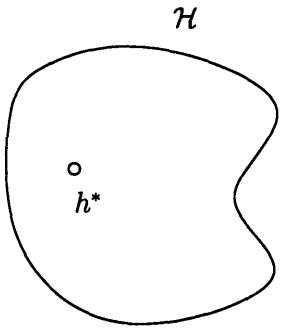



Figure 5-1: Realizable setting.

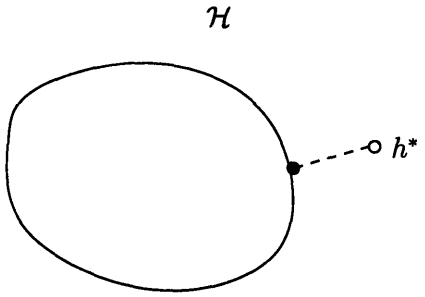



Figure 5-2: Single minimum of expected error.

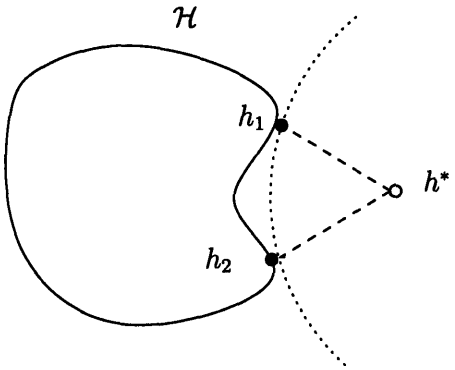



Figure 5-3: Finite number of minimizers of expected error.

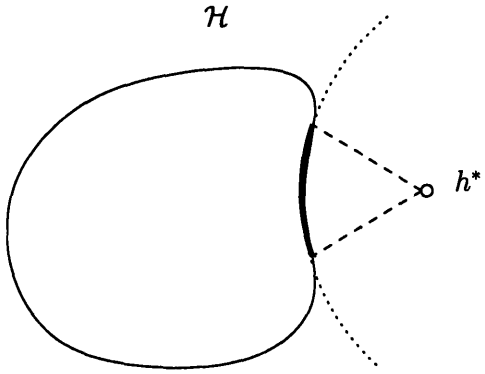



Figure 5-4: Infinitely many minimizers of expected error.

The simplest situation, that of *realizable setting*, occurs when the target function $h^*$ (the function achieving the zero expected error) belongs to the class $\mathcal{H}$. In this case, one can show that the empirical minimizers converge to this function in $L_2(P)$ distance and, by the triangle inequality, the distance between two empirical minimizers over different sets converges to zero in probability. In fact, one can upper-bound the rate of this decrease. Similar behavior can be shown for non-realizable settings when there is a single (Figure 5-2) best function in the given class. For a finite number (Figure 5-3) of best functions in the given class, one can show, based on the binomial result, that $\Omega(\sqrt{n})$ changes is enough to induce a large jump of the empirical minimizer. Nevertheless, in this case empirical minimization is still stable with respect to $o(\sqrt{n})$ changes. Hence, $n^{-1/2}$ is the rate defining the transition between stability and instability in the case of the finite number of minimizers of the expected error. Once

the number of minimizers is infinite (Figure 5-4), the problem of showing closeness of two empirical minimizers is difficult, and this is the situation addressed by this Chapter. Of course, we do not expect the distance between empirical minimizers over completely different sets to decay to zero. For example, if one is searching for the least (or most) dense region in an interval, where the data is drawn from the uniform distribution, one expects the least (most) dense regions to be different for completely different samples (see Section 5.7). Nevertheless, from the results of this Chapter it follows that the $n^{-1/2}$ rate again defines the transition between stability and instability. We stress that the situation of infinitely-many minimizers of expected error is not artificial since the measure $P$ defining these "distances" is unknown. Furthermore, we prove results for very general classes of functions, not necessarily obtained by composing the square loss with a function class.

The proof of Theorem 5.3.1 relies on the *almost sure representation theorem* [71, Thm. 1.10.4]. Here we state the theorem applied to $\nu_n$ and $\nu$.

**Proposition 5.3.1.** *Suppose $\mathcal{F}$ is $P$-Donsker. Let $\nu_n : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process. There exist a probability space $(\mathcal{Z}', \mathcal{G}', P')$ and maps $\nu'_n : \mathcal{Z}' \mapsto \ell^\infty(\mathcal{F})$ such that*

1. $\nu'_n \xrightarrow{au} \nu'$
2. $\mathbb{E}^* f(\nu'_n) = \mathbb{E}^* f(\nu_n)$ *for every bounded* $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ *for all* $n$.

Lemma 5.3.1 is the main preliminary result used in the proof of Theorem 5.3.1 (and Theorem 5.5.1 in Section 5.5). We postpone its proof to Appendix A.

**Lemma 5.3.1.** *Let $\nu_n : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process. Fix $n$ and assume that there exist a probability space $(\mathcal{Z}', \mathcal{G}', P')$ and a map $\nu'_n : \mathcal{Z}' \mapsto \ell^\infty(\mathcal{F})$ such that $\mathbb{E}^* f(\nu'_n) = \mathbb{E}^* f(\nu_n)$ for every bounded $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$. Let $\nu'$ be a $P$-Brownian bridge defined on $(\mathcal{Z}', \mathcal{G}', P')$. Fix $C > 0$, $\epsilon = \min(C^3/128, C/4)$ and suppose $\delta \geq \xi\sqrt{n}$ for a given $\xi > 0$. Then, if $\mathcal{F}$ is $P$-Donsker, the following inequality holds*

$$\Pr^* \left( \mathrm{diam} \mathcal{M}_S^\xi > C \right) \leq \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)^2 \left( \frac{128\delta}{C^3} + \Pr^* \left( \sup_{\mathcal{F}} |\nu'_n - \nu'| \geq \delta/2 \right) \right)$$

We are now ready to prove the main result of this section.

*Theorem 5.3.1.* Lemma 1.9.3 in [71] shows that when the limiting process is Borel measurable, almost uniform convergence implies convergence in outer probability. Therefore, the first implication of Proposition 5.3.1 states that for any $\delta > 0$

$$\Pr^* \left( \sup_{\mathcal{F}} |\nu'_n - \nu'| > \delta \right) \to 0.$$

By Lemma 5.3.1,

$$\Pr^* \left( \mathrm{diam} \mathcal{M}_S^{\xi(n)} > C \right) \leq \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)^2 \left( \frac{128\delta}{C^3} + \Pr^* \left( \sup_{\mathcal{F}} |\nu'_n - \nu'| \geq \delta/2 \right) \right)$$

for any $C > 0$, $\epsilon = \min(C^3/128, C/4)$, and any $\delta \geq \xi(n)\sqrt{n}$. Since $\xi(n) = o(n^{-1/2})$, $\delta$ can be chosen arbitrarily small, and so $\Pr^* \left( \mathrm{diam} \mathcal{M}_S^{\xi(n)} > C \right) \to 0$. □

The following corollary, whose proof is given in Appendix A, extends the above result to $L_2$ (and thus $L_1$) diameters.

**Corollary 5.3.1.** *The result of Theorem 5.3.1 holds if the diameter is defined with respect to the $L_2(P)$ norm.*

## 5.4 Stability of almost-ERM

The main result of this section, Corollary 5.4.1, shows *$L_2$-stability* of almost-ERM on Donsker classes. It implies that, in probability, the $L_2$ (and thus $L_1$) distance between almost-minimizers on similar training sets (with $o(\sqrt{n})$ changes) goes to zero when $n$ tends to infinity.

This result provides a partial answer to the questions raised in the machine learning literature by [45, 55]: is it true that when one point is added to the training set, the ERM algorithm is less and less likely to jump to a far (in the $L_1$ sense) hypothesis? In fact, since binary-valued function classes are uniform Donsker if and only if the VC dimension is finite, Corollary 5.4.1 proves that almost-ERM over binary VC classes possesses $L_1$-stability. For the real-valued classes, uniform Glivenko-Cantelli

property is weaker than uniform Donsker property, and therefore it remains unclear if almost-ERM over uGC but not uniform Donsker classes is stable in the $L_1$ sense.

The use of $L_1$-stability goes back to [22], who showed that this stability is sufficient to bound the difference between the leave-one-out error and the expected error of a learning algorithm. In particular, Devroye and Wagner show that nearest-neighbor rules possess $L_1$-stability [see also 20]. Our Corollary 5.4.1 implies $L_1$-stability of ERM (or almost-ERM) algorithms on Donsker classes.

In the following $[n]$ denotes the set $\{1, 2, \ldots, n\}$ and $A \triangle B$ is the symmetric difference of sets $A$ and $B$.

**Corollary 5.4.1.** *Assume $\mathcal{F}$ is $P$-Donsker and uniformly bounded with envelope $F \equiv 1$. For $I \subset \mathbb{N}$, define $S(I) = (Z_i)_{i \in I}$. Let $I_n \subset \mathbb{N}$ such that $M_n := |I_n \triangle [n]| = o(n^{1/2})$. Suppose $f_n \in \mathcal{M}^{\xi(n)}_{S([n])}$ and $f'_n \in \mathcal{M}^{\xi'(n)}_{S(I_n)}$ for some $\xi(n) = o(n^{-1/2})$ and $\xi'(n) = o(n^{-1/2})$. Then*

$$\|f_n - f'_n\| \xrightarrow{P^*} 0.$$

*The norm $\|\cdot\|$ can be replaced by $L_2(P)$ or $L_1(P)$ norm.*

*Proof.* It is enough to show that $f'_n \in \mathcal{M}^{\xi''(n)}_{S([n])}$ for some $\xi''(n) = o(n^{-1/2})$ and result follows from the Theorem 5.3.1.

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]} f'_n(Z_i) &\le \frac{M_n}{n} + \frac{1}{n}\sum_{i\in I_n} f'_n(Z_i) \\
&\le \frac{M_n}{n} + \frac{|I_n|}{n}\left(\xi'(n) + \inf_{g\in\mathcal{F}} \frac{1}{|I_n|}\sum_{i\in I_n} g(Z_i)\right) \\
&\le \frac{M_n}{n} + \frac{|I_n|}{n}\xi'(n) + \frac{1}{n}\sum_{i\in I_n} f_n(Z_i) \\
&\le 2\frac{M_n}{n} + \frac{|I_n|}{n}\xi'(n) + \frac{1}{n}\sum_{i\in[n]} f_n(Z_i) \\
&\le 2\frac{M_n}{n} + \frac{|I_n|}{n}\xi'(n) + \xi(n) + \inf_{g\in\mathcal{F}} \frac{1}{n}\sum_{i\in[n]} g(Z_i)
\end{aligned}$$

Define

$$\xi''(n) := 2\frac{M_n}{n} + \frac{|I_n|}{n}\xi'(n) + \xi(n).$$

Because $M_n = o(n^{\frac{1}{2}})$, it follows that $\xi''(n) = o(n^{-1/2})$. Corollary 5.3.1 implies convergence in $L_2(P)$, and, therefore, in $L_1(P)$ norm. □

Let us now generalize the above result to functions with a bounded Lipschitz constant. The Lipschitz assumption allows us to study sensitivity of functions with respect to perturbations of the data points in the space instead of complete removals or additions.

More precisely, assume that the space $\mathcal{Z}$ is equipped with a distance metric $d$. Furthermore, suppose $\mathcal{Z}$ is compact: say, $\mathcal{Z} \subset B_d(0, R)$. Suppose that $\mathcal{F}$ consists of functions with a bounded Lipschitz constant $L$:

$$\forall z, z' \in \mathcal{Z},\ \forall f \in \mathcal{F},\ |f(z) - f(z')| \leq Ld(z, z').$$

Define the "distance" between two sets $S, T$ as follows. For sets $S = \{z_1, \ldots, z_n\}$, $T = \{z'_1, \ldots, z'_n\}$ of equal size

$$d(S, T) = \inf_{\pi} \frac{1}{n} \sum_{i=1}^{|S|} d(z_i, z'_{\pi(i)}).$$

If $|S| < |T|$,

$$d(S, T) = \inf_{\pi} \frac{1}{n} \sum_{i=1}^{|S|} d(z_i, z'_{\pi(i)}) + 2R(|T| - |S|)$$

and if $|S| > |T|$,

$$d(S, T) = \inf_{\pi} \frac{1}{n} \sum_{i=1}^{|T|} d(z_{\pi(i)}, z') + 2R(|S| - |T|).$$

In other words, the "distance" between two sets is defined as the best way to pair up points from one set with points from the other set, and paying the constant (diameter of the ball) for each unmatched point.

**Corollary 5.4.2.** *Assume $\mathcal{F}$ is $P$-Donsker and uniformly bounded with envelope $F \equiv 1$. Suppose $\mathcal{F}$ consists of functions with a bounded Lipschitz constant $L$ and $\mathcal{Z} \subset B_d(0, R)$. Suppose $f_S \in \mathcal{M}_S^{\xi(n)}$ and $f_T \in \mathcal{M}_T^{\xi'(n)}$ for some $\xi(n) = o(n^{-1/2})$ and*

$\xi'(n) = o(n^{-1/2})$. *If* $d(S,T) = o(n^{-1/2})$, *then*

$$\|f_S - f_T\| \xrightarrow{P^*} 0.$$

*The norm* $\|\cdot\|$ *can be replaced by* $L_2(P)$ *or* $L_1(P)$ *norm.*

*Proof.* Without loss of generality suppose that $n = |S| < |T|$. Similarly to the proof of Corollary 5.4.1,

$$\begin{aligned}\frac{1}{n}\sum_{z\in T} f_S(z) - \frac{1}{n}\sum_{z\in T} f_T(z) = &\frac{1}{n}\sum_{z\in T} f_S(z) - \frac{1}{n}\sum_{z\in S} f_S(z) \\ &+ \frac{1}{n}\sum_{z\in S} f_S(z) - \frac{1}{n}\sum_{z\in S} f_T(z) \\ &+ \frac{1}{n}\sum_{z\in S} f_T(z) - \frac{1}{n}\sum_{z\in T} f_T(z) \\ &\leq \frac{2L}{n}\sum_{i=1}^{n} d(z_i, z_i') + 2R(|T| - |S|) + \xi(n)\end{aligned}$$

In fact, since we can permute $T$ in any suitable way, we have

$$\frac{1}{n}\sum_{z\in T} f_S(z) - \frac{1}{n}\sum_{z\in T} f_T \leq 2Ld(S,T) + \xi(n).$$

Since both terms decaying faster than $n^{-1/2}$ and $f_T \in \mathcal{M}_T^{\xi'(n)}$, we have that

$$f_S \in \mathcal{M}_T^{\xi''(n)}$$

for $\xi''(n) = o(n^{-1/2})$. We apply Theorem 5.3.1 to obtain the result.

□

## 5.5 Rates of Decay of $\mathrm{diam}\mathcal{M}_S^{\xi(n)}$

The statement of Lemma 5.3.1 reveals that the rate of the decay of the diameter $\mathrm{diam}\mathcal{M}_S^{\xi(n)}$ is related to the rate at which $\Pr^*(\sup_{\mathcal{F}} |\nu - \nu_n| \geq \delta) \to 0$ for a fixed $\delta$. A number of papers studied this rate of convergence, and here we refer to the notion of

*Komlos-Major-Tusnady class* (KMT class), as defined by [42]. Let $\nu'_n : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process defined on the probability space $(\mathcal{Z}', \mathcal{G}', P')$.

**Definition 5.5.1.** *$\mathcal{F}$ is called a Komlos-Major-Tusnady class with respect to $P$ and with the rate of convergence $\tau_n$ ($\mathcal{F} \in KMT(P; \tau_n)$) if $\mathcal{F}$ is $P$-pregaussian and for each $n \geq 1$ there is a version $\nu^{(n)}$ of $P$-Brownian bridge defined on $(\mathcal{Z}', \mathcal{G}'^{\text{"""""""}}, P')$ such that for all $t > 0$,*

$$\Pr^* \left( \sup_{\mathcal{F}} |\nu^{(n)} - \nu'_n| \geq \tau_n(t + K \log n) \right) \leq \Lambda e^{-\theta t}$$

*where $K > 0$, $\Lambda > 0$ and $\theta > 0$ are constants, depending only on $\mathcal{F}$.*

Sufficient conditions for a class to be $KMT(P; n^{-\alpha})$ have been investigated in the literature; some results of this type can be found in [42, 62] and [25], Section 9.5(B).

The following theorem shows that for KMT classes fulfilling a suitable entropy condition, it is possible to give explicit rates of decay for the diameter of ERM almost-minimizers.

**Theorem 5.5.1.** *Assume $\mathcal{F}$ is $P$-Donsker and $\mathcal{F} \in KMT(P; n^{-\alpha})$ for some $\alpha > 0$. Assume $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \left(\frac{A}{\epsilon}\right)^V$ for some constants $A, V > 0$. Let $\xi(n)\sqrt{n} = o(n^{-\eta})$, $\eta > 0$. Then*

$$n^\gamma \text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$$

*for any $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$.*

*Proof.* The result of Lemma 5.3.1 is stated for a fixed $n$. We now choose $C$, $\xi$, and $\delta$ depending on $n$ as follows. Let $C(n) = Bn^{-\gamma}$, where $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$ and $B > 0$ is an arbitrary constant. Let $\xi = \xi(n)$. Let $\delta(n) = n^{-\beta}$, where $\beta = \frac{1}{2}(\min(\alpha, \eta) + 3(2V+1)\gamma)$. When $\beta$ is defined this way, we have

$$\min(\alpha, \gamma) > \beta > 3(2V+1)\gamma$$

because $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$ by assumption. In particular, $\beta < \eta$ and, hence, eventually $\delta(n) > \xi(n)\sqrt(n) = o(n^{-\eta})$.

Since $C(n)$ decays to zero and $\epsilon(n) = \min(C(n)^3/128, C(n)/4)$, eventually $\epsilon(n) = C(n)^3/128 = n^{-3\gamma}B^3/128$.

Since $\mathcal{F} \in KMT(P; n^{-\alpha})$,

$$\Pr^*\left(\sup_{\mathcal{F}} |\nu^{(n)} - \nu_n| \geq n^{-\alpha}(t + K\log n)\right) \leq \Lambda e^{-\theta t}$$

for any $t > 0$, choosing $t = n^\alpha \delta(n)/2 - K\log n$ we obtain

$$\Pr^*\left(\sup_{\mathcal{F}} |\nu^{(n)} - \nu_n| \geq \delta(n)/2\right) \leq \Lambda e^{-\theta(n^{\alpha-\beta}/2 - K\log n)}.$$

Lemma 5.3.1 then implies

$$\begin{aligned}
&\Pr^*\left(\mathrm{diam}\mathcal{M}_S^{\xi(n)} > C(n)\right) \leq \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)^2 \left(\frac{128\delta}{C(n)^3} + \Pr^*\left(\sup_{\mathcal{F}} |\nu_n' - \nu'| \geq \delta/2\right)\right) \\
&\leq \left(\frac{128A}{B^3} n^{3\gamma}\right)^{2V} \frac{128}{B^3} n^{-\beta} n^{3\gamma} + \left(\frac{128A}{B^3} n^{3\gamma}\right)^{2V} \Lambda e^{-\theta(n^{\alpha-\beta}/2 - K\log n)} \\
&= \left(\frac{128A}{B^3}\right)^{2V} \frac{128}{B^3} n^{3\gamma(2V+1)-\beta} + \Lambda \left(\frac{128A}{B^3}\right)^{2V} n^{k\theta+6\gamma V} e^{-\frac{\theta}{2}n^{\alpha-\beta}}
\end{aligned}$$

Since $\alpha > \beta > 3\gamma(2V+1)$, both terms above go to zero, i.e.

$$\Pr^*\left(n^\gamma \mathrm{diam}\mathcal{M}_S^{\xi(n)} > B\right) \to 0 \text{ for any } B > 0.$$

□

The entropy condition in Theorem 5.5.1 is clearly verified by VC-subgraph classes of dimension $V$. In fact, since $L_2$ norm dominates $\|\cdot\|$ seminorm, upper bounds on $L_2$ covering numbers of VC-subgraph classes induce analogous bounds on $\|\cdot\|$ covering numbers. Corollary 5.5.1 is a an application of Theorem 5.5.1 to this important family of classes. It follows in a straight-forward way from the remark above.

**Corollary 5.5.1.** *Assume $\mathcal{F}$ is a VC-subgraph class with VC-dimension $V$, and for some $\alpha > 0$ $\mathcal{F} \in KMT(P, n^{-\alpha})$. Let $\xi(n)\sqrt{n} = o(n^{-\eta})$, $\eta > 0$. Then*

$$n^\gamma \mathrm{diam}\mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$$

*for any* $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$.

## 5.6 Expected Error Stability of almost-ERM

In the previous section, we proved bounds on the rate of decay of the diameter of almost-minimizers. In this section, we show that given such a bound, as well as some additional conditions on the class, the differences between *expected errors* of almost-minimizers decay faster than $n^{-1/2}$. This implies a form of *strong expected error stability* for ERM.

The proof of Theorem 5.6.1 relies on the following ratio inequality of [59].

**Proposition 5.6.1.** *Let* $\mathcal{G}$ *be a uniformly bounded function class with the envelope function* $G \equiv 2$. *Assume* $\mathcal{N}(\gamma, \mathcal{G}) = \sup_Q \mathcal{N}(2\gamma, \mathcal{G}, L_1(Q)) < \infty$ *for* $0 < \gamma \leq 1$ *and* $Q$ *ranging over all discrete probability measures. Then*

$$\Pr^* \left( \sup_{\mathcal{G}} \frac{|P_n f - Pf|}{\epsilon(P_n|f| + P|f|) + 5\gamma} > 26 \right) \leq 32 \mathcal{N}(\gamma, \mathcal{G}) \exp(-n\epsilon\gamma)$$

The next theorem gives explicit rates for expected error stability of ERM over VC-subgraph classes fulfilling a KMT type condition.

**Theorem 5.6.1.** *If* $\mathcal{F}$ *is a VC-subgraph class with VC-dimension* $V$, $\sqrt{n}\xi(n) = o(n^{-\eta})$, *and* $\mathcal{F} \in KMT(P; n^{-\alpha})$, *then for any* $\kappa < \min\left(\frac{1}{6(2V+1)} \min(\alpha, \eta), 1/2\right)$

$$n^{1/2+\kappa} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \xrightarrow{P^*} 0.$$

## 5.7 Applications

We now apply the results of the previous sections to the unsupervised setting. Suppose we are given $n$ i.i.d. samples $X_1, \ldots, X_n$ from the unknown $P$ and we are interested in finding out something about $P$. Clustering and density estimation are two such tasks. However, let us first consider a simpler example.

### 5.7.1 Finding the Least (or Most) Dense Region

Suppose we are interested in finding the most dense (or least dense) contiguous region of a fixed size (see Figure 5-5). A natural question is: what is the stability of such a procedure? Note that finding the most dense contiguous region can be phrased as an empirical risk minimization procedure, as described in the next paragraph. If the underlying density has a single mode, we expect that the most dense region will be located at that mode for large $n$ and will not be significantly shifting with perturbations of the data. This corresponds to the single-minimizer setting, discussed in Section 5.3 and depicted in Figure 5-2. However, if there are two equal modes in the density, we expect the most dense region to jump between the modes with the addition of $\Omega(n^{1/2})$ points. This situation corresponds to Figure 5-3. If the underlying density is uniform, the setting corresponds to the one depicted in Figure 5-4, as any region of a fixed size is equally good (equally bad) with respect to the uniform density. The stability of the latter case is difficult to analyze, and we employ the result of Theorem 5.3.1 for this purpose.

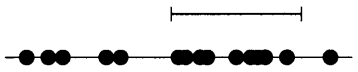


Figure 5-5: The most dense region of a fixed size.

Suppose for simplicity that $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$. Define

$$\mathcal{H}_c = \{h_{[a,b]} : 0 \leq a < b \leq 1,\ b - a \leq c,\ h_{[a,b]}(x) = I(x \notin [a,b])\}$$

for some fixed $c > 0$. In other words, the function class consists of binary functions taking the value 0 on an interval of length smaller than $c$ and the value 1 everywhere else. Let $P(Y = y|X) = \delta_{y=0}$, i.e. the data has the $y$-label zero. Let $\ell(h(X), Y) = h(X) - Y = h(X)$, i.e. the functions in $\mathcal{H}_c$ are the losses. Note that $h_{[a,b]}$ makes mistakes on the data $X_1, \ldots, X_n$ whenever $X_i \notin [a,b]$. Hence, to minimize the number

of mistakes, $[a, b]$ has to contain the most number of points out of $X_1, \ldots, X_n$, which is the most dense region.

Note that the VC dimension of $\mathcal{H}_c$ is 2 for any $c > 0$. Indeed, no three points $x_1 < x_2 < x_3$ can be shattered, as there is no way to assign 1 to $x_2$ and 0 to $x_1, x_2$ with a function from $\mathcal{H}_c$.

Hence, $\mathcal{H}_c$ is uniform Donsker and we obtain the following result.

**Corollary 5.7.1.** *Let $I_n = \{X_1, \ldots, X_n\}$ and $J_n = \{X'_1, \ldots, X'_n\}$, $X_i, X'_i \in [0, 1]$ are i.i.d. according to $P$. Suppose that $M_n := |I_n \triangle J_n| = o(n^{1/2})$. Let $[a_I, b_I]$ and $[a_J, b_J]$ be most dense regions of size $c > 0$ in $I_n$ and $J_n$. Then*

$$|a_I - a_J| + |b_I - b_J| \xrightarrow{P^*} 0.$$

The example extends naturally to $d$-dimensional axis-parallel boxes or other finite-VC classes.

**Corollary 5.7.2.** *Let $I_n = \{X_1, \ldots, X_n\}$ and $J_n = \{X'_1, \ldots, X'_n\}$, $X_i, X'_i \in [0, 1]^d$ are i.i.d. according to $P$. Suppose that $M_n := |I_n \triangle J_n| = o(n^{1/2})$. Let $[a_I^1, b_I^1] \times \ldots \times [a_I^d, b_I^d]$ and $[a_J^1, b_J^1] \times \ldots \times [a_J^d, b_J^d]$ be most dense regions in $I_n$ and $J_n$ such that $a^i - b^i \leq c_i$. Then for any $1 \leq i \leq d$*

$$|a_I^i - a_J^i| + |b_I^i - b_J^i| \xrightarrow{P^*} 0.$$

**Remark 5.7.1.** *The following extensions are straightforward:*

- *The size of the d-dimensional boxes can be restricted in many other ways, depending on the problem at hand.*
- *The same results hold for the* least dense region *problem.*
- *The results hold for k most dense (or least dense) disjoint regions.*

### 5.7.2 Clustering

In the previous section we discussed the connection between the problem of finding the most (or least) dense region and empirical risk minimization. Furthermore, we showed

that the underlying density and the function class determine one of the settings depicted in Figures 5-1 through 5-4. The same reasoning holds for clustering, where an objective function determines the quality of the clustering (such as the within-point scatter for $K$-means). If there is only one best clustering (i.e. the minimum of the objective function is unique), the situation is represented in Figure 5-2, and we expect stability of the minimizers of the objective function with respect to complete changes of the dataset. However, if there are finite number of minimizers, the binomial result tells us that we expect stability with respect to $o(\sqrt{n})$ changes of points, while no stability is expected for $\Omega(\sqrt{n})$ changes. Again, the case of infinitely-many minimizers cannot be resolved by similar arguments, and we employ the result of Theorem 5.3.1.

Let $Z_1, \ldots, Z_n \in \mathbb{R}^m$ be a sample of points. A partition function $C : \mathcal{Z} \mapsto \{1, \ldots, K\}$ assigns to each point $Z$ its "cluster identity". The quality of $C$ on $Z_1, \ldots, Z_n$ is measured by the "within-point scatter" (see [33])

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j:C(Z_i)=C(Z_j)=k} \|Z_i - Z_j\|^2. \tag{5.1}$$

Because the similarity of samples is the Euclidean square distance, the within-point scatter can be rewritten as

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i:C(Z_i)=k} \|Z_i - \bar{Z}_k\|^2 \tag{5.2}$$

where $\bar{Z}_k$ is the mean of the $k$-th cluster based on the assignment $C$ (see Figure 5-6).

The $K$-means clustering algorithm is an alternating procedure minimizing the within-point scatter $W(C)$. The centers $\{\bar{Z}_k\}_{k=1}^K$ are computed in the first step, following by the assignment of each $Z_i$ to its closest center $\bar{Z}_k$; the procedure is repeated. The algorithm can get into a local minima, and various strategies, such as starting with several random assignments, are employed.

The problem of minimizing $W(C)$ can be phrased as an empirical minimization procedure. The $K$-means algorithm is an attempt at finding the minimizer in practice

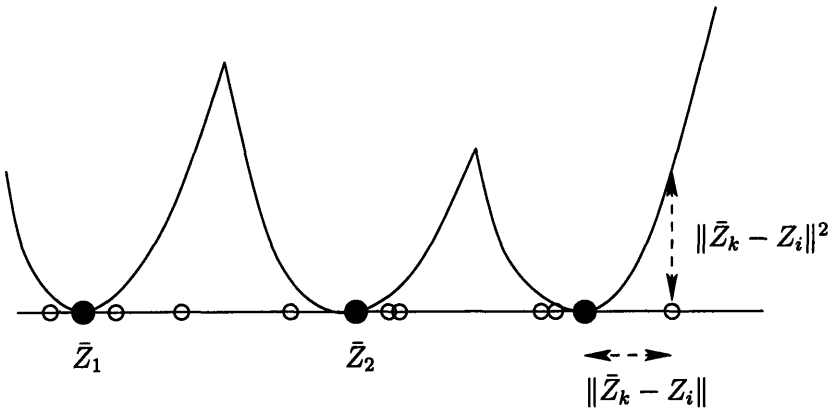



Figure 5-6: The clustering objective is to place the centers $\bar{Z}_k$ to minimize the sum of squared distances from points to their closest centers.

by an alternating minimization procedure, but the convergence to the minimizer is not guaranteed.

Let

$$\mathcal{H}_K = \{h_{z_1,\ldots,z_k}(z) = \|z - z_i\|^2,\ i = \underset{j\in\{1\ldots K\}}{\operatorname{argmin}} \|z - z_j\|^2 : z_1, \ldots, z_K \in \mathcal{Z} \subset R^m\}. \quad (5.3)$$

Functions $h_{z_1,\ldots,z_k}(z)$ in $\mathcal{H}_K$ can also be written as

$$h_{z_1,\ldots,z_k}(z) = \sum_{i=1}^{K} \|z - z_i\|^2 I(z \text{ is closest to } z_i),$$

where ties are broken in some reasonable way.

Hence, functions $h_{z_1,\ldots,z_K} \in \mathcal{H}_K$ are $K$ parabolas glued together with centers at $z_1, \ldots, z_K$, as shown in Figure 5-6. We claim that

$$\min_C W(C) = \min_{h\in\mathcal{H}_K} \sum_{i=1}^{n} h(Z_i).$$

Moreover, the $C^*$ clustering minimizing the left-hand side has to assign each point to the closest cluster center; hence, if the minimum is unique, $C^*$ has to coincide with the assignment of $h_{z_1^*,\ldots,z_K^*}$ minimizing the right-hand side. Unfortunately, the minimum

of the empirical average $\min_{h\in\mathcal{H}} \sum_{i=1}^{n} h(Z_i)$ might not be unique, corresponding to the scenario depicted in Figures 5-3 and 5-4. In this case, it is interesting to address the question of stability of this minimization problem.

Let $B_2(R)$ denote an $L_2$-ball of radius $R$, centered at the origin. If $\mathcal{Z} \subset B_2(R)$ for some constant $R$, the functions in $\mathcal{H}_K$ are bounded above by $4R^2$. Hence, for a fixed $K$, the class $\mathcal{H}_K$ is Donsker. We can apply the result of Corollary 5.4.1 to deduce the following result.

**Theorem 5.7.1.** *Suppose $\mathcal{Z} \subset B_2(R)$. Let $h_{a_1,\ldots,a_K}$ and $h_{b_1,\ldots,b_K}$ be minimizers of*

$$\min_{h\in\mathcal{H}_K} \sum_{i=1}^{n} h(Z_i)$$

*over the sets $S$ and $T$, respectively. Here $\mathcal{H}_K$ is defined as in 5.3. Suppose that $|S \triangle T| = o(\sqrt{n})$. Then*

$$\|h_{a_1,\ldots,a_K} - h_{b_1,\ldots,b_K}\|_{L_1(P)} \xrightarrow{P} 0.$$

Stability of $h_{a_1,\ldots,a_K}$ implies stability of the centers of the clusters with respect to perturbation of the data $Z_1, \ldots, Z_n$.

**Definition 5.7.1.** *Suppose $\{a_1, \ldots, a_K\}$ and $\{b_1, \ldots, b_K\}$ are centers of two clusterings. Define a "distance" between these clusterings as*

$$d_{max}(\{a_1, \ldots, a_K\}, \{b_1, \ldots, b_K\}) := \max\left(\max_i \min_j \|a_i - b_j\|, \max_j \min_i \|a_i - b_j\|\right)$$

**Lemma 5.7.1.** *Assume the measure $P$ is bounded away from 0, i.e. $P(z) > c$ for some $c > 0$. Assume further that $\mathcal{Z} \subset B_2(0, R)$ for some $R < \infty$. Suppose*

$$\|h_{a_1,\ldots,a_K} - h_{b_1,\ldots,b_K}\|_{L_1(P)} \leq \varepsilon.$$

*Then*

$$d_{max}(\{a_1, \ldots, a_K\}, \{b_1, \ldots, b_K\}) \leq \left(\frac{\varepsilon}{c_{c,m,R}}\right)^{1/m}$$

*where $c_{c,m,R}$ depends only on c, m, and R.*

*Proof.* Without loss of generality, assume that $d_{\max}(\{a_1,\ldots,a_K\},\{b_1,\ldots,b_K\})$ is attained at $a_1$ and $b_1$ such that $b_1$ is the closest center to $a_1$ out of $\{b_1,\ldots,b_K\}$. Suppose $d_{\max} = \|a_1 - b_1\| = d$. Consider $B = B_2(a_1, d/2)$, a ball of radius $d/2$ centered at $a_1$. Since any point $z \in B$ is closer to $a_1$ than to $b_1$, we have

$$\|z - a_1\|^2 \leq \|z - b_1\|^2.$$

Refer to Figure $5-7$ for the pictorial representation of the proof.

Note that $b_j \notin B$ for any $j \in \{1 \ldots K\}$. Also note that for any $a_i$,

$$\|z - a_1\|^2 \geq \sum_{i=1}^{K} \|z - a_i\|^2 I(a_i \text{ is closest to } z).$$

Indeed, trivially, if $\|z - a_i\| \leq \|z - a_1\|$ for some $i$, then $\|z - a_1\|^2 \geq \|z - a_i\|^2$.

Combining all the information,

$$\begin{aligned}
\|h_{a_1,\ldots,a_K} - h_{b_1,\ldots,b_K}\|_{L_1(P)} &= \int |h_{a_1,\ldots,a_K}(z) - h_{b_1,\ldots,b_K}(z)|\, dP(z) \\
&\geq \int_B |h_{a_1,\ldots,a_K}(z) - h_{b_1,\ldots,b_K}(z)|\, dP(z) \\
&= \int_B \left|h_{a_1,\ldots,a_K}(z) - \|z - b_1\|^2\right| dP(z) \\
&= \int_B \left(\|z - b_1\|^2 - h_{a_1,\ldots,a_K}(z)\right) dP(z) \\
&= \int_B \left(\|z - b_1\|^2 - \sum_{i=1}^{K} \|z - a_i\|^2 I(a_i \text{ is closest to } z)\right) dP(z) \\
&\geq \int_B \left(\|z - b_1\|^2 - \|z - a_1\|^2\right) dP(z) \\
&\geq \int_B \left((d/2)^2 - \|z - a_1\|^2\right) dP(z) \\
&\geq c \cdot \text{vol}(B_2(0,R)) \cdot \left((d/2)^2 \frac{\pi^{m/2}(d/2)^m}{\Gamma(m/2+1)} - \frac{2\pi^{m/2}(d/2)^{m+2}}{(m+2)\Gamma(m/2)}\right) \\
&= c_{c,m,R} \cdot d^{m+2}
\end{aligned}$$

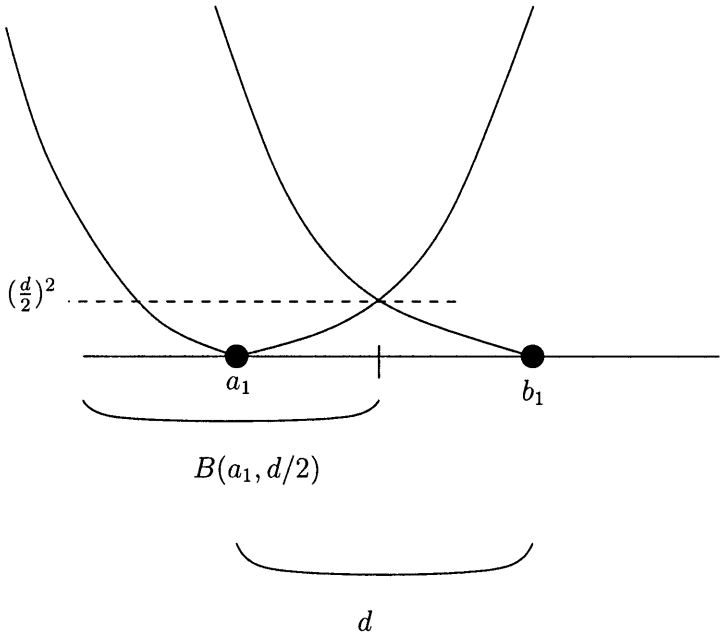



Figure 5-7: To prove Lemma 5.7.1 it is enough to show that the shaded area is upperbounded by the $L_1$ distance between the functions $h_{a_1,\ldots,a_K}$ and $h_{b_1,\ldots,b_K}$ and lower-bounded by a power of $d$. We deduce that $d$ cannot be large.

Since, by assumption,

$$\|h_{a_1,\ldots,a_K} - h_{b_1,\ldots,b_K}\|_{L_1(P)} \leq \varepsilon,$$

we obtain

$$d \leq \left(\frac{\varepsilon}{c_{c,m,R}}\right)^{1/m}.$$

□

From the above lemma, we immediately obtain the following Theorem.

**Theorem 5.7.2.** *Suppose $\mathcal{Z} \subset B_2(0,R)$ for some $R < \infty$. Assume the measure $P$ is bounded away from 0, i.e. $P(z) > c$ for some $c > 0$. Let $a_1, \ldots, a_K$ and $b_1, \ldots, b_K$ be centers minimizing the within-point scatter $W(C)$ (Equation 5.2) over the sets $S$ and $T$, respectively. Suppose that $|S \triangle T| = o(\sqrt{n})$. Then*

$$d_{max}(\{a_1, \ldots, a_K\}, \{b_1, \ldots, b_K\}) \xrightarrow{P} 0.$$

Hence, the minimum of the within-point scatter is stable with respect to perturbations of $o(\sqrt{n})$ points. Similar results can be obtained for other procedures which optimize some function of the data by applying Corollary 5.4.1.

## 5.8 Conclusions

We have presented some new results establishing stability properties of ERM over certain classes of functions. This study was motivated by the question, raised by some recent papers, of $L_1$-stability of ERM under perturbations of a single sample [55, 45, 60]. We gave a partially positive answer to this question, proving that, in fact, ERM over Donsker classes fulfills $L_2$-stability (and hence also $L_1$-stability) under perturbations of $o(n^{\frac{1}{2}})$ among the $n$ samples of the training set. This property follows directly from the main result which shows decay (in probability) of the diameter of the set of solutions of almost-ERM with tolerance function $\xi(n) = o(n^{-\frac{1}{2}})$. We stress that for classification problems (i.e. for binary-valued functions) no generality is lost in assuming the Donsker property, since for ERM to be a sound algorithm, the equivalent Glivenko-Cantelli property has to be assumed anyway. On the other hand, in the real-valued case, many complexity-based characterizations of Donsker property are available in the literature.

In the perspective of possible algorithmic applications, we have analyzed some additional assumptions implying uniform rates of the decay of the $L_1$ diameter of almost-minimizers. It turned out that an explicit rate of this type can be given for VC-subgraph classes satisfying a suitable Komlos-Major-Tusnady type condition. For this condition, many independent characterizations are known. Using a suitable ratio inequality, we showed how $L_1$-stability results can induce strong forms of expected error stability, providing a further insight into the behavior of the empirical risk minimization algorithm.

As in the case of empirical risk minimization, where the geometry of the class and the underlying measure determine the stability of the minimizers, robustness of clustering is also related to the number of minimizers of the objective function (i.e.

best clusters); we applied our result on the $L_1$-stability of ERM to clustering and the problem of finding the most/least dense region.

# Chapter 6

# Concentration and Stability

In Chapter 3, we proved probabilistic bounds on $\widetilde{\mathcal{R}}_{\text{emp}}$ and $\widetilde{\mathcal{R}}_{\text{loo}}$ in terms of stability conditions. The key tools were various deviation and concentration inequalities stated in Chapter 2.

Recall that the variance of a function of $n$ random variables is small if the function is not sensitive to changes of each coordinate alone. This corresponds naturally to the idea of algorithmic stability: the concept learned by the algorithm should not be sensitive to a change of a training sample. Chapter 3 made the connections between deviation inequalities and algorithmic stability precise. In the present Chapter, we provide some further theoretical results on the concentration of functions.

## 6.1 Concentration of Almost-Everywhere Smooth Functions

Consider a probability space $(\mathcal{Z}, \mathcal{G}, \mu)$ and the product space $(\mathcal{Z}^n, \mu^n)$. Let

$$T_n : \mathcal{Z}^n \mapsto [-1, 1].$$

We are interested in the connection between the concentration of $T_n$ around its expectation and the smoothness properties of $T_n$.

We start with the following definitions.

**Definition 6.1.1** (Kutin and Niyogi [43]). *We say that $T_n : \mathcal{Z}^n \mapsto [-1,1]$ is strongly difference-bounded by $(\beta, \delta)$ if there is a subset $B \subset \mathcal{Z}^n$ of measure $\mu^n(B) \leq \delta$ such that for any $1 \leq k \leq n$, if $\omega, \omega' \in \mathcal{Z}^n$ differ only in the $k$-th coordinate, then*

$$|T_n(\omega) - T_n(\omega')| \leq \beta \text{ whenever } \omega \notin B. \tag{6.1}$$

*We will call $B$ the bad set.*

**Remark 6.1.1.** *While in the literature on the concentration of measure phenomenon the term "concentration" is used in conjunction with exponential bounds on the probability, we will use this term to denote any convergence of $|T_n - \mathbb{E}T_n|$ to zero in probability.*

**Definition 6.1.2.** *We will say that $T_n$ concentrates if for any $\varepsilon > 0$*

$$\mathbb{P}\left(|T_n - \mathbb{E}T_n| > \varepsilon\right) \to 0$$

*as $n \to \infty$.*

Assume that $T_n$ is a bounded function and recall McDiarmid's inequality (Theorem 2.4.3).

**Theorem 6.1.1** (McDiarmid [54]). *If $T_n : \mathcal{Z}^n \mapsto [-1,1]$ is strongly difference-bounded by $(\beta, 0)$, then*

$$\mathbb{P}\left(|T_n - \mathbb{E}T_n| > \varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2}{n\beta^2}\right)$$

*for any $\varepsilon > 0$.*

Hence, $T_n$ concentrates whenever it is strongly difference-bounded by $(\beta_n, 0)$ and $\beta_n = o(n^{-1/2})$. We are interested in extensions of McDiarmid's inequality to non-zero bad sets, i.e. to functions which are strongly difference-bounded by $(\beta_n, \delta_n)$.

The following extension has been proved by Kutin and Niyogi [43].

**Theorem 6.1.2.** *If $T_n : \mathcal{Z}^n \mapsto [-1,1]$ is strongly difference-bounded by $(\beta_n, \delta_n)$, then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(|T_n - \mathbb{E}T_n| > \varepsilon\right) \leq 2\left(\exp\left(\frac{-\varepsilon^2}{8n\beta_n^2}\right) + \frac{2n\delta_n}{\beta_n}\right). \tag{6.2}$$

Hence, if $T_n$ is strongly difference-bounded by $(\beta_n, \delta_n)$ such that $\beta_n = o(n^{-1/2})$ and $\delta_n = o(\beta_n/n)$, then $T_n$ concentrates.

Another straightforward calculation assures that $T_n$ is concentrated under weaker conditions on $T_n$:

**Proposition 6.1.1.** *If $T_n : \mathcal{Z}^n \mapsto [-1,1]$ is strongly difference-bounded by $(\beta_n, \delta_n)$, then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(|T_n - \mathbb{E}T_n| > \varepsilon\right) \leq \frac{n\beta_n^2 + n\delta_n}{2\varepsilon^2}.$$

*Proof.* Denote

$$\Gamma = T_n(Z_1, \ldots, Z_n)$$

and

$$\Gamma_i' = T_n(Z_1, \ldots, Z_i', \ldots, Z_n).$$

By Efron-Stein's inequality,

$$\begin{aligned}
\mathrm{Var}(T_n) &\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(\Gamma - \Gamma_i')^2 \\
&= \frac{1}{2}\mathbb{E}\left[I_{(Z_1,\ldots,Z_n)\notin B}\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2 + I_{(Z_1,\ldots,Z_n)\in B}\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2\right] \\
&\leq \frac{1}{2}\left(n\beta_n^2 + n\delta_n\right).
\end{aligned}$$

The result follows from Chebyshev's inequality. □

Hence, if $T_n$ is strongly difference-bounded by $(\beta_n, \delta_n)$ such that $\beta_n = o(n^{-1/2})$ and $\delta_n = o(n^{-1})$, then $T_n$ concentrates.

The bound in Proposition 6.1.1 uses the second moment to upper-bound the probability of the deviation. Similarly, we can use powerful moment inequalities, recently

developed by Boucheron et al [15], to bound the $q$-th moment of $T_n$. Moreover, $q$ can be optimized to get the tightest bounds[1].

Define random variables $V_+$ and $V_-$ as

$$V_+ = \mathbb{E}\left[\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2 I_{\Gamma \geq \Gamma_i'} | Z_1, \ldots, Z_n\right], \; V_- = \mathbb{E}\left[\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2 I_{\Gamma < \Gamma_i'} | Z_1, \ldots, Z_n\right].$$

Further, for a random variable $W$, define

$$\|W\|_q = \left(\mathbb{E}\left[|W|^q\right]\right)^{1/q}$$

for $q > 0$.

**Theorem 6.1.3** (Boucheron et al [15]). *For $T_n : \mathcal{Z}^n \mapsto \mathbb{R}$, let $\Gamma = T_n(Z_1, \ldots, Z_n)$. For any $q \geq 2$,*

$$\|(\Gamma - \mathbb{E}\Gamma)_+\|_q \leq \sqrt{2\kappa q}\|\sqrt{V_+}\|_q, \text{ and } \|(\Gamma - \mathbb{E}\Gamma)_-\|_q \leq \sqrt{2\kappa q}\|\sqrt{V_+}\|_q,$$

*where $x_+ = \max(0, x)$ and $\kappa \approx 1.271$ is a constant.*

This result leads to the following theorem:

**Theorem 6.1.4.** *Assume $T_n : \mathcal{Z}^n \mapsto \mathbb{R}$ satisfies the bounded difference condition (6.1) on a set of measure $1 - \delta_n$. Then for any $q \geq 2$ and $\varepsilon > 0$,*

$$\mathbb{P}\left(T_n - \mathbb{E}T_n > \varepsilon\right) \leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q},$$

*where $\kappa \approx 1.271$.*

*Proof.* Note that

$$\mathbb{E}V_+^{q/2} = \mathbb{E}\{I_G V_+^{q/2} + I_{\bar{G}} V_+^{q/2}\} \leq (n\beta_n^2)^{q/2} + (nq(2M)^2)^{q/2}\delta_n.$$

[1]Thanks to Gábor Lugosi for suggesting this method.

By Theorem 6.1.3,

$$\begin{aligned}\mathbb{E}(\Gamma - \mathbb{E}\Gamma)_+^q &\leq (2\kappa q)^{q/2}\mathbb{E}V_+^{q/2} \\ &\leq (n\beta_n^2 q 2\kappa)^{q/2} + (n(2M)^2)^{q/2}\delta_n.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\left(T_n - \mathbb{E}T_n > \varepsilon\right) &\leq \frac{\mathbb{E}(\Gamma - \mathbb{E}\Gamma)_+^q}{\varepsilon^q} \\ &\leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q}.\end{aligned}$$

□

The bound of Theorem 6.1.4 holds for any $q \geq 2$. To clarify the asymptotic behavior of the bound, assume $\beta_n = n^{-\gamma}$ for some $\gamma > 1/2$, and let

$$q = \varepsilon^2\beta_n^{-2}n^{-2\gamma+\eta} = \varepsilon^2 n^\eta$$

for some $\eta$ to be chosen later such that $2\gamma - 1 > \eta > 0$. Assume $\delta_n = \exp(n^{-\theta})$ for some $\theta > 0$. The bound of Theorem 6.1.4 becomes

$$\begin{aligned}\mathbb{P}\left(T_n - \mathbb{E}T_n > \varepsilon\right) &\leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q} \\ &\leq \left(\frac{2\kappa nq\beta_n^2}{\varepsilon^2}\right)^{q/2} + \delta_n\left(\frac{4M^2nq}{\varepsilon^2}\right)^{q/2} \\ &\leq \left(2\kappa n^{1+\eta-2\gamma}\right)^{\frac{\varepsilon^2}{2}n^\eta} + \left(4M^2n^{1+\eta}\right)^{\frac{\varepsilon^2}{2}n^\eta}\exp(-n^\theta) \\ &\leq \exp\left((1 + (1+\eta-2\gamma)\log n)n^\eta\frac{\varepsilon^2}{2}\right) \\ &\quad + \exp\left((2\log(2M) + (1+\eta)\log n)\,n^\eta\frac{\varepsilon^2}{2} - n^\theta\right). \qquad (6.3)\end{aligned}$$

Since $1 + \eta - 2\gamma < 0$, the first term is decaying exponentially with $n$. We can now choose $\eta < \min(\theta, 2\gamma - 1)$ for the second term to decay exponentially. In particular, let us compare our result to the result of Theorem 6.1.2. With $\delta_n = \exp(n^{-\theta})$ the

bound in Equation (6.2) becomes

$$\mathbb{P}\left(T_n - \mathbb{E}T_n > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{8} n^{2\gamma-1}\right) + \exp\left(\left(\log M + (\gamma+1)\log n\right) - n^{\theta}\right). \tag{6.4}$$

Depending on whether $\theta < 2\gamma - 1$ or not, the first or second term dominates convergence to zero, which coincides exactly with the asymptotic behavior of our bound. In fact, one can verify that the terms in the exponents of bounds (6.3) and (6.4) have the same order.

*We have therefore recovered the result*[2] *of Theorem 6.1.2 for the interesting case* $\delta_n = exp(-n^{\theta})$ *by using moment inequality of Boucheron et al [15].* Note that the result of Theorem 6.1.4 is very general and different ways of picking $q$ might prove useful. For instance, if $\delta_n = 0$, i.e. the bounded difference condition (6.1) holds over the whole $\mathcal{Z}^n$, we can choose

$$q = \frac{\varepsilon^2}{4n\beta_n^2}$$

to recover McDiarmid's inequality.

## 6.2 The Bad Set

The results of the previous section provide guarantees for the concentration of $T_n$ in terms of $\delta_n$ and $\beta_n$. However, not every rate of decay of $\delta_n$ and $\beta_n$ implies that $T_n$ is concentrated. In fact, this is not due to a weakness in our approach, but rather due to an apparent phase transition between concentration and non-concentration for functions of $n$ random variables.

The next example is a negative result: there exists a function $T_n$ with $\beta_n = 0$ and $\delta_n = \Omega(n^{-1/2})$ which does not concentrate.

**Example 4.** *Let* $\mathcal{Z} = \{-1, 1\}$. *Let*

$$T_n(Z_1, \ldots, Z_n) = majority(Z_1, \ldots, Z_n).$$

[2]This gives an answer to the open question 6.2 in [43].

*In other words, $T_n$ takes the value 1 if the majority (or exactly half) of the coordinates are 1, and takes the value $-1$ otherwise. Note that $T_n$ changes the value on the boundary between 1's and $-1$'s. Hence, $\mu^n(B) = \Omega(n^{-1/2})$. Clearly, $T_n$ does not concentrate, as it takes values $\pm 1$ while $\mathbb{E}T_n = 0$.*

While Proposition 6.1.1 assures that functions with $\mu^n(B) = o(n^{-1})$ do concentrate, the above example shows that the rate $\mu^n(B) = \Omega(n^{-1/2})$ is too slow for concentration.

Can something be said about the concentration of a function with the size of the bad set decreasing faster than $1/\sqrt{n}$ but slower than $1/n$? It appears that methods of the kind used in the proof of Theorem 6.1.4 will not be able to answer this question due to the way the bad and the good sets are combined together.

In this Chapter we show how the question of the size of the bad set can be phrased geometrically in terms of the size of certain boundaries. We use isoperimetry and classical work on the size of extremal sets to derive sharp results connecting the concentration of functions and the size of bad sets.

### 6.2.1 Main Result

Consider Example 4 on the discrete cube $\{0,1\}^n$. More precisely, for $x = (x_1, ..., x_n) \in \{0,1\}^n$, let $f(x) = 1$ if the majority of $x_i$ are 1's, and $f(x) = -1$ otherwise. The cube is partitioned into two regions $F_{+1}^n$ and $F_{-1}^n$ according to the value of $f$. When changing one coordinate of $x$, the change of the value of $f(x)$ occurs exactly at the boundary between these two regions, which is at the $n/2$ Hamming distance from the origin. This boundary contains $\Omega(2^n/\sqrt{n})$ vertices. The reader will notice that the boundary is exactly the "bad set", i.e. points such that a change of one coordinate results in a large jump of the value of $f$.

Now consider an arbitrary $\{-1,1\}$-valued function on the cube. Assume uniform measure $\mu$ on the vertices of the cube. Notice that $f$ is concentrated around its mean if and only if $\mathbb{E}f = 1$ or $\mathbb{E}f = -1$. Assume that $f$ is not concentrated around its

mean and that $\mathbb{E}f = 0$[3] . Then, clearly,

$$\mu(F_{+1}^n) = \mu(F_{-1}^n) = 1/2.$$

An example of such a function $f$ is depicted in Figure 6-1 below.

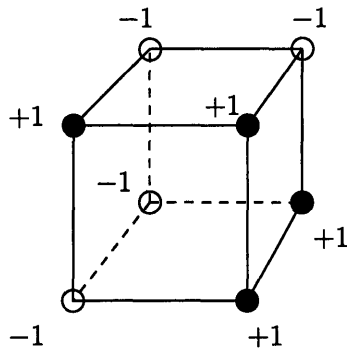



Figure 6-1: Function $f$ defined at the vertices as $-1$ or 1 such that $\mathbb{E}f = 0$.

As in the previous example, the value of $f(x)$ changes only at the boundary between $F_{+1}^n$ and $F_{-1}^n$, but this boundary can be more complex than the one in the previous example (see Figure 6-2). Moreover, by the isoperimetric result of Harper [31], the extremal set[4] of measure 1/2 is exactly the set

$$\{x \in \{0,1\}^n : \sum_{i=1}^{n} x_i \leq n/2\}$$

i.e. $F_{-1}^n$ of the "majority" example (see Figure 6-3). For more information on extremal sets see [47], page 31.

Hence, the boundary between two sets of vertices of measure 1/2 has measure $\Omega(1/\sqrt{n})$. We therefore have the following theorem:

**Theorem 6.2.1.** *If $f : \{0,1\}^n \to \{-1,+1\}$ does not concentrate, then the measure of the bad set is $\Omega(1/\sqrt{n})$.*

[3]Throughout this Section we assume, for simplicity, that $f$ is zero-mean, although the proofs are the same for any non-zero constant mean.

[4]A set is called extremal if it has the smallest boundary out of sets with the given measure.

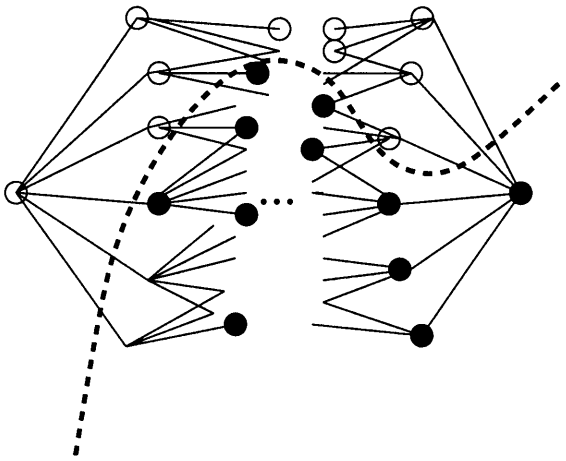


Figure 6-2: $n$-dimensional cube with a $\{-1, 1\}$-valued function defined on the vertices. The dashed line is the boundary separating the set of $-1$'s from the set of 1's. The points at the boundary are the "bad set".

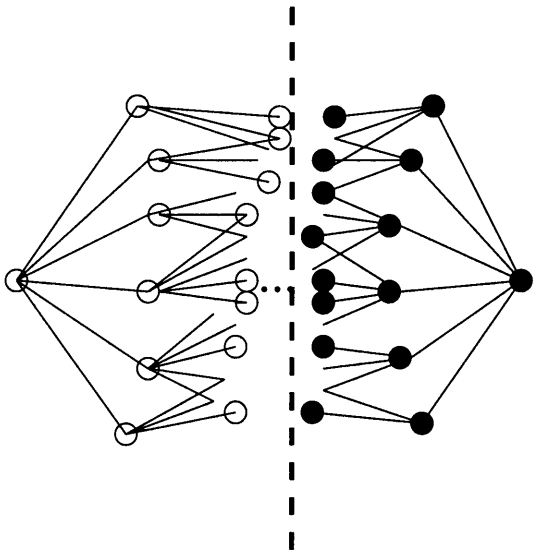


Figure 6-3: The boundary is smallest when the cube is cut in the middle. The extremal set is the set of points at most $n/2$-Hamming distance away from the origin.

We will now extend the above result to $[-1, 1]$-valued zero-mean functions on the

binary cube. While for the binary-valued functions, the notion of a "jump" of the function value was clear, for the $[-1,1]$-valued functions we need to set the scale. In particular, we will say that the "bad set" consists of points such that a change of a coordinate leads to a jump by more than a multiple of $n^{-1/2}$. Let $x^i \in \{0,1\}^n$ be the point obtained by flipping the $i$-th coordinate of $x \in \{0,1\}^n$.

**Definition 6.2.1.** *Define*

$$G_L^n(f) = \{x \in \{0,1\}^n : \forall i,\ |f(x) - f(x^i)| < \frac{L}{\sqrt{n}}\}$$

*and let the*

$$B_L^n(f) = \{0,1\}^n \setminus G_L^n(f)$$

*be the complement of* $G_L^n(f)$.

Consider the sets

$$F_{+c}^n = \{f \geq c\} = \{(x_1, ..., x_n) \in \{0,1\}^n : f(x_1, ..., x_n) > c\}$$

and

$$F_{-c}^n = \{f \leq -c\} = \{(x_1, ..., x_n) \in \{0,1\}^n : f(x_1, ..., x_n) < -c\}.$$

Assume that $f$ does not concentrate around 0, i.e. $\mu(|f| > c)$ does not tend to zero for some fixed $c > 0$. Define

$$a = \sup\{c > 0 : \exists \delta \text{ s.t. } \mu(F_{+c}^n) > \delta \text{ for infinitely many } n\}$$

and

$$b = \sup\{c > 0 : \exists \delta' \text{ s.t. } \mu(F_{-c}^n) > \delta' \text{ for infinitely many } n\}$$

where $\delta, \delta'$ are constants. In other words, $a$ is the largest positive level such that there is a constant measure of points with function values above this level. Similarly, $-b$ is the largest negative level. Note that it cannot happen that both of these suprema do not exist because $f$ is not concentrated. It also cannot happen that one of these

levels is positive while the other supremum does not exist since $\mathbb{E}f = 0$ and $f$ is bounded. Thus, $a > 0$ and $b > 0$. Let $c = \min(a, b)$. By the definition, $\mu(F^n_{+c}) > \delta_+$ and $\mu(F^n_{-c}) > \delta_-$, where $\delta_+$ and $\delta_-$ are constants that depend on $c$ only. Choosing $\delta_c = \min\{\delta_+, \delta_-\}$, we obtain

$$\mu(F^n_{+c}) > \delta_c \text{ and } \mu(F^n_{-c}) > \delta_c.$$

The set $\{x : -c < f(x) < c\}$ has measure at most $1 - 2\delta_c$. Consider slices of $[-c, c]$:

$$C_t = \left\{ x : f(x) \in \left[ -c + t\frac{L}{\sqrt{n}}, -c + (t+1)\frac{L}{\sqrt{n}} \right] \right\}$$

for $t = 0, ..., \frac{\sqrt{n}(2c)}{L} - 1$. These are sets of points on which $f$ takes values within a $L/\sqrt{n}$ window. Hence, there exists an interval $t_0$ such that

$$\mu(C_{t_0}) \leq \frac{1 - 2\delta_c}{\frac{\sqrt{n}(2c)}{L}} = \frac{L(1 - 2\delta_c)}{\sqrt{n}(2c)}.$$

Consider a new (binary-valued) function $g : \{0,1\}^n \to \{-1, 1\}$ obtained from $f$ as follows:

$$g(x) = \begin{cases} 1, & \text{if } f(x) > -c + (t_0 + 1/2)\frac{L}{\sqrt{n}} \\ -1, & \text{otherwise} \end{cases}$$

Consider a point $x \notin C_{t_0}$. If a change of any of the coordinates of $x$ results in a change of the value of $g$, then the same change of the coordinate results in a change of the value of $f$ by more than $L/\sqrt{n}$. Therefore, the size of the "bad set" of $g$ is smaller than the size of the bad set of $f$ plus the size of $C_{t_0}$:

$$\mu(B^n_2(g)) \leq \mu(B^n_L(f)) + \mu(C_{t_0}) \leq \mu(B^n_L(f)) + \frac{L(1 - 2\delta_c)}{\sqrt{n}(2c)}$$

The result then follows from the Theorem 6.2.1:

**Theorem 6.2.2.** *If $f : \{0,1\}^n \to [-1, +1]$ does not concentrate, then there exists an absolute constant $L$ such that the size of the bad set $B^n_L$ is $\Omega(1/\sqrt{n})$.*

### 6.2.2 Symmetric Functions

In this section we give an alternate proof for the special case of symmetric functions. Although the proof of Theorem 6.2.2 is simpler, the following provides some insight into the geometry of symmetric functions on the binary cube.

*The reasoning in this section will be as follows. Assuming that $f$ does not concentrate, we will find two sets of constant measure, on which $f$ is "large" positive and "large" negative, respectively. Next we will show that at least half of each set has to be within const/$\sqrt{n}$ Hamming distance from the main diagonal of the cube. This will imply that $\Omega(\sqrt{n})$ flips is enough to change a point on which $f$ is large positive into one on which $f$ is large and negative. In order for this to happen, one of the steps must be large and that's what we will call a part of the "bad set". Due to the nature of the symmetric functions on the cube, this will imply that a large portion of points is contained in this "bad set". This establishes the connection between the concentration of $f$ and the size of the "bad set".*

Note that the symmetric function $f(x_1, ..., x_n)$ can take only $n$ values and these values are determined by the number of $1's$ in the bit-string $x_1, ..., x_n$. Let

$$S_i = \{(x_1, ..., x_n) \in \{0,1\}^n : \sum x_j = i\}$$

and note that $f$ is constant on $S_i$. Since $f$ is symmetric, both $F^n_{+c}$ and $F^n_{-c}$ are unions of $S_j$'s. Assuming uniform measure, $\mu(S_i) = |S_i|/2^n$. The size $|S_i|$ is exactly $\binom{n}{i}$, while $\sum_{i=1}^n |S_i| = 2^n$ and $\mu(S_{n/2}) = \Omega(1/\sqrt{n})$. For $x \in \{0,1\}^n$ let $|x|$ denote the number of 1's in $x$ (equivalently, Hamming distance from the origin). By definition, $x \in S_{|x|}$.

**Theorem 6.2.3.** *Assume $f : \{0,1\}^n \to [-1,1]$ is symmetric. If $f$ does not concentrate around its mean, there exists an absolute constant $L$ such that the size of the bad set $B^n_L$ is at least $\Omega(1/\sqrt{n})$.*

*Proof.* We define $F^n_{+c}$ and $F^n_{+c}$ as in the previous section and recall that $\mu(F^n_{+c}) > \delta_c$, $\mu(F^n_{-c}) > \delta_c$. First, we would like to say that at least half of the set $F^n_{+c}$ is within

$r_c\sqrt{n}$ Hamming distance from the "main diagonal" $S_{n/2}$, where $r_c$ is some constant. From Talagrand's inequality [47] it follows that

$$\mu\left(\bigcup_{i=1}^{r_c\sqrt{n}} S_i\right) = \mu\left(\left\{x : d_H(x, \bigcup_{j=n/2}^{n} S_j) > r_c\sqrt{n}\right\}\right) \leq \frac{1}{2}e^{-r_c^2}.$$

So,

$$\mu\left(\left\{\bigcup_{i=1}^{r_c\sqrt{n}} S_i\right\} \bigcup \left\{\bigcup_{i=n/2+r_c\sqrt{n}}^{n} S_i\right\}\right) \leq e^{-r_c^2}.$$

The message of the above inequality is that most of the mass in the cube is concentrated around the main diagonal $S_{n/2}$. Choosing $r_c > \sqrt{\log \frac{2}{\delta_c}}$,

$$\mu\left(\{\bigcup_{i=1}^{r_c\sqrt{n}} S_i\} \bigcup \left\{\bigcup_{i=n/2+r_c\sqrt{n}}^{n} S_i\right\}\right) \leq \frac{\delta_c}{2}$$

and therefore at least half of $F_{+c}^n$ and at least half of $F_{-c}^n$ are within $r_c\sqrt{n}$ Hamming distance from $S_{n/2}$. Denote these subsets by

$$H_{+c}^n = F_{+c}^n \bigcap \left\{x \in \{0,1\}^n : |x| \in [n/2 - r_c\sqrt{n}, n/2 + r_c\sqrt{n}]\right\}$$

and

$$H_{-c}^n = F_{-c}^n \bigcap \left\{x \in \{0,1\}^n : |x| \in [n/2 - r_c\sqrt{n}, n/2 + r_c\sqrt{n}]\right\}.$$

By the above argument, $\mu(H_{+c}^n) \geq \frac{\delta_c}{2}$ and $\mu(H_{-c}^n) \geq \frac{\delta_c}{2}$.

Both $H_{+c}^n$ and $H_{-c}^n$ are unions of $S_j$'s and $j \in [n/2 - r_c\sqrt{n}, n/2 + r_c\sqrt{n}]$. Thus, there must be two indices $i, j \in [n/2 - r_c\sqrt{n}, n/2 + r_c\sqrt{n}]$ such that $S_i \subset H_{+c}^n$ and $S_j \subset H_{-c}^n$ and by construction, $|i - j| \leq 2r_c\sqrt{n}$.

Pick a point $x \in S_i \subset H_{+c}^n$ and change $|i - j|$ coordinates to arrive at some $y \in S_j \subset H_{-c}^n$. Thus, there exists a path of at most $2r_c\sqrt{n}$ steps from $x \in H_{+c}^n$ to some $y \in H_{-c}^n$. Note that $f(x) > c$ and $f(y) < -c$ by definition. Therefore, there is at least one change of a coordinate on this path which results in a jump of the function value by at least $\frac{2c}{2r_c\sqrt{n}}$. Assume this jump occurs between some $w$

and $w^k$ on the path between $x$ and $y$, i.e. $|f(w) - f(w^k)| > \frac{c}{r_c\sqrt{n}}$ for some $k$. Then $w \in B^n_{\frac{c}{r_c}}$. Since $f$ is symmetric, the whole set $S_{|w|}$ belongs to $B^n_{\frac{c}{r_c}}$. By construction, $|w| \in [n/2 - r_c\sqrt{n}, n/2 + r_c\sqrt{n}]$. It then follows that $\mu(S_{|w|}) = \Omega(1/\sqrt{n})$ and thus $\mu(B^n_{\frac{c}{r_c}})$ is at least $\Omega(1/\sqrt{n})$. □

## 6.3 Concentration of Measure: Application of Inequality of Bobkov-Ledoux

Following Bobkov and Ledoux [12], consider a probability measure $\mu$ on a metric space $(\mathcal{Z}, d)$ and a product measure $\mu^n$ on $\mathcal{Z}^n$. For $g : \mathcal{Z}^n \mapsto \mathbb{R}$, define

$$|\nabla g|(x) = \limsup_{y \to x} \frac{|g(x) - g(y)|}{d(x, y)} \tag{6.5}$$

and let $|\nabla_i g|$ denote the gradient with respect to the $i$th coordinate. We say that $\mu$ satisfies a Poincaré inequality with constant $\lambda$ if, for every $g$ such that $\int g^2 d\mu < \infty$ and $\int |\nabla g|^2 d\mu < \infty$,

$$\lambda \mathrm{Var}_\mu(g) \leq \int |\nabla g|^2 d\mu.$$

**Theorem 6.3.1** (Corollary 3.2, [12]). *Assume that $\mu$ satisfies 6.5 with $\lambda > 0$. Then for every bounded function $g$ on $\mathcal{Z}^n$ such that*

$$\sum_{i=1}^n |\nabla_i g|^2 \leq \alpha^2 \quad \textit{and} \quad \max_{i \leq i \leq n} |\nabla_i g| \leq \beta$$

*$\mu$-a.e., and for every $t \geq 0$*

$$\mu^n \left( g \geq \int f d\mu^n + t \right) \leq \exp\left( -\frac{1}{K} \min\left( \frac{t}{\beta}, \frac{t^2}{\alpha^2} \right) \right)$$

*where $K > 0$ and only depends on $\lambda > 0$.*

For $S = \{Z_1, \ldots, Z_n\}$, let $f_S$ be an (approximate) empirical minimizer over $\mathcal{F}$. Similarly, $f_T$ is an (approximate) empirical minimizer for the set $T = \{Z'_1, ..., Z'_k\}$.

We will apply the concentration result of Theorem 6.3.1 to

$$g(Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^{n} f_S(z_i).$$

To do so, we need to show that $g$ is smooth.

Note that if $S$ and $T$ differ only in the $i$th sample,

$$\left| \frac{1}{n} \sum_{z \in S} f_S(z) - \frac{1}{n} \sum_{z \in T} f_T(z) \right| \leq \frac{L}{n} d(z_i, z_i')$$

by a proof similar to that of Corollary 5.4.2. Hence,

$$|\nabla_i g| \leq \frac{L}{n}.$$

Now, assume $\mu$ satisfies the Poincaré inequality with $\lambda > 0$. We now apply Theorem 6.3.1 to the function $g : \mathcal{Z}^n \mapsto [-2, 2]$ with $\alpha^2 = \frac{L^2}{n}$ and $\beta = \frac{L}{n}$. We obtain

$$\mathbb{P}\left(g \geq \mathbb{E}g + t\right) \leq \exp\left(-\frac{1}{K} \min\left(\frac{nt}{L}, \frac{nt^2}{L^2}\right)\right).$$

By applying the concentration inequality to $-g$ we can obtain the two-sided inequality:

$$\mathbb{P}\left(|g - \mathbb{E}g| \geq t\right) \leq 2 \exp\left(-\frac{1}{K} \min\left(\frac{nt}{L}, \frac{nt^2}{L^2}\right)\right).$$

Note that $\mathbb{E}g$ is a constant which depends on the problem.

**Theorem 6.3.2.** *Assume $\mu$ satisfies the Poincaré inequality with $\lambda > 0$ and that all functions $f \in \mathcal{F}$ are Lipschitz with a constant $L$. For any $\delta > 0$ and $n > K \log \frac{2}{\delta}$,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} f_S(z_i) - \phi_n \right| \leq \frac{KL \log \frac{2}{\delta}}{n}$$

*with probability at least $1 - \delta$. Here $\phi_n = \mathbb{E}_S f_S(z_1)$ is a data-independent quantity and $K$ depends on $\lambda$.*

The Poincaré condition on the measure $\mu$ is fairly restrictive, as it allows such

tight concentration of empirical errors. In particular, it follows that for independent draws of $S$ and $T$,

$$\left| \frac{1}{n} \sum_{z \in S} f_S(z) - \frac{1}{n} \sum_{z \in T} f_T(z) \right| \leq \frac{KL \log \frac{2}{\delta}}{n}$$

with probability at least $1 - 2\delta$. Hence, the empirical errors of empirical minimizers over different samples are very close to each other with high probability. Such behavior has been observed by Boucheron et al [14] (Theorem 19) and others, although under different complexity conditions on the function class.

# Appendix A

# Technical Proofs

In this appendix we derive some results presented in Section 5.3. In particular, we prove Lemma 5.3.1, which was used in the proof of Theorem 5.3.1, and Corollary 5.3.1. Let us start with some technical Lemmas.

**Lemma A.0.1.** *Let* $f_0, f_1 \in \mathcal{F}$, $\|f_0 - f_1\| \geq C/2$, $\|f_1\| \leq \|f_0\|$. *Let* $h : \mathcal{F} \to \mathbb{R}$ *be defined as* $h(f') = \frac{\langle f', f_0 \rangle}{\|f_0\|^2}$. *Then for any* $\epsilon \leq \frac{C^3}{128}$

$$\inf_{\mathcal{B}(f_0,\epsilon)} h - \sup_{\mathcal{B}(f_1,\epsilon)} h \geq \frac{C^2}{16}.$$

*Proof.*

$$\begin{aligned}
\Delta &:= \inf_{\mathcal{B}(f_0,\epsilon)} h - \sup_{\mathcal{B}(f_1,\epsilon)} h \\
&= h(f_0) - h(f_1) + \inf\{h(f' - f_0) + h(f_1 - f'') | f' \in \mathcal{B}(f_0, \epsilon), f'' \in \mathcal{B}(f_1, \epsilon)\} \\
&\geq h(f_0) - h(f_1) - \frac{2\epsilon}{\|f_0\|} \geq h(f_0) - h(f_1) - \frac{8\epsilon}{C},
\end{aligned}$$

since $\|f_0\| \geq C/4$.

Finally

$$2\langle f_0 - f_1, f_0 \rangle = \|f_0 - f_1\|^2 - \|f_1\|^2 + \|f_0\|^2 \geq \|f_0 - f_1\|^2 \geq \frac{C^2}{4},$$

then

$$h(f_0) - h(f_1) \geq \frac{C^2}{8\|f_0\|^2} \geq \frac{C^2}{8},$$

which proves that

$$\Delta \geq \frac{C^2}{8} - \frac{8\epsilon}{C} \geq \frac{C^2}{16}.$$

□

The following Lemma is an adaptation of Lemma 2.3 of [37].

**Lemma A.0.2.** *Let $f_0, f_1, h$ be defined as in Lemma A.0.1. Suppose $\epsilon \leq \frac{C^3}{128}$. Let $\nu_\mu$ be a Gaussian process on $\mathcal{F}$ with mean $\mu$ and covariance $\mathrm{cov}(\nu_\mu(f), \nu_\mu(f')) = \langle f, f' \rangle$.*

*Then for all $\delta > 0$*

$$\Pr^*\left(|\sup_{\mathcal{B}(f_0,\epsilon)} \nu_\mu - \sup_{\mathcal{B}(f_1,\epsilon)} \nu_\mu| \leq \delta\right) \leq \frac{64\delta}{C^3}.$$

*Proof.* Define the Gaussian process $Y(\cdot) = \nu_\mu(\cdot) - h(\cdot)\nu_\mu(f_0)$. Since

$$\mathrm{cov}(Y(f'), \nu_\mu(f_0)) = \langle f', f_0 \rangle - h(f')\|f_0\|^2 = 0,$$

$\nu_\mu(f_0)$ and $Y(\cdot)$ are independent.

We now reason conditionally with respect to $Y(\cdot)$. Define

$$\Gamma_i(z) = \sup_{\mathcal{B}(f_i,\epsilon)} \{Y(\cdot) + h(\cdot)z\} \qquad \text{with } i = 0, 1.$$

Notice that

$$\Pr^*\left(|\sup_{\mathcal{B}(f_0,\epsilon)} \nu_\mu - \sup_{\mathcal{B}(f_1,\epsilon)} \nu_\mu| \leq \delta | Y\right) = \Pr^*\left(|\Gamma_0(\nu_\mu(f_0)) - \Gamma_1(\nu_\mu(f_0))| \leq \delta\right).$$

Moreover $\Gamma_0$ and $\Gamma_1$ are convex and

$$\inf \partial_- \Gamma_0 - \sup \partial_+ \Gamma_1 \geq \inf_{\mathcal{B}(f_0,\epsilon)} h - \sup_{\mathcal{B}(f_1,\epsilon)} h \geq \frac{C^2}{16},$$

by Lemma A.0.1. Then $\Gamma_0 = \Gamma_1$ in a single point $z_0$ and

$$\Pr^* \left( |\Gamma_0(\nu_\mu(f_0)) - \Gamma_1(\nu_\mu(f_0))| \leq \delta \right) \leq \Pr^* \left( \nu_\mu(f_0) \in [z_0 - \Delta, z_0 + \Delta] \right),$$

with $\Delta = 16\delta/C^2$.

Furthermore,

$$\Pr^* \left( \nu_\mu(f_0) \in [z_0 - \Delta, z_0 + \Delta] \right) \leq \frac{32\delta}{C^2 \sqrt{2\pi \mathrm{var}(\nu_\mu(f_0))}},$$

and $\mathrm{var}(\nu_\mu(f_0)) = \|f_0\|^2 \geq C^2/16$, which completes the proof. □

The reasoning in the proof of the next lemma goes as follows. We consider a finite cover of $\mathcal{F}$. Pick any two almost-minimizers which are far apart. They belong to two covering balls with centers far apart. Because the two almost-minimizers belong to these balls, the infima of the empirical risks over these two balls are close. This is translated into the event that the suprema of the shifted empirical process over these two balls are close. By looking at the Gaussian limit process, we are able to exploit the covariance structure to show that the suprema of the Gaussian process over balls with centers far apart are unlikely to be close.

*Lemma 5.3.1.* Consider the $\epsilon$-covering $\{f_i | i = 1, \ldots, \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)\}$. Such a covering exists because $\mathcal{F}$ is totally bounded in $\|\cdot\|$ norm [see page 89, 71]. For any $f, f' \in \mathcal{M}_S^\xi$ s.t. $\|f - f'\| > C$, there exist $k$ and $l$ such that $\|f - f_k\| \leq \epsilon \leq C/4$, $\|f' - f_l\| \leq \epsilon \leq C/4$. By triangle inequality it follows that $\|f_k - f_l\| \geq C/2$.

Moreover

$$\inf_{\mathcal{F}} P_n \leq \inf_{\mathcal{B}(f_k,\epsilon)} P_n \leq P_n f \leq \inf_{\mathcal{F}} P_n + \xi$$

and

$$\inf_{\mathcal{F}} P_n \leq \inf_{\mathcal{B}(f_l,\epsilon)} P_n \leq P_n f' \leq \inf_{\mathcal{F}} P_n + \xi.$$

Therefore,

$$\left| \inf_{\mathcal{B}(f_k,\epsilon)} P_n - \inf_{\mathcal{B}(f_l,\epsilon)} P_n \right| \leq \xi.$$

The last relation can be restated in terms of the empirical process $\nu_n$:

$$\left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)} \{-\nu_n - \sqrt{n}P\} \right| \leq \xi\sqrt{n} \leq \delta.$$

Now,

$$\Pr^*\left(\mathrm{diam}\mathcal{M}_S^\xi > C\right) = \Pr^*\left(\exists f, f' \in \mathcal{M}_S^\xi, \|f - f'\| > C\right) \leq$$
$$\Pr^*\left(\exists l, k \quad \text{s.t.} \ \|f_k - f_l\| \geq C/2, \ \left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)} -\{\nu_n - \sqrt{n}P\} \right| \leq \delta\right).$$

By union bound

$$\Pr^*\left(\mathrm{diam}\mathcal{M}_S^\xi > C\right)$$
$$\leq \sum_{\substack{k,l=1 \\ \|f_k - f_l\| \geq C/2}}^{\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|)} \Pr^*\left(\left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)} \{-\nu_n - \sqrt{n}P\} \right| \leq \delta\right).$$

We now want to bound the terms in the sum above. Assuming without loss of generality that $\|f_k\| \geq \|f_l\|$, we obtain

$$\Pr^*\left(\left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)} \{-\nu_n - \sqrt{n}P\} \right| \leq \delta\right)$$
$$= \Pr^*\left(\left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu'_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)} \{-\nu'_n - \sqrt{n}P\} \right| \leq \delta\right)$$
$$= \Pr^*\left(\left| \sup_{\mathcal{B}(f_k,\epsilon)} \{-\nu' - \sqrt{n}P + \nu' - \nu'_n\} - \sup_{\mathcal{B}(f_l,\epsilon)} \{-\nu' - \sqrt{n}P + \nu' - \nu'_n\} \right| \leq \delta\right)$$

$$
\begin{aligned}
&\leq \Pr^*\left(\left|\sup_{\mathcal{B}(f_k,\epsilon)}\{-\nu' - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l,\epsilon)}\{-\nu' - \sqrt{n}P\}\right| \leq 2\delta\right) \\
&+ \Pr^*\left(\sup_{\mathcal{F}}|\nu'_n - \nu'| \geq \delta/2\right) \\
&\leq \frac{128\delta}{C^3} + \Pr^*\left(\sup_{\mathcal{F}}|\nu'_n - \nu'| \geq \delta/2\right),
\end{aligned}
$$

where the first inequality results from a union bound argument while the second one results from Lemma A.0.2 noticing that $-\nu' - \sqrt{n}P$ is a Gaussian process with covariance $\langle f, f' \rangle$ and mean $-\sqrt{n}P$, and since by construction $\epsilon \leq C^3/128$.

Finally, the claimed result follows from the two last relations. □

We now prove, Corollary 5.3.1, the extension of Theorem 5.3.1 to $L_2$ diameters. The proof relies on the observation that a $P$-Donsker class is also Glivenko-Cantelli.

*Corollary 5.3.1.* Note that

$$
\|f - f'\|_{L_2}^2 = \|f - f'\|^2 + (P(f - f'))^2 .
$$

The expected errors of almost-minimizers over a Glivenko-Cantelli (and therefore over Donsker) class are close because empirical averages uniformly converge to the expectations.

$$
\begin{aligned}
&\Pr^*\left(\exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t. } \|f - f'\|_{L_2} > C\right) \\
&\leq \Pr^*\left(\exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t. } |Pf - Pf'| > C/\sqrt{2}\right) + \Pr^*\left(\text{diam}\mathcal{M}_S^{\xi(n)} > C/\sqrt{2}\right).
\end{aligned}
$$

The first term can be bounded as

$$
\begin{aligned}
&\Pr^*\left(\exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t. } |Pf - Pf'| > C/\sqrt{2}\right) \\
&\leq \Pr^*\left(\exists f, f' \in \mathcal{F}, |P_n f - P_n f'| \leq \xi(n), |Pf - Pf'| > C/\sqrt{2}\right) \\
&\leq \Pr^*\left(\sup_{f,f' \in \mathcal{F}} |(P_n - P)(f - f')| > |C/\sqrt{2} - \xi(n)|\right)
\end{aligned}
$$

which goes to 0 because the class $\{f - f' | f, f' \in \mathcal{F}\}$ is Glivenko-Cantelli. The second

term goes to 0 by Theorem 5.3.1. □

We now report the proof of Theorem 5.6.1 stated in Section 5.6. We first need to derive a preliminary lemma.

**Lemma A.0.3.** *Let $\mathcal{F}$ be P-Donsker class with envelope function $G \equiv 1$. Assume $\mathcal{N}(\gamma, \mathcal{F}) = \sup_Q \mathcal{N}(\gamma, \mathcal{F}, L_1(Q)) < \infty$ for $0 < \gamma \leq 1$ and $Q$ ranging over all discrete probability measures. Let $\mathcal{M}_S^{\xi(n)}$ be defined as above with $\xi(n) = o(n^{-1/2})$ and assume that for some sequence of positive numbers $\lambda(n) = o(n^{1/2})$*

$$\lambda(n) \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} P|f - f'| \xrightarrow{P^*} 0. \tag{A.1}$$

*Suppose further that for some $1/2 < \rho < 1$*

$$\lambda(n)^{2\rho-1} - \log \mathcal{N}(\frac{1}{2} n^{-1/2} \lambda(n)^{\rho-1}, \mathcal{F}) \to +\infty. \tag{A.2}$$

*Then*

$$\Pr^* \left( \sqrt{n} \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \leq \sqrt{n}\xi(n) + 131\lambda(n)^{\rho-1} \right) \to 0.$$

*Proof.* Define $\mathcal{G} = \{f - f' : f, f' \in \mathcal{F}\}$ and $\mathcal{G}' = \{|f - f'| : f, f' \in \mathcal{F}\}$. By Example 2.10.7 of [71], $\mathcal{G} = (\mathcal{F}) + (-\mathcal{F})$ and $\mathcal{G}' = |\mathcal{G}| \subseteq (\mathcal{G} \wedge 0) \vee (-\mathcal{G} \wedge 0)$ are Donsker as well. Moreover, $\mathcal{N}(2\gamma, \mathcal{G}) \leq \mathcal{N}(\gamma, \mathcal{F})^2$ and the envelope of $\mathcal{G}$ is $G \equiv 2$. Applying Proposition 5.6.1 to the class $\mathcal{G}$, we obtain

$$\Pr^* \left( \sup_{f,f' \in \mathcal{F}} \frac{|P_n(f - f') - P(f - f')|}{\epsilon(P_n|f - f'| + P|f - f'|) + 5\gamma} > 26 \right) \leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\epsilon\gamma).$$

The inequality therefore holds if the sup is taken over a smaller (random) subclass $\mathcal{M}_S^{\xi(n)}$.

$$\Pr^* \left( \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} \frac{|P(f - f')| - \xi(n)}{\epsilon(P_n|f - f'| + P|f - f'|) + 5\gamma} > 26 \right) \leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\epsilon\gamma).$$

Since $\sup_x \frac{A(x)}{B(x)} \geq \sup_x \frac{A(x)}{\sup_x B(x)} = \frac{\sup_x A(x)}{\sup_x B(x)}$,

$$\Pr^* \left( \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} (|P(f-f')| - \xi(n)) > 26 \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} (\epsilon(P_n|f-f'| + P|f-f'|) + 5\gamma) \right) \tag{A.3}$$

$$\leq 32 \mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\epsilon\gamma).$$

By assumption,

$$\lambda(n) \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} P|f-f'| \xrightarrow{P^*} 0.$$

Because $\mathcal{G}'$ is Donsker and $\lambda(n) = o(n^{1/2})$,

$$\lambda(n) \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} |P_n|f-f'| - P|f-f'|| \xrightarrow{P^*} 0.$$

Thus,

$$\lambda(n) \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} P_n|f-f'| + P|f-f'| \xrightarrow{P^*} 0.$$

Letting $\epsilon = \epsilon(n) := n^{-1/2}\lambda(n)^\rho$, this implies that for any $\delta > 0$, there exist $N_\delta$ such that for all $n > N_\delta$,

$$\Pr^* \left( \sqrt{n} \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} 26\epsilon(n) \left(P_n|f-f'| + P|f-f'|\right) > \lambda(n)^{\rho-1} \right) < \delta.$$

Now, choose $\gamma = \gamma(n) := n^{-1/2}\lambda(n)^{\rho-1}$ (note that since $\rho < 1$, eventually $0 < \gamma(n) < 1$), the last inequality can be rewritten in the following form

$$\Pr^* \left( \sqrt{n} \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} 26 \left(\epsilon(n) \left(P_n|f-f'| + P|f-f'|\right) + 5\gamma(n)\right) > 131\lambda(n)^{\rho-1} \right) < \delta.$$

Combining the relation above with Eqn. A.3,

$$\Pr^*\left(\sqrt{n}\sup_{f,f'\in\mathcal{M}_S^{\xi(n)}}|P(f-f')|\leq\sqrt{n}\xi(n)+131\lambda(n)^{\rho-1}\right)$$
$$\geq 1-32\mathcal{N}\left(\frac{1}{2}n^{-1/2}\lambda(n)^{\rho-1},\mathcal{F}\right)^2\exp(-\lambda(n)^{2\rho-1})-\delta.$$

The result follows by the assumption on the entropy and by arbitrariness of $\delta$.

□

We are now ready to prove Theorem 5.6.1.

*Theorem 5.6.1.* By Corollary 5.5.1,

$$n^\gamma \text{diam}\mathcal{M}_S^{\xi(n)}\xrightarrow{P^*}0$$

for any $\gamma<\min\left(\frac{1}{3(2V+1)}\min(\alpha,\eta),1/2\right)$. Let $\lambda(n)=n^\gamma$ and note that $\lambda(n)=o(\sqrt{n})$, which is a condition in Lemma A.0.3. First, we show that a power decay of the $\|\cdot\|$ diameter implies the same rate of decay of the $L_1$ diameter, hence verifying condition (A.1) in Lemma A.0.3. Proof of this fact is very similar to the proof of Corollary 5.3.1, except that $C$ is replaced by $C\lambda(n)^{-1}$.

$$\Pr^*\left(\exists f,f'\in\mathcal{M}_S^{\xi(n)}\quad\text{s.t. }\|f-f'\|_{L_2}>C\lambda(n)^{-1}\right)$$
$$\leq\Pr^*\left(\exists f,f'\in\mathcal{M}_S^{\xi(n)}\quad\text{s.t. }|Pf-Pf'|>C\lambda(n)^{-1}/\sqrt{2}\right)$$
$$+\Pr^*\left(\text{diam}\mathcal{M}_S^{\xi(n)}>C\lambda(n)^{-1}/\sqrt{2}\right).$$

The second term goes to zero since $\lambda(n)\text{diam}\mathcal{M}_S^{\xi(n)}\xrightarrow{P^*}0$. Moreover, since $\lambda(n)=$

$o(\sqrt{n})$ and $\mathcal{G}$ is Donsker, the first term can be bounded as

$$\begin{aligned}
&\Pr^*\left(\exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t. } |Pf - Pf'| > C\lambda(n)^{-1}/\sqrt{2}\right) \\
&\leq \Pr^*\left(\exists f, f' \in \mathcal{F}, |P_n f - P_n f'| \leq \xi(n), |Pf - Pf'| > C\lambda(n)^{-1}/\sqrt{2}\right) \\
&\leq \Pr^*\left(\sup_{f,f' \in \mathcal{F}} |P(f-f') - P_n(f-f')| > \left|\frac{C}{\sqrt{2}}\lambda(n)^{-1} - \xi(n)\right|\right) \\
&= \Pr^*\left(\lambda(n)\sup_{g \in \mathcal{G}} |Pg - P_n g| > \left|\frac{C}{\sqrt{2}} - \xi(n)\lambda(n)\right|\right) \to 0,
\end{aligned}$$

proving condition (A.1) in Lemma A.0.3.

We now verify condition (A.2) in Lemma A.0.3. Since $\mathcal{F}$ is a VC-subgraph class of dimension $V$, its entropy numbers $\log \mathcal{N}(\epsilon, \mathcal{F})$ behave like $V \log \frac{A}{\epsilon}$ ($A$ is a constant), that is

$$\log \mathcal{N}\left(\frac{1}{2}n^{-1/2}\lambda(n)^{\rho-1}, \mathcal{F}\right) \leq const + \frac{1}{2}V \log n + (1-\rho)V \log \lambda(n).$$

Condition (A.2) of Lemma A.0.3 will therefore hold whenever $\lambda(n)$ grows faster than $(\log n)^{\frac{1}{2\rho-1}}$, for any $1 > \rho > \frac{1}{2}$. In our problem, $\lambda(n)$ grows polynomially, so condition (A.2) is satisfied for any fixed $1 > \rho > 1/2$.

Hence, by Lemma A.0.3

$$\Pr^*\left(\sqrt{n}\sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} |P(f-f')| \leq \sqrt{n}\xi(n) + 131n^{\gamma(\rho-1)}\right) \to 0.$$

Choose any $0 < \kappa < \gamma/2$ and multiply both sides of the inequality by $n^\kappa$. We obtain

$$\Pr^*\left(n^\kappa\sqrt{n}\sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} |P(f-f')| \leq \sqrt{n}\xi(n)n^\kappa + 131n^{\gamma(\rho-1)+\kappa}\right) \to 0. \qquad \text{(A.4)}$$

Now fix a $\rho$ such that $1/2 < \rho < 1 - \kappa/\gamma$. Because $0 < \kappa < \gamma/2$, there is always such a choice of $\rho$. Furthermore, $1 > \rho > 1/2$ so that the above convergence holds. Our choice of $\rho$ implies that $\gamma(\rho-1) + \kappa < 0$ and so $n^{\gamma(\rho-1)+\kappa} \to 0$. Since $\kappa < \gamma/2 < \eta$,

$\sqrt{n}\xi(n)n^{\kappa} \to 0$. Hence,

$$n^{1/2+\kappa} \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \xrightarrow{P^*} 0$$

for any $\kappa < \min\left(\frac{1}{6(2V+1)} \min(\alpha, \eta), 1/2\right)$. □

# Bibliography


[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergnce, and learnability. *J. of the ACM*, 44(4):615–631, 1997.

[2] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. Hardness of approximate optima in lattices, codes, and linear systems. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

[3] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transaction on Information Theory*, 39(3):930–945, May 1993.

[4] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

[5] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2005. To appear.

[6] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 2005. To appear.

[7] Peter L. Bartlett, Shahar Mendelson, and PetraPhilips. Local complexities for empirical risk minimization. In *Proceedings of the 17th Annual Conference on Computational Learning Theory (COLT2004)*, volume 3120, pages 270–284. Springer, 2004.

[8] Peter L. Bartlett, Shahar Mendelson, and PetraPhilips. Optimal sample-based estimates of the expectation of the empirical minimizer. Submitted., 2005.

[9] S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[10] G. Bennett. Probability inequalities for sums of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[11] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[12] S. Bobkov and M. Ledoux. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107:383–400, 1997.

[13] S. Boucheron, O. Bousquet, and G. Lugosi. *Concentration Inequalities*, pages 208–240. springer, 2004.

[14] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31:1583–1614, 2003.

[15] Stephane Boucheron, Olivier Bousquet, Gabor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.

[16] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[17] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[18] A. Caponnetto and A. Rakhlin. Some properties of empirical risk minimization over Donsker classes. Available at http://cbcl.mit.edu/people/rakhlin/erm.pdf. Submitted to JMLR, 2005.

[19] A. Caponnetto and A. Rakhlin. Some properties of empirical risk minimization over donsker classes. AI Memo 2005-018, Massachusetts Institute of Technology, May 2005.

[20] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Number 31 in Applications of mathematics. Springer, New York, 1996.

[21] L. Devroye and T. Wagner. Distribution-free probability inequalities for the deleted and holdout estimates. *IEEE Transactions on Information Theory*, 25:202–207, 1979.

[22] L.P. Devroye and T.J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

[23] L.P. Devroye and T.J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics*, 8:231–239, 1980.

[24] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[25] Richard M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

[26] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.

[27] Y. Freund and R. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[28] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989, 1984.

[29] E. Giné and J. Zinn. Gaussian characterization of uniform Donsker classes of functions. *The Annals of Probability*, 19:758–782, 1991.

[30] L. S. Gottfredson. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24(1):13–23, 1997.

[31] L.H. Harper. Optimal numberings and isoperimetric problems on graphs. *Journal of Combinatorial Theory*, 1:385–393, 1966.

[32] T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1990.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, 2002. HAS t 01:1 1.Ex.

[34] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[35] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for Projection Pursuit Regression and neural network training. *The Annals of Statistics*, 20(1):608–613, March 1992.

[36] M. J. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory (COLT)*, pages 152–162. ACM Press, New York, 1997.

[37] J. Kim and D. Pollard. Cube root asymptotics. *Annals of Statistics*, 18:191–219, 1990.

[38] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. Technical report, University of New Mexico, August 2003.

[39] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459, 2000.

[40] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[41] V. Koltchinskii, D. Panchenko, and F. Lozano. Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. *Annals of Applied Probability*, 13(1):213–252, 2003.

[42] Vladimir I. Koltchinskii. Komlós-Major-Tusnády approximation for the general empirical process and Haar expansion of classes of functions. *Journal of Theoretical Probability*, 7:73–118, 1994.

[43] S. Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. Technical report TR-2002-04, University of Chicago, 2002.

[44] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical report TR-2003-03, University of Chicago, 2002.

[45] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI*, pages 275–282, 2002.

[46] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[47] Michael Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, 2001.

[48] 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.

[49] J. Li and A. Barron. Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural information processings systems 12*, San Mateo, CA, 1999. Morgan Kaufmann Publishers.

[50] Jonathan Q. Li. *Estimation of Mixture Models.* PhD thesis, The Department of Statistics. Yale University, 1999.

[51] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30–55, 2004.

[52] Shie Mannor, Ron Meir, and Tong Zhang. The consistency of greedy algorithms for classification. In *COLT*, pages 319–333, 2002.

[53] Shie Mannor, Ron Meir, and Tong Zhang. Greedy algorithms for classification – consistency, convergence rates, and adaptivity. *Journal of Machine Learning Research*, 4:713–741, 2003.

[54] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, London Math. Soc. Lect. Note Series 141*, pages 148–188, 1989.

[55] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is necessary and sufficient for consistency of empirical risk minimization. CBCL Paper 2002-023, Massachusetts Institute of Technology, December 2002 [January 2004 revision].

[56] P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics*, 10:51–80, 1999.

[57] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990.

[58] T. Poggio, T. Rifkin, R. Mukherjee S., and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, pages 419–422, 2004.

[59] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.

[60] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–419, 2005.

[61] A. Rakhlin, D. Panchenko, and S. Mukherjee. Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229, 2005.

[62] E. Rio. Strong approximation for set-indexed partial sum processes via KMT constructions I. *The Annals of Probability*, 21(2):759–790, 1993.

[63] W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

[64] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*. To appear.

[65] R. Schapire. A brief introduction to boosting. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406, 1999.

[66] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[67] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22:20–76, 1994.

[68] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, Washington, D.C., 1977.

[69] S.A. van de Geer. Rates of convergence for the maximum likelihood estimator in mixture models. *Nonparametric Statistics*, 6:293–310, 1996.

[70] S.A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[71] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.

[72] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[73] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–99, 1999.

[74] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[75] V.N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for the uniform convergence of averages to their expected values. *Teoriya Veroyatnostei i Ee Primeneniya*, 26(3):543–564, 1981.

[76] V.N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.

[77] G. Wahba. *Splines Models for Observational Data.* Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[78] W.H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates for sieve mles. *Annals of Statistics*, 23:339–362, 1995.

[79] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[80] Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.