# Data-Dependent Analysis of Learning Algorithms

**Petra Philips**

A thesis submitted for the degree of Doctor of Philosophy
at The Australian National University

May 2005

Typeset in Computer Modern by $\TeX$ and $\LaTeX\,2_\varepsilon$.

Except where otherwise indicated, this thesis is my own original work.

The results in this thesis were produced under the supervision of Shahar Mendelson and Bob Williamson, and partly in collaboration with Peter Bartlett. The main contribution of this thesis are two related parts. The main technical results in the first part on random subclass bounds appeared as a journal paper with Shahar Mendelson [1], and an earlier conference paper [2]. The results were discussed with my supervisors Shahar Mendelson and Bob Williamson, who gave me advice and direction. The results on the data-dependent estimation of localized complexities for the Empirical Risk Minimization algorithm appeared as part of a conference paper with Peter Bartlett and Shahar Mendelson [3], and the optimality results are work in progress and contained in an unpublished manuscript with Peter Bartlett and Shahar Mendelson [4]. This second part of the thesis is based on intensive discussions and technical advice from Shahar Mendelson and Peter Bartlett.

**List of Publications:**

[1] S. Mendelson and P. Philips. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

[2] S. Mendelson and P. Philips. Random subclass bounds. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory, COLT 2003*, pages 329–343. Springer, 2003.

[3] P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 270–284. Springer, 2004.

[4] P. L. Bartlett, S. Mendelson, and P. Philips. Work in progress, 2005.

Petra Philips
8 May 2005

Ce n'est pas une image juste, c'est juste une image.
*(Jean-Luc Godard)*

# Acknowledgements

First of all, I would like to express my special gratitude to both of my supervisors Bob Williamson and Shahar Mendelson for the continuous support they provided me with while working on this thesis. Bob Williamson introduced me to statistical learning theory, and was the inspiration for and the reason that I came to the ANU. I would like to thank Bob for his trust and understanding, and his valuable guidance, support, and advice throughout the years. I will also never forget the generosity of Bob, Angharad, and Myvanmy Williamson in being such lovely hosts during my first visit to Australia, many years before starting this Ph.D. project.

A most special thank you goes to Shahar Mendelson, who is an outstanding teacher, mentor, and friend. I feel privileged to have had the opportunity to work so closely with him, and I am very grateful for his constant support, availability, patience, for his intensive teaching, comments, constructive critics and discussions. His enthusiasm and sharpness encouraged, motivated, and inspired me immensely, and I treasure much our many friendly conversations sparkling with humour and wit.

I would also like to record my debt of gratitude to Peter Bartlett for being my advisor, and for giving me the unforgettable opportunity to work and learn from him while visiting UC Berkeley. My special thanks to Olivier Bousquet, Ralf Herbrich, Gábor Lugosi, and Bernhard Schölkopf for hosting me kindheartedly while visiting their institutes. Thank you to all of them for stimulating technical and very pleasant personal discussions.

I was lucky to have charming and helpful colleagues and motivating partners for general discussions, among them Olivier Buffet, Cheng Soon Ong, Gunnar Rätsch, Alex Smola, Vishy Vishvanathan. Thanks especially to Omri Guttman, Evan Greensmith, and Tim Sears for a great time while sharing an office (and heaps of chocolates).

There are many other people who inspired me professionally at some point throughout these years. They are Shai Ben-David, Nicolò Cesa-Bianchi, Stephane Boucheron, André Elisseeff, Matthias Hein, Vladimir Koltchinskii, Risi Kondor, John Langford, Phil Long, Ulrike von Luxburg, Shie Mannor, Mario Marchand, Sayan Mukherjee, Dmitry Panchenko, Alexander Rakhlin, Matthias Seeger, Karim Seghouane, John Shawe-Taylor, Alexandre Tsybakov, Manfred Warmuth, Tong Zhang, and Joel Zinn.

Thanks to Cheng Soon Ong, and especially to Alexander Rakhlin, for proof-reading background parts of this thesis.

My very heartfelt thank you goes to my friends who shared a lot of laughter, secrets, mishaps, debates, ideas, an extraordinary house, terrace, and garden, cooking evenings, weddings, but who, above all, proved their loyalty, trustworthiness, and their support in times of disappointment, personal pain, and despair — especially in the last weeks of work on this thesis. Lisa Batten, Bina D'Costa, Dave Kilham, Emily Kilham, Tao Kong, Stephanie Lee, and Torsten Juelich were lovely house-mates. Bina and Dave are wonderful friends, whose kindness, integrity, and open-mindedness give them a special place in my heart. The "whole Kilhams" are magicians of delicious dinners and were a warm and welcoming family and the kindest of hosts. Emily became very fast from "Dave's sister" an exceptionally dear, delicate, and warmhearted friend. Thanks to Steph for her caring, enthusiastic, and lively presence, for animated discussions, and for the daily challenge of climbing Canberra's Black Mountain. Thanks to Bina and Dave, and also to my long-time friends in Europe Ina Ambela and Frank Abegg, for establishing that I am perfectly qualified to be bridesmaid in both a Bengoli-Australian and a Greek-German wedding – I still feel touched by the great trust and affection they showed me. Thanks to Lilach Zac for patience and advice when I needed it, and to Uwe Zimmer for numerous movie nights in his "film club".

And finally, deepest felt thanks to my parents for their unconditional love.

# Abstract

This thesis studies the generalization ability of machine learning algorithms in a statistical setting. It focuses on the data-dependent analysis of the generalization performance of learning algorithms in order to make full use of the potential of the actual training sample from which these algorithms learn.

First, we propose an extension of the standard framework for the derivation of generalization bounds for algorithms taking their hypotheses from random classes of functions. This approach is motivated by the fact that the function produced by a learning algorithm based on a random sample of data depends on this sample and is therefore a random function. Such an approach avoids the detour of the worst-case uniform bounds as done in the standard approach. We show that the mechanism which allows one to obtain generalization bounds for random classes in our framework is based on a "small complexity" of certain random coordinate projections. We demonstrate how this notion of complexity relates to learnability and how one can explore geometric properties of these projections in order to derive estimates of rates of convergence and good confidence interval estimates for the expected risk. We then demonstrate the generality of our new approach by presenting a range of examples, among them the algorithm-dependent compression schemes and the data-dependent luckiness frameworks, which fall into our random subclass framework.

Second, we study in more detail generalization bounds for a specific algorithm which is of central importance in learning theory, namely the Empirical Risk Minimization algorithm (ERM). Recent results show that one can significantly improve the high-probability estimates for the convergence rates for empirical minimizers by a direct analysis of the ERM algorithm. These results are based on a new localized notion of complexity of subsets of hypothesis functions with identical expected errors and are therefore dependent on the underlying unknown distribution. We investigate the extent to which one can estimate these high-probability convergence rates in a data-dependent manner. We provide an algorithm which computes a data-dependent upper bound for the expected error of empirical minimizers in terms of the "complexity" of data-dependent local subsets. These subsets are sets of functions of empirical errors of a given range and can be determined based solely on empirical data. We then show that recent direct estimates, which are essentially sharp estimates on the high-probability convergence rate for the ERM algorithm, can not be recovered universally from empirical data.

# Contents

Chapter 1

# General Overview, Background, and Notation

## 1.1 Introduction and Motivation

In this thesis we study the problem of bounding the generalization error of learning algorithms. *Learning algorithms* are algorithms that learn functions based on a finite sample of empirical data with the goal of finding the functions which reflect relationships in data and thus best explain unseen data.

An example for a machine learning task can be to design an algorithm for a system which automatically recognizes whether a medical image of a skin mole is indicating skin cancer or not. The input of such a system is the multidimensional vector of pixel values of an image. Each input is associated with an output label which is either “yes” or “no”. A huge database of manually annotated images is available to train the system. Some image examples contain skin spots which look like skin cancer although they are harmless and vice versa. The accuracy of the system should be tuned to correctly diagnose the pictures which contain a cancer image, since the health risk for a false label “no” is much higher than for a false “yes”. The goal in designing the algorithm is to “learn” a map from images to labels such that one can predict, given any new image, with high accuracy its right label and with the least risk of damage on average over all possible images.

Another example of a machine learning task is that of speech recognition, where one aims to map an acoustic signal to a phonetic sequence. In this case, the output labels are arbitrary phonetic sequences of arbitrary length, and not just binary values “yes” and “no”, and the input space are sequences of acoustic features. Many more examples of learning problems arise in bioinformatics, where one goal is to use available genome sequences in order to detect regularities and to predict gene functionality, as well as in data mining, for example for text categorization and topic detection.

The starting point in this thesis is a probabilistic model for learning tasks. The main assumption in this model is that the relationship between the input and the output space (denoted here by $\mathcal{X}$ and $\mathcal{Y}$) can be quantified through a joint probability measure $\mu$ on $\mathcal{X} \times \mathcal{Y}$. For example, such a measure quantifies how likely it is to have each of the labels “yes” and “no” associated to a skin image. This measure is, however, *unknown*, and the goal is to approximate this measure from the available training data. For that, we assume that the available training data $((X_1, Y_1), \ldots, (X_n, Y_n))$ consists of $n$ independent samples generated according to the unknown distribution $\mu$. We define a loss function $l : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ , where $l(v,t)$ quantifies the damage or risk when the true label, for a given input, is $t$ whereas the system outputs the label $v$ (for example, the loss when labelling with “no” when the true label is “yes”). With this, the goal of the learning task is to find, based on the training sample, a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which has a small expected loss $\mathbb{E}_\mu l(f(X), Y)$ with respect to the measure $\mu$.

So far the only freedom we have in the model is the choice of the loss function. Given a fixed loss, one can easily find a function $f$ which has a small average loss *on the training data*. However, this function might perform poorly on new unseen data. The ability of a function to achieve a small error on unseen samples is called *generalization*. *Statistical learning theory* is concerned with the analysis, within a statistical framework of the performance of learning algorithms by studying the generalization ability of the function they produce from an empirical sample.

In order to avoid functions which learn the training sample “by heart” and do not generalize well on unseen samples, one usually imposes constraints on the function $f$. The key ingredient in the design of a learning algorithm is thus to define these constraints and calibrate them according to the amount and quality of existing training data. One way to impose constraints is by restricting the possible choices of functions to a fixed class of functions from which the learning algorithm chooses its hypothesis. Much of the existing work in statistical learning theory assumes either implicitly or explicitly that the only significant choice one has in the design of a learning algorithm is the choice of this class of functions. This function class is called the *hypothesis class*. Given a fixed hypothesis class $H$, the goal of a learning algorithm is thus to choose the hypothesis function $h^*$ in $H$ which has the smallest expected error on data drawn according to the underlying probability measure,

$$h^* = \operatorname*{argmin}_{h \in H} \mathbb{E}_\mu l(f(X), Y) \,.$$

Unfortunately, since this probability measure is unknown, it is impossible to compute the expected error and thus this best function. In the standard setting of statistical learning theory, the only information available about the expected error is through the error on the finite empirical sample. One can precisely characterize function classes in which empirical errors converge to expected errors *uniformly* – both with respect to the measure and with respect to the functions. Such classes have been the object of study in empirical process theory, where they are called *uniform Glivenko-Cantelli classes*.

**Definition 1.1 (uniform Glivenko-Cantelli class)** *Let $F$ be a class of real-valued functions defined on a measurable space $\Omega$. $F$ is a uniform Glivenko-Cantelli class (uGC) if, for every $t > 0$,*

$$\lim_{n \longrightarrow \infty} \sup_{\mu} Pr_{\mathbf{X}} \Big\{ \sup_{f \in F} \big| \mathbb{E}_{\mu} f - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \big| \geq t \Big\} = 0 \, ,$$

*where $\mu$ is any measure $\mu$ on $\Omega$ and $\mathbf{X} = (X_1, ..., X_n)$ and $X_i$ are independent random variables distributed according to $\mu$.*

Clearly, if for a given hypothesis class $H$ and a fixed loss function $l$, the class of functions

$$H_l = \{ l_h : (\mathcal{X} \times \mathcal{Y}) \longrightarrow \mathbb{R} \, : \, h \in H, \, l_h(x,y) = l(h(x), y) \}$$

is a Glivenko-Cantelli class, then any hypothesis $\hat{h}$ with a small *empirical error* $\frac{1}{n} \sum_{i=1}^{n} l_{\hat{h}}(X_i, Y_i)$ on the sample will have, for *any* distribution $\mu$ and in the limit of an infinite sample size, also a small expected error and thus will be a good approximation of $h^*$. In statistical learning theory, classes $H_l$ in which the convergence of empirical to expected errors is guaranteed are the *learnable* (simple) classes, as opposed to the *non-learnable* (complex) ones. However, since the empirical training data is finite, this characterization is not useful in a practical setting and one needs an analysis for the case of finite samples.

It turns out that one can also provide conditions under which functions produced by a learning algorithm from the *finite* sample of data are likely to have a small expected error. Moreover, one can quantify the trade-off between generalization ability, error on the sample and sample size through probabilistic bounds for the deviation of expected and empirical error. The additional mathematical quantities which enter these bounds characterize, from the point of view of the generalization performance, the *complexity* of the learning problem. Since the ability of a function to achieve a small error on unseen samples is called *generalization*, these bounds are called *generalization bounds*.

Such bounds are important for practice for the following reasons:

- Since they give an estimate on the number of samples needed for a good performance and allow one to formulate estimates for the expected error of the solution, they can assure us that a learning algorithm does something meaningful, and not just produce random outputs.

- They give an intuition about the quantities and structural properties which are essential for a learning process and therefore about which problems are inherently easier than others.

- They quantify the influence of parameters and indicate what prior knowledge is relevant in a learning setting and therefore they guide the analysis, design, and improvement of learning algorithms.

Thus, a measure of complexity in learning theory should reflect which learning problems are inherently easier than others. It should ideally reflect which parameters are relevant and how they influence the generalization ability of an algorithm. In physics and mathematics, for example, simple models are often those which can be described by only a few parameters. For instance, the class of all sinusoids is simple since it can be fully parameterized by amplitude, frequency and phase only. However, for a learning problem, it is intuitively clear that one can guess the right function from its values on an empirical sample only when there are not too many functions in the class which are similar on finite samples. The class of all sinusoids is therefore very complex since different functions can be fitted to have the same values on a finite set. A different notion of complexity from that of the number of parameters is necessary to characterize learning problems which quantifies how empirical errors converge to the expected ones.

The standard approach in statistical theory is to define the complexity of the learning problem through some notion of "richness", "size", "capacity" of the hypothesis class. One arrives at these notions of complexity based on the analysis of *uniform* deviations of the expectation from the empirical mean through probabilistic bounds of the form

$$Pr\left\{\sup_{f\in H_l}\left|\mathbb{E}_\mu f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right|\geq t\right\}\leq\delta\,. \tag{1.1}$$

While ignoring the choice of hypotheses from within the class made by a specific learning algorithm, statements of the form (1.1) allow one to bound the difference between the error on the sample and the expected error on unseen samples simultaneously for *all* functions in the hypothesis class and are therefore called *uniform bounds*. The bounds obtained hold thus for *any* function in the class, in particular for the one produced by a learning algorithm that chooses functions from this class. The *complexity* of the function class $H$ is the mathematical quantity through which the function $t=\gamma(n,\delta,H)$, for a given $\delta$, depends on $H$.

Since there is no universal measure of the "complexity" of a function class, much effort has been put into finding an adequate measure of complexity for a learning setting. Clearly, the complexity should distinguish between learnable and non-learnable classes. More important than the mere characterization of learnability are, however, the *rates of the convergence* of the empirical to the expected errors, that is $t = \gamma(n, \delta, H)$ as a function of the number of samples. They allow one to precisely compare the complexity of different models and to formulate, in a sample-dependent way, for each $\delta$ *confidence intervals* for the expected error of functions in $H_l$ (and as we will see later on, also of the hypothesis functions) in terms of the empirical error and the complexity of the class. They also lead to estimates on the *sample complexity* which is the number of samples needed to learn with a given accuracy $t$ and confidence $\delta$.

Learnable and non-learnable classes were already fully characterized, in the case when they are binary-valued, by Vapnik and Chervonenkis (1971), and for real-valued classes by Alon et al. (1993). The complexity measure proposed in Vapnik and Chervonenkis (1971), the *Vapnik-Chervonenkis (VC) dimension*, as well as the scale-sensitive *pseudo-dimension* in Pollard (1984) are *combinatorial* measures of the richness of classes of functions when evaluated on samples. A more accurate scale-sensitive combinatorial measure for classes of real-valued functions, the *fat-shattering dimension*, was presented by Pajor (1985), and shown to characterize learnability in Alon et al. (1993). [1] VC-dimension, pseudo-dimension, and fat-shattering dimension are independent of the underlying probability measure and of the particular sample, and hence are worst-case estimates with regard to these quantities.

Subsequently, one could show that VC-dimension and fat-shattering dimension can be employed to upper bound a complexity measure which depends on the distribution according to which the data is drawn, called the *metric entropy*, and which characterizes uniform Glivenko-Cantelli classes (e.g., Pajor 1985; Dudley 1987). Metric entropy is defined in terms of scale-dependent capacity notions for metric spaces of functions, the *covering numbers* . [2]

However, it is only recently that it has been understood how to define distribution-dependent complexity measures which lead not only to the tightest uniform confidence intervals and best sample complexities so far, but, more importantly and unlike for combinatorial and metric complexities, allow us to conceptually separate the influence of the complexity of the class, which guarantees learnability, from the influence of parameters which determine the confidence intervals and the sample complexity. These new complexity measures, called *Rademacher averages*, are obtained through symmetrization techniques from empirical process theory.

Unfortunately, the distribution of the data on which the Rademacher averages depend on is unknown. In order to make these estimates practically useful, one can approximate the unknown distribution with the empirical distribution and compute *empirical* versions of Rademacher complexities. Such empirical complexities are called

[1] Note that the fat-shattering dimension appears also – implicitly – in Talagrand (1987).

[2] Note that the first distribution-dependent complexity measure in learning theory, the *VC-entropy*, was proposed already in Vapnik and Chervonenkis (1971).

*data-dependent* because they can be computed, if the function class is known, from the values of the functions on the training sample only. They continue a whole line of research of empirical estimates of uniform complexity measures.

However, bounds in terms of uniform and empirical versions of uniform complexities are too pessimistic as they do not take into account the way the algorithm explores the function space and interacts with the actual sample. They allow one to bound the deviation of the empirical error on the observed sample and the expected error simultaneously for *any* function in the class, whereas one is actually only interested in the deviation for the function which a *particular* learning algorithm produces from *actual data*. Since this function depends on the actual sample it is itself *data-dependent*, and thus one cannot use directly the uniform bound approach for a "small" hypothesis class depending only on this single function. It turns out that it is often possible to improve the uniform bounds by using additional knowledge about the specific algorithm or the problem at hand. Especially, one can make use of prior knowledge to obtain tighter confidence intervals for the expected error by *conditioning* on the data. This is a fact known for a long time in the statistics community, where it is argued and exemplified that universally valid confidence intervals can be very bad for particular data sets. In such cases, prior knowledge about the distribution can be employed to improve the estimates by conditioning on data (Casella 1988; Robinson 1979; Berger 1985; Kiefer 1977).

In statistical machine learning a range of non-uniform complexities which make use of additional knowledge to quantify the complexity of the learning problem were proposed. For example, knowledge about the way certain algorithms choose their hypothesis was used to bound the generalization error directly for the function produced by these learning algorithm. Notable are the bounds for *compression schemes* (Littlestone and Warmuth 1986; Floyd and Warmuth 1995) and for *stable algorithms* (Kearns and Ron 1997; Devroye and Wagner 1979; Bousquet and Elisseeff 2002). Instead of the "size" of the function class, properties of the learning algorithms such as the *compression size* and the *stability constant* were employed to bound their generalization error. Other examples are Support Vector Machines (SVMs) and kernel methods, which are machine learning algorithms designed to maximize the size of the margin on the sample. It was observed that in certain settings the SVM tends to choose sparse solutions dependent only on a few support vectors (see, among others, Boser et al. 1992; Schölkopf 1997; Steinwart 2003; Bartlett and Tewari 2004). It was also shown that data-dependent complexities like the size of the margin or the number of support vectors for linear classifiers (Shawe-Taylor et al. 1998), or the empirical margin distribution for convex combinations of classifiers (Schapire et al. 1998) can be employed to bound the generalization error. [3] Recently, new complexities for voting classifiers

[3] The latter became particularly popular since it is an easy-to-compute and intuitive quantity, and

which measure the sparsity of the weights of convex combinations and clustering properties of the base functions were introduced in Koltchinskii et al. (2003); Koltchinskii and Panchenko (2005). Generalization bounds based on a notion of *diversity* of the base classifiers for boosting were derived in Long (2002); Dasgupta and Long (2003). Other data-dependent complexities were given for micro-choice algorithms and self-bounding algorithms (Freund 1998; Langford and Blum 1999), and for set covering machines (Marchand and Shawe-Taylor 2002). A data-dependent complexity governing the performance of online algorithms, the *online statistic*, was presented in Cesa-Bianchi et al. (2004).

Although these non-uniform complexities are not necessary conditions for learnability, their appeal is that they are easy to check and that they give an easy and intuitive quantification of prior knowledge about the learning problem. The data-dependent complexity notions like margin or clustering properties are very easy-to-compute and, being dependent on the actual sample, can capture directly the performance of the data-dependent hypothesis, seemingly without the assumption of a “small” hypothesis class. Whereas the mechanism which allows one to derive uniform complexity measures are better understood, when looking at non-uniform complexities it is still open which underlying mechanisms are involved in obtaining the non-uniform performance guarantees and which assumptions are intrinsically different from each other.

One systematic framework was proposed to explain the mechanism for some of these data-dependent results in form of a “luckiness theory” (see Section 4.3.4). The luckiness framework subsumes results in terms of some empirical uniform complexity measures, as well as margin, compression, and sparsity bounds. This framework was originally introduced in Shawe-Taylor et al. (1998) and further extended in Herbrich and Williamson (2003). It introduces a formal way to take advantage of prior knowledge about the link between the function class or the function learned by the algorithm and particular samples. It is based on metric entropy complexities and on a technically intricate condition which allows one to condition on the data and take advantage of the specific structure of the sample.

Another example for the way in which one can use additional knowledge about a specific algorithm is to take into account that specific algorithms are more likely to choose functions from “small” subclasses of the hypothesis class. An algorithm which takes a special place in learning theory is the Empirical Risk Minimization algorithm (ERM). ERM produces the function with the smallest error on the sample, and one can show that this function is very likely to have a small variance. This idea has led to tighter bounds on the generalization error of empirical risk minimizers in terms of *local complexity measures* (Koltchinskii and Panchenko 2000; Massart 2000b; Bous-

especially since it allows one to explain the good performance of practically successful algorithms like SVMs, kernel methods, neural networks, and voting algorithms like boosting and bagging.

quet 2002b; Bousquet et al. 2002; Bartlett et al. 2004a; Koltchinskii 2003; Lugosi and Wegkamp 2004; Bartlett and Mendelson 2005; Bartlett et al. 2004b). Local complexities measure the capacity of subclasses of functions with a small expectation or variance, as opposed to *global complexities* which measure the size of the entire function class. The size of the small "local subset", leading to optimal bounds, can be determined theoretically and, recently, Bartlett and Mendelson (2005) proposed a new and tighter notion of localization in terms of "belts" of functions of a given expectation. Since these complexity notions are distribution-dependent, an open question is whether it is possible to derive empirical estimates of these complexities.

**Introductory Bibliography:** There exist a number of textbooks covering topics of statistical learning theory and machine learning, among them Vapnik (1982, 1995, 1998); Devroye et al. (1996); Vidyasagar (1997); Anthony and Bartlett (1999); Christianini and Shawe-Taylor (2001); Duda et al. (2000); Herbrich (2002); Schölkopf and Smola (2002); Györfi et al. (2002). Notable surveys on learning theory are Kulkarni et al. (1998); Devroye and Lugosi (1995); Vapnik (1999); Herbrich and Williamson (2002); Cucker and Smale (2002); Anthony (2002); Mendelson (2003, 2005); Boucheron et al. (2004b); Bousquet et al. (2004); Massart (2003).

## 1.2 Contribution of the Thesis

This thesis contains two contributions on data-dependent generalization bounds.

First, we explore notions of complexities that allow one to derive non-uniform generalization bounds for *random* subclasses of hypothesis functions, that is, for function classes which depend on the sample. This is motivated by the fact that the function produced by a learning algorithm depends on the sample and is therefore a random function. We will then show that this framework is general enough to capture many of the previous frameworks which make use of additional knowledge about the learning algorithm and the training data, like compression, sparsity, and luckiness frameworks, and show the general underlying mechanisms which make these frameworks work.

Second, we continue existing investigations on data-dependent complexities for the Empirical Risk Minimization algorithm in terms of local complexity measures. We show how one can compute an empirical version of the complexities proposed in Bartlett and Mendelson (2005) and we then investigate the optimality of these empirical estimates.

## 1.3 Overview of the Thesis

This thesis is organized as follows. In Chapter 2 we present the general theoretical setting for the analysis of learning algorithms common in statistical learning theory. We first present the probabilistic model for the learning problem (Section 2.1), show how one can analyze the performance of learning algorithms by bounding their generalization error (Section 2.2), and present the main ideas in deriving generalization bounds based on results for uniform deviations of expectations and empirical averages (Section 2.3). We also present some examples for learning algorithms (Section 2.5). In Chapter 3 we then present the techniques used in deriving generalization bounds, namely concentration inequalities and symmetrization techniques. The next two chapters contain the contributions of this thesis. In Chapter 4 we develop the random subclass framework and present the examples which fall into this framework. Chapter 5 studies the data-dependent complexities for the Empirical Risk Minimization algorithm in terms of local complexity measures. Finally, Chapter 6 contains the conclusion of this thesis.

## 1.4 Notation and General Definitions

We end this chapter with some notation which will be used throughout the thesis. A glossary of symbols can be found at the end of the thesis.

### Sets and Vectors

If $S$ is a set, we denote its complement by $S^c$ and its power set (that is, the set of all subsets of $S$) by $\mathcal{P}(S)$ . The *indicator function* of the set $S$ is defined as $I_S(x) = 1$ , if $x \in S$ and 0 otherwise. For two sets $A, B \subseteq \mathbb{R}^n$ we denote by $A + B = \{a + b \ : a \in A,\ b \in B\}$. The *absolute convex hull* of a set $A$ is defined as

$$\text{absconv}(A) = \{\sum_{i=1}^{n} c_i a_i \ : \ n \in \mathbb{N},\, a_i \in A,\, \sum_{i=1}^{n} |c_i| \leq 1\}.$$

If $\mathbf{X} = (X_1, ..., X_n)$ is a vector then for $1 \leq j \leq m \leq n$ we denote by $\mathbf{X}|_{i=j}^{m}$ the vector $(X_j, ..., X_m)$ .

### Probabilities and Expectations

Let $\Omega$ be a measurable space [4] and let $\mu$ be a probability measure on $\Omega$. $\Omega^n$ denotes the product space $\Omega \times \cdots \times \Omega$ endowed with the product measure $\mu^n$. $Pr_\mu$ and $\mathbb{E}_\mu$ will denote the probability and the expectation with respect to $\mu$. We will use capital letters $X, Y, Z, \ldots$ for random variables and lower-case letters $x, y, z, \ldots$ for their observed values in a particular instance. For random vectors, we will use bold letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots$, and lower-case bold letters $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$ will denote particular instances of these random vectors.

[4] For an introduction to probability theory, see, for example, Shiryayev (1984).

In general, for any random variable $X$, $Pr_X$ and $\mathbb{E}_X$ will denote the probability and the expectation with respect to the distribution of $X$. Especially for conditional expectations, the index will indicate that we are conditioning on all remaining random variables, for example $\mathbb{E}_X f(X,Y) = \mathbb{E}[f(X,Y)|Y]$ . Sometimes, in order to keep notation simple, and when the probability measure is clear from the context, we will omit the subscripts. $Pr\{A\}$ will then denote the probability of the event $A$, where the probability is taken over all random variables in $A$. $\mathbb{E}(Z)$ will denote the expectation of the random variable $Z$ with respect to all random variables occurring in $Z$.

Let $F$ be a class of real-valued functions defined on $\Omega$ which take values in $[-b,b]$. For any vector $\mathbf{x} \in \Omega^n$,

$$F/\mathbf{x} = \{\big(f(x_1),\ldots,f(x_n)\big) : f \in F\}$$

is called the *coordinate projection* of the set $F$ onto the set of coordinates $\mathbf{x}$.

Let $X, X_1, \ldots, X_n$ be independent random variables distributed according to $\mu$. $Pr_{\mathbf{X}}$ and $\mathbb{E}_{\mathbf{X}}$ denote the probability and the expectation with respect to the random vector $\mathbf{X} = (X_1,\ldots,X_n)$ (and therefore with respect to $\mu^n$). $\mathbb{E}_\mu(f)$ is the expectation and $\mathrm{Var}(f)$ is the variance of the random variable $f(X)$. $\mu_n(\mathbf{X})$ denotes the random *empirical probability measure* supported on $(X_1,\ldots,X_n)$, that is,

$$\mu_n(\mathbf{X}) := n^{-1}\sum_{i=1}^{n} \delta_{X_i},$$

where $\delta_{X_i}$ denote the Dirac measures at $X_i$ in $\Omega$, that is, $\delta_{X_i} A = 1$ if $X_i \in A$ and 0 otherwise. Whenever $\mathbf{X}$ is clear from the context, we will denote the empirical measure by $\mu_n$.

We will denote the empirical expectation of $f$ on $(X_1,\ldots,X_n)$ by

$$\mathbb{E}_n(f) = \mathbb{E}_{\mu_n}(f) := n^{-1}\sum_{i=1}^{n} f(X_i).$$

Note that $\mathbb{E}_n(f)$ is a random variable, as it depends on the random variables $X_1,\ldots,X_n$.

### Stochastic Processes

Let $V$ be a subset of $\mathbb{R}^n$, let $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)$ be a vector of independent Rademacher variables (i.e. $Pr_{\mathbf{X}}\{\varepsilon_i = -1\} = Pr_{\mathbf{X}}\{\varepsilon_i = 1\} = 1/2$), and let $\mathbf{g} = (g_1,\ldots,g_n)$ be a vector of independent standard Gaussian random variables. The collection of random variables $\{\sum_{i=1}^n \varepsilon_i v_i : \mathbf{v} \in V\}$ and $\{\sum_{i=1}^n g_i v_i : \mathbf{v} \in V\}$ is called the *Rademacher (Gaussian) process indexed by the set* $V$. The random variables $\sup_{v\in V}\big|\sum_{i=1}^n \varepsilon_i v_i\big|$ and $\sup_{v\in V}\big|\sum_{i=1}^n g_i v_i\big|$ are called the *supremum of the Rademacher (Gaussian) process* *indexed by the set $V$.*

### Metric Spaces

The *Hamming distance* on $\mathbb{R}^n$ is defined as

$$d_H(x,y) = |\{i \,:\, 1 \leq i \leq n,\, x_i \neq y_i\}| \, .$$

Set $\ell_p^n$ to be $\mathbb{R}^n$ with the norm

$$\|x\|_p := \big(\sum_{i=1}^{n} |x_i|^p\big)^{1/p}$$

and put $B_p^n$ to be the unit ball of $\ell_p^n$. $\ell_\infty^n$ is $\mathbb{R}^n$ endowed with the norm

$$\|x\|_\infty := \sup_{1 \leq i \leq n} |x_i| \, .$$

Let $L_\infty(\Omega)$ be the set of bounded functions on $\Omega$ with respect to the norm

$$\|f\|_\infty := \sup_{\omega \in \Omega} |f(\omega)|$$

and denote its unit ball by $B\big(L_\infty(\Omega)\big)$. For a probability measure $\mu$ on a measurable space $\Omega$ and $1 \leq p < \infty$, let $L_p(\mu)$ be the space of measurable functions on $\Omega$ with a finite norm

$$\|f\|_{L_p(\mu)} := (\int |f|^p d\mu)^{1/p} \, .$$

For an introduction to normed function spaces and functional analysis, see, for instance, Reed and Simon (1980).

Let $(Y,d)$ be a metric space. If $F \subset Y$ then for every $\epsilon > 0$, $N(\epsilon, F, d)$ is the minimal number of open balls (with respect to the metric $d$) needed to cover $F$. A corresponding set $\{y_1, \ldots, y_m\} \subset Y$ of minimal cardinality such that for every $f \in F$ there is some $y_i$ with $d(f, y_i) < \epsilon$ is called an *$\epsilon$-cover of $F$*. The *$\epsilon$-entropy of $F$* is denoted by $H(\epsilon, F, d)$ and is defined as

$$H(\epsilon, F, d) := \log N(\epsilon, F, d) \, .$$

For $1 \leq p < \infty$, denote by $N\big(\epsilon, F, L_p(\mu_n)\big)$ the *covering number* of $F$ at scale $\epsilon$ with respect to the $L_p(\mu_n)$ norm. Similarly, one can define the *packing number* at scale $\epsilon$, which is the maximal cardinality of a set $\{y_1, \ldots, y_k\} \subset F$ such that for every $i \neq j$ , $d(y_i, y_j) \geq \epsilon$ . Denote the *$\epsilon$-packing numbers* by $M(\epsilon, F, d)$ and note that, for every $\epsilon > 0$,

$$N(\epsilon, F, d) \leq M(\epsilon, F, d) \leq N(\epsilon/2, F, d)\,.$$

### Measures of Complexity for Function Classes

If $F$ is a binary-valued class of functions, the *VC-dimension* of the function class $F$ will be denoted by $VC(F)$ and the *VC-entropy* by $H_{VC}\,(F)$. For general classes of functions, $\mathrm{fat}_\epsilon\,(F)$ denotes the *fat-shattering dimension at scale* $\epsilon$ (their exact definitions are given in Section 2.4.1).

Given a random vector $\mathbf{X}$, $F/\mathbf{X}$ denotes the random set

$$F/\mathbf{X} = \{\big(f(X_1), \ldots, f(X_n)\big) : f \in F\},$$

and is called the *coordinate projection* of the set $F$ onto the random set of coordinates $\mathbf{X}$. The empirical VC-dimension will be denoted by $\widehat{VC}\,(F, \mathbf{X})$, the empirical fat-shattering dimension by $\widehat{\mathrm{fat}}_\epsilon\,(F, \mathbf{X})$, and the empirical VC-entropy by $\widehat{H}_{VC}\,(F, \mathbf{X})$.

For every $f \in F$, we denote by $R_n\,(f, \mathbf{X}, \boldsymbol{\varepsilon})$ the *Rademacher sum* for $f$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent *Rademacher random variables*, that is, symmetric, $\{-1, 1\}$-valued random variables (see Definition 2.16). $R_n\,(f, \mathbf{X})$ is therefore a random variable that depends on the random variables $X_1, \ldots, X_n$ and on the Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$. For simplicity, we will sometimes write $R_n f$ instead of $R_n\,(f, \mathbf{X}, \boldsymbol{\varepsilon})$. The supremum $\sup_{f \in F} R_n\,(f, \mathbf{X}, \boldsymbol{\varepsilon})$ is called the *Rademacher penalty* of $F$. The *empirical Rademacher average*, the (global) *Rademacher average*, and the *uniform Rademacher average* of the class of functions $F$ are denoted by $\widehat{R}_n\,(F, \mathbf{X})$, $R_n\,(F)$, and $\overline{R}_n\,(F)$ respectively, and their exact definitions can be found in Section 2.4.1, page 26. Note that the empirical Rademacher average $\widehat{R}_n\,(F)$ is a random variable which depends on $X_1, \ldots, X_n$.

### Constants

Finally, throughout this thesis all absolute constants are denoted by $c$, $C$, or $K$. Their values may change from line to line, or even within the same line.

Chapter 2

# Preliminaries

## 2.1 The Learning Problem

We call a *learning problem* the task of finding a general rule which explains a given set of data. Such learning tasks arise naturally in many fields of science like engineering, physics, economics, biology, evolutionary science or cognitive science. In order to be able to infer from empirical training data to future unseen data and to find a solution for a learning problem, we have to assume that training data and unseen data represent the same underlying phenomenon. In statistical learning theory, the relationship between training and unseen data is modelled through a common underlying probability measure. This probabilistic model is presented in the next section.

### The Probabilistic Model

In this thesis, we will consider the following *probabilistic learning model* (Vapnik and Chervonenkis 1971; Valiant 1984; Haussler 1992). It relies on the fundamental assumption that both seen and future data are generated by the same fixed *underlying probability measure*, which, although unknown, allows us to infer from present data to future data and therefore to generalize.

**The probabilistic learning model**: Let $\Omega = \mathcal{X} \times \mathcal{Y}$ be a measurable space, and let $\mu$ be an unknown probability distribution on $\Omega$. The set $\mathcal{X}$ is called the input space, the set $\mathcal{Y}$ the output space. Let $((X_1, Y_1), \ldots, (X_n, Y_n)) \in \Omega^n$ be a finite training sample, where each pair $(X_i, Y_i)$ is generated independently [1] according to $\mu$. The goal of a learning algorithm is to find, based on this sample, a function $h : \mathcal{X} \longrightarrow \mathcal{Y}$ which predicts the most likely value of $Y \in \mathcal{Y}$ given $X \in \mathcal{X}$.

[1] Note that, besides assuming a probabilistic dependence between training and test data, we additionally make the assumption that each data sample is *independent*. This assumption is not always fulfilled in real-world applications. For instance, for time-series predictions such as stock market predictions, each new data point depends on the values of previous data points. If the dependence can be modelled by a Markov process, one can extend the classical probabilistic learning model to Markov chains, as for example in Gamarnik (1999).

When $\mathcal{Y}$ is finite the learning task is called *classification*, whereas when $Y$ is a subset of the real space $\mathbb{R}^n$ it is called *regression*.

### The Learning Algorithm and the Hypothesis Class

Let $\mathcal{Y}^{\mathcal{X}}$ denote the set of all functions from the input space $\mathcal{X}$ into the output space $\mathcal{Y}$, and let $\mathcal{P}(\mathcal{Y}^{\mathcal{X}})$ be the power set of $\mathcal{Y}^{\mathcal{X}}$. *A learning algorithm* $\boldsymbol{\mathcal{A}}$ is defined as a mapping from the set of all finite samples $Z \in \Omega^n$ to $\mathcal{Y}^{\mathcal{X}}$. In order to infer from the training data which function predicts best the relationship between input and output, we have to make additional assumptions within our model (see, e.g., Duda et al. 2000, "No Free Lunch"). Since machine learning algorithms usually restrict, implicitly or explicitly, the set of functions which they explore, we will here additionally assume that the functions produced by the learning algorithm belong to a fixed function class $H \subset \mathcal{Y}^{\mathcal{X}}$ specified in advance. This function class is called the *hypothesis class*. The selection of a *specific* hypothesis class is a part of modelling the learning problem *at hand*. For example, it is common to assume that the functions to be learned are linear or that they are convex combinations of a given base class of functions. The choice of a hypothesis class, through its "complexity", turns out to be crucial for analyzing the generalization properties of learning algorithms. To make the dependence of the learned function on the hypothesis class explicit, we define formally a learning algorithm as a mapping

$$\boldsymbol{\mathcal{A}} : \bigcup_{n=1}^{\infty} \Omega^n \times \mathcal{P}(\mathcal{Y}^{\mathcal{X}}) \longrightarrow \mathcal{Y}^{\mathcal{X}} .$$

For a given sample $\mathbf{z}$ and a given hypothesis class $H$, the function produced by algorithm $\boldsymbol{\mathcal{A}}$ is therefore $\boldsymbol{\mathcal{A}}(\mathbf{z}, H) \in H$. [2]

Statistical learning theory addresses the question of how to design learning algorithms which are, given a probabilistic model and a fixed hypothesis class, likely to produce "good" functions. A possible quantification of "good" is the topic of the next two sections.

### The Loss Function and the Loss Function Class

In order to make the most accurate choice for a function from the hypothesis class, we have to be able to assess its quality. In our learning model, we assume that the quantitative measure of the discrepancy between the predicted value and the actual value is given by a *loss function* $l : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$. The loss function is a choice of the

[2] Different from this model is the one used in the *PAC-Bayesian* framework, as the function produced by the algorithm does not belong directly to the hypothesis class $H$. The output of the algorithm is a distribution over the hypothesis class and the produced classifier is a weighted majority classifier computed according to this distribution, and therefore $\boldsymbol{\mathcal{A}}(\mathbf{z}, H) \notin H$.

modeller. For example, for classification, a common loss function is the *0-1 loss* defined by

$$l_{0-1}(r,s) = \begin{cases} 0 & \text{if } r = s, \\ 1 & \text{if } r \neq s. \end{cases}$$

Another common example is the *square-loss* for regression tasks:

$$l_2(r,s) = (r-s)^2.$$

The selection of an appropriate loss function is crucial in the design of a learning algorithm, and determines both the quality of the solution as well as the computational tractability. Some recent results on the impact of the loss function on quality and robustness in learning problems are presented in Christmann and Steinwart (2004); Bartlett et al. (2003); Zhang (2004b). In this thesis, we will not address these issues.

In what follows, we will always make one more additional assumption on the loss function, namely that the loss function is *bounded*, $l : \mathcal{Y}^2 \longrightarrow [-b,b]$ , for some $b$. This is a technical assumption which is required to hold for all the results presented in this thesis. [3] Strictly speaking, such an assumption is violated even for relatively simple cases, as for example in linear regression with the square-loss in a case where the true linear model is perturbed by additive Gaussian noise. In practical situations, however, since the observed data lies only in a limited range, it is sufficient to define the loss such that it is bounded in this range.

For every $h \in H$ we can now define the associated loss function

$$l_h : (\mathcal{X} \times \mathcal{Y}) \longrightarrow [-b,b], \qquad l_h(x,y) = l(h(x),y)$$

and denote by

$$H_l = \{l_h : (\mathcal{X} \times \mathcal{Y}) \longrightarrow [-b,b] \ : \ h \in H\}$$

the *loss class* associated with the learning problem. In the analysis of a learning problem, as we will see, it is often easier to think in terms of the associated loss class $H_l$ instead of the original hypothesis class $H$.

### Loss Minimization and Model Selection

The next step in modelling a learning problem is to choose a measure of the overall quality of a function. In statistical learning theory, the *risk* of a function (also called

[3] It is worth noting that it is possible to obtain results without this assumption but only with weaker probability estimates.

*expected error*) is defined as the expected loss:

$$\mathcal{R}(h) = \mathbb{E}_{\mu}\, l_h(X,Y)\,,$$

where the expectation is taken with respect to the probability measure $\mu$ on the data.

The best estimate $t^* \in \mathcal{Y}^{\mathcal{X}}$ is therefore the one for which the expected loss is as small as possible, that is,

$$t^* = \operatorname*{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(h)\ .$$

The function $t^*$ is called the *target hypothesis*. However, since our model restricts the functions produced by the algorithms to the hypothesis class $H$, the best the algorithm can do is to estimate the $h^* \in H$,

$$h^* = \operatorname*{argmin}_{h \in H} \mathcal{R}(h)\ .$$

We will assume in the following that such an $h^*$ exists.

The choice of a hypothesis class $H$, the *model selection*, determines the quality of the solution and is governed by the *approximation-estimation (or bias-variance) dilemma* (e.g., Duda et al. 2000, page 466). To see why we have a dilemma, we decompose

$$\mathcal{R}(h) - \mathcal{R}(t^*) = (\mathcal{R}(h) - \mathcal{R}(h^*)) + (\mathcal{R}(h^*) - \mathcal{R}(t^*))\ .$$

The second term on the right-hand side, $\mathcal{R}(h^*) - \mathcal{R}(t^*)$, is called *approximation error*. It is independent of the function $h$, and depends solely on the class $H$. Clearly, the larger the class $H$ is, the better $h^*$ can approximate $t^*$ and the approximation error is therefore smaller. The remaining term on the right-hand side, the *estimation error* $\mathcal{R}(h) - \mathcal{R}(h^*)$, depends on $h$. As will be shown in the sequel, the estimation error for the function produced by an algorithm $\mathcal{A}$ will be likely to increase if the "complexity" of the function class increases. A larger class will therefore imply a larger estimation error. The optimal choice of $H$ corresponding to a minimal $\mathcal{R}(\mathcal{A}(\mathbf{z}, H)) - \mathcal{R}(t^*)$ is therefore the result of a trade-off between approximation and estimation error.

It is possible to perform model selection by automatically trading off the complexity of the function class against the approximation error. In order to fully automatize the process of model selection, data-dependent complexities are of advantage, as one can compute the complexities entirely from the sample. Results in statistical learning theory regarding model selection based on data-dependent complexities can be found, for example, in Vapnik (1982); Kearns et al. (1997); Shawe-Taylor et al. (1998); Lugosi and Nobel (1999); Massart (2000b); Koltchinskii (2001); Bartlett et al. (2002); Cucker and Smale (2002); Massart (2003); Lugosi and Wegkamp (2004); Kääriäinen et al. (2004); Fromont (2004); Vito et al. (2005). The model selection process itself will not be discussed in detail in this thesis. We will focus here on the complexity measures which bound the estimation error and on ways to obtain these data-dependent complexities.

Our object of study is the estimation term $\boldsymbol{\mathcal{R}}(h) - \boldsymbol{\mathcal{R}}(h^*) = \mathbb{E}_\mu(l_h - l_{h^*})$ and therefore it is convenient to define a shifted loss class, the *excess loss class*,

$$H_l^* = \{l'_h = l_h - l_{h^*} : h \in H\}.$$

Note that all functions in $H_l^*$ have a nonnegative expectation, though they can take negative values, and that $0 \in H_l^*$. The minimizers $h^*$ correspond to the functions $l'_h \in H_l^*$ with minimal expectation $\mathbb{E}_\mu l'_h = 0$. Unfortunately, since the measure $\mu$ is unknown, it is not possible to determine $h^*$ and $H_l^*$. Learning algorithms therefore try to estimate the function $h^*$ on the basis of empirical data.

**Remark:** As a matter of fact, the function produced by a learning algorithm obviously depends also on the loss function $l$, as $h^*$ will depend on $l$. To be precise, we should therefore write $\boldsymbol{\mathcal{A}}(\mathbf{z}, H, l)$ instead of $\boldsymbol{\mathcal{A}}(\mathbf{z}, H)$. Because the influence of the loss function on the solution is not a focus of this thesis, and in order to keep notation simple, we will keep the notation $\boldsymbol{\mathcal{A}}(\mathbf{z}, H)$ and remind the reader whenever necessary about the dependence of the hypothesis on the loss.

### Analysis of Algorithms

With the above preliminaries, we can now assert that a good learning algorithm is one which, when presented with a random training sample $\mathbf{Z}$ and restricted to hypotheses from the class $H$, approximates $h^*$ well and thus produces a function with a small risk, $\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)) \approx \boldsymbol{\mathcal{R}}(h^*)$. Recall that the expected errors of the functions in $H$ cannot be computed because the probability distribution $\mu$ is unknown. In the standard learning model, the only information available about the expected errors is through the errors $l_h(X_i, Y_i)$ on the finite empirical sample. Learning algorithms approximate $h^*$ by looking only at the sample, and we would like to know how good this approximation is. Statistical learning theory is concerned with guarantees on a good behaviour of the risk $\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H))$.

The first type of guarantee comes from analyzing the *consistency* of a learning algorithm. A learning algorithm is *consistent* if the expected error of its solution converges in probability, in the limit of an infinite sample, to the expected error of the target hypothesis $t^*$ (for precise definitions of different variants of consistency, see, e.g., Devroye et al. 1996, Chapter 6, page 92). Vapnik and Chervonenkis (1971) gave a complete characterization for the consistency of the Empirical Risk Minimization algorithm in case of binary-valued classes (see Section 2.5, page 31) in terms of the "complexity" of the hypothesis class measured by the VC-dimension. Consistency for other algorithms was shown, for example, for k-Nearest-Neighbour methods (Devroye et al. 1996), spectral clustering (von Luxburg et al. 2005), kernel methods (Steinwart 2002, 2005), boosting algorithms (Mannor et al. 2002; Lugosi and Vayatis 2004; Zhang 2004b; Jiang 2004), and for classifiers using a convex loss function (Bartlett et al. 2003). Consistency is a property one would like an algorithm to have – however, it is a weak property as it does not say anything about the behaviour of the algorithm when presented with finite samples.

Stronger guarantees for the behaviour of the risk $\mathcal{R}(\mathcal{A}(\mathbf{Z}, H))$ and therefore for the performance of a learning algorithm can be obtained through probabilistic *finite sample generalization bounds*. Finite sample bounds depend on the sample size and can give an a-priori estimate on the number of samples needed for a given performance and thus allow one to estimate the value range for the expected error of the solution. From finite sample generalization bounds one can often derive consistency results. These bounds and the parameters which appear in them are the topic of this thesis. In the next chapter, we will present how generalization bounds can be employed as a measure of performance for learning machines. [4]

## 2.2 Generalization Bounds as Performance Measure for Algorithms

In statistical learning theory, one quantifies the behaviour of a learning algorithm through the value of the expected loss $\mathcal{R}(\mathcal{A}(\mathbf{Z}, H))$, where $\mathbf{Z} \in \Omega^n$ is a random sample of $n$ independently and identically distributed (i.i.d.) random variables. Because the loss $\mathcal{R}(\mathcal{A}(\mathbf{Z}, H))$ depends on random data, it is itself a random variable, and the statements which can be made about $\mathcal{R}(\mathcal{A}(\mathbf{Z}, H))$ are probabilistic. The performance of learning algorithms can be evaluated through *generalization bounds*. Generalization bounds are probabilistic statements of the following form:

$$Pr_{\mathbf{Z}}\{\, \mathcal{R}(\mathcal{A}(\mathbf{Z}, H)) \, \geq \, \gamma \,\} \, \leq \, \rho \, . \tag{2.1}$$

For a fixed probability of generalization error $\rho$, $\gamma$ is a function (which has to be determined) and the probability $Pr_{\mathbf{Z}}$ is taken with respect to the random vector $\mathbf{Z} \in \Omega^n$ and therefore with respect to the measure $\mu^n$. The quantities appearing in $\gamma$, apart from $n$ and $\rho$, are called the *complexity measures* for the generalization performance of the algorithm $\mathcal{A}$.

[4] It is worth mentioning that the model presented here is just one way to assess performance for learning algorithms. Other theoretical models were proposed, which define the success of learning algorithms differently. Mistake bounds (Littlestone and Warmuth 1986; Littlestone 1989) are one notable example of a non-probabilistic performance measure in which the learner is evaluated by the total number of mistakes it makes before it converges to the correct hypothesis. It is possible to design algorithms for specific tasks which have a guaranteed performance in this sense (Littlestone 1989).

In general, statistical learning theory studies the following aspects regarding $\gamma$:

- *functional form of* $\gamma$, in the sense that the "right" quantities which characterize the difficulty of the learning problem appear in the bound;
- *rates of convergence* as a function of sample size $n$;
- *sample complexity*, that is, how many samples are necessary to learn with a given accuracy $\gamma$ and a given confidence $\rho$;
- *relation* between various complexity terms.

Statements of the form (2.1) allow us to build *confidence intervals* $(-\infty; \gamma]$ for the expected loss $\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H))$. Confidence intervals for the expected value of a function are studied in a branch of statistics called theory of interval estimation (e.g., Stuart et al. 1999). A confidence interval gives an estimated range of values which is likely to include an unknown parameter of a distribution, the estimated range being calculated from a given set of data sampled from this distribution.

**Definition 2.1 (confidence interval)** *Let $X$ be a random variable depending on a real-valued parameter $\theta$. A confidence interval for $\theta$ for some confidence $0 < \alpha < 1$ is the smallest interval $C(X, \alpha)$ which depends on the variable $X$ and which satisfies*

$$\forall \theta : \quad P_X[\theta \in C(X, \alpha)] \geq 1 - \alpha \,,$$

*(e.g., Stuart et al. 1999).*

In learning theory, the estimated parameter $\theta$ is the conditional expectation of the random variable $l_{\boldsymbol{\mathcal{A}}(\mathbf{Z},H)}(\mathbf{Z})$. For each algorithm and hypothesis class, a good bound of the form (2.1) is one in which the function $\gamma$ gives a tight confidence interval for $\theta$.

In the following, we will sketch the standard way of deriving generalization bounds in statistical learning theory. Recall that we would like to compute an expectation – that of the loss of a function – from empirical data without knowing the distribution of the data. The Law of Large Numbers suggests to approximate the expected loss through the average loss on an empirical sample and to try to quantify how far away from each other these two quantities are. If $\mathbf{Z} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is a random training sample, then the average error of a hypothesis function $h$ on the given sample $\mathbf{Z}$ is called the *empirical error* and will be denoted by

$$\widehat{\boldsymbol{\mathcal{R}}}(h, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} l_h(X_i, Y_i) \,.$$

## 2.3 Uniform Bounds and Suprema of Empirical Processes

In a number of "classical" generalization bounds, $\gamma$ will depend on the algorithm $\boldsymbol{\mathcal{A}}$ and the sample $\mathbf{Z}$ *only via the empirical error* $\widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{Z})$. Their derivation, as proposed first in Vapnik and Chervonenkis (1971), follows the following "recipe":

- $\boldsymbol{\mathcal{R}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H))$ is rewritten as a sum of the empirical error and the deviation between expected error and empirical error,

$$\boldsymbol{\mathcal{R}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)) = \widehat{\boldsymbol{\mathcal{R}}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H), \mathbf{Z}) + \big[\boldsymbol{\mathcal{R}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)) - \widehat{\boldsymbol{\mathcal{R}}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H), \mathbf{Z})\big];$$

- the empirical error $\widehat{\boldsymbol{\mathcal{R}}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H), \mathbf{Z})$ can be computed from the specific sample at hand; the object of study remains the quantity $\boldsymbol{\mathcal{R}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)) - \widehat{\boldsymbol{\mathcal{R}}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H), \mathbf{Z})$ (see Figure 2.1);
- the quantity $\boldsymbol{\mathcal{R}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)) - \widehat{\boldsymbol{\mathcal{R}}}\,(\boldsymbol{\mathcal{A}}(\mathbf{Z}, H))$ is replaced by its worst case estimate, that is, by the *uniform* estimate over the whole set of hypothesis functions $\sup_{h\in H} \boldsymbol{\mathcal{R}}\,(h) - \widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{Z})$ ;
- a *uniform bound* of the following form is derived:

$$Pr_{\mathbf{Z}}\big\{ \sup_{h\in H} \big|\boldsymbol{\mathcal{R}}\,(h) - \widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{Z})\big| \;\geq\; \gamma(n, \rho, H, l) \,\big\} \;\leq\; \rho\,, \tag{2.2}$$

  where $\gamma$ depends only on the confidence $\rho$, the sample size $n$, and on the loss class $H_l$.

The bound obtained holds for *any* hypothesis function $h \in H$, in particular for $\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)$.

How can we derive a uniform bound? Since $\rho$ does not depend on the random variable $\mathbf{Z}$, we can rewrite the statement (2.2) to get an "inverse" form

$$Pr_{\mathbf{Z}}\big\{ \sup_{h\in H} \big|\boldsymbol{\mathcal{R}}\,(h) - \widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{Z})\big| \;\geq\; t \,\big\} \;\leq\; \rho(n, t, H, l)\,,$$

where $t > 0$, and $\rho$ is a function which has to be determined. Recalling the definition of the expected and empirical error and of the loss class, this is equivalent to

$$Pr_{\mathbf{Z}}\big\{ \sup_{f\in H_l} \big|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\big| \geq t \,\big\} \leq\; \rho(n, t, H, l)\,, \tag{2.3}$$

where $\mathbf{Z} = (Z_1, ..., Z_n)$ . As we will see in Section 2.3, statements of the form (2.3) are called *tail probabilities* of a random variable. Random variables of the specific form as in (2.3) are called *empirical processes*, and it will be possible to tackle (2.3) with tools from empirical process theory (namely symmetrization and concentration inequalities). The quantities which will appear in $\rho$, as they only depend on the model

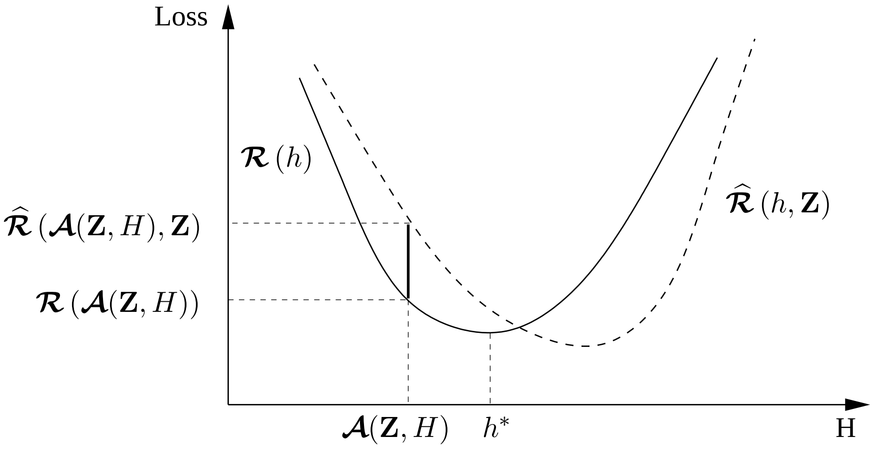

Loss
$\mathcal{R}(h)$
$\widehat{\mathcal{R}}(\mathcal{A}(\mathbf{Z},H),\mathbf{Z})$
$\mathcal{R}(\mathcal{A}(\mathbf{Z},H))$
$\widehat{\mathcal{R}}(h,\mathbf{Z})$
$\mathcal{A}(\mathbf{Z},H)$
$h^*$
H

**Figure 2.1:** Expected and empirical error for the function produced by an algorithm $\mathcal{A}$ for a given sample $\mathbf{Z}$ from a given hypothesis class $H$ (picture as in Bousquet et al. 2004). One way to study the behaviour of the random variable $\mathcal{R}(\mathcal{A}(\mathbf{Z},H))$ is by comparison with its empirical approximation $\widehat{\mathcal{R}}(\mathcal{A}(\mathbf{Z},H),\mathbf{Z})$. This is usually done by characterizing the behaviour of their difference $|\mathcal{R}(\mathcal{A}(\mathbf{Z},H)) - \widehat{\mathcal{R}}(\mathcal{A}(\mathbf{Z},H),\mathbf{Z})|$.

through the loss class $H_l$, are called the complexity measures (the "size" or "capacity") of $H_l$. The dependency on $n$ will determine both the rate of convergence and the sample complexity. If $\rho(n,t,H,l)$ goes to 0 as $n$ goes to infinity, this implies *learnability.*

A variation in the derivation of uniform bounds is to employ *relative* or *reweighted* tail inequalities, which allow one to derive bounds of the form

$$Pr_{\mathbf{Z}}\Big\{ \sup_{h\in H} \frac{\mathcal{R}(h) - \widehat{\mathcal{R}}(h,\mathbf{Z})}{\omega(h)} \geq t \Big\} \leq \rho(n,t,H,l)\,. \tag{2.4}$$

For some of these, one can derive empirical versions of relative inequalities, in which $\omega(h)$ is replaced by a sample-dependent $\hat{\omega}(h,\mathbf{Z})$. In these inequalities, $\omega$ is usually a function related to the variance $\mathrm{Var}\,(h)$ or the expectation $\mathbb{E}\,(h)$. Such bounds allow one to get tighter confidence intervals for classes in which functions have small variances or small expectations, which supports the intuition that such functions have less variation on samples and are therefore easier to learn (see also the results on localized complexities for ERM from Section 4.3.5 and Section 5.3).

## 2.4 Uniform Complexity Measures

### 2.4.1 Uniform Complexity Measures – Definitions

Let $F$ be a class of real-valued functions defined on a probability space $(\Omega,\mu)$. In this section, we will present the definition of a range of complexity measures from statistical learning theory which reflect the "size" or "richness" of the class $F$. First, we define the coordinate projections of a class of functions onto a set.

**Definition 2.2 (coordinate projections)** *The set of all patterns realized by the function class $F$ on a set $\mathbf{x} = \{x_1, \ldots, x_n\}$ are the coordinate projections,*

$$F/\mathbf{x} = \big\{ (f(x_1), \ldots, f(x_n)) \,:\, f \in F \big\} \ .$$

*If $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector, we denote by $F/\mathbf{x}$ the coordinate projections of $F$ onto the set $\{x_1, \ldots, x_n\}$ .*

### Combinatorial Complexity Measures for Binary-Valued Classes of Functions

Let $F$ be a set of binary-valued functions from $\Omega$ to $\{0,1\}$. For binary-valued classes, the cardinality of the coordinate projections is the *shattering coefficient.*

**Definition 2.3 (shattering coefficient)**

$$S_n(F, \mathbf{x}) = |F/\mathbf{x}|.$$

Vapnik and Chervonenkis (1971) introduced, for binary-valued classes, the first combinatorial notion of dimension based on the "richness" of the coordinate projections of $F$. A class is defined as "rich" when its projections on samples contain entire combinatorial cubes $\{0,1\}^n$ of large dimensions. A sample set $\mathbf{x}$ is said to be *shattered* by $F$ if the projection $F/\mathbf{x}$ is the entire combinatorial cube.

**Definition 2.4 (shattered set)** *The set $\mathbf{x} \in \Omega^n$ is shattered by the binary-valued class $F$ if $S_n(F, \mathbf{x}) = 2^n$.*

One can now define a combinatorial dimension for $F$ by looking for the worst case, that is, for the largest combinatorial cube contained in any of the coordinate projections of $F$. The VC-dimension is the largest dimension of a coordinate projection which is the entire combinatorial cube.

**Definition 2.5 (Vapnik-Chervonenkis dimension (VC-dimension))** *The VC-dimension is the size of the largest set shattered by $F$.*

$$\begin{aligned} VC(F) &= \max \left\{ n \in \mathbb{N} \ : \ \max_{\mathbf{x} \in \Omega^n} S_n(F, \mathbf{x}) = 2^n \right\} \\ &= \max \{ n \in \mathbb{N} \ : \ \exists \mathbf{x} \in \Omega^n \text{ such that } \mathbf{x} \text{ is shattered by } F \} \ . \end{aligned}$$

The VC-dimension ignores the underlying measure on the data and is therefore distribution-independent. [5]

[5] A more detailed account on the VC-dimension and its value for specific classes are given, for example, in Anthony and Bartlett (1999); Dudley (1999).

Let $\mathbf{X} = (X_1, ..., X_n)$ denote an i.i.d. sample distributed according to $\mu^n$. For each sample $\mathbf{X}$, we can define an empirical counterpart of the VC-dimension, which has the following two advantages. First, it is computable entirely from the sample at hand, and it can therefore be estimated for many classes for which the VC-dimension is hard to compute. Second, the empirical VC-dimension is smaller or equal to the worst-case distribution-independent VC-dimension.

**Definition 2.6 (empirical VC-dimension)** *The empirical VC-dimension is the VC-dimension of $F$ restricted to the domain* $\mathbf{X}$.

$$\widehat{VC}\,(F, \mathbf{X}) = \max\left\{|\mathbf{x}| \ : \ \mathbf{x} \subseteq \mathbf{X},\, |F/\mathbf{x}| = 2^{|\mathbf{x}|}\right\}.$$

The *VC-entropy* is the expectation of the logarithm of the shattering coefficient.

**Definition 2.7 (VC-entropy)**

$$H_{VC}\,(F) = \mathbb{E}_{\mathbf{X}}\,\left(\log S_n(F, \mathbf{X})\right)\,.$$

Although the VC-entropy is data-independent, it is *distribution-dependent* since it depends on the underlying distribution generating $\mathbf{X}$. The empirical version of the *VC-entropy* is defined as

**Definition 2.8 (empirical VC-entropy)**

$$\widehat{H}_{VC}\,(F, \mathbf{X}) = \log S_n(F, \mathbf{X})\,.$$

If $\mathbf{X}$ is a random variable, then the empirical VC-dimension and the empirical VC-entropy are random variables. These random variable were shown in Boucheron et al. (2000) to be concentrated around their mean (see also Theorem 3.13, page 46). Empirical VC-entropy and VC-entropy are thus, with high probability, similar.

The shattering coefficient can be upper bounded through the VC-dimension, as stated by the following theorem (e.g., Vapnik and Chervonenkis 1971):

**Theorem 2.9 (Sauer-Shelah lemma)**

$$S_n(F, \mathbf{X}) \leq \sum_{i=1}^{\widehat{VC}(F,\mathbf{X})} \binom{n}{i} \leq \sum_{i=1}^{VC(F)} \binom{n}{i}$$

Its significance is that it shows that a finite VC-dimension $VC(F) = d$ implies polynomial growth of order $O(n^d)$, rather than the potential exponential growth $O(2^n)$, of the size of projections, since $\sum_{i=1}^{d} \binom{n}{i} \leq (en/d)^d \leq (n+1)^d$ .

A direct application of the Sauer-Shelah lemma and the definition of the empirical VC-dimension allow us to obtain the following relation between shattering coefficient and empirical VC-dimension:

**Corollary 2.10 (Comparison shattering coefficient and empirical VC)** *For any class $F$ and any sample $\mathbf{X} \in \Omega^n$,*

$$2^{\widehat{VC}(F,\mathbf{X})} \leq S_n(F,\mathbf{X}) \leq (n+1)^{\widehat{VC}(F,\mathbf{X})}.$$

### Combinatorial Complexity Measures for General Classes of Functions

Let $F$ be a set of functions from $\Omega$ to $\mathbb{R}$. One can generalize the definition of a shattered set to a scale-sensitive version for real-valued functions. Unlike in the binary-valued case, where we look for combinatorial cubes in the projection, we now require that the real-valued projections "contain" cubes of a given size $\epsilon$. Sets in $\Omega^n$ are called *$\epsilon$-shattered*, if there are some real numbers $s_1, s_2, \ldots, s_n$ (defining the center of the cube) such that the functions in $F$ can realize all possible "above/below by at least $\epsilon$" combinations around the $s_i$.

**Definition 2.11 ($\epsilon$-shattered set)** *The set $\mathbf{x} \in \Omega^n$ is $\epsilon$-shattered by $F$ if there are real numbers $s_1, s_2, \ldots, s_n$ such that for each $\mathbf{I} \in \{0,1\}^n$ there is some $f_{\mathbf{I}} \in F$ for which*

$$f_{\mathbf{I}}(x_i) \begin{cases} \geq s_i + \epsilon & \text{if } \mathbf{I}_i = 1, \\ \leq s_i - \epsilon & \text{if } \mathbf{I}_i = 0, \end{cases}$$

*for $i = 1, \ldots, n$.*

One can then extend the combinatorial dimension to the real-valued case by defining the complexity to be the largest dimension of a projection containing a real-valued cube of a given size. The shattering coefficient quantifies the relationship between the size of the cube and the dimension of the largest projections containing a full cube.

**Definition 2.12 (fat-shattering dimension)**

$$\mathrm{fat}_\epsilon(F) = \max\{n \in \mathbb{N} \ : \ \exists \mathbf{x} \in \Omega^n \text{ such that } \mathbf{x} \text{ is } \epsilon\text{-shattered by } F\}\ .$$

The empirical version is defined as

**Definition 2.13 (empirical fat-shattering dimension)**

$$\widehat{\mathrm{fat}}_\epsilon(F,\mathbf{X}) = \max\{n \in \mathbb{N} \ : \ \exists \mathbf{x} \in \mathbf{X},\ |x| = n,\ \text{ such that } \mathbf{x} \text{ is } \epsilon\text{-shattered by } F\}\ .$$

Like VC-dimension and VC-entropy, it was shown in Boucheron et al. (2000) that the fat-shattering dimension is highly concentrated around its expectation (see also Theorem 3.13, page 46).

### Metric Complexities

It is possible to define complexity measures of a class of functions $F$ which are related to the metric structure of the function class. Given a random sample $\mathbf{X}$, recall that $\mu_n$ denotes the random empirical probability measure supported on $\mathbf{X}$ and that $L_p(\mu_n)$ is the metric which corresponds to the $L_p$ norm of the coordinate projections $F/\mathbf{X}$.

**Definition 2.14 (metric entropy, covering numbers)** *The metric entropy of $F$ is defined as the logarithm of the covering numbers of $F$ with respect to the $L_p(\mu_n)$ metric,*

$$H(\epsilon, F, L_p(\mu_n)) = \log N(\epsilon, F, L_p(\mu_n)) \,.$$

These complexities are random variables because $\mu_n$ is supported on random samples. Uniform versions of covering numbers and metric entropies which are worst-case with respect to samples are defined as

**Definition 2.15 (uniform metric entropy, uniform covering numbers)**

$$N_p(\epsilon, F, n) = \sup_{\mu_n} N(\epsilon, F, L_p(\mu_n)); \qquad\qquad H_p(\epsilon, F, n) = \log N_p(\epsilon, F, n) \,.$$

### Rademacher Averages

**Definition 2.16 (Rademacher sums)** *Let $\varepsilon_1, \ldots, \varepsilon_n$ denote independent Rademacher random variables (i.e. $Pr_{\mathbf{X}}\Big\{\varepsilon_i = -1\Big\} = Pr_{\mathbf{X}}\Big\{\varepsilon_i = 1\Big\} = 1/2$). The Rademacher sum for $f \in F$ and $\mathbf{X} = (X_1, \ldots, X_n) \in \Omega^n$ is the absolute value of the Rademacher process indexed by $(f(X_1), \ldots f(X_n))$ defined as*

$$R_n\left(f, \mathbf{X}, \boldsymbol{\varepsilon}\right) = \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \,.$$

**Definition 2.17 (Rademacher penalties)** *The supremum of the Rademacher process indexed by the coordinate projections of $F$ onto $\mathbf{X}$,*

$$\sup_{\mathbf{v} \in F/\mathbf{X}} \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| = \sup_{f \in F} R_n\left(f, \mathbf{X}, \boldsymbol{\varepsilon}\right)$$

*is called the Rademacher penalty of $F$.*

Note that Rademacher sums and Rademacher penalties are random variables that depend on both the random variables $X_1, \ldots, X_n$ and on the Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$.

We define the *empirical Rademacher averages* as expectations over $\varepsilon_1, \ldots, \varepsilon_n$ of Rademacher penalties:

**Definition 2.18 (empirical Rademacher averages)**

$$\widehat{R}_n(F, \mathbf{X}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in F} R_n(f, \mathbf{X}, \boldsymbol{\varepsilon}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \left( \sup_{f \in F} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right).$$

$\widehat{R}_n(F, \mathbf{X})$ are therefore random variables which depend on $X_1, \ldots, X_n$.

**Definition 2.19 (Rademacher averages)** *The Rademacher averages of the class $F$ are defined as*

$$R_n(F) = \mathbb{E}_{\mathbf{X}} \left( \widehat{R}_n(F, \mathbf{X}) \right) = \mathbb{E}_{\mathbf{X}, \boldsymbol{\varepsilon}} \left( \sup_{f \in F} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right).$$

*In the last equality the expectation is taken with respect to all random variables $X_i$ and $\varepsilon_i$.*

(Note that in the literature Rademacher averages are often defined to be normalized by $1/n$ or $1/\sqrt{n}$.) Rademacher penalties, empirical Rademacher averages, and Rademacher averages are likely to be similar for large sample sizes $n$, as they are concentrated around their means (e.g., Koltchinskii 2001; Mendelson 2003).

An analogous measure are the Gaussian averages, where $\varepsilon_i$ are replaced by Gaussian random variables. (Again, we define them here without normalizing by $1/n$ or $1/\sqrt{n}$.)

**Definition 2.20 (Gaussian averages)** *The Gaussian averages of the class $F$ are defined as*

$$G_n(F) = \mathbb{E}_{\mathbf{X}, \mathbf{g}} \left( \sup_{f \in F} \left| \sum_{i=1}^{n} g_i f(X_i) \right| \right),$$

*where $g_i$ are standard Gaussian random variables, and the expectation is taken with respect to all random variables $X_i$ and $g_i$.*

One can also define uniform versions of the Rademacher averages which are worst-case instead of averaged with respect to the sample (e.g., Mendelson 2002b,c, 2005).

**Definition 2.21 (uniform Rademacher averages)**

$$\overline{R}_n(F) = \sup_{\mathbf{X} \in \Omega^n} \widehat{R}_n(F, \mathbf{X}).$$

### 2.4.2 Bounds with Uniform Complexities

In the following we assume that $F = H_l$ is a class of real-valued functions which arises as the loss class associated with a hypothesis class $H$ and a bounded loss function $l$. Without loss of generality, we will assume in the following that the loss is bounded by 1. This implies that functions in $F$ take only values in $[-1, 1]$.

In this section we will present the standard results in terms of uniform complexities, and show how they relate to each other. The exposition in this section follows closely the one in Mendelson (2005).

#### Bounds with Metric and Combinatorial Complexities

The standard approach in learning theory for deriving bounds for uniform deviations of expectations from empirical averages was developed in Vapnik and Chervonenkis (1971) and is based on symmetrization techniques (see Section 3.2) and concentration results for *single* functions (i.e. Hoeffding's or Bernstein's inequality, see Section 3.1). The basic idea in this approach is to replace the possibly infinite set of functions $F = H_l$ with a finite set of functions which "approximates" $F$. One combines then the concentration results for each of the functions in the finite set through the union bound. Thus, the cardinality of this finite set will reflect the complexity of $F$. As we will show later, such an approach is suboptimal, because the functions in the set can behave similarly and by treating their deviation separately the union bound is potentially loose.

The first result we present is a bound involving covering number estimates of the class $F$ with respect to the $L_p(\mu_n)$ metric. The classical VC-type bounds in terms of the VC-dimension and fat-shattering shattering dimension follow from this bound, since one can upper bound covering numbers through combinatorial dimensions.

If $F$ has finite covering numbers with respect to any $L_p(\mu_n)$ metric, one can approximate $F$ by a cover in $L_p(\mu_n)$ and obtain the following theorem. Its proof can be found, for example, in Anthony and Bartlett (1999), page 143.

**Theorem 2.22 (Covering bound)** *Let $F$ be a class of real-valued functions defined on a measurable space $\Omega$ and which take values in $[-1, 1]$. Then, for any probability measure $\mu$, any $p > 1$, every $0 < t < 1$, and any $n \geq 8/t^2$,*

$$Pr_{\mathbf{X}} \left\{ \sup_{f \in F} \ \big| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \big| \geq t \right\} \leq 8 \mathbb{E}_\mu \left\{ N(t/8, F, L_p(\mu_n)) \right\} e^{-\frac{nt^2}{128}},$$

*where $\mathbf{X} = (X_1, ..., X_n)$ is an i.i.d. sample distributed according to $\mu^n$ .*

The following confidence interval and sample complexity estimates follow directly:

**Corollary 2.23** *With $F$ defined as above, for any probability measure $\mu$, every $0 < t < 1$, and any $n \geq 8/t^2$, with probability at least $1 - \delta$,*

$$\sup_{f \in F} \ \Big| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \Big| < \sqrt{\frac{128}{n} \left( H_p(t/8, F, n) + \log \frac{8}{\delta} \right)} \, .$$

**Corollary 2.24** *With $F$ defined as above, for any probability measure $\mu$, any $0 < t < 1$, and any $\delta < 1$, if*

$$n \geq \frac{128}{t^2} \left( H_p \left( t/8, F, n \right) + \log \frac{8}{\delta} \right) ,$$

*then $Pr_{\mathbf{X}} \left\{ \sup_{f \in F} \ \left| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \right| \geq t \right\} \leq \delta$ .*

As shown in Mendelson (2005), one can often relate the uniform entropy of the loss class $F$ directly to that of the hypothesis class $H$. For example when $H$ contains only functions taking values in $[0, 1]$ and $l$ is the square-loss, then $N_p(t, F, n) \leq N_p(t/4, H, n)$. For classes of binary-valued functions this result can be combined with the following theorem due to Dudley which shows that one can control the uniform covering numbers through the VC-dimension.

**Theorem 2.25 (Uniform covering numbers and VC-dimension)** *For any $1 \leq p < \infty$, there are constants $c_p$ which satisfy that for any set of binary functions $F$ with $VC(F) = d$, any $0 < \epsilon < 1$, and any $n$,*

$$N_p(\epsilon, F, n) \leq \left( c_p \log \frac{2}{\epsilon} \right)^d \epsilon^{-pd} \, .$$

Note that the estimate from Theorem 2.25 was further improved by Haussler (1995) to an estimate where the $\log(1/\epsilon)$ factor is removed. As a consequence, one recovers the original results from Vapnik and Chervonenkis (1971) in terms of the VC-dimension.

**Corollary 2.26 (VC bound)** *There exists a fixed constant $c$ such that, for any set of binary functions $F$ with $VC(F) = d$, and any probability measure $\mu$, with probability at least $1 - \delta$,*

$$\sup_{f \in F} \ \Big| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \Big| < \sqrt{\frac{c}{n} \left( d + \log \frac{1}{\delta} \right)} \, .$$

One can also bound the uniform entropy numbers in terms of the fat-shattering dimension (though the proofs are far more involved). Estimates on the $L_\infty$ covering numbers were first given in Alon et al. (1997) and sharpened in Rudelson and Vershynin (2005), and on $L_p$ covering numbers in Pajor (1985); Talagrand (1992); Mendelson and Vershynin (2003). For $L_p$ covering numbers we can state the following theorem:

**Theorem 2.27** *For any $1 \leq p < \infty$, there are constants $c_p$ and $k_p$ which satisfy that for any set of real-valued functions $F$ bounded by $[-1, 1]$, any probability measure $\mu$, any $0 < \epsilon < 1$, and any $n$,*

$$N_p(\epsilon, F, n) \leq \left(\frac{2}{\epsilon}\right)^{k_p \mathrm{fat}_{c_p \epsilon}(F)} .$$

These estimates recover the standard estimates in terms of the fat-shattering dimension, as found for example in Anthony and Bartlett (1999).

### Bounds with Rademacher Averages

A much sharper concentration result can be derived by combining symmetrization with Talagrand's concentration inequality (Theorem 3.12) directly for the suprema of empirical processes. The following theorem shows the confidence interval and sample complexity estimates in terms of Rademacher averages.

**Theorem 2.28 (Rademacher bound)** *Let $F$ be a class of real-valued functions defined on a measurable space $\Omega$ and which take values in $[-1, 1]$, and set $\mu$ to be a probability measure on $\Omega$. Let $\sigma^2 = \sup_{f \in F} \mathrm{Var}\,(f)$ . Then there is a constant $C$ such that, for any $0 < t$ and every $\delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{f \in F} \ \Big| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \Big| < \frac{4 R_n\,(F)}{n} + C \left( \sigma \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right) .$$

*In particular, if*

$$n \geq \frac{C}{t^2} \max \left\{ (R_n\,(F))^2, \log \frac{1}{\delta} \right\} ,$$

*then $Pr_{\mathbf{X}} \left\{ \sup_{f \in F} \ \left| \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \right| \geq t \right\} \leq \delta$ .*

Note that we obtain estimates in terms of the Rademacher averages of the loss class $F$. Using Lipschitz properties of the loss function and the contraction principle (Theorem A.2, page 120), these estimates can be replaced (up to constants) by the same estimates in terms of Rademacher averages of the initial hypothesis class $H$.

We will now show that the Rademacher bound can be used to recover the results in terms of combinatorial complexities since one can relate Rademacher averages and metric entropy estimates. The following theorem reflects the relationship between the Gaussian averages and the metric entropy of a class.

**Theorem 2.29** *There are positive constants $c$ and $C$ such that for every class $F$, any* *integer $n$, and any sample $\mathbf{X} \in \Omega^n$,*

$$\sqrt{n} \sup_{u>0} u\sqrt{\log N(u, F, L_2(\mu_n))} \leq \mathbb{E}_{\mathbf{g}} \left( \sup_{f \in F} \left| \sum_{i=1}^{n} g_i f(X_i) \right| \right)$$
$$\leq C\sqrt{n} \int_0^\infty \sqrt{\log N(u, F, L_2(\mu_n))}\, du\,.$$

It follows directly from Dudley's entropy integral (Theorem A.3, page 120) and Sudakov's minoration (Theorem A.5, page 121). The additional normalization factor $\sqrt{n}$ arises from the fact that the unit ball in $L_2(\mu_n)$ is $\sqrt{n}\, B_2^n$.

Since the Rademacher averages are bounded by Gaussian averages (Theorem A.1, page 120), it follows directly that

$$\overline{R}_n\,(F) \leq C\sqrt{n} \int_0^\infty \sqrt{\log N(u, F, L_2(\mu_n))}\, du\,. \tag{2.5}$$

Using (2.5), we can relate Rademacher averages to combinatorial dimensions. This enables us to estimate Rademacher averages for particular classes with small combinatorial dimensions, namely classes with finite VC dimension and classes with polynomial fat-shattering dimension.

**Theorem 2.30 (Rademacher averages and VC-dimension)** *There is an absolute constant $c$ such that, if $F$ is a binary-valued class of functions with finite VC dimension $VC(F) = d$ then, for every $n$, $\overline{R}_n\,(F) \leq c\sqrt{dn}$.*

The proof follows from (2.5) and Dudley's result (Theorem 2.25) which relates the metric entropy and the VC-dimension.

For classes with polynomial fat-shattering dimension, where $\mathrm{fat}_\epsilon\,(F) \leq \gamma/\varepsilon^{-p}$ , one can obtain estimates for Rademacher averages of the order $O(\sqrt{n})$ when $0 < p < 2$ , $O(\sqrt{n \log^3 n})$ when $p = 2$, and $O(n^{1-1/p} \log^{1/p} n)$ when $p > 2$ (Mendelson 2003, Theorem 3.16). In particular, for $p = 1$, we state the following theorem which will be used later on.

**Theorem 2.31 (Rademacher averages and fat-shattering dimension)** *If $F$ is a class of functions taking values in $[-b, b]$, and if there is some $\gamma > 1$ such that for any $\varepsilon > 0$, $\mathrm{fat}_\epsilon\,(F) \leq \gamma/\varepsilon$, then there is an absolute constant $c$ such that, for every $n$, $R_n(F) \leq c\sqrt{\gamma n}$.*

The proof, based also on Dudley's entropy integral (Theorem A.3, page 120), can again be found in Mendelson (2003).

It is easy to show that the sample complexity estimates obtained through the Rademacher bound are better than the ones through the covering bound, by considering the example of a class with $\operatorname{fat}_{\epsilon}(F) \leq \gamma\varepsilon^{-2}$. For such a class, the sample complexity estimates obtained through Corollary 2.24 and Theorem 2.27 are of the order $c/t^4 \log(1/t)$, whereas the Rademacher bound leads to a significantly better sample complexity estimate of the order $c/t^2$ (Mendelson 2003, Corollary 3.17).

### 2.4.3 Characterization of Uniform Glivenko-Cantelli Classes

Recall the definition of uGC classes (Definition 1.1, page 3). The following theorem gives a characterization of uGC classes in terms of uniform complexity measures. Note that, although the bounds in terms of the metric and combinatorial complexities are loose in comparison to these in terms of Rademacher averages, metric entropy and VC-dimension do both also characterize uGC classes.

**Theorem 2.32 (Characterization of uGC classes)** *A binary-valued class of functions $F$ is a uGC class if and only if the VC-dimension $VC(F)$ is finite (Vapnik and Chervonenkis 1971).*

*A uniformly bounded class of functions $F$ is a uGC class if and only if any of the following holds:*

*1. There is some $p \geq 1$ such that for every $t > 0$,*

$$\lim_{n \longrightarrow \infty} \frac{N_p(\epsilon, F, n)}{n} = 0$$

*(Dudley et al. 1991).*

*2. For every $\varepsilon > 0$, the fat-shattering dimension* $\operatorname{fat}_{\epsilon}(F)$ *is finite (Alon et al. 1997).*

*3.*

$$\lim_{n \longrightarrow \infty} \frac{\overline{R}_n(F)}{n} = 0\,,$$

*(Dudley et al. 1991).*

## 2.5 Examples of Learning Algorithms

**Empirical Risk Minimization**

Recall that a good algorithm is one which produces a function with smallest expected error $\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{A}}(\mathbf{z}, H)) \approx \boldsymbol{\mathcal{R}}(h^*)$, but that the expected error can only be approximated from the empirical sample. As the Law of Large Numbers suggests to approximate the expectation through the empirical mean, this leads to the first choice for a learning algorithm, *Empirical Risk Minimization (ERM).* Instead of $h^*$, the algorithm chooses the hypothesis function from $H$ which has the smallest average error *on the sample.* The minimizer function produced by the Empirical Risk Minimization

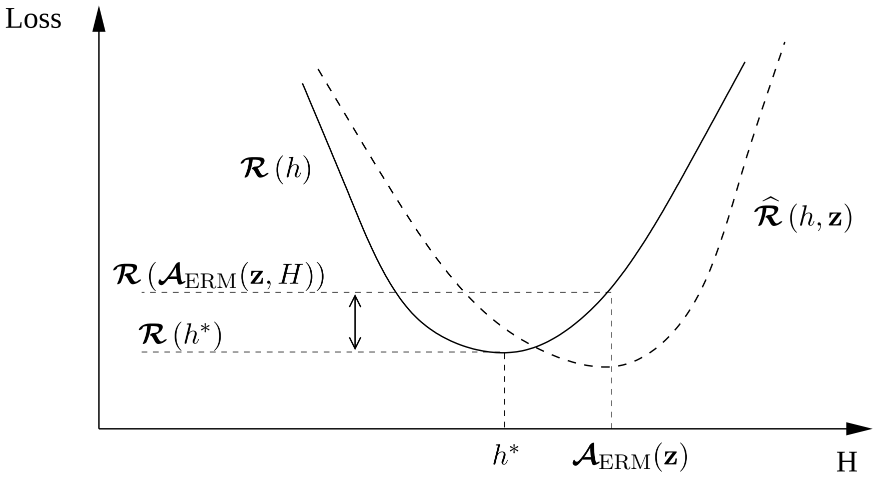


**Figure 2.2:** Expected and empirical error for a function class $H$ and a given sample $\mathbf{z}$. The empirical minimizer $\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H)$, unlike the minimizer $h^*$, depends on the sample. Of interest is the difference $\mathcal{R}(\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H)) - \mathcal{R}(h^*)$ (picture as in Bousquet et al. 2004).

algorithm $\mathcal{A}_{\mathrm{ERM}}$,

$$\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H) = \operatorname*{argmin}_{h \in H} \widehat{\mathcal{R}}(h, \mathbf{z}) \ , \tag{2.6}$$

is called an *empirical minimizer of the class* $H$. Since $\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H)$ depends on the training sample, it will usually be a different function for each new training sample.

One question of statistical learning theory is to understand and quantify the guarantees which can be given on the expected error of the empirical minimizer. As we will see in the sequel, it is possible to quantify conditions on the hypothesis class which guarantee that $\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H)$ has a small expected error. The first results of this type, upper bounding $\mathcal{R}(\mathcal{A}_{\mathrm{ERM}}(\mathbf{z}, H)) - \mathcal{R}(h^*)$, were presented in the pioneering paper by Vapnik and Chervonenkis (1971), and were later elaborated on in the now classical textbooks Vapnik (1982, 1995, 1998). These bounds involve the VC-dimension (see definition 2.5, page 22), and a related quantity called the VC-entropy (see definition 2.7, page 23) and show that for the ERM algorithm, learnability is equivalent to consistency. An illustration of the expected loss for a function class $H$, the minimizer $h^*$, the empirical loss for a given sample $\mathbf{z}$, and the corresponding empirical minimizer $\mathcal{A}_{\mathrm{ERM}}(\mathbf{z})$ are given in Figure 2.2.

Bounds for ERM in terms of Rademacher complexities (see definition 2.19, page 26) were derived in Koltchinskii (2001); Bartlett et al. (2002); Mendelson (2002a). Lately, tighter bounds in terms of Rademacher complexities of “local subsets” were proposed in Koltchinskii and Panchenko (2000); Massart (2000b); Bousquet (2002b); Bousquet et al. (2002); Bartlett et al. (2004a); Koltchinskii (2003); Lugosi and Wegkamp (2004); Bartlett and Mendelson (2005); Bartlett et al. (2004b).

In practice, however, ERM is computationally infeasible because computing or even approximating the empirical minimizer is an NP-hard problem. For example, a strong negative result was proved in Arora et al. (1997), showing that even for simple hypothesis classes consisting of linear classifiers, the task of learning halfspaces achieving a constant ratio of misclassifications in comparison to the best classifier is NP-hard. Similar hardness results for other hypothesis classes are given in Ben-David et al. (2003).

Because of this computational aspect, other methods to efficiently choose a function based on empirical data were developed, and a main question is to find among these the ones which can guarantee a small expected error.

**Boosting algorithms**

Boosting algorithms, introduced by Freund and Schapire (1999) (see also Duda et al. (2000); Schapire (2002); Meir and Rätsch (2003)), choose their hypothesis as a linear combination of simpler functions from a fixed base class with the goal to form a classifier with "boosted" performance. Boosting methods allow to combine "weak" classifiers (that is, classifiers which are only slightly better than random guessing) into "strong" combined classifiers with a very high accuracy. AdaBoost, for example, greedily adds classifiers to the resulting combination. At each step, the newly added classifier is chosen such that it improves performance on the samples misclassified in previous combinations. Boosting methods have proved successful in practice and therefore much work has been done in explaining their good performance. A first theoretical explanation of the good performance of boosting algorithms through generalization bounds was first given in Schapire et al. (1998) (see also Anthony and Bartlett (1999)). Schapire et al. (1998) showed that a quantity called *empirical margin distribution*, besides the VC-dimension of the base class, governs the generalization bound for weighted voting algorithms like boosting, and that boosting algorithms tend to optimize these bounds by maximizing the confidence of separation of classes (the "margin"). This result was further developed in Koltchinskii and Panchenko (2002) in terms Rademacher averages and extended in Koltchinskii et al. (2003); Koltchinskii and Panchenko (2005) in terms of additional complexities which quantify sparsity of the weights or clustering properties of the base class. The diversity or independence of errors of the base classifiers was shown to imply good generalization ability in Long (2002); Dasgupta and Long (2003). The work by Andonova Jaeger (2004) generalizes these results to convex combinations of random base classes. Consistency results for boosting algorithms were presented for example in Mannor et al. (2002); Blanchard et al. (2003); Zhang (2004b,a); Lugosi and Vayatis (2004); Jiang (2004).

**Kernel methods**

Kernel methods, such as *Support Vector Machines (SVMs)*, originate in the work of Vapnik and Chervonenkis (1971); Boser et al. (1992) and Wahba (1969); Craven and Wahba (1979). They have been applied successfully, for example, in computational biology (Noble 2004) and for document classification (Joachims 2002). The hypothesis class for the kernel algorithms are subsets of a *reproducing kernel Hilbert space (RKHS)*. They typically minimize a combination of the empirical error on the sample and a regularization term involving the squared (Hilbert space) norm of the function,

$$\boldsymbol{\mathcal{A}}_{\mathrm{SVM}}(\mathbf{z}, H) = \mathrm{argmin}_{h \in H} \widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{z}) + C\|h\|^2_{\mathrm{RKHS}} \,.$$

Any RKHS is generated by the span of a symmetric positive definite *kernel function* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$. The *Representer Theorem* states that the hypotheses produced by kernel algorithms are finite linear combinations of the base class $H = \{k(x, \cdot) | x \in \mathbb{R}\}$ of the form

$$\boldsymbol{\mathcal{A}}_{\mathrm{SVM}}(\mathbf{z}, H) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\,, \tag{2.7}$$

for some real coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$. For surveys on kernel methods see, for example, Vapnik (1982); Christianini and Shawe-Taylor (2001); Schölkopf and Smola (2002); Herbrich (2002). Note that equation (2.7) implies that the hypothesis function $A_{\mathrm{SVM}}(\mathbf{z}, H)$ belongs to the *data-dependent* subclass of functions $H_{\mathrm{SVM}}(\mathbf{z}) \subseteq H$,

$$H_{\mathrm{SVM}}(\mathbf{z}) = \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \;:\; (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n \right\} .$$

Generalization bounds for kernel methods in terms of margin (Vapnik 1982; Herbrich 2002), (empirical) eigenvalues of the kernel operator (Mendelson 2002a), and based on stability properties (Bousquet and Elisseeff 2002) were derived. Consistency and general properties of kernel classes are studied in Cucker and Smale (2002); Steinwart (2002, 2003). The approximation-estimation trade-off for kernel methods is studied in Cucker and Smale (2002); Smale and Zhou (2003); Vito et al. (2005).

Chapter 3

# Tools

## 3.1 Deviation and Concentration Inequalities

Deviation and concentration inequalities are basic tools used to derive generalization bounds in statistical learning theory. They quantify how a random variable deviates from its mean value. Unlike in classical probability, where the Law of Large Numbers and the Central Limit Theorem give asymptotic characterizations for the behaviour of sums of independent random variables with identical distributions, deviation inequalities bound tail probabilities *nonasymptotically.* For example, Hoeffding's and Bernstein's exponential inequalities (Theorems 3.3 and 3.6) give two nonasymptotic characterizations for the tail probabilities of sums of bounded independent random variables (one as a sub-Gaussian, whereas the other as a mixture of a sub-Gaussian and a sub-exponential tail, see the remark after Theorem 3.6, page 38). They are called *exponential inequalities* since they state that the probability of the deviation of the average of independent random variables from their expectation is exponentially small in the number of variables. Exponential inequalities of this type can be derived to bound tail probabilities not only for *sums* of independent random variables but also for general functions of independent random variables which satisfy, for instance, some type of Lipschitz conditions, as well as for martingales (see, e.g. Ledoux 2001).

This phenomenon is called *concentration of measure*. "The concentration of measure phenomenon is an elementary, yet non-trivial observation. It is often a high dimensional effect, or a property of a large number of variables, for which functions with small local oscillations are almost constant" (Ledoux 2001) [1].

The basic probabilistic results in this section can be found in various textbooks on probability [2]. The exposition here follows closely the ones in Devroye et al. (1996); Lugosi (2003) and Boucheron et al. (2004a). All results in this section hold for *independent*

[1] Books on the concentration of measure phenomenon are, for example, Ledoux and Talagrand (1991) and Ledoux (2001). Recent surveys on concentration inequalities with a focus on their applications in statistical learning theory are Massart (2000b); Lugosi (2003); Boucheron et al. (2004a, 2005).

[2] See, for example, Shiryayev (1984); Billingsley (1986); Dudley (1989); Durrett (1996); Feller (1971).

random variables.

### 3.1.1 Nonexponential Inequalities

We will first state two simple *nonexponential* tail inequalities: Markov's inequality for nonnegative random variables and Chebyshev's inequality for random variables with a finite variance. Markov's inequality is the basis for the proofs of many other inequalities, especially of the powerful exponential inequalities which we will present in the sequel. By comparing Markov's and Chebyshev's inequalities, we will illustrate an important idea which recurs in this thesis: information on the variance can potentially improve tail bounds.

**Theorem 3.1 (Markov's inequality)** *Let $X$ be a nonnegative random variable. Then, for every $t > 0$,*

$$Pr\left\{X \geq t\right\} \leq \frac{\mathbb{E}X}{t}\,.$$

Markov's inequality is the simplest inequality which relates tail probabilities to expectations.

The next inequality, called Chebyshev's inequality, follows directly from Markov's inequality for the random variable $Y = (X - \mathbb{E}X)^2$. It is a concentration inequality as it shows how $X$ is concentrated around its mean value.

**Theorem 3.2 (Chebyshev's inequality)** *Let $X$ be a random variable with finite variance. Then, for every $t > 0$,*

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

Unlike Markov's bound, which uses only the value of the expectation of $X$, Chebyshev's inequality makes use of the "shape" of the distribution through the variance. Similar bounds can be obtained by applying Markov's inequality to $\phi(|X - \mathbb{E}X|)$, where $\phi$ is any nonnegative convex function (see, e.g., Durrett 1996, page 15), which leads in particular to concentration results in terms of any higher-order moments by setting $\phi(x) = x^q$, for any integer $q \geq 3$.

In learning theory, the variables whose tail we want to bound are the empirical averages. We will show in the following that for these variables one can obtain bounds which decay *exponentially* fast with the sample size $n$.

### 3.1.2 Exponential Inequalities for Sums of Independent Random Variables

A generalization of the procedure in the previous section is to make use of the moment generating function (mgf) of $X$ (which is the Laplace transform of the density of $X$),

$\mathbb{E}e^{sX}$, by applying Markov's inequality to $Y = e^{sX}$. This idea is called *Chernoff's bounding method*. The method is particularly useful for bounding the tail of sums of independent random variables, since the mgf of a sum of independent random variables is the product of the individual mgf's. This technique leads, for example, to Hoeffding's and Bernstein's *exponential* inequalities.

**Theorem 3.3 (Hoeffding's inequality (Hoeffding 1963))** *Let $X_1, \ldots, X_n$ be independent bounded random variables such that $X_i$ fall in the interval $[a_i, b_i]$ with probability one and let $S_n = \sum_{i=1}^n X_i$. Then, for every $t > 0$,*

$$Pr\Big\{\big|\mathbb{E}S_n - S_n\big| \geq t\Big\} \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}.$$

*In particular, if the independent random variables $X_1, \ldots, X_n$ fall in the same interval $[a, b]$ with probability one and have the same expectation $\mathbb{E}X$, then, for every $t > 0$,*

$$Pr\Big\{\big|\mathbb{E}X - \frac{1}{n}\sum_{i=1}^n X_i\big| \geq t\Big\} \leq 2e^{-\frac{2nt^2}{(b-a)^2}}.$$

The proof is standard, and can be found, for instance, in Hoeffding (1963) or in Devroye et al. (1996), page 122. We will sketch briefly the proof for the second inequality to show later on why one can not use this inequality directly to bound the tail probabilities for functions produced by learning algorithms. We will also need the following lemma upper bounding the mgf of $X$, which we will state without proof (Devroye et al. 1996, Lemma 8.1, page 122):

**Lemma 3.4** *Let $X$ be a bounded random variable with $\mathbb{E}X = 0$, $a \leq X \leq b$. Then for any $s > 0$,*

$$\mathbb{E}e^{sX} \leq e^{s^2(b-a)^2/8}.$$

**Proof of Theorem 3.3:** Let $X = n^{-1}\sum_{i=1}^n X_i$ and let $s > 0$ be a fixed number. Since the exponential function is monotone, $Pr\left\{X \geq t\right\} = Pr\left\{e^{sX} \geq e^{st}\right\}$ and we can apply Markov's inequality to $Y = e^{s(X - \mathbb{E}X)}$. We obtain

$$\begin{aligned} Pr\left\{X - \mathbb{E}X \geq t\right\} &= Pr\left\{e^{s(X-\mathbb{E}X)} \geq e^{st}\right\} \\ &\leq e^{-st}\mathbb{E}e^{s(X-\mathbb{E}X)} \\ &= e^{-snt}\mathbb{E}e^{s\sum_{i=1}^n (X_i - \mathbb{E}X_i)}. \end{aligned}$$

Since $e^{s(X_i - \mathbb{E}X_i)}$ are independent,

$$\mathbb{E}e^{s\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)} = \prod_{i=1}^{n} \mathbb{E}e^{s(X_i - \mathbb{E}X_i)}. \tag{3.1}$$

From Lemma 3.4, observing that $a - \mathbb{E}X_i \leq X_i - \mathbb{E}X_i \leq b - \mathbb{E}X_i$ , it follows that

$$\prod_{i=1}^{n} \mathbb{E}e^{s(X_i - \mathbb{E}X_i)} \leq \prod_{i=1}^{n} e^{s^2(b-a)^2/8} = e^{s^2 n(b-a)^2/8},$$

and hence, by setting $s = 4t/(b-a)^2$ , we have

$$Pr\left\{X - \mathbb{E}X \geq t\right\} \leq e^{-\frac{2nt^2}{(b-a)^2}}.$$

A similar proof shows that $Pr\left\{X - \mathbb{E}X \leq -t\right\} \leq e^{-\frac{2nt^2}{(b-a)^2}}$ , and the claim of the theorem follows from the union bound. ■

Thus, Hoeffding's inequality provides a sub-Gaussian tail bound for sums of bounded i.i.d. variables with zero mean [3]. Note that the fact that the variables are bounded implies that their sum is Lipschitz with respect to the $\ell_\infty^n$ metric.

The next two inequalities, due to Bernstein and Bennett, do take into account the variance of the sum of bounded variables (see, for instance, Devroye et al. 1996, Theorem 8.2, page 124). Their proof is similar to that of Hoeffding's inequality.

**Theorem 3.5 (Bennett's inequality (Bennett 1962))** *Let $X_1, \ldots, X_n$ be independent real-valued random variables such that $X_i$ fall in the interval $[a,b]$ with probability one and have the same expectation $\mathbb{E}X$. Let*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left(X_i\right).$$

*Then, for every $t > 0$,*

$$Pr\Big\{\big|\mathbb{E}X - \frac{1}{n}\sum_{i=1}^{n} X_i\big| \geq t\Big\} \leq 2e^{\frac{-n\sigma^2}{(b-a)^2}h\left(\frac{(b-a)t}{n\sigma^2}\right)}$$

*where $h(u) = (1+u)\log(1+u) - u$ , for $u \geq 0$.*

By observing that for $u \geq 0$, $h(u) \geq u^2/(2 + 2u/3)$ , one can derive the following inequality due to Bernstein.

[3] A result of similar flavour is true for sub-Gaussian processes, and in particular for the Rademacher process $\sum_{i=1}^{n} \varepsilon_i X_i$ (see, e.g., van der Vaart and Wellner 1996, page 100).

**Theorem 3.6 (Bernstein's inequality (Bernstein 1946))** *Let $X_1, \ldots, X_n$ be independent real-valued random variables such that $X_i$ fall in the interval $[a, b]$ with probability one and have the same expectation $\mathbb{E}X$. Let*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\,(X_i)\,.$$

*Then, for every $t > 0$,*

$$Pr\Big\{\big|\mathbb{E}X - \frac{1}{n}\sum_{i=1}^{n} X_i\big| \geq t\Big\} \leq 2e^{-\frac{nt^2}{2\sigma^2+2t(b-a)/3}}\,.$$

**Remark:** Observe that Bernstein's inequality is an improvement in comparison to Hoeffding's inequality whenever the variance $\sigma^2$ is much smaller than $(b-a)^2/4$. For $t$ such that $t \geq \sigma^2$, Bernstein's inequality provides a sub-exponential tail ($e^{-nt}$) rather than a sub-Gaussian tail ($e^{-nt^2}$) bound rate as given by Hoeffding's inequality. For $t < 1$ and large $n$ (the range we are interested in learning theory) this is a substantial improvement.

In general, for any fixed function $f$ bounded by $[a, b]$, Hoeffding's inequality for the random variable $Z = f(X)$ implies that

$$Pr\Big\{\big|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t\Big\} \leq 2e^{-\frac{2nt^2}{(b-a)^2}}\,,$$

or in other words, with probability at least $1-\delta$ over the random draw of samples $(X_1, \ldots, X_n)$,

$$\big|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \leq (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\,. \tag{3.2}$$

Bernstein's inequality for the same variable yields

$$Pr\Big\{\big|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t\Big\} \leq 2e^{-\frac{nt^2}{2\mathrm{Var}(f)+2t(b-a)/3}}\,, \tag{3.3}$$

which implies that with probability at least $1-\delta$ over the random draw of samples $(X_1, \ldots, X_n)$,

$$\big|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \leq \sqrt{\frac{2\mathrm{Var}\,(f)\log(2/\delta)}{n}} + \frac{2(b-a)\log(2/\delta)}{3n} \tag{3.4}$$

(where in order to obtain the above form (3.4), we solve the quadratic equation in $t$, $t^2 = 2\mathrm{Var}\,(f)\log(2/\delta)/n + t\,2(b-a)\log(2/\delta)/3n$ , and then use that $\sqrt{A} + \sqrt{B} \geq \sqrt{A+B}$ for any $A, B \geq 0$). For a fixed $n$ and a function whose variance satisfies that

$\mathrm{Var}\,(f) \ll \log(1/\delta)/n$ , the first term in (3.4) is small and the deviation is thus of the order of $O(\log(1/\delta)/n)$ . The order of the deviation, for a given confidence $\delta$, becomes $O(c\log(1/\delta)/n)$ rather than $O(c\log(1/\delta)/\sqrt{n})$ in (3.2), and the deviation tail bound in (3.3) is exponential ($e^{-nt}$) rather than Gaussian ($e^{-nt^2}$).

As discussed in Section 2.2, for the analysis of learning algorithms, we are interested in bounding deviations of means from expectations for the functions produced by a learning algorithm. For any *fixed* function, we can apply Hoeffding's and Bernstein's bounds. In equation (2.3), however, the function $\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)$ depends on the sample and is therefore itself *random.* Because the variables $X_i = \boldsymbol{\mathcal{A}}(\mathbf{Z}, H)(Z_i)$ are no longer independent, the argumentation of line (3.1) in the proof of Hoeffding's inequality does not hold for this case.

In order to avoid the technical difficulties posed by the randomness of the function $\boldsymbol{\mathcal{A}}(\mathbf{Z}, H)$, one method to attack the problem is to upper bound (see (2.2), page 20) the deviation of the expected from the empirical loss of this random function with a worst-case uniform deviation of the form $\sup_{f\in F} \left|\boldsymbol{\mathcal{R}}\,(f) - \widehat{\boldsymbol{\mathcal{R}}}\,(f, \mathbf{Z})\,\right|$ (where $F$ is either the loss or the excess loss class). For finite classes $F$, a crude approach to bound such a deviation is to use a concentration inequality for a single function (e.g., Hoeffding's or Bernstein's inequality) and a union bound argument to combine them (e.g., Anthony and Bartlett 1999, page 21). However, the union bound can be potentially very loose, especially if the events in the union are statistically dependent (and this is easily the case, for example, if the functions in $F$ are correlated). For infinite classes $F$, one can make use of the metric structure of $F$ — for example in the spaces $L_p(\mu_n)$ — and approximate the class $F$ with a finite cover at a scale which is of the same order of magnitude as the deviation. The loss of replacing $F$ with its cover can be quantified and one can proceed as in the case of finite classes (e.g., Anthony and Bartlett 1999, page 143); again, the union bound is potentially loose. With such a covering approach, in order to be able to prove learnability, the degree of concentration has to "beat" the metric complexity of the class (as measured by the covering numbers or equivalently metric entropy). Fortunately, as we will see in the next section, it is possible to take a different approach and to derive concentration inequalities directly for the supremum of empirical processes. These more sophisticated results simplify the earlier versions for the proof and avoid the union bound and thus its potential looseness. Furthermore, as presented in Section 4.2, they also allow one to separate the influence of the two key properties — the degree of concentration and the complexity of the hypothesis class — which play a role in the analysis of the generalization ability.

In the next section we present concentration inequalities with exponential tails which can be applied directly to random variables of the form $Z = \sup_{f\in F} |\mathbb{E}f - \mathbb{E}_n\, f|$.

### 3.1.3 Concentration Inequalities for General Functions

In this section, we present several results which allow one to derive concentration results for general functions (as opposed to sums) of independent random variables. They differ in the assumptions on which they are based. All of these assumptions are different ways of ensuring that the functions are Lipschitz with respect to various metrics. We will also emphasize how the different assumptions lead to potentially different degrees of concentration.

**Bounded differences and Rademacher averages for bounded sets**

The first result generalizes Hoeffding's inequality for general functions of independent random variables. It is due to McDiarmid (McDiarmid 1989) and therefore also referred to as *McDiarmid's inequality*. It follows as a special case of a more general result called *Azuma's inequality* (Azuma 1967) based on martingale methods (see, e.g., Ledoux 2001, Chapter 4.1).

**Theorem 3.7 (The bounded differences inequality (McDiarmid 1989))** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\Omega$ and assume that $f : \Omega^n \longrightarrow \mathbb{R}$ satisfies the bounded difference condition, that is, for every $1 \leq i \leq n$,*

$$\sup_{x_1,\ldots,x_n,x'_i \in \Omega} \left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n) \right| \leq c_i \,.$$

*Then, for every $t > 0$,*

$$Pr\left\{ \left| f(X_1, \ldots, X_n) - \mathbb{E} f(X_1, \ldots, X_n) \right| \geq t \right\} \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}} \,.$$

This theorem states that, whenever a function of $n$ independent random variables satisfies the condition that a change in each coordinate separately is bounded by a constant, the function is highly concentrated. As a special case, for bounded i.i.d. random variables and $f$ being the sum, we obtain Hoeffding's inequality. The bounded difference condition is one way of ensuring a Lipschitz condition with respect to the Hamming metric, namely,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq \max_i c_i \, d_H(\mathbf{x}, \mathbf{y}) \,,$$

which leads to concentration of measure.

In order to compare Theorem 3.7 to Hoeffding's inequality, we will apply it to the supremum of an empirical process. The only assumption we require in this case is *boundedness* of the functions. We obtain the following corollary:

**Corollary 3.8** *Let $F$ be a class of functions which take values in $[a,b]$, let $X_1, ..., X_n$ be independent random variables distributed according to $\mu$, and let $Z = \sup_{f \in F} \left| \mathbb{E} f - \frac{1}{n}\sum_{i=1}^n f(X_i) \right|$. Then for every $\delta > 0$, with probability at least $1-\delta$ over the random choice of samples,*

$$Z \leq \mathbb{E} Z + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}} .$$

By comparing this result to (3.2), we observe that we obtained a Hoeffding-type bound which holds *uniformly* over the class $F$. The proof can be found for example in Boucheron et al. (2004a) and relies on the fact that, because the functions in $F$ are bounded, the variable $Z$ satisfies the bounded difference condition with constants $c_i = (b-a)/n$.

Another direct corollary of Theorem 3.7 is the following useful concentration result for the suprema of Rademacher processes indexed by a subset of a ball in $\ell_\infty^n$. Recall that the supremum of a Rademacher process indexed by a set $V \subset \mathbb{R}^n$ is defined as $\sup_{\mathbf{v} \in V} \left|\sum_{i=1}^n \varepsilon_i v_i\right|$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are independent Rademacher variables (see Appendix A, page 119). The following result can be found, for example, in van der Vaart and Wellner (1996), page 101.

**Corollary 3.9 (Concentration of suprema of Rademacher processes)** *For every set $V \subset bB_\infty^n$, let $Z = \sup_{\mathbf{v} \in V} \left| \sum_{i=1}^n \varepsilon_i v_i \right|$. Then, for every $t > 0$,*

$$Pr_{\boldsymbol{\varepsilon}} \left\{ |Z - \mathbb{E}_{\boldsymbol{\varepsilon}} Z| \geq t \right\} \leq e^{-\frac{t^2}{2b^2 n}} ,$$

*where the probability is taken with respect to the Rademacher random variables $\boldsymbol{\varepsilon}$.*

This corollary immediately applies to Rademacher sums of functions bounded by $b$, and its proof is similar to the proof of the previous corollary.

**Proof:** Define $h(\varepsilon_1, \ldots, \varepsilon_n) := \sup_{\mathbf{v} \in V} \left|\sum_{i=1}^n \varepsilon_i v_i\right|$. By the triangle inequality, for every $1 \leq i \leq n$,

$$\sup_{\{\varepsilon_1, \ldots, \varepsilon_n, \tilde{\varepsilon_i}\}} \left| h(\varepsilon_1, \ldots, \varepsilon_n) - h(\varepsilon_1, \ldots, \varepsilon_{i-1}, \tilde{\varepsilon_i}, \varepsilon_{i+1}, \ldots, \varepsilon_n) \right| \leq 2b,$$

and the claim follows directly from McDiarmid's bounded differences inequality for $h$. ■

Thus, the fact that $V$ is bounded with respect to the supremum norm $\ell_\infty^n$ is already sufficient to ensure concentration of the suprema of Rademacher averages.

### Convexity and Rademacher averages

The last corollary we presented, showing that suprema of Rademacher processes are concentrated, is based on the boundedness of the set $V$ and follows from concentration results using martingale methods. However, Talagrand has shown that the *convexity* of a function of a random variable, together with a Lipschitz condition, can lead to potentially tighter concentration results. This result follows from the famous *convex distance inequality* (Talagrand 1995) (see also Ledoux (2001), Chapter 4.2).

For convex Lipschitz functions, we will present here one concentration result based on Talagrand's convex distance inequality which can be found in Ledoux (2001), Corollary 4.10, page 77:

**Theorem 3.10 (Corollary of Talagrand's convex distance inequality for convex functions)** *For every probability measure $\mu$ on $[0,1]^n$, and every convex function $F$ on $\mathbb{R}^n$ which is Lipschitz with respect to the $\ell_2^n$ metric with Lipschitz constant $L$, let $\mathbf{X}$ be distributed according to $\mu$ and denote by $M_\mu F(\mathbf{X})$ the median of $F$. Then, for every $t \geq 0$,*

$$Pr\left\{|F(\mathbf{X}) - M_\mu F(\mathbf{X})| \geq t\right\} \leq 4e^{-t^2/4L^2}\,.$$

Talagrand's convex distance inequality refers to concentration around the median rather than the expectation of $Z$. Since concentration implies that median and expectation are close (within a constant), one can derive a result for the tail deviation of $F$ from its expectation and replace the median by the mean up to universal constants.

Observe that, if $V \subset \ell_2^n$, one can define a norm on $\mathbb{R}^n$ given by $|||\mathbf{x}||| := \sup_{\mathbf{v}\in V} |\langle \mathbf{v}, \mathbf{x}\rangle|$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\ell_2^n$. Hence, the function $\mathbf{x} \mapsto |||\mathbf{x}|||$ is convex. In particular, the function $\boldsymbol{\varepsilon} \mapsto \sup_{\mathbf{v}\in V}\left|\sum_{i=1}^n \varepsilon_i v_i\right|$ is convex, as it is a norm because $\sup_{\mathbf{v}\in V}\left|\sum_{i=1}^n \varepsilon_i v_i\right| = |||\sum_{i=1}^n \varepsilon_i e_i|||$, where $\{e_1,\ldots,e_n\}$ is the standard orthonormal basis in $\ell_2^n$. It is easy to see that, if $V$ is bounded in $\ell_2^n$, then the function $\mathbf{x} \mapsto |||\mathbf{x}|||$ is also Lipschitz with Lipschitz constant $\sup_{\mathbf{v}\in V}\|\mathbf{v}\|_2$. Indeed, by the triangle inequality and Cauchy-Schwartz inequality, it follows that

$$\left|\,|||\mathbf{x}||| - |||\mathbf{x}'|||\,\right| \leq \sup_{\mathbf{v}\in V}\left|\langle \mathbf{v}, \mathbf{x}-\mathbf{x}'\rangle\right| \leq \sup_{\mathbf{v}\in V}\|\mathbf{v}\|_2\|\mathbf{x}-\mathbf{x}'\|_2\,.$$

Thus, one can apply the corollary of Talagrand's convex distance inequality to suprema of Rademacher processes. The resulting corollary can be also found, for example, in Ledoux (2001), page 76 (see also Ledoux (2001), Chapter 7).

**Theorem 3.11 (Corollary of Talagrand's convex distance inequality for suprema of Rademacher processes)** *For every set $V \subset vB_2^n$, let $Z = \sup_{\mathbf{v}\in V}\left|\sum_{i=1}^n \varepsilon_i v_i\right|$ and denote by $M_{\boldsymbol{\varepsilon}} Z$ the median of $Z$. Then, for every $t > 0$,*

$$Pr_{\boldsymbol{\varepsilon}}\left\{|Z - M_{\boldsymbol{\varepsilon}} Z| \geq t\right\} \leq 4e^{-\frac{t^2}{4v^2}}\,.$$

*Furthermore,*

$$|\mathbb{E}_{\boldsymbol{\varepsilon}} Z - M_{\boldsymbol{\varepsilon}} Z| \leq 4\pi v \quad \textit{and} \quad \mathrm{Var}\,(Z) \leq 16v^2\,.$$

Note that if $V$ is bounded with respect to the $\ell_\infty^n$ metric, one can recover the bound from Corollary 3.9, by observing that for any $\mathbf{v} \in V \subseteq \mathbb{R}^n$, $\|\mathbf{v}\|_2 \leq \sqrt{n}\|\mathbf{v}\|_\infty$. However, if more information on the set $V$ is available which ensures that $V$ has a much smaller diameter in $\ell_2^n$ than $\sqrt{n}\,b$ (where $b$ is the diameter of $V$ in $\ell_\infty^n$), then the corollary of Talagrand's convex distance inequality gives an improvement on the bound in Corollary 3.9.

### Control of the variance

A difficult and long-open question has been the derivation of a Bernstein-type counterpart which holds uniformly over a function class and which allows one to take advantage of functions with small variances.

Note that the fact that a bound on the maximal variance of a function class is connected to stronger concentration is already exhibited in Theorem 3.11. Indeed, if $F$ is a bounded function class, then by symmetrization techniques (see Section 3.2, Corollary 3.19), the tail probabilities of $\sup_{f \in F} |\mathbb{E}f - \mathbb{E}_n f|$ are controlled by those of $\sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(X_i)|$, and therefore, following Theorem 3.11, by the maximal $\ell_2^n$ norm of the coordinate projections of $F$ on the random sample. For large $n$, $\mathbb{E}f^2 \sim \sum_{i=1}^n f^2(X_i)/n$, and thus $\mathrm{Var}\,(f) \sim \|(f(X_1), \ldots, f(X_n))\|_2^2/n$. Hence , control of the $\ell_2^n$ norm of the projections can be connected to control of the variances of the functions in the class.

Historically, a range of results existed for *limit* theorems ($n$ going to $\infty$) for the suprema of weighted empirical processes (see Giné et al. (2004) and references therein). For the finite sample case, Vapnik and Chervonenkis (1971) were the first to prove, for binary-valued classes of functions (a case in which the variance is the same as the expectation), uniform bounds for the weighted empirical process $Z = \sup_{f \in F} \left(|\mathbb{E}f - \mathbb{E}_n f| / \sqrt{\mathbb{E}f}\right)$. Their proof is based on a union bound argument combined with Hoeffding's inequality, a proof very similar to early proofs for the unweighted case. In the statistical learning theory context, variations, generalizations to real-valued functions, and improvements were given for example by Haussler (1992); Anthony and Shawe-Taylor (1993); Lee et al. (1996); Bartlett and Lugosi (1999); Anthony and Bartlett (1999) (see also Bousquet (2002b) and references therein).

A breakthrough occurred with Talagrand's inequality, based on concentration in product spaces (Talagrand 1995). Talagrand's general result allows one to get directly a Bernstein's (Bennett's) type inequality for suprema of empirical processes [4]. It was originally proved in Talagrand (1994, 1996b). A convenient version is due to Massart:

[4] See Talagrand's induction method, Talagrand (1995, 1996b,c); Ledoux (2001); Panchenko (2001, 2002, 2003) and also Massart (2000b); Boucheron et al. (2000); Bousquet (2002b); Boucheron et al. (2003); Bousquet (2002a); Massart (2003); Boucheron et al. (2004a).

**Theorem 3.12 (Talagrand's concentration inequality)** *Let $F$ be a class of functions defined on a measurable space $(\Omega, \mu)$, such that for every $f \in F$, $\|f\|_\infty \le b$ , and such that $\mathbb{E}f = 0$. Let $X_1, ..., X_n$ be independent random variables distributed according to $\mu$ and set $\sigma^2 = n \sup_{f\in F} \mathrm{Var}\,(f)$ . Let*

$$Z = \sup_{f \in F} \sum_{i=1}^n f(X_i),$$
$$\bar{Z} = \sup_{f \in F} \big| \sum_{i=1}^n f(X_i) \big|.$$

*Then there is an absolute constant $c > 0$ such that, for every $t > 0$ and every $\rho > 0$, the following holds:*

$$Pr\left\{Z \ge (1+\rho)\mathbb{E}Z + \sigma\sqrt{ct} + c(1+\rho^{-1})bt\right\} \le e^{-t},$$
$$Pr\left\{Z \le (1-\rho)\mathbb{E}Z - \sigma\sqrt{ct} - c(1+\rho^{-1})bt\right\} \le e^{-t},$$

*and the same inequalities hold for $\bar{Z}$.*

The inequality for $\bar{Z}$ is due to Massart (2000a). The one sided versions were shown by Rio (2001) and Klein (2002). For $b = 1$, the best estimates on the constants in all cases are due to Bousquet (2002b). As we can see, the concentration inequalities depend on a combination of the maximal variance of the functions in the class and the maximal $\ell_\infty^n$ norm, allowing us to take advantage in cases in which one can ensure a priori that the variances of all functions are small.

Very recently, Bernstein-like inequalities for *general* functions (as opposed to suprema of empirical processes) satisfying a specific Lipschitz condition were derived in Boucheron et al. (2000, 2003); Bousquet (2002a); Boucheron et al. (2005). They are based on the idea (similar to the one in the Efron-Stein inequality, (see, e.g., Lugosi 2003) to bound the variance in terms of "partial variances". The Lipschitz condition is different from that used in McDiarmid's bounded differences inequality: whereas in McDiarmid's inequality, functions have to be only within a constant when *replacing* any one variable from the sample, the self-bounding condition restricts the values of the function when compared to another function which depends on less variables. The proof for the results uses the entropy method based on logarithmic Sobolev inequalities (see, e.g., Ledoux 2001, Chapter 5). Such Bernstein-type inequalities can be applied to prove the concentration of empirical complexities as shown in Boucheron et al. (2000, 2003) (see the following Lemma 3.15).

Here we state one result for functions satisfying the *self-bounding property* (conditions 1 and 2 in Theorem 3.13), a generalization of the bounded differences property. One can show that the self-bounding property implies that $\mathrm{Var}\,(f) \le \mathbb{E}f$ (e.g., Lugosi

2003), and thus is a property which allows one to control the variance.

**Theorem 3.13 (Concentration of self-bounding functions)** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\Omega$, and set $Z = f(X_1, \ldots, X_n)$ to be a random variable. Assume that there is a function $g : \Omega^{n-1} \longrightarrow \mathbb{R}$ satisfying the following for all $\{x_1, \ldots, x_n\} \subset \Omega$ :*

*1.* $0 \leq f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$ *for every* $1 \leq i \leq n$ *,*

*2.* $\sum_{i=1}^{n} \left(f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\right) \leq f(x_1, \ldots, x_n)$ *.*

*Then, for every $t > 0$,*

$$Pr\{Z \geq \mathbb{E}Z + t\} \leq e^{-\frac{t^2}{\mathbb{E}Z+2t/3}},$$

$$Pr\{Z \leq \mathbb{E}Z - t\} \leq e^{-\frac{t^2}{2\mathbb{E}Z}}.$$

Observe that the expectation appears in the exponent of the bound, and thus the degree of concentration is dependent on the value of the expectation. Indeed, whenever the expectation of a self-bounding function is of the order of the deviation $t$ or smaller, $\mathbb{E}Z \leq ct$, Theorem 3.13 gives a tail bound of the order $e^{-c't}$ (as opposed to $e^{-c't^2}$ in the bounded differences inequality, Theorem 3.7).

In the following, we will state two lemmata with regard to self-bounding functions which will be useful later on. The first one states that self-bounding symmetric functions evaluated on a double sample do not grow too fast in comparison to their value on a single sample.

**Lemma 3.14** *Let $(f_n)_{n\in\mathbb{N}}$ be a series of functions such that, for any $n$, the following three conditions hold:*

*1.* $f_n : \Omega^n \longrightarrow \mathbb{R}$*,*

*2.* $f_n$ *satisfies the self-bounding property (conditions 1 and 2 in Theorem 3.13) with respect to* $g = f_{n-1}$*,*

*3.* $f_n$ *is symmetric (that is, it is invariant with respect to permutations of its variables).*

*Let $\mu$ be a probability measure on $\Omega$. Then, for any $n$ and $X_1, \ldots, X_{2n}$ distributed i.i.d. according to $\mu$,*

$$\mathbb{E}_{\mu^{2n}} f_{2n}(X_1, \ldots, X_{2n}) \leq 2\mathbb{E}_{\mu^n} f_n(X_1, \ldots, X_n)\,. \tag{3.5}$$

*Furthermore, there exist absolute constants $c, c' > 0$ such that, for any $t' > 0$, if $\mathbb{E}_{\mu^n} f_n(X_1, \ldots, X_n) = d$ then with probability at least $1 - e^{-cd} - e^{-c't'^2/d}$,*

$$f_{2n}(X_1, \ldots, X_{2n}) \leq 4f_n(X_1, \ldots, X_n) + t'\,.$$

**Proof:** For any $n$, by the symmetry of $f_n$, and taking the expectation in condition 2 implies that

$$n(\mathbb{E}_{\mu^n} f_n(x_1,\ldots,x_n) - \mathbb{E}_{\mu^{n-1}} f_{n-1}(x_1,\ldots,x_{n-1})) \leq \mathbb{E}_{\mu^n} f(x_1,\ldots,x_n)\,, \tag{3.6}$$

and thus

$$\mathbb{E}_{\mu^n} f_n(x_1,\ldots,x_n) \leq \frac{n}{n-1}\mathbb{E}_{\mu^{n-1}} f_{n-1}(x_1,\ldots,x_{n-1})\,.$$

Applying the last inequality to $n+1,\ldots,2n$ proves thus claim (3.5).

In order to obtain (3.6) we apply Theorem 3.13 to both

$$Z_1 = f_n(X_1,\ldots,X_n) \quad \text{and} \quad Z_2 = f_{2n}(X_1,\ldots,X_{2n})\,.$$

Let $d = \mathbb{E}Z_1$. From (3.5), $\mathbb{E}Z_2 \leq 2d$ . By Theorem 3.13, for any $t, t' > 0$,

$$Pr\left\{4Z_1 \leq 4d - 4(t+t')\right\} = Pr\left\{Z_1 \leq d-(t+t')\right\} \leq e^{-(t+t')^2/2d} \leq e^{-t'^2/2d}$$

and

$$Pr\left\{Z_2 \geq 2d+t\right\} \leq Pr\left\{Z_2 \geq \mathbb{E}Z_2+t\right\} \leq e^{-t^2/(\mathbb{E}Z_2+2t/3)} \leq e^{-t^2/(2d+2t/3)}\,.$$

Observe that for $t = 2d/5$, $4d-4t = 2d+t$, and therefore $4Z_1+4t' \geq 4d-4t$ and $Z_2 \leq 2d+t$ implies $4Z_1+4t' \geq Z_2$. Thus,by the union bound,

$$\begin{aligned} Pr\left\{4Z_1+4t' \geq Z_2\right\} &\geq Pr\left\{4Z_1+4t' \geq 4d-4t \text{ and } Z_2 \leq 2d+t\right\} \\ &\geq 1 - Pr\left\{4Z_1+4t' < 4d-4t\right\} - Pr\left\{Z_2 > 2d+t\right\} \\ &\geq 1 - e^{-t'^2/2d} - e^{-t^2/(2d+4d/15)} \geq 1 - 2e^{-4d/75} - e^{-t'^2/2d}\,. \end{aligned}$$

■

Thus, the self-bounding property for symmetric functions implies, besides a strong concentration for any $n$, that these functions do not grow “too much” *as a function of* $n$, since $\mathbb{E}f_{2n}/\mathbb{E}f_n \leq 2$. For large values of $d$, with very high probability, $f_{2n} \leq cf_n$. In particular, this applies to empirical complexities of a function class $F$, since the empirical VC-dimension, empirical VC-entropy, empirical fat-shattering dimension, and empirical Rademacher averages of $F$ are self-bounding, symmetric functions, as shown in the following corollary.

**Lemma 3.15** *Let $F$ be a class of binary-valued functions. Then the empirical VC-dimension and the empirical VC-entropy are self-bounding, symmetric functions, which satisfy, for each $n$, conditions 1,2, and 3 in Lemma 3.14. If $F$ is a class of functions bounded by $[-1,1]$, the same holds for the empirical fat-shattering dimension and the* *empirical Rademacher averages of $F$.*

In order to prove the self-bounding property, define

$$\begin{aligned} g_{\widehat{VC}}(x_1,\dots,x_{n-1}) &= \widehat{VC}\left(F,(x_1,\dots,x_{n-1})\right) , \\ g_{\widehat{H}_{VC}}(x_1,\dots,x_{n-1}) &= \widehat{H}_{VC}\left(F,(x_1,\dots,x_{n-1})\right) , \\ g_{\widehat{R}_n}(x_1,\dots,x_{n-1}) &= \widehat{R}_n\left(F,(x_1,\dots,x_{n-1})\right) , \end{aligned}$$

and, for any fixed $\epsilon$, let

$$g_{\widehat{\mathrm{fat}}_\epsilon}(x_1,\dots,x_{n-1}) = \widehat{\mathrm{fat}}_\epsilon\left(F,(x_1,\dots,x_{n-1})\right) ,$$

and check that conditions 1 and 2 are satisfied in each case. Detailed proofs can be found in Boucheron et al. (2004a). The symmetry is clear from the definition.

Since the empirical complexities in Lemma 3.15 are self-bounding functions, it follows from Theorem 3.13 that they are highly concentrated around their expectation. Additionally, as a function of the sample size $n$, they do not grow "too much", since with high probability $f_{2n}/f_n \le c$. This is a strong constraint if we consider that the cardinality of the coordinate projections of $F$ onto random samples of size $2n$ compared to that onto random samples of size $n$ could grow exponentially with $n$, that is $f_{2n}/f_n \sim 2^n$.

We will state explicitly the concentration result for the empirical Rademacher averages following from Theorem 3.13 in a form which will be convenient for later results in this thesis. This corollary shows how the Rademacher averages of a class can be upper bounded by the empirical Rademacher averages of this class. The following formulation can be found in Bartlett et al. (2004a). Note that, for general bounded functions, in order to apply Lemma 3.15 one has to first scale the functions such that they only take values in $[-1, 1]$.

**Corollary 3.16** *Let $F$ be a class of bounded functions defined on $\Omega$ taking values in $[a,b]$, and let $\mu$ be a probability measure on $\Omega$. Then, for any $0 \le \alpha \le 1$ and any $t > 0$, with probability at least $1 - e^{-t}$,*

$$R_n\left(F\right) \le \frac{1}{1-\alpha}\widehat{R}_n\left(F\right) + \frac{(b-a)t}{4\alpha(1-\alpha)} .$$

We conclude this section by noting that, although we apply the inequalities presented in this section only in our specific machine learning setting, their usefulness is far reaching beyond machine learning applications [5].

[5] They were used for results in analysis of algorithms, discrete and combinatorial mathematics (graph theory), geometry, functional analysis and infinite-dimensional integration, complexity theory, and probability theory (McDiarmid 1998; Molloy 1998; Steele 1997; Boucheron et al. 2004a).

## 3.2 Symmetrization

The symmetrization technique was first considered by Kahane (1968), and further developed by Hoffmann-Jørgensen (1977). The following variants of symmetrization results for probabilities and expectations can be found, together with their proofs, in van der Vaart and Wellner (1996), Chapter 2.3.

Let $F$ be a class of real-valued functions defined on a measurable space $\Omega$ and which take values in $[-1,1]$, and set $\mu$ to be a probability measure on $\Omega$. Let $X_1,...,X_n$ be independent random variables distributed according to $\mu$. The main idea of symmetrization is to introduce an additional i.i.d. sample $(X'_1,\ldots,X'_n)$ (the "ghost sample") which is independent of $(X_1,\ldots,X_n)$. The symmetrization technique makes it possible to relate the deviation of a mean from the expectation to the deviation of the empirical means evaluated on two different samples. This is a key procedure for obtaining data-dependent generalization error bounds.

**Theorem 3.17 (Symmetrization by a ghost sample)** *Let $F$ be defined as above. Then, for any probability measure $\mu$ and every $t>0$,*

$$\left(1-\frac{4}{nt^2}\sup_{f\in F}\mathrm{Var}\,(f)\right) Pr_{\mathbf{X}}\left\{\sup_{f\in F}\ \Big|\sum_{i=1}^n f(X_i)-\mathbb{E}_\mu f\Big|\geq t\right\}$$
$$\leq Pr_{\mathbf{X},\mathbf{X}'}\left\{\sup_{f\in F}\Big|\sum_{i=1}^n\big(f(X_i)-f(X'_i)\big)\Big|\geq\frac{t}{2}\right\},$$

*where $\mathbf{X}=(X_1,...,X_n)$ is an i.i.d. sample distributed according to $\mu^n$ , and $\mathbf{X}'=(X'_1,...,X'_n)$ is an independent copy of $\mathbf{X}$.*

*Also,*

$$\mathbb{E}_{\mathbf{X}}\left\{\sup_{f\in F}\Big|\sum_{i=1}^n f(X_i)-\mathbb{E}_\mu f\Big|\right\}\leq\mathbb{E}_{\mathbf{X},\mathbf{X}'}\left\{\sup_{f\in F}\Big|\sum_{i=1}^n\big(f(X_i)-f(X'_i)\big)\Big|\right\}.$$

The proof is standard and can be found for example in van der Vaart and Wellner (1996), pages 108 and 112. The first inequality is based on estimating the probability that, for any given function $f$, both independent events $\big|\sum_{i=1}^n f(X_i)-\mathbb{E}_\mu f\big|\geq t$ and $\big|\sum_{i=1}^n\left(f(X'_i)-\mathbb{E}_\mu f\right)\big|\leq t/2$ hold simultaneously. The probability of the second event can be lower bounded with Chebyshev's inequality for the complementary event (Theorem 3.2, page 36), which leads to the factor $\big(1-\frac{4}{nt^2}\sup_{f\in F}\mathrm{Var}\,(f)\big)$ on the left-hand side. The second inequality follows directly from the triangle inequality.

**Theorem 3.18 (Symmetrization by random signs)** *Let $F$ be defined as above.*

*For every $t > 0$,*

$$Pr_{\mathbf{X},\mathbf{X}'}\left\{\sup_{f\in F}\big|\sum_{i=1}^{n}\big(f(X_i)-f(X_i')\big)\big|\geq\frac{t}{2}\right\}\leq 2Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\left\{\sup_{f\in F}\big|\sum_{i=1}^{n}\varepsilon_i f(X_i)\big|\geq\frac{t}{4}\right\},$$

*where $\mathbf{X}=(X_1,...,X_n)$ is an i.i.d. sample distributed according to $\mu^n$, $\mathbf{X}'=(X_1',...,X_n')$ is an independent copy of $\mathbf{X}$, , and $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)$ are independent Rademacher variables.*

*Also,*

$$\mathbb{E}_{\mathbf{X},\mathbf{X}'}\left\{\sup_{f\in F}\big|\sum_{i=1}^{n}\big(f(X_i)-f(X_i')\big)\big|\right\}\leq 2\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}\left\{\sup_{f\in F}\big|\sum_{i=1}^{n}\varepsilon_i f(X_i)\big|\right\}.$$

A proof can be found in van der Vaart and Wellner (1996), pages 109 and 112, and follows directly from the triangle inequality, the convexity of the supremum, and the fact that $\mathbf{X}$ and $\mathbf{X}'$ have the same distribution.

The following two corollaries follow directly by combining the results for the probabilities and expectations in Theorems 3.17 and 3.18. They relate the tail probability and the expectation of $\sup_{f\in F}|\mathbb{E}f-\mathbb{E}_n f|$ to the Rademacher penalties and Rademacher averages of $F$.

**Corollary 3.19 (Symmetrization for probabilities)** *Let $F$ be a class of bounded functions defined on a probability space $(\Omega,\mu)$, let $X_1,...,X_n$ be independent random variables distributed according to $\mu$, and denote by $\mathbb{E}_n f$ the empirical average of $f\in F$ on $X_1,...,X_n$. Then, for any $t>0$ and for any $n\geq 8\sup_{f\in F}\mathrm{Var}\left(f\right)/t^2$ ,*

$$Pr\left\{\sup_{f\in F}|\mathbb{E}f-\mathbb{E}_n f|\geq t\right\}\leq 2Pr\left\{\sup_{f\in F}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\geq\frac{t}{4}\right\}.$$

**Corollary 3.20 (Symmetrization for expectations)** *Let $F$ be a class of functions defined on a probability space $(\Omega,\mu)$, and let $X_1,...,X_n$ be independent random variables distributed according to $\mu$. Then,*

$$\mathbb{E}\sup_{f\in F}|\mathbb{E}f-\mathbb{E}_n f|\leq 2\frac{R_n\left(F\right)}{n},$$

*where $R_n\left(F\right)$ are the Rademacher averages of $F$.*

Chapter 4

# A General Framework for Data-Dependent Generalization Bounds

## 4.1 Motivation and Overview

In this chapter, we present a new framework to derive generalization bounds for learning algorithms which extends the standard approach from a fixed hypothesis class to *random* classes of hypothesis functions, that is, to function classes which depend on the sample. Recall that the standard approach is based on the analysis of *uniform* deviations of empirical averages from their expectations by deriving estimates for the probability

$$Pr_{\mathbf{X}}\Big\{\sup_{f\in F}\big|\mathbb{E}_{\mu}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\big|\geq t\Big\}\,, \tag{4.1}$$

where $F$ is the entire loss or excess loss class and $\mathbf{X}=(X_1,...,X_n)$ is an i.i.d. sample of size $n$ distributed according to a probability measure $\mu^n$. The results obtained hold therefore for *any* hypothesis function in the hypothesis class, and are usually pessimistic when applied to the *specific* function produced by a *specific* algorithm. Thus, it is desirable to get an estimate for the tail probability of the deviation of empirical averages from their expectations *only* for the hypothesis function produced by the specific learning algorithm when presented with the actual training sample.

However — and this is the motivation to explore bounds for random classes of functions — the function produced by the algorithm based on a random sample depends on this sample and is therefore a *random* function. For example, for support vector machines this function is the hyperplane having a maximum margin *on the actual sample.* The difficulty which arises in analyzing its generalization ability with the usual concentration tools is that the loss of this function is also random and could change with the sample. As presented in Section 3.1, Hoeffding's or Bernstein's exponential inequalities rely on the fact that all the summands of the sum $\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)$ are independent (see equation (3.1) in the proof of Hoeffding's inequality). Applied to the random variables $f(X_i)$, if $f$ is a random function, this independence does not hold. Hence, it is not possible to use the classical generalization bounds which are based on Hoeffding's or Bernstein's inequalities (Theorem 3.3 and 3.6) in order to directly bound the tail of the deviation of the empirical average from the expectation for this single function.

Furthermore, it can be useful to derive generalization bounds not only for the loss of the *one* function produced by an algorithm, but for the losses of a larger *random subset* of the function class in which the output function will be contained (however, this random subset could still be significantly "smaller" than the whole hypothesis class). For example, in case of noisy data, it is useful to relax the support vector optimization constraint and allow the algorithm to produce a hyperplane having a margin which differs by at most $\varepsilon$ from the margin of the maximum margin hyperplane. In this case, one is interested in a bound which holds uniformly for all losses corresponding to the hyperplanes $\varepsilon$ away from the maximum margin solution. Another random subset of functions which is interesting for kernel algorithms is the data-dependent class $K = \{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n\}$ (see page 34), as this class is potentially much smaller than the set of *all* linear combinations of kernel functions $\operatorname{span}(\{k(x,\cdot) : x \in \Omega\})$ which form the hypothesis class.

In the following, we will show that it is possible to modify the classical proof for deriving uniform tail bounds for (4.1) in order to obtain generalization bounds for *random classes of functions*. The tools we use are modified symmetrization techniques and concentration results (Section 4.2). We will particularly show that the mechanism which allows one to obtain generalization bounds for random subclasses is based on the following two principles: The first, which guarantees learnability, is a "small complexity" of random classes measured by Rademacher averages of certain random coordinate projections. The second, which determines the confidence intervals and the sample complexity, is the degree of concentration of suprema of Rademacher processes indexed by these same random coordinate projections.

In order to demonstrate the generality of the proposed framework for random subclass bounds, we will then give different examples of results which follow as special cases from the random subclass framework (Section 4.3). We will first show that the classical Glivenko-Cantelli results fall easily within our framework (Section 4.3.1), as well as some previously proposed bounds for random classes of functions derived in Gat (1999) and Cannon et al. (2002) (Section 4.3.2).

Furthermore, a range of seemingly different frameworks like compression schemes (Littlestone and Warmuth 1986; Floyd and Warmuth 1995), sparsity (Herbrich and

Williamson 2002), and the luckiness bounds (Shawe-Taylor et al. 1998; Herbrich and Williamson 2002) also fall into our framework. These frameworks were derived based on a different philosophy, as their starting point is to use, besides the standard assumptions of a probabilistic model, prior knowledge about properties of the specific algorithm or about the distribution. The complexity notions defined in these frameworks reflect these different assumptions and are thus different from the uniform complexity notions in statistical learning theory. Nevertheless, as we will show in Sections 4.3.3 and 4.3.4, the assumption made in the compression, sparsity, and luckiness frameworks are intrinsically the same as in our random subclass framework (and thus also the same as in the classical framework), namely, they are just different ways of ensuring small coordinate projections paired with a sufficient condition for concentration.

Finally, we will demonstrate that the reason why one can derive sharper bounds on the sample complexity for ERM is also a combination of small coordinate projections and a strong concentration result. In Section 4.3.5, we will show that the crucial property allowing us to derive these sharper bounds on the sample complexity for ERM is the fact that one can control the variance of the functions in the class by the expectation and that this control of the variance allows us to obtain this stronger degree of concentration.

The results presented in this chapter were published (in a slightly different form) in Mendelson and Philips (2003, 2004).

## 4.2 Random Subclass Bounds

Let $F$ be a class of bounded real-valued functions defined on a measurable space $\Omega$ with underlying measure $\mu$ and taking values in $[-b, b]$. Recall that, in the learning setting, such a class arises as a loss or excess loss class associated with a fixed hypothesis class and a bounded loss function. For every integer $n$, let $F_n$ denote a set-valued map which assigns to each $\mathbf{x} \in \Omega^n$ a subset of $F$. The quantity we want to bound in the sequel is the probability

$$Pr_{\mathbf{X}}\Big\{ \sup_{f \in F_n(\mathbf{X})} \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t \Big\}, \tag{4.2}$$

where $\mathbf{X} = (X_1, ..., X_n)$ is a random sample drawn according to $\mu^n$.

Many results of statistical learning theory are for non-random classes of functions, $F_n(\mathbf{X}) = F$, where one obtains a “worst-case” bound on the generalization error which holds uniformly over the entire loss class associated with the hypothesis class. The other extreme is to provide a bound only for the singleton containing the loss of the function produced by a specific learning algorithm from $\mathbf{X}$.

In this section, we will derive generalization bounds for random classes of functions. As we will show, the complexity notion which determines the generalization bound for random classes are random coordinate projections of certain symmetric sets, similar to the coordinate projections in the non-random case.

The derivation is analogous to the non-random case (see Section 3.2). It is based on two symmetrization steps followed by application of a concentration inequality.

### 4.2.1 Symmetrization

The main result we present in this section, Theorem 4.5, is a random symmetrization theorem which will enable us to bound the quantity in equation (4.2) in terms of suprema of Rademacher processes associated with sets of coordinate projections. Following the original proof of the Glivenko-Cantelli case, we employ a symmetrization procedure in two steps: first, a symmetrization by a ghost sample which relates the deviation of the expectation from the empirical average of functions in a random class to the deviation of the empirical averages evaluated on two different samples; second, a symmetrization by signs which relates the latter deviation to the probability of having "large" Rademacher sums.

The first symmetrization step is a variation of the standard symmetrization step for probabilities (Theorem 3.17). The proof follows closely the proof for non-random function classes given in van der Vaart and Wellner (1996) (Lemma 2.3.7, page 112).

**Lemma 4.1 (Symmetrization by a ghost sample)** *Let $F_n$ be defined as above. For any probability measure $\mu$, and every $t > 0$,*

$$\Big(1 - \frac{4n}{t^2} \sup_{f \in F} \mathrm{Var}\,(f)\Big) Pr_{\mathbf{X}}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\sum_{i=1}^{n}\big(f(X_i) - \mathbb{E}_\mu f\big)\big| \geq t\Big\}$$

$$\leq Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\big| \geq \frac{t}{2}\Big\},$$

*where $\mathbf{X} = (X_1, ..., X_n)$ is an i.i.d. sample distributed according to $\mu^n$ and $\mathbf{X}' = (X_1', ..., X_n')$ is an independent copy of $\mathbf{X}$.*

**Proof:** Fix $t > 0$. Let

$$\beta := \inf_{f \in F} Pr_{\mathbf{X}}\Big\{\big|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}_\mu f\right)\big| \leq t/2\Big\}$$

and define the subset $A \subseteq \Omega^n$ as

$$A := \Big\{\mathbf{x} \in \Omega^n :\ \exists f \in F_n(\mathbf{x}),\ \big|\sum_{i=1}^{n}\left(f(x_i) - \mathbb{E}_\mu f\right)\big| \geq t\Big\}.$$

By the definition of $A$, for every element $\mathbf{x} \in A$ there is some $f_{\mathbf{x}} \in F_n(\mathbf{x})$ such that $|\sum_{i=1}^{n}(f_{\mathbf{x}}(x_i)-\mathbb{E}_\mu f_{\mathbf{x}})|\geq t$. Fix this function $f$ and observe that by the triangle inequality,

$$\text{if}\quad |\sum_{i=1}^{n}\big(f_{\mathbf{x}}(X_i')-\mathbb{E}_\mu f_{\mathbf{x}}\big)|\leq t/2\qquad\text{then}\quad |\sum_{i=1}^{n}\big(f_{\mathbf{x}}(x_i)-f_{\mathbf{x}}(X_i')\big)|\geq t/2\,. \tag{4.3}$$

Since $\mathbf{X}'$ is an independent copy of $\mathbf{X}$,

$$\begin{aligned}
\beta &\leq Pr_{\mathbf{X}'}\Big\{|\sum_{i=1}^{n}\big(f_{\mathbf{x}}(X_i')-\mathbb{E}_\mu f_{\mathbf{x}}\big)|\leq\frac{t}{2}\Big\} && \text{(definition of } \beta)\\
&\leq Pr_{\mathbf{X}'}\Big\{|\sum_{i=1}^{n}\big(f_{\mathbf{x}}(x_i)-f_{\mathbf{x}}(X_i')\big)|\geq\frac{t}{2}\Big\} && \text{(by (4.3))}\\
&\leq Pr_{\mathbf{X}'}\Big\{\exists f\in F_n(\mathbf{x}),\ |\sum_{i=1}^{n}\big(f(x_i)-f(X_i')\big)|\geq\frac{t}{2}\Big\}=:\alpha(\mathbf{x})\,. && (4.4)
\end{aligned}$$

Note that $\beta$ is independent of the specific selection of $f\in F$, and the extreme right side of this inequality is independent of the specific selection of $f\in F_n(\mathbf{x})$. Inequality (4.4) holds on the whole set $A$, and therefore we can integrate both extreme sides of the inequality with respect to $\mathbf{x}$ on $A$. First, recall the basic fact that for any set $A\subseteq\Omega^n$, $\int_A d\mu_n(\mathbf{X})=\int_{\Omega^n}I_A(\mathbf{X})\,d\mu_n(\mathbf{X})=Pr_{\mathbf{X}}\{A\}$. Therefore, by integrating inequality (4.4) we obtain $\beta Pr_{\mathbf{X}}\{A\}\leq\int_A\alpha(\mathbf{x})\,d\mu_n(\mathbf{X})$. We can write $\alpha(\mathbf{x})$ in terms of the indicator function of the set $A'(\mathbf{X})$ as $\alpha(\mathbf{x})=\int_{\Omega^n}I_{A'(\mathbf{X})}(\mathbf{X}')\,d\mu_n(\mathbf{X}')$, where

$$A'(\mathbf{x}):=\Big\{\mathbf{x}'\in\Omega^n:\ \exists f\in F_n(\mathbf{x}),\ |\sum_{i=1}^{n}\big(f(x_i)-f(x_i')\big)|\geq\frac{t}{2}\Big\}.$$

Since this quantity is always positive, the integral over the set $A$ is upper bounded by the integral over the whole set $\Omega^n$, and therefore $\int_A\alpha(\mathbf{x})\,d\mu_n(\mathbf{X})$ is upper bounded by

$$\int_{\Omega^n}\Big(\int_{\Omega^n}I_{A'(\mathbf{X})}(\mathbf{X}')\,d\mu_n(\mathbf{X}')\Big)\,d\mu_n(\mathbf{X})=\int_{\Omega^{2n}}I_{A'(\mathbf{X})}(\mathbf{X}')\,d\mu_{2n}(\mathbf{X},\mathbf{X}')\,.$$

It follows that

$$\begin{aligned}
&\beta Pr_{\mathbf{X}}\Big\{\exists f\in F_n(\mathbf{X}),\ |\sum_{i=1}^{n}(f(x_i)-\mathbb{E}_\mu f)|\geq t\Big\}\\
&\qquad\qquad\leq Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f\in F_n(\mathbf{X}),\ |\sum_{i=1}^{n}\big(f(X_i)-f(X_i')\big)|\geq\frac{t}{2}\Big\}.
\end{aligned}$$

Finally, to estimate $\beta$, note that by Chebyshev's inequality (Theorem 3.2, page 36),

$$Pr_{\mathbf{X}}\Big\{ \big| \sum_{i=1}^{n} (f(X_i) - \mathbb{E}_\mu f) \big| \geq \frac{t}{2} \Big\} \leq \frac{4n}{t^2} \mathrm{Var}\,(f)$$

for every $f \in F$, and thus, $\beta \geq 1 - (4n/t^2) \sup_{f \in F} \mathrm{Var}\,(f)$. ■

The second symmetrization step relates the tail behaviour of deviations of empirical averages evaluated on two samples to the behaviour of the supremum of a Rademacher process. Its importance lies in the fact that we can study the behaviour of suprema of Rademacher processes by studying their expectation, since suprema of Rademacher processes are concentrated around their expectation (cf. Corollary 3.9 and Theorem 3.11). This symmetrization step requires an additional property of the random subclass, namely, that it is invariant under permutations of the sample.

**Definition 4.2 (symmetric function)** *Let $\Phi_n$ be a set-valued map from $\Omega^n$ to subsets of $F$. We say that the function $\Phi_n$ is symmetric, if for every $\mathbf{x} \in \Omega^n$ and every permutation $\pi(\mathbf{x})$ of $\mathbf{x}$, $\Phi_n(\mathbf{x}) = \Phi_n(\pi(\mathbf{x}))$.*

**Lemma 4.3 (Symmetrization by random signs)** *Let $F_{2n}^{\mathrm{sym}}$ be a symmetric map from subsets of $\Omega^{2n}$ to subsets of $F$. Then, for any probability measure $\mu$ on $\Omega$ and every $t > 0$,*

$$\begin{aligned} &Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\big| \geq t\Big\} \\ &\qquad \leq 2Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \big|\sum_{i=1}^{n}\varepsilon_i f(X_i)\big| \geq \frac{nt}{2}\Big\}, \end{aligned}$$

*where $\mathbf{X} = (X_1, ..., X_n)$ is an i.i.d. sample distributed according to $\mu^n$, $\mathbf{X}' = (X_1', ..., X_n')$ is an independent copy of $\mathbf{X}$, , and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are independent Rademacher variables.*

**Proof:** By the symmetry of $F_{2n}^{\mathrm{sym}}$ it follows that for every $(\varepsilon_1, ..., \varepsilon_n) \in \{-1, 1\}^n$,

$$\begin{aligned} &Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\big| \geq t\Big\} \\ &\qquad = Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(f(X_i) - f(X_i')\big)\big| \geq t\Big\}. \end{aligned}$$

Taking the expectation with respect to the random signs (that is, with respect to the Rademacher random variables $\boldsymbol{\varepsilon}$), the proof follows from the triangle inequality and the fact that $(X_1, ..., X_n)$ has the same distribution as $(X_1', ..., X_n')$. ■

In order to make use of the two symmetrization results, we need to find an appropriate *symmetric* set-valued map $F_{2n}^{\mathrm{sym}}$ . The following definition relating the set-valued maps $F_n$ and $F_{2n}^{\mathrm{sym}}$ will be needed later:

**Definition 4.4 ($\delta$-symmetric condition)** *Let $F_n$ and $F_{2n}^{\mathrm{sym}}$ be set-valued maps. We say that $F_n$ and $F_{2n}^{\mathrm{sym}}$ satisfy the $\delta$-symmetric condition, if there exists a constant $\delta > 0$ such that, for every $t > 0$,*

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\big| \geq t\Big\} \leq$$

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\big| \geq t\Big\} + \delta\,, \tag{4.5}$$

*where $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{X}' = (X_1', ..., X_n')$.*

The $\delta$-symmetric condition quantifies that by replacing the original random subset of hypotheses with another symmetric random subset dependent on the double-sample — and which is therefore invariant under permutations of this double-sample — the probability of having large deviations of empirical means evaluated on the sample and ghost sample increases by at most $\delta$. Thus, by giving up only a constant probability, we can proceed with the symmetrization by random signs which reduces the estimation of a bound on the generalization ability for random classes of functions to the study of Rademacher processes indexed by random sets.

Note that (4.5) holds trivially with a constant $\delta = 0$ if for every double-sample $(\mathbf{X},\mathbf{X}')$, $F_n(\mathbf{X}) \subseteq F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$. An extreme case occurs when both set-valued maps are the constant set-valued map, $F_n(\mathbf{X}) = F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')) = F$ .

Given $F_n$, one can always define a symmetric mapping $F_{2n}^{\mathrm{sym}}$ to satisfy (4.5) with $\delta = 0$ as the *symmetric extension of $F_n$*. The symmetric extension $\mathrm{Sym}_{(x_1,...,x_{2n})}(F_n)$ is defined to be the union of all subsets corresponding to the first half of permutations of the double-sample $(x_1, \ldots, x_{2n})$,

$$\mathrm{Sym}_{(x_1,...,x_{2n})}(F_n) := \bigcup_{\pi \in S_{2n}} F_n\big(\pi(x_1, \ldots, x_{2n})|_{i=1}^{n}\big)\,, \tag{4.6}$$

where $S_{2n}$ is the set of all permutations of the numbers $1, \ldots, 2n$, and $\pi(x_1, \ldots, x_{2n}) = (x_{\pi(1)}, \ldots, x_{\pi(2n)})$ . If for every double-sample $(\mathbf{X},\mathbf{X}')$, $F_{2n}^{\mathrm{sym}}(\mathbf{X},\mathbf{X}') := \mathrm{Sym}_{\mathbf{X},\mathbf{X}'}(F_n)$ , then $F_{2n}^{\mathrm{sym}}$ and $F_n$ satisfy a 0-symmetric condition, since $F_n(\mathbf{X}) \subseteq F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ .

Although we can always find a symmetric mapping $F_{2n}^{\mathrm{sym}}$ by a symmetric extension of $F_n$, Definition 4.4 allows us to replace the original subset $F_n(\mathbf{X})$ with a symmetric subset $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ potentially smaller than $\mathrm{Sym}_{\mathbf{X},\mathbf{X}'}(F_n)$ as long as the change in probabilities can be controlled (as done in the luckiness frameworks). The best ("smallest") symmetric set would thus be one with the smallest corresponding Rademacher process. Unfortunately, there is no general way of finding such a "better" symmetric map (in the same sense in which there is no general way to find good luckiness functions, see Section 4.3.4).

The main result of this section is the following symmetrization theorem obtained by combining Lemma 4.1, the $\delta$-symmetric condition (4.5), and Lemma 4.3.

**Theorem 4.5 (Symmetrization theorem)** *Let $F$ be a class of real-valued functions defined on a measurable space $(\Omega, \mu)$. Let $\mathbf{X} = (X_1, ..., X_n)$ be an i.i.d. sample distributed according to $\mu^n$, and let $\mathbf{X}' = (X_1', ..., X_n')$ be an independent copy of $\mathbf{X}$. Let $0 \leq \delta \leq 1$ and let $F_n$ and $F_{2n}^{\mathrm{sym}}$ denote set-valued maps which satisfy the $\delta$-symmetric condition and such that $F_{2n}^{\mathrm{sym}}$ is symmetric. Then, for every $t > 0$,*

$$\Big(1 - \frac{4}{nt^2} \sup_{f \in F} \mathrm{Var}\,(f)\Big) \cdot Pr_{\mathbf{X}}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\big| \geq t\Big\}$$
$$\leq 2Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\Big\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \ \big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \geq \frac{nt}{4}\Big\} + 2\delta\,,$$

*where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are independent Rademacher variables.*

Theorem 4.5 reduces the analysis of the uniform deviation of expectation and empirical averages of functions in $F_n$ to the analysis of the supremum of the empirical process

$$\sup_{f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| ,$$

which is, conditioned on $(\mathbf{X}, \mathbf{X}')$, a Rademacher process, and we can thus use analysis techniques and employ concentration results for empirical processes to study this quantity.

To sum up, the quantity which determines the generalization bounds for random subclasses of functions is the supremum of the Rademacher process

$$\boxed{\sup_{\mathbf{v} \in V(\mathbf{X},\mathbf{X}')} \left|\sum_{i=1}^n \varepsilon_i v_i\right|} \tag{4.7}$$

indexed by the coordinate projection $V(\mathbf{X}, \mathbf{X}') \subseteq \mathbb{R}^n$ of the random class $F_{2n}^{\mathrm{sym}}((\mathbf{X}, \mathbf{X}'))$ onto $\mathbf{X} = (X_1, ..., X_n)$ , where

$$V(\mathbf{X}, \mathbf{X}') := F_{2n}^{\mathrm{sym}}((\mathbf{X}, \mathbf{X}'))/\mathbf{X} = \Big\{\big(f(X_1), ..., f(X_n)\big) :\ f \in F_{2n}^{\mathrm{sym}}((\mathbf{X}, \mathbf{X}'))\Big\}\,.$$

We will see in the following how control of the geometry of the set $V(\mathbf{X}, \mathbf{X}')$ allows one to obtain sharp tail bounds for the probability of the deviation

$$Pr_{\mathbf{X}}\Big\{ \sup_{f \in F_n(\mathbf{X})} \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t \Big\} . \tag{4.8}$$

We define

$$Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) := \sup_{\mathbf{v} \in V(\mathbf{X},\mathbf{X}')} \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| , \tag{4.9}$$

and recall that by Corollary 3.9 the variable $Z_{\mathbf{X},\mathbf{X}'}$ is concentrated around its expectation whenever $V(\mathbf{X},\mathbf{X}')$ are bounded in $\ell_\infty^n$. Recall also that, by the definition of empirical Rademacher averages (Definition 2.18, page 26),

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \widehat{R}_n \left( F_{2n}^{\mathrm{sym}}\left((\mathbf{X},\mathbf{X}')\right), \mathbf{X} \right) .$$

With this, we are ready to state the concentration results in the next section.

### 4.2.2 Concentration

**Corollary 4.6 (Concentration I)** *With the notation above, for every $t > 0$, for all $n > 0$,*

$$\begin{aligned}\Big(1 - \frac{4}{nt^2} \sup_{f \in F} \mathrm{Var}\,(f)\Big) \cdot Pr_{\mathbf{X}}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t\Big\} \\ \leq 2\Big( Pr_{\mathbf{X},\mathbf{X}'}\left\{\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) > nt/8\right\} \\ + Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\Big\{\big|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\big| \geq \frac{nt}{8}\Big\}\Big) + 2\delta .\end{aligned} \tag{4.10}$$

*In particular, for bounded functions taking values only in $[-b, b]$, for any $t$ and $n \geq 8 \sup_{f \in F} \mathrm{Var}\,(f)/t^2$,*

$$\begin{aligned}Pr_{\mathbf{X}}\Big\{\exists f \in F_n(\mathbf{X}),\ \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq t\Big\} \\ \leq 4 Pr_{\mathbf{X},\mathbf{X}'}\left\{\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) > nt/8\right\} + c e^{-\frac{c' nt^2}{b^2}} + 4\delta ,\end{aligned} \tag{4.11}$$

*where $c, c' > 0$ are absolute constants.*

**Proof:** By Theorem 4.5, for each $t$, the left-hand side of (4.10) is bounded from above by the probability of the random event

$$B_t := \Big\{ (\mathbf{X}, \mathbf{X}', \boldsymbol{\varepsilon}) \,:\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) \geq \frac{nt}{4} \Big\}$$

plus $2\delta$. Let

$$A_t := \left\{(\mathbf{X}, \mathbf{X}') : \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) \le nt/8\right\}. \tag{4.12}$$

By the union bound, if $A_t^c$ denotes the complement of the set $A_t$, since $B_t \cap A_t^c \subseteq A_t^c$ and $A_t^c$ is independent of $\boldsymbol{\varepsilon}$,

$$\begin{aligned} Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\{B_t\} &= Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\{B_t \cap A_t^c\} + Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\{B_t \cap A_t\} \\ &\le Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\{A_t^c\} + Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\left\{Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) \ge \frac{nt}{4} \text{ and } \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) \le \frac{nt}{8}\right\} \\ &\le Pr_{\mathbf{X},\mathbf{X}'}\{A_t^c\} + Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\left\{\left|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\right| \ge \frac{nt}{8}\right\}, \end{aligned}$$

which proves the first claim. For bounded functions, we can apply Corollary 3.9 to $Z_{\mathbf{X},\mathbf{X}'}$ to obtain (4.11). ■

Thus, in order to obtain, for given $t > 0$ and $n > 0$, a bound for (4.8) under the $\delta$-symmetric condition (4.5), it is sufficient to ensure that the following additional conditions are satisfied:

1. **Small coordinate projection:** With high probability over random samples,

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) \le nt/8\,; \tag{4.13}$$

2. **Concentration:** With high probability $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ is concentrated around its mean,

$$\left|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\right| \le nt/8\,. \tag{4.14}$$

The first condition is a "complexity condition" on the random class $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$, which quantifies that the coordinate projections $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ are "small". In order to ensure *learnability*, one would have to show that for each $t > 0$, and for $n \ge n_0$, this condition is satisfied, which is equivalent to the expectation of the Rademacher process indexed by $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ being $o(n)$. Such a condition can be ensured by restrictions on the geometry of the random coordinate projections $V(\mathbf{X},\mathbf{X}') = F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ . One extreme (and bad) example is when $V(\mathbf{X},\mathbf{X}')$ is close to the unit cube (unit ball in $\ell_\infty^n$) for "many" $(\mathbf{X},\mathbf{X}')$, in which case $\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = n$ ; [1] a strong restriction on the geometry occurs when $V(\mathbf{X},\mathbf{X}')$ is a subset of the unit ball in $\ell_2^n$, which implies that $\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = O(\sqrt{n})$ and that equation (4.13) is satisfied.

The second condition controls the degree of concentration. Clearly, in order to just ensure that $Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\left\{\left|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\right| \ge nt/8\right\}$ goes to 0 as $n$ tends to

[1] This is the reason why cubic structures in VC theory are bad. A finite VC-dimension ensures precisely that, for large $n$, the projections of the whole non-random class $F/\mathbf{X}$ are much smaller than the full combinatorial cube $\{-1,1\}^n$. Here, we have the real-valued analogue of this fact.

infinity, a finite variance of the random variable $Z_{\mathbf{X},\mathbf{X}'}$, together with Chebysheff's inequality would be sufficient. However, already for bounded functions, one obtains easily a stronger degree of concentration as given in equation (4.11). For $\delta = 0$, and since $ce^{-\frac{c'nt^2}{b^2}} \longrightarrow 0$ for any fixed $t$, one has thus to study only the average behaviour $F_{2n}^{\text{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ in order to determine, for a given confidence $\rho$, exponential bounds on the rate of convergence $t = \gamma(n,\rho)$ as a function of the sample size $n$ (see Section 2.2, especially equation (2.1), page 18). Hence, for every fixed $\rho$, condition 1 determines the *rate of convergence* and ensures *learnability.*

In order to obtain a better degree of confidence and to ensure that the confidence intervals for the expected risk of the function produced by an algorithm tend to 0 with growing sample size, we require a stronger degree of concentration for the variable $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$. One way to obtain a stronger concentration result, as was presented in Section 3.1, is through control of the diameter of $V(\mathbf{X},\mathbf{X}')$ in $\ell_2^n$. Indeed, recall that, for any fixed $(\mathbf{X},\mathbf{X}')$, the function $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ is convex in $\boldsymbol{\varepsilon}$ (since $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ conditioned on $(\mathbf{X},\mathbf{X}')$ is the norm of $\boldsymbol{\varepsilon} = \sum_{i=1}^n \varepsilon_i e_i$ in the dual space of the space defined by the absolute convex hull of $V(\mathbf{X},\mathbf{X}')$; see Appendix A). By the convex distance inequality, such a function is concentrated around its expectation $\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$, as long as the diameter of $V(\mathbf{X},\mathbf{X}')$ in $\ell_2^n$ is finite (cf. Theorem 3.11, page 43), and the degree of concentration is determined by the diameter of $V(\mathbf{X},\mathbf{X}')$ in $\ell_2^n$. Thus, we can study the behaviour of the random variable $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ by studying its expectation and the diameter of $V(\mathbf{X},\mathbf{X}')$ in $\ell_2^n$. Theorem 4.7 states this formally. Furthermore, as we will see, different assumptions allow one to get different bounds on the diameter of $V(\mathbf{X},\mathbf{X}')$ and thus different levels of concentration.

For example, the fact that the functions in $F$ are bounded, a standard assumption in the statistical learning setting, implies that the set $V(\mathbf{X},\mathbf{X}')$ is contained in a ball of finite radius in $\ell_\infty^n$ and thus of radius $O(\sqrt{n})$ in $\ell_2^n$. Therefore, the centered random variable $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ exhibits a sub-Gaussian tail $e^{-cnt^2}$, which can lead to a sample complexity as good as $c/t^2$ (up to logarithmic terms in $\rho$), or equivalently to convergence rates of $O(1/\sqrt{n})$, as stated already in Corollary 4.6. As we will see in the following Section 4.3, most of the standard results from learning theory ensure precisely that this condition is satisfied. Moreover, these conditions are also already sufficient to show that the *data-dependent* frameworks based on sparsity, compression schemes, and luckiness conditions work to the extent they do.

In order to recover *sub-exponential* concentration bounds for $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$, a stronger control of the diameter in $\ell_2^n$ is necessary. We will give an example of such a case, and show in Section 4.3.5 that such a condition can be imposed by control of the variance of all functions in the class $F$, as is done in a *Bernstein class of functions*, where the variance for every function is bounded by a power of its expectation uniformly over the class (see also Definition 5.1, page 88). We will show that the Bernstein condition restricts the $\ell_2^n$ diameter of projections onto the sample. Such a condition, since it leads to sharper concentration, recovers the tighter results for the generalization performance of the ERM algorithm for Bernstein classes of functions proved in Mendelson (2002b, 2003), which can lead to an improved sample complexity as good as $c/t$ (up to logarithmic factors), and to rates of convergence of $O(1/n)$.

Let us state our main and most general concentration result for the random variable $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ defined in equation (4.9), which follows directly from the corollary of Talagrand's convex distance inequality for suprema of Rademacher processes (Theorem 3.11) and is based on the fact that Rademacher processes are convex functions in $\boldsymbol{\varepsilon}$.

**Theorem 4.7 (Concentration II)** *Let $F$ be a class of functions defined on a measurable space $(\Omega, \mu)$. Let $\mathbf{X} = (X_1, ..., X_n)$ be an i.i.d. sample distributed according to $\mu^n$, let $\mathbf{X}' = (X_1', ..., X_n')$ be an independent copy of $\mathbf{X}$, and let $F_n$ and $F_{2n}^{\mathrm{sym}}$ be defined as above and satisfying the $\delta$-symmetric condition* (4.5)*. Then there exist absolute constants $c, c' > 0$, such that for every $t, v > 0$ and $n \geq 36\pi v/t$,*

$$\left(1 - \frac{4}{nt^2}\sup_{f\in F}\mathrm{Var}\,(f)\right)\cdot Pr_{\mathbf{X}}\left\{\exists f \in F_n(\mathbf{X}),\ \left|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right| \geq t\right\}$$
$$\leq 2\left(Pr_{\mathbf{X},\mathbf{X}'}\left\{\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) > nt/8\right\} + ce^{-\frac{c'n^2t^2}{v^2}}\right.$$
$$\left. + Pr_{\mathbf{X},\mathbf{X}'}\{v_2(\mathbf{X},\mathbf{X}') > v\}\right) + 2\delta\,,$$

*with*

$$V(\mathbf{X},\mathbf{X}') := F_{2n}^{\mathrm{sym}}(\mathbf{X},\mathbf{X}')/\mathbf{X}\,, \qquad v_2(\mathbf{X},\mathbf{X}') := \sup_{\mathbf{v}\in V(\mathbf{X},\mathbf{X}')}\|\mathbf{v}\|_2\,,$$

*and*

$$Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) := \sup_{\mathbf{v}\in V(\mathbf{X},\mathbf{X}')}\left|\sum_{i=1}^{n}\varepsilon_i v_i\right|\,.$$

**Proof:** By Theorem 3.11, if $v_2(\mathbf{X},\mathbf{X}') < v$ then for every $t$,

$$Pr_{\boldsymbol{\varepsilon}}\left\{\left|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\right| \geq nt/8\right\}$$
$$\leq\ Pr_{\boldsymbol{\varepsilon}}\left\{\left|Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) - M_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})\right| \geq nt/8 - 4\pi v\right\} \leq ce^{-\frac{c'n^2t^2}{v^2}}\,,$$

and thus, by the union bound and from Corollary 4.6, equation (4.11), it follows directly that

$$Pr_{\mathbf{X},\mathbf{X}',\epsilon}\left\{\exists f \in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\ \left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| > \frac{nt}{4}\right\}$$
$$\leq Pr_{\mathbf{X},\mathbf{X}'}\left\{\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) > nt/8\right\} + ce^{-\frac{c'n^2t^2}{v^2}} + Pr_{\mathbf{X},\mathbf{X}'}\{v_2(\mathbf{X},\mathbf{X}') \geq v\}\,.$$

■

The parameter $v$ trades off the $\ell_2^n$ radius of a ball centered at the origin in which the projections are contained with high probability against the degree of concentration. For bounded functions, the result from Corollary 4.6 can be recovered directly, by observing that in this case, $\|\mathbf{v}\|_2 \leq c\sqrt{n}$ with probability 1. Thus, for bounded functions, condition (4.14) is trivially satisfied with a degree of concentration given by a sub-Gaussian tail of the form $e^{-cnt^2}$. As we see from this Theorem, if $F$ is a bounded class of functions, the $\delta$-symmetric condition (4.5) is sufficient to guarantee a high probability bound for the generalization error of a learning algorithm drawing its hypotheses from the random set $F_n(\mathbf{X})$ of the order $e^{-cnt^2}$, whenever condition (4.13) is satisfied. Hence, it suffices to guarantee that the empirical Rademacher averages associated with the projection of $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ onto $\mathbf{X}$ are "small" (of size $o(n)$) with high probability.

## 4.3 Examples

We are now ready to show that Glivenko-Cantelli conditions, as well as compression schemes, sparsity, and luckiness conditions are just different ways to ensure that the random coordinate projections of $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ are small, that is, that

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \widehat{R}_n\left(F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\mathbf{X}\right) = o(n)\,, \tag{4.15}$$

and hence the condition (4.13) is satisfied. As we will show, the sample complexities obtained in these frameworks can be recovered within the random subclass framework.

### 4.3.1 Uniform Glivenko-Cantelli Classes

If $F$ is a uniform GC class, one can recover the optimal deviation estimates by selecting the constant set-valued maps

$$F_n(\mathbf{X}) = F_n^{\mathrm{sym}}(\mathbf{X}) := F\,.$$

The $\delta$-symmetric condition (4.5) is therefore satisfied with $\delta = 0$ and

$$V(\mathbf{X},\mathbf{X}') := F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X} = \{(f(X_1),...,f(X_n)) : f \in F\}$$

for every double-sample $(\mathbf{X},\mathbf{X}')$. Then

$$Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \sup_{\mathbf{v}\in V(\mathbf{X},\mathbf{X}')} \Big|\sum_{i=1}^{n}\varepsilon_i v_i\Big| = \sup_{f\in F}\Big|\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\,,$$

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in F} \Big| \sum_{i=1}^{n} \varepsilon_i f(X_i) \Big| = \widehat{R}_n (F, \mathbf{X}) \ .$$

The fact that the coordinate projections are small and that condition (4.13) is satisfied follows from the characterization of uniform Glivenko-Cantelli classes (Theorem 2.32). Assuming that the functions in the class are uniformly bounded, it remains only to show that the empirical Rademacher averages are "small".

Recall that a class of uniformly bounded functions is a uniform Glivenko-Cantelli class if and only if

$$\lim_{n \longrightarrow \infty} \frac{\overline{R}_n (F)}{n} = 0 \, ,$$

where $\overline{R}_n (F)$ denotes the uniform Rademacher averages of $F$ (Theorem 2.32, page 31). Since the uniform Rademacher averages are defines as worst-case with respect to any sample, this implies that $\widehat{R}_n (F, \mathbf{X}) = o(n)$ and thus that the probability of the set $\{(\mathbf{X}, \mathbf{X}') : \widehat{R}_n (F, \mathbf{X}) > nt/8\}$ tends to 0 as $n$ goes to infinity. To see that, let $F$ be a class of functions bounded by $[-b, b]$ . For every $t > 0$ let $n_0$ be such that for every $n \geq n_0$, $\overline{R}_n (F) \leq nt/8$ which implies $\widehat{R}_n (F, \mathbf{X}) \leq nt/8$. Since $\sup_{f \in F} \mathrm{Var}\,(f) \leq b^2$, for any $n \geq 8b^2/t^2$ it holds that $1 - 4 \sup_{f \in F} \mathrm{var}(f)/nt^2 \geq 1/2$ . Therefore, by Corollary 4.6, it follows that for every integer $n > \max\{8b^2/t^2, n_0\}$ and for any probability measure $\mu$,

$$Pr_{\mathbf{X}} \Big\{ \sup_{f \in F} \big| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \big| \geq t \Big\} \leq c e^{-\frac{c' n t^2}{b^2}} \, ,$$

which are the optimal (up to constants) deviation bounds for uniform Glivenko-Cantelli classes obtainable without any further assumptions besides boundedness of the loss class.

In cases where one has a priori estimates on the size of the class (e.g., the shattering dimension or the uniform entropy), one can recover from these better uniform deviation results. In particular, if $VC(F) = d$, then by Theorem 2.30 $\overline{R}_n (F) \leq C\sqrt{dn}$ where $C$ is an absolute constant, and thus again $\widehat{R}_n (F, \mathbf{X}) = o(n)$ . For classes with a polynomial shattering dimension, by applying the bounds on $\overline{R}_n (f)$ from Mendelson (2003) (see also Theorem 2.31 for a special case), one gets estimates for $R_n (F)$ which are either logarithmic in $n$ or of the order $o(\sqrt{n})$. It follows again that $R_n (F) = o(n)$, and thus with high probability $\widehat{R}_n (F, \mathbf{X}) = o(n)$ .

Hence, in Glivenko-Cantelli classes,

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \widehat{R}_n (F, \mathbf{X}) = o(n) \, ,$$

and condition (4.13) is satisfied.

Clearly, if one knows specific rates on how $R_n (F) / n \longrightarrow 0$, then one can also recover better estimates for the probability of deviation (4.8) through Theorem 4.7.

### 4.3.2 Data-Dependent Class Bounds of Gat (1999) and Cannon et al. (2002)

Gat (1999) and Cannon et al. (2002) formulated data-dependent class bounds for sample-dependent hypothesis classes and binary losses using a similar symmetrization argument to that presented in Section 4.2.1. The subset $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ is defined to be the symmetric extension of $F_n$ (see (4.6)),

$$F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')) := \mathrm{Sym}_{\mathbf{X},\mathbf{X}'}(F_n)\,.$$

Thus, by construction, $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ is symmetric and $F_n(\mathbf{X}) \subseteq F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ , satisfying a 0-symmetric condition (4.5).

Gat (1999) assumed that the cardinality of $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ is bounded by a function which is sample-independent and only depends on the sample size $n$,

$$|F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))| \le d(n)\,.$$

With that, he proposed the following generalization bound based on concentration for single functions and a union bound argument:

$$Pr\left\{\sup_{f\in F_n(\mathbf{X})} \left|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^n f(X_i)\right| \ge t\right\} \le 2d(n)e^{-nt^2-2t}.$$

Thus, his result ensures learnability, whenever $d(n) = o(2^n)$, and exhibits a rate of convergence of $O(\sqrt{\log d(n)/n})$.

One can easily show that a small $d(n)$ is just a way of guaranteeing a small complexity of $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ in the sense of equation (4.15) and therefore, by Corollary 4.6, learnability conditions and generalization bounds in terms of $d(n)$. Indeed, since the class $F$ consists of functions bounded by 1,

$$V(\mathbf{X},\mathbf{X}') := F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X} \subset B_\infty^n \subset \sqrt{n}B_2^n\,.$$

Thus, by a comparison argument (Theorem A.1 and Corollary A.4, page 120), there is an absolute constant $C$ such that

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v}\in V} \left|\sum_{i=1}^n \varepsilon_i v_i\right| \le C\sqrt{n}\sqrt{\log|V|}\,.$$

Hence, for any $n$ and for any sample $(\mathbf{X},\mathbf{X}')$ of length $2n$,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v}\in V(\mathbf{X},\mathbf{X}')} \left|\sum_{i=1}^n \varepsilon_i v_i\right| \le C\sqrt{\log\left|F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}\right|}\,\sqrt{n} \le C\sqrt{\log d(n)}\sqrt{n}\,.$$

Thus, whenever $d(n) = o(2^n)$, Corollary 4.6 ensures learnability. Also, there is an absolute positive constant $c''$, such that for any $t \geq c'' \max\{1/\sqrt{n}, \sqrt{\log d(n)/n}\}$, it holds that $\widehat{R}_n\left(F_{2n}^{\mathrm{sym}}\left((\mathbf{X},\mathbf{X}')\right),\mathbf{X}\right) \leq nt/8$. Note that the rates of convergence are the same as in Gat (1999). This is due to the fact that the upper bound through the comparison Corollary A.4 (which only uses finiteness of the set $V$) is effectively like a union bound.

The result in Cannon et al. (2002) is similar; their bounds are in terms of the fat-shattering dimension rather than the cardinality of the set $F_{2n}^{\mathrm{sym}}\left((\mathbf{X},\mathbf{X}')\right)$, and one can recover them analogously to the proof presented above.

### 4.3.3 Compression Schemes

Littlestone and Warmuth (1986) and Floyd and Warmuth (1995) formulated generalization bounds for a particular class of binary classification learning algorithms called compression schemes. These algorithms have the property that they can reconstruct the hypothesis produced from a given training sample by using only a small "compressed" subset of the training sample. We denote by $C(\mathbf{X})$ the size of the smallest compressed sample by which the compression scheme algorithm $\mathcal{A}$, when presented with the training sample $\mathbf{X}$, can reconstruct the hypothesis that generated the labels. A *sample compression scheme of compression size* $\mathcal{K}$ is one for which $C(\mathbf{X}) \leq \mathcal{K}$ for every sample $\mathbf{X}$. The sample complexity bounds due to Littlestone and Warmuth (1986); Floyd and Warmuth (1995) are for the case in which the target function has empirical error equal to 0. Without any additional assumption like the one of a small empirical error, the same arguments can only lead to the following statement (see, e.g., Herbrich 2002, page 181): Let $\mathcal{A}(\mathbf{X})$ denote the function produced by a compression scheme algorithm with constant compression size $\mathcal{K}$ from a training sample $\mathbf{X}$. Then, for each $0 \leq \delta \leq 1$, there is an $n_0$ such that for any $n \geq n_0$, with probability at least $1-\delta$,

$$\left|\boldsymbol{\mathcal{R}}\left(\mathcal{A}(\mathbf{X})\right) - \widehat{\boldsymbol{\mathcal{R}}}\left(\mathcal{A}(\mathbf{X}),\mathbf{X}\right)\right| \leq c\sqrt{\frac{\ln\binom{n}{\mathcal{K}} + \ln\left(\frac{n}{\delta}\right)}{n-\mathcal{K}}}\,, \tag{4.16}$$

where $\boldsymbol{\mathcal{R}}$ and $\widehat{\boldsymbol{\mathcal{R}}}$ denote the expected and empirical risk, that is the expected and empirical average of the 0-1 loss function associated with the hypothesis $\mathcal{A}(\mathbf{X})$. (The original results in Littlestone and Warmuth (1986); Floyd and Warmuth (1995), for the case in which the target function has empirical error equal to 0, exhibit a rate of convergence of $O(\ln\binom{n}{\mathcal{K}} + \ln(n/\delta)/(n-\mathcal{K}))$.)

We set

$$\tilde{F}_n(\mathbf{X}) := \{\mathcal{A}(\mathbf{X})\} \quad \text{and} \quad F_n(\mathbf{X}) := \{l_{\mathcal{A}(\mathbf{X})}\}\,,$$

where $l_h$ is the 0-1 loss associated with $h$, and let $\tilde{F}_{2n}^{\mathrm{sym}}\left((\mathbf{X},\mathbf{X}')\right)$ be the symmetric extension of $\tilde{F}_n(\mathbf{X})$, and $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ is the loss class associated with $\tilde{F}_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$,

$$\tilde{F}_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')) := \mathrm{Sym}_{\mathbf{X},\mathbf{X}'}(\tilde{F}_n) \quad \text{and} \quad F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')) := \{l_f \ : \ f \in \tilde{F}_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))\}\,.$$

As before, the $\delta$-symmetric condition (4.5) is satisfied with $\delta = 0$, since $F_n(\mathbf{X}) \subseteq F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$.

We will show that a fixed small compression coefficient $\mathcal{K}$ guarantees small Rademacher averages with respect to $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ and thus, by Corollary 4.6, learnability, as proved by Littlestone and Warmuth (1986); Floyd and Warmuth (1995), is guaranteed. The number of loss functions in $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$, being at most the number of hypotheses in $\tilde{F}_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$, is upper bounded by the maximal number of functions which can be reproduced from a training sample of size at most $\mathcal{K}$. For binary classification functions taking values only in $\{-1,1\}$, the number of functions which can be reproduced from a sample of size $i$ is less than $2^i$, and so, for every $\mathbf{X},\mathbf{X}'$,

$$|F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))| \le |\tilde{F}_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))| \le \sum_{i=0}^{\mathcal{K}} \binom{2n}{i} 2^i \le 2^{\mathcal{K}} \left(\frac{2en}{\mathcal{K}}\right)^{\mathcal{K}} = \left(\frac{cn}{\mathcal{K}}\right)^{\mathcal{K}},$$

where we used the fact that $\sum_{i=0}^{d}\binom{n}{i} \le \left(\frac{en}{d}\right)^d$ (see, e.g., Anthony and Bartlett 1999, page 40). From the comparison theorem for finite sets (Corollary A.4, page 121) it follows immediately that

$$\mathbb{E}_\varepsilon \sup_{\mathbf{v}\in F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))/\mathbf{X}} \Big|\sum_{i=1}^{n} \varepsilon_i v_i\Big| \le C\sqrt{\mathcal{K}n(\log n - \log \mathcal{K})}\,,$$

where $C$ is an absolute constant. Again, this implies that for any $n$ and any $t \ge 8C\sqrt{\mathcal{K}\log n/n}$, the complexity associated with the coordinate projections of the set $F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))$ onto $\mathbf{X}$, as defined in equation (4.13), are small, that is $Pr_{\mathbf{X},\mathbf{X}'}\{\widehat{R}_n(F_{2n}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}')),\mathbf{X}) > nt/8\} = 0$. One can apply Corollary 4.6 for $t = \sqrt{\ln\binom{n}{\mathcal{K}} + \ln(n/\delta)/(n-\mathcal{K})}$ to recover the rates of convergence from equation (4.16).[2] Then, for each $0 \le \delta \le 1$ by Corollary 4.6,

$$Pr\left\{\left|\mathbb{E}_\mu l_{\mathcal{A}(\mathbf{X})} - \frac{1}{n}\sum_{i=1}^{n} l_{\mathcal{A}(\mathbf{X})}(X_i)\right| \ge \sqrt{\frac{\ln\binom{n}{\mathcal{K}} + \ln\left(\frac{n}{\delta}\right)}{n-\mathcal{K}}}\right\} \le ce^{-c'n\left(\frac{\ln\binom{n}{\mathcal{K}}+\ln\left(\frac{n}{\delta}\right)}{n-\mathcal{K}}\right)},$$

and since the right-hand side of this equation goes to 0 as $n$ goes to infinity, there is

[2] A data-dependent bound in terms of $C(\mathbf{X})$ instead of $\mathcal{K}$ can be recovered analogously, see "Sparsity luckiness" in Mendelson and Philips (2003).

an $n_0$ such that for all $n \geq n_0$,

$$Pr\left\{\left|\mathbb{E}_\mu l_{\mathcal{A}(\mathbf{X})} - \frac{1}{n}\sum_{i=1}^{n} l_{\mathcal{A}(\mathbf{X})}(X_i)\right| \leq \sqrt{\frac{\ln\binom{n}{\mathcal{K}} + \ln\left(\frac{n}{\delta}\right)}{n-\mathcal{K}}}\right\} \geq 1-\delta\,.$$

### 4.3.4 Luckiness

#### Classical Luckiness

In the classical luckiness framework introduced in Shawe-Taylor et al. (1998), bounds on the generalization error of functions are formulated a-priori but evaluated *a-posteriori*, after having seen a sample $\mathbf{X}$. The bounds are given in terms of an upper bound on some empirical, computable quantity dependent on the sample. The main idea driving the luckiness proof is that one wants to exploit additional knowledge regarding the actual sample and the function class. This knowledge is quantified through a luckiness function which allows one to decompose the function class in data-dependent sets. The luckiness framework shows that a certain property of the luckiness function, $\omega$-smallness (with respect to an appropriately defined function $\omega$), allows one to obtain generalization bounds conditioned on the sample.

In the following, we will introduce the luckiness framework following the presentation in Herbrich and Williamson (2003). Let $\mathcal{H}$ be an hypothesis function class and $l$ a loss function bounded by 1. The *luckiness function* is a function $L : \mathcal{H} \times \cup_k \Omega^k \longrightarrow \mathbb{R}$ which is invariant under permutations of the sample, that is, it depends only on the set $\{x_1, ..., x_k\}$. Using the luckiness function one can construct sample dependent subsets of $\mathcal{H}$, called *lucky sets* in the following manner: for every sample $\mathbf{X}$ and $f \in \mathcal{H}$, the lucky set consists of all the functions luckier on this sample than the given function, that is,

$$H(f, \mathbf{X}) := \{g \in \mathcal{H} : L(g, \mathbf{X}) \geq L(f, \mathbf{X})\}.$$

Denote by $H_l(f, \mathbf{X})$ the loss class associated with $H(f, \mathbf{X})$,

$$H_l(f, \mathbf{X}) := \{l_f \ : \ f \in H(f, \mathbf{X})\}\,.$$

Observe that the luckiness function imposes a structure of increasing subsets of $\mathcal{H}$, because $H(g, \mathbf{X}) \subseteq H(f, \mathbf{X})$ if and only if $L(g, \mathbf{X}) \geq L(f, \mathbf{X})$ .

The second ingredient in the luckiness framework is the *$\omega$-function*, $\omega : \mathbb{R} \times \mathbb{N} \times (0, 1] \to \mathbb{N}$.

The third ingredient is the *$\omega$-smallness condition*, which is a joint property of the luckiness and $\omega$ functions. It states that for every $n \in \mathbb{N}$, every $\delta \in (0, 1]$ and every probability measure $\mu$, for $(\mathbf{X}, \mathbf{X}')$ distributed according to $\mu^{2n}$,

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in \mathcal{H} : M\big(\tfrac{1}{2n}, H_l(f, (\mathbf{X}, \mathbf{X}')), L_1(\mu_{2n})\big) > \omega\big(L(f, \mathbf{X}), n, \delta\big)\Big\} < \delta\,. \tag{4.17}$$

Intuitively, this condition states that the "size" of the loss functions corresponding to the lucky set $H(f, (\mathbf{X}, \mathbf{X}'))$ of any hypothesis function $f$ with respect to the double-sample $(\mathbf{X}, \mathbf{X}')$ — measured by packing numbers — is bounded by a function of the luckiness of the same function $f$ *on the sample.* The original definition in Herbrich and Williamson (2003) uses covering numbers instead of packing numbers. Since covering numbers and packing numbers are closely related, $N(\varepsilon, \mathcal{H}, d) \leq M(\varepsilon, \mathcal{H}, d) \leq N(\varepsilon/2, \mathcal{H}, d)$, we will use here this modified version employing packing numbers which is more convenient for our proofs. In fact, in the original luckiness framework in Shawe-Taylor et al. (1998) which was formulated only for binary-valued classes of functions, the notion of size used for the $\omega$-smallness condition is the shattering coefficient. The fact that the "size" in Herbrich and Williamson (2003) is measured by packing (covering) numbers has also historical reasons because covering numbers were the tightest notion of size known at the time when it was formulated.

As an example, because $H(f, (\mathbf{X}, \mathbf{X}')) \subseteq \mathcal{H}$, any empirical complexity notion from Section 2.4.1 which upper bounds the covering numbers of $\mathcal{H}$ is a suitable luckiness function, and one can formulate an $\omega$-function by considering how this complexity grows with the sample size (see discussion at the end of this section).

The following is the main result of the luckiness framework:

**Theorem 4.8 (Luckiness bound)** *Let $\mathcal{H}$ be a hypothesis function class and $l$ a loss function bounded by* 1*. Let $L$ and $\omega$ be functions satisfying the $\omega$-smallness condition* (4.17)*. Then, for every probability measure $\mu$, every $d \in \mathbb{N}$ and every $\delta \in (0, 1]$, there is a set of probability larger than $1 - 12\delta$ such that if $\omega\big(L(f, \mathbf{X}), n, \delta\big) \leq 2^d$, then*

$$\Big|\mathbb{E}_\mu l_f - \frac{1}{n}\sum_{i=1}^{n} l_f(X_i)\Big| \leq C\sqrt{\frac{d}{n}\log\frac{1}{\delta}}\,,$$

*where $C$ is an absolute constant.*

Its original proof can be found in Shawe-Taylor et al. (1998), and we will prove it here subsequently using the general result of Corollary 4.6.

Examples of truly data-dependent luckiness functions given in Shawe-Taylor et al. (1998); Herbrich (2002); Herbrich and Williamson (2003) are the empirical VC-dimension of a binary function class with respect to a sample — in this case the luckiness function is independent on the particular hypothesis function and all lucky sets are equal to the whole set $\mathcal{H}$ — and the margin of linear classifiers. Their corresponding $\omega$-functions can be found in Shawe-Taylor et al. (1998). Although the luckiness framework gives a unified proof for existing generalization bounds, finding pairs of luckiness and $\omega$-functions seems in general to be difficult because of the quite technical $\omega$-smallness condition.

In the following, we will show that the luckiness framework is a special case of the random subclass framework and that the $\omega$-smallness condition is just a way of ensuring that a random coordinate projection of the random set is "small" in the sense of Corollary 4.6. In particular, we will show that both the $\delta$-symmetric condition (4.5) and condition (4.13) follow from the $\omega$-smallness condition. Since the functions in the loss class of $\mathcal{H}$ are bounded, we will show that the $\omega$-smallness condition in fact allows one to control the "size" of the coordinate projections of a random symmetric set onto random samples, in the sense of (4.13). This will link the luckiness framework directly to approaches using Rademacher averages as a notion of size.

In order to define $F_n(\mathbf{X})$, we use the luckiness and associated $\omega$-function satisfying (4.17) in the following way: for any luckiness function $L$ and any $\omega$-function, for any fixed integer $d$ and $\delta \in (0,1]$, define

$$F_{n,d}(\mathbf{X}) := \left\{ l_f \ : \ f \in \mathcal{H}, \ \omega\big(L(f,\mathbf{X}), n, \delta\big) \le 2^d \right\}. \tag{4.18}$$

Let $n$ be a fixed sample size, $d$ be a given fixed integer, and set $\delta \in (0,1]$. We first need the following lemma which states that for each sample there is a unique set containing all functions with lucky sets of "size" $2^d$. This unique set will allow us to define $F^{\text{sym}}_{2n,d}$. Let $\mathbf{X}$ be a sample of size $n$ and let $A_{n,d}(\mathbf{X}) \subseteq \mathcal{H}$ be the set of all functions with lucky sets of size smaller than or equal to $2^d$,

$$A_{n,d}(\mathbf{X}) := \{ f \in \mathcal{H} \ : \ M\big(\tfrac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \le 2^d \},$$

and define

$$H_{n,d}(\mathbf{X}) := \bigcup_{f \in A_{n,d}(\mathbf{X})} H_l(f,\mathbf{X}).$$

**Lemma 4.9** *For every integer $d$ and sample $\mathbf{X}$ of size $n$, if $\mu_n$ is the empirical measure supported on $\mathbf{X}$, then $H_{n,d}(\mathbf{X})$ defined above is the unique set with the following properties:*

1. $M\big(\tfrac{1}{n}, H_{n,d}(\mathbf{X}), L_1(\mu_n)\big) \le 2^d$.

2. *If $f \in F$ satisfies $M\big(\tfrac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \le 2^d$, then $l_f \in H_{n,d}(\mathbf{X})$.*

The proof can be found in Appendix B.1.

For every double-sample $(\mathbf{X},\mathbf{X}')$, set

$$F^{\text{sym}}_{2n,d}\left((\mathbf{X},\mathbf{X}')\right) := H_{2n,d}((\mathbf{X},\mathbf{X}')), \tag{4.19}$$

and observe that this random class is permutation invariant because the luckiness function is, by definition, invariant under permutations of the sample, implying that $F^{\text{sym}}_{2n,d}$ is symmetric as required.

The following result shows that the $\omega$-smallness of $L$ ensures that the $\delta$-symmetric condition (4.5) holds.

**Lemma 4.10** *For any positive integers $n$ and $d$, and $\delta \in (0,1]$, let $F_{n,d}$ and $F^{\text{sym}}_{2n,d}$ be defined as in* (4.18) *and* (4.19)*. If a luckiness function $L$ and an $\omega$-function satisfy the $\omega$-smallness condition* (4.17)*, then for every $t > 0$,*

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{n,d}(\mathbf{X}),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\Big| \ge t\Big\}$$
$$\le Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')),\ \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\Big| \ge t\Big\} + \delta\,.$$

The proof can be found in Appendix B.1.

We have shown so far how to define, for each $d$, $F_{n,d}$ and $F^{\text{sym}}_{2n,d}$ for the luckiness framework. Now, we are ready to show that with high probability, $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ is sufficiently small and that the generalization bound for the luckiness framework in Theorem 4.8 follows from the random subclass Corollary 4.6. The main part of the proof is to show that the definition of $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ ensures that the covering numbers and therefore the Rademacher averages of $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ are small. Therefore, the $\omega$-smallness condition is just a way of requiring that $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ has, with high probability, small random coordinate projections, implying thus that condition (4.13) holds.

**Proof (of Theorem 4.8 via Corollary 4.6):** Let $n, d$ be arbitrary but fixed integers, and let $F_{n,d}$ and $F^{\text{sym}}_{2n,d}$ be defined as above, and observe that

$$M\big(\tfrac{1}{2n}, F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \le 2^d \tag{4.20}$$

for every $(\mathbf{X},\mathbf{X}')$. By Corollary 4.6 we have to estimate

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\mathbb{E}_\varepsilon \sup_{f\in F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))}\Big|\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big| > \frac{nt}{8}\Big\}\,,$$

where $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{X}' = (X_1', ..., X_n')$. Let $V_d \subset \ell_2^n$ be the coordinate projections

$$V_d = F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))/\mathbf{X} = \Big\{\big(f(X_1), ..., f(X_n)\big) : f \in F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))\Big\}\,,$$

put $\mu_{2n}$ to be the empirical measure supported on $(\mathbf{X},\mathbf{X}')$ and set $\nu_n$ to be the empirical measure supported on $\mathbf{X}$. Note that for every $f,g$, $\mathbb{E}_{\mu_{2n}}|f-g| \geq \mathbb{E}_{\nu_n}|f-g|/2$ . Thus, every $1/2n$-cover of $F^{\mathrm{sym}}_{2n,d}\left((\mathbf{X},\mathbf{X}')\right)$ in $L_1(\mu_{2n})$ is a $1/n$-cover of the same set in $L_1(\nu_n)$. In particular, if $A$ is a maximal $1/2n$-packing of $F^{\mathrm{sym}}_{2n,d}\left((\mathbf{X},\mathbf{X}')\right)$ in $L_1(\mu_{2n})$ , it is a $1/n$ cover of that set in $L_1(\nu_n)$. It is easy to verify that $B\big(L_1(\nu_n)\big)$ is isomorphic to $nB_1^n$, and in particular, up to isomorphism,

$$V_d \subset A + \frac{1}{n}\cdot nB_1^n = A + B_1^n\,,$$

where

$$A+B = \{\mathbf{a}+\mathbf{b}\ :\mathbf{a}\in A,\ \mathbf{b}\in B\}\,.$$

By the triangle inequality,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v}\in V_d} \Big|\sum_{i=1}^n \varepsilon_i v_i\Big| = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{a}\in A,\,\mathbf{b}\in B_1^n} \Big|\sum_{i=1}^n \varepsilon_i (a_i+b_i)\Big| \tag{4.21}$$

$$\leq \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{a}\in A} \Big|\sum_{i=1}^n \varepsilon_i a_i\Big| + \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{b}\in B_1^n} \Big|\sum_{i=1}^n \varepsilon_i b_i\Big|\,. \tag{4.22}$$

The first term of (4.22) can be bounded by the comparison theorem (Corollary A.4, page 121). Since our class consists of functions bounded by $b$, then $V_d \subset B_\infty^n \subset \sqrt{n}B_2^n$ and since the Rademacher averages are upper bounded (up to an absolute constant) by the Gaussian ones (cf. Theorem A.1), then

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{a}\in A} \Big|\sum_{i=1}^n \varepsilon_i a_i\Big| \leq C\mathbb{E}_{\mathbf{g}} \sup_{\mathbf{a}\in A} \Big|\sum_{i=1}^n g_i a_i\Big| \leq C\sqrt{\log|A|}\sqrt{n} \leq C\sqrt{nd}\,,$$

where the final inequality holds because $|A| \leq 2^d$ by (4.20).

In order to estimate the second term in equation (4.22), one can apply the triangle inequality to show that

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{b}\in B_1^n} \Big|\sum_{i=1}^n \varepsilon_i b_i\Big| \leq 1\,.$$

In conclusion,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \leq C\sqrt{nd}\,. \tag{4.23}$$

In order to complete the proof, apply Corollary 4.6 for $t = C\sqrt{\frac{d}{n}\log(1/\delta)}$. ■

Note that in obtaining this result we only used boundedness of the loss function and therefore of the functions in the loss class. Any additional constraint on the variance of functions in the loss class which would restrict the $\ell_2^n$ radius of the projections further could conceivably lead to tighter error bounds.

**Discussion:** A direct consequence of the reasoning given above is a generalization of the luckiness framework which replaces the metric complexity with Rademacher averages. Such a "Rademacher-luckiness" avoids the potential looseness of the union bound used in approaches using the metric complexities. Indeed, for each $d > 0$ and each $n$, let $A_{n,d}(\mathbf{X}) \subseteq \mathcal{H}$ be the set of all functions with lucky sets of size smaller than or equal to $\gamma_n(d)$, (where $\gamma_n(d) \leq C\sqrt{nd}$ with the constant $C$ from equation (4.23)),

$$A_{n,d}(\mathbf{X}) := \{f \in \mathcal{H} \,:\, R_{2n}\left(H_l(f, (\mathbf{X}, \mathbf{X}'))\right) \leq \gamma_n(d)\}\,,$$

and let $H_{n,d}(\mathbf{X}, \mathbf{X}')$ be

$$H_{n,d}(\mathbf{X}, \mathbf{X}') := \bigcup_{f \in A_{n,d}(\mathbf{X})} H_l(f, (\mathbf{X}, \mathbf{X}'))\,.$$

Assume that there is an $f' \in \mathcal{H}$ such that $f' = \operatorname{argmin}_{f_k \in A_{n,d}(\mathbf{X})} L(f_k, (\mathbf{X}, \mathbf{X}'))$. [3] In this case, $H_{n,d}(\mathbf{X}, \mathbf{X}') = H_l(f', (\mathbf{X}, \mathbf{X}'))$ and thus $H_{n,d}(\mathbf{X}, \mathbf{X}')$ is the unique set satisfying that $R_{2n}\left(H_{n,d}(\mathbf{X}, \mathbf{X}')\right) \leq \gamma_n(d)$ and such that $R_{2n}\left(H_l(f, (\mathbf{X}, \mathbf{X}')), \mathbf{X}\right) \leq \gamma_n(d)$ implies that $f \in H_{n,d}(\mathbf{X}, \mathbf{X}')$. Setting

$$F_{n,d}(\mathbf{X}) := \left\{l_f \;:\; f \in \mathcal{H},\; \omega\big(L(f, \mathbf{X}), n, \delta\big) \leq \gamma_n(d)\right\}$$

and

$$F_{2n,d}^{\mathrm{sym}}\left((\mathbf{X}, \mathbf{X}')\right) := H_{n,d}(\mathbf{X}, \mathbf{X}')\,,$$

the following $\omega$-smallness condition in terms of Rademacher averages ensures, analogously to the proofs for the "classical" luckiness, that condition (4.13) holds:

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in \mathcal{H} : R_{2n}\left(H_l(f, (\mathbf{X}, \mathbf{X}'))\right) > \omega\big(L(f, \mathbf{X}), n, \delta\big)\Big\} < \delta\,. \tag{4.24}$$

Note that uniqueness of the set $F_{2n,d}^{\mathrm{sym}}\left((\mathbf{X}, \mathbf{X}')\right)$ is required, analogously to Lemma 4.9, in order to ensure that whenever we measure a small enough $\omega\big(L(f, \mathbf{X}), n, \delta\big)$ for a given function $f$ and a given sample $\mathbf{X}$, the deviation for this function $f$ is controlled by the deviation of functions in $F_{2n,d}^{\mathrm{sym}}\left((\mathbf{X}, \mathbf{X}')\right)$.

Thus, the $\omega\big(L(f, \mathbf{X}), n, \delta\big)$ function is any (hopefully data-dependent) upper bound on the complexity of the lucky sets corresponding to double-samples. Let us consider

[3] This assumption is restrictive but it is sufficient in order to illustrate the point we want to make in the following. Note that it is possible to prove the same results without this assumption with a more sophisticated proof (Bousquet, personal communication), by replacing the definition of Rademacher averages as being the supremum of Rademacher averages of all finite subsets of $F$ (see, e.g., Ledoux (1994), page 69), and by using limit arguments.

the case when one defines a luckiness function which is independent of the hypothesis function and only depends on the sample (and thus all lucky sets are equal to $\mathcal{H}$). In such a case, a way to interpret the $\omega$-smallness condition is that it replaces the concentration condition for complexities with a condition which ensures that coordinate projections of the loss class on double-samples do not grow “too much” (as a function of the sample size $n$) compared to these on the single sample. For example, in the proof presented by Herbrich (2002), page 293, showing that the empirical VC-dimension is a special case of luckiness, the crucial step is the one which shows that, for any binary-valued class $F$, with high probability, the empirical VC-dimension on samples of size $2n$ is not much larger than a constant times the empirical VC-dimension on the sample of size $n$. Such a result follows also from Corollary 3.14 based on the fact that the empirical VC-dimension has the self-bounding property. The self-bounding property is also the property which allows one to derive the concentration result for the empirical VC-dimension around its expectation (Boucheron et al. 2000). Similarly, one can define the (data-dependent but hypothesis-independent) luckiness function as being any empirical complexity notion from Section 2.4.1 which upper bounds the empirical Rademacher averages or upper bounds the Rademacher averages and has the self-bounding property (as proved in Bartlett et al. (2002)). Indeed, by concentration, with high probability the Rademacher averages are similar to the empirical Rademacher averages, $R_{2n}\left(\mathcal{H}_l\right) \sim \widehat{R}_{2n}\left(\mathcal{H}_l, (\mathbf{X}, \mathbf{X}')\right)$, and by Lemma 3.14, since the empirical Rademacher averages satisfy the self-bounding property, with high probability $\widehat{R}_{2n}\left(\mathcal{H}_l, (\mathbf{X}, \mathbf{X}')\right) \leq c\widehat{R}_{2n}\left(\mathcal{H}_l, \mathbf{X}\right)$.

Clearly, the $\omega$-smallness condition is more general: it allows, in principle, to encode additional prior knowledge about the sample *and* the hypothesis if one finds data-and-hypothesis-dependent complexities (like the margin) which upper bound the complexity of the lucky sets. However, there is no general recipe to find new pairs of luckiness and $\omega$-functions; it is up to the intuition of the designer of a machine learning algorithm to find new data-dependent complexities like the luckiness function to bound the Rademacher complexity of the lucky sets.

### Algorithmic Luckiness

In the algorithmic luckiness framework (Herbrich and Williamson 2002), an extension of the luckiness framework, the generalization error bound is also evaluated a-posteriori, after having seen a sample. It differs from the luckiness framework because it gives bounds on the generalization error of the function learned by the learning algorithm from the sample at hand. Again, the bound is in terms of a computable quantity dependent on the sample and the algorithm.

In a similar fashion to the luckiness framework, an algorithmic luckiness function and an $\omega$-function are introduced in order to define the random subsets $F_n$ and $F_{2n}^{\mathrm{sym}}$

for the given sample. The two functions satisfy a joint smallness condition which ensures that the random symmetrization condition is satisfied and that the size of the symmetric subset $F_{2n}^{\mathrm{sym}}$ is small enough to result in tight generalization error bounds.

As we did before, let $n$ be a fixed sample size, $d$ is a given fixed integer, and set $\delta \in (0,1]$. Recall that $\mathcal{H}$ is a hypothesis class and $l$ is a loss function bounded by 1. Denote by $\mathcal{A}$ a fixed permutation invariant learning algorithm, by $\mathcal{A}(\mathbf{x})$ the function produced by the algorithm from the sample $\mathbf{x}$, and set $\mathcal{A}(\mathcal{H}) = \{f = \mathcal{A}(\mathbf{x}) : \ \mathbf{x} \in \Omega^n\}$.

Three concepts are used in the algorithmic luckiness framework: The first is the algorithmic luckiness function which is a function $L : \mathcal{A}(\mathcal{H}) \longrightarrow \mathbb{R}$. Using the algorithmic luckiness function one can construct sample dependent subsets of $\mathcal{H}$, called *lucky sets* in the following manner: For every sample $\overline{\mathbf{X}}$ of size $2n$, the *lucky set* $G(\overline{\mathbf{X}})$ is defined as the subset of functions learned by the algorithm on the first half of the sample, when permuting the whole sample, as long as the function the algorithm produced on the first half of the permuted sample is "luckier" than on the original one. Define the lucky set as

$$G(\overline{\mathbf{X}}) := \Big\{\mathcal{A}\big(\pi(\overline{\mathbf{X}})|_{i=1}^{n}\big) : \ L\big(\mathcal{A}(\pi(\overline{\mathbf{X}})|_{i=1}^{n})\big) \geq L\big(\mathcal{A}(\overline{\mathbf{X}}|_{i=1}^{n})\big), \ \pi \in S_{2n}\Big\}. \tag{4.25}$$

If $G_{\mathcal{A}}(\overline{\mathbf{X}})$ is the subset of hypothesis functions corresponding to functions learned by $\mathcal{A}$ on the first half of all the permutations of the double-sample $\overline{\mathbf{X}}$, then $G(\overline{\mathbf{X}}) \subset G_{\mathcal{A}}(\overline{\mathbf{X}})$, and clearly, $|G_{\mathcal{A}}(\overline{\mathbf{X}})| \leq (2n)! < \infty$. Therefore, we can order the functions in $G_{\mathcal{A}}(\overline{\mathbf{X}})$ in decreasing order according to their luckiness. Define the ordered set

$$G_{\mathcal{A}}(\overline{\mathbf{X}}) := \Big[\underbrace{f_1, f_2, f_3, \ldots, f_{k-1}, f_k}_{G(\overline{\mathbf{X}})}, f_{k+1}, \ldots, f_m\Big], \tag{4.26}$$

and for the sake of simplicity, assume that for every $i < j$, $L(f_i) > L(f_j)$. Only a small modification is required in the general case, where some functions might have the same luckiness.

Set $f_k = \mathcal{A}(\overline{\mathbf{X}}|_{i=1}^{n})$ and let $G_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}})$ be the subset consisting of the first $\ell$ functions in $G_{\mathcal{A}}(\overline{\mathbf{X}})$, i.e. $G_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}}) = \{f_1, f_2, f_3, \ldots, f_\ell\}$. Let $F_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}})$ be the loss class associated with $G_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}})$,

$$F_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}}) = \{l_f \ : \ f \in G_{\mathcal{A}}^{\ell}(\overline{\mathbf{X}})\}\,.$$

For any integer $d$ put $k_d^*$ to be the largest integer such that

$$M\big(\tfrac{1}{n}, F_{\mathcal{A}}^{k_d^*}((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \leq 2^d \text{ and } M\big(\tfrac{1}{n}, F_{\mathcal{A}}^{k_d^*+1}((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) > 2^d. \tag{4.27}$$

Then, for the double-sample $(\mathbf{X},\mathbf{X}')$, by setting

$$F_{2n,d}^{\mathrm{sym}}\,((\mathbf{X},\mathbf{X}')) := F_{\mathcal{A}}^{k_d^*}((\mathbf{X},\mathbf{X}')) \tag{4.28}$$

it follows that $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ is symmetric, since the learning algorithm is permutation invariant.

The second ingredient, the $\omega$-function, $\omega : \mathbb{R} \times \mathbb{N} \times (0,1] \longrightarrow \mathbb{N}$ is used to define $F_{n,d}(\mathbf{X})$ for each $d$. Let

$$F_{n,d}(\mathbf{X}) := \begin{cases} \{l_{\mathcal{A}(\mathbf{X})}\} & \text{if } \omega\big(L(\mathcal{A}(\mathbf{X})), n, \delta\big) \leq 2^d \\ \emptyset & \text{otherwise,} \end{cases} \tag{4.29}$$

and note that $|F_{n,d}(\mathbf{X})| \leq 1$.

The third ingredient is the $\omega$-smallness condition, which is a joint property of the algorithmic luckiness and $\omega$ functions. It states that for every integer $n$, every $\delta \in (0,1]$, and every probability measure $\mu$,

$$P_{\mathbf{X},\mathbf{X}'}\Big\{M\big(\tfrac{1}{n}, G_l((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \geq \omega\big(L(\mathcal{A}(\mathbf{X})), n, \delta\big)\Big\} < \delta\,, \tag{4.30}$$

where $G_l((\mathbf{X},\mathbf{X}'))$ is the loss class associated with $G((\mathbf{X},\mathbf{X}'))$.

The following results show that the $\omega$-smallness of $L$ ensures that Assumption 4.4 holds, and that, with high probability, $F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))/\mathbf{X}$ and thus the random coordinate projections are sufficiently small and the condition 4.13 is satisfied. Their proofs can be found in Appendix B.1 and are very similar to the proofs in the classical luckiness case.

**Lemma 4.11** *Let $\mathcal{A}$ be a permutation-invariant learning algorithm, fix an integer $d$ and some $\delta \in (0,1]$, and let $F_{n,d}$ and $F^{\text{sym}}_{2n,d}$ be as in (4.29) and (4.28). If an algorithmic luckiness function $L$ and an $\omega$-function satisfy the $\omega$-smallness condition (4.30), then for every $t > 0$*

$$\begin{aligned} &Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F_{n,d}(\mathbf{X}) : \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\Big| \geq t\Big\} \\ &\qquad \leq Pr_{\mathbf{X},\mathbf{X}'}\Big\{\exists f \in F^{\text{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')) : \Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f(X_i')\big)\Big| \geq t\Big\} + \delta. \end{aligned}$$

Now, we are ready to formulate the generalization bound for the algorithmic luckiness framework which recovers the main result of Herbrich and Williamson (2002).

**Theorem 4.12** *Let $\mathcal{A}$ be a permutation-invariant learning algorithm, denote by $\mathcal{A}(\mathbf{X})$ the function produced by the algorithm from the sample $\mathbf{X}$, and assume that the loss function takes values in $[-1,1]$. Let $L$ and $\omega$ be functions satisfying the $\omega$-smallness condition (4.30). Then, for every probability measure $\mu$, every $d \in \mathbb{N}$ and every $\delta \in (0,1]$, there is a set of probability at least $1 - 12\delta$ such that if $\omega\big(L(\mathcal{A}(\mathbf{X})), n, \delta\big) \leq 2^d$,* *then*

$$\left|\mathbb{E}_\mu l_{\mathcal{A}(\mathbf{X})} - \frac{1}{n}\sum_{i=1}^{n} l_{\mathcal{A}(\mathbf{X})}(X_i)\right| \leq C\sqrt{\frac{d}{n}\log\frac{1}{\delta}},$$

*where $C$ is an absolute constant.*

Thus, again, boundedness of the loss functions leads to a convergence rate of $O(1/\sqrt{n})$, and it is conceivable to obtain faster rates up to $O(1/n)$ by imposing additional constraints on the variance of the functions (or that the empirical error of $\mathcal{A}(\mathbf{X})$ is 0).

Similarly to the classical luckiness case, an algorithmic luckiness framework which uses Rademacher complexities in order to avoid the looseness of the union bound can be easily derived. Such a formulation in terms of Rademacher complexities exhibits more clearly that the $\omega$-smallness condition is a way of restricting coordinate projections of lucky sets by using prior knowledge encoded in a pair of luckiness and $\omega$ functions.

### 4.3.5 Sharper Bounds through Control on the Variance

In this section, we will give an example of a specific result which proves rates of convergence potentially as good as $O(1/n)$ and which can be explained by using the full potential of Theorem 4.7 and explicit control on the variance. Our example is the derivation of error bounds for the function produced by the Empirical Risk Minimization algorithm based on results in Mendelson (2002b, 2003); Bartlett et al. (2004a).

Mendelson (2002b, 2003) and Bartlett et al. (2004a) proved that one can obtain generalization bounds for star-shaped sets of functions satisfying an additional constraint on the variance. For these function sets, it was shown that the dominating bounding term is the "size" of subsets of the function class containing functions with a variance of the same order of magnitude as the deviation of the mean from the expected mean.

Let $F$ be an excess squared-loss class. Mendelson (2003) proved tighter bounds for the "almost empirical minimizer" by bounding the following probability:

$$Pr_{\mathbf{X}}\left\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^{n} f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t\right\}.$$

This is the probability that an empirical minimizer of the loss functional (or more generally, an "almost empirical minimizer") will have a relatively large expectation. These bounds were proved under two additional assumptions on the class $F$, namely, first, that $F$ is star-shaped around 0 (i.e, for every $f \in F$ and $0 \leq t \leq 1$, $tf \in F$); the second is that there is some $B > 0$ such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$ (see Section 5.2 for details on these assumptions). Under these two assumptions, it was shown that for every $t > 0$,

$$\begin{aligned} &Pr_{\mathbf{X}}\Big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^{n} f(X_i) \leq t,\ \mathbb{E}_\mu f \geq 2t\Big\} \\ &\leq 2Pr_{\mathbf{X}}\Big\{\sup_{f\in F,\ \mathbb{E}_\mu f^2 \leq Bt} \Big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\Big| \geq t\Big\}. \end{aligned} \tag{4.31}$$

Thus, the complexity of the problem is governed by the complexity of the function class $F$ intersected with a ball of variance of the order of the deviation $t$.

A generalization of this result was formulated by Bartlett et al. (2004a). They showed that the fixed point of any function which upper bounds the Rademacher averages of a subclass of small variance (or an upper bound thereof) governs the generalization ability and allows the derivation of error bounds of the order as good as $e^{-cnt}$. Since, in the learning setting, we are interested in deviations $t < 1$, and for large sample sizes $n$, these exponential bounds are a strong improvement on the sub-Gaussian bounds of the order of $e^{-cnt^2}$.

We will show in this section how one can recover the results from Mendelson (2003) and Bartlett et al. (2004a) in terms of the "local" subsets of small variance through Theorem 4.7. For this, let $F_n$ and $F_{2n}^{\mathrm{sym}}$ be constant set-valued maps, defined as

$$F_n(\mathbf{X}) = F_{2n}^{\mathrm{sym}}\left((\mathbf{X}, \mathbf{X}')\right) := \{f \in F,\ \mathbb{E}_\mu f^2 \leq Bt\}.$$

These sets, like in the Glivenko-Cantelli example (Section 4.3.1) are not data-dependent, and the purpose in this section is to illustrate how control of the variance leads to tighter concentration and better rates of convergence through Theorem 4.7.

For the sake of simplicity we present our proof for functions bounded by 1 ($b = 1$) and with $B = 1$, which is the case if $F$ consists of nonnegative functions taking values in $[0, 1]$. The general case follows an identical path. Denote by

$$\bar{F}_t := \{f \in F\ :\ \mathbb{E}_\mu f^2 \leq t\}.$$

Thus,

$$Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \sup_{f\in\bar{F}_t} \Big|\sum_{i=1}^{n} \varepsilon_i f(X_i)\Big|,$$

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon}) = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in\bar{F}_t} \Big|\sum_{i=1}^{n} \varepsilon_i f(X_i)\Big| = \widehat{R}_n\left(\bar{F}_t, \mathbf{X}\right).$$

We will show in the following that what is hidden in the proofs for these localized results is the fact that the Bernstein condition, together with an a priori control of the complexity of the class $\bar{F}_t$ (that is, that $R_n\left(\bar{F}_t\right) = o(n)$ ), implies that the coordinate projections $\bar{F}_t/\mathbf{X}$ are contained in a small ball in $\ell_2^n$ of radius $O(\sqrt{nt})$ (as opposed to a radius of $O(\sqrt{n})$ given by simple boundedness of functions). Thus, the control of the variance in Bernstein classes allows one to obtain through Talagrand's convex distance inequality (as in Theorem 4.7) — if the function class is not too "complex" — a stronger degree of concentration of the order of $e^{-cnt}$ for $Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ around its expectation $\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ . The same condition on the complexity of the projections $\bar{F}_t/\mathbf{X}$, namely that $R_n\left(\bar{F}_t\right) = o(n)$, controls the expectation $\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$. Specifically, if $R_n\left(\bar{F}_t\right) \sim nt$, then the concentration result for self-bounding functions Theorem 3.13 implies that $Pr_{\mathbf{X},\mathbf{X}'}\left\{\widehat{R}_n\left(\bar{F}_t,\mathbf{X}\right) > nt/8\right\} \leq e^{-cnt}$ . Thus, both terms in the right side of Theorem 4.7 are of the order of $e^{-cnt}$. Since $F_n = F_{2n}^{\mathrm{sym}} := \bar{F}_t$, the $\delta$-symmetric condition (4.5) holds trivially with $\delta = 0$ and by applying Theorem 4.7 we recover the desired tail bounds.

**Theorem 4.13** *There are absolute constants $K$, $c$ and $c_1$ for which the following holds. Let $F \subset B\big(L_\infty(\Omega)\big)$ be star-shaped around $0$ such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq \mathbb{E}_\mu f$. If $t \geq c_1/n$ satisfies that*

$$R_n\left(\bar{F}_t\right) = \mathbb{E}\sup_{f\in\bar{F}_t}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \leq \frac{nt}{16}, \tag{4.32}$$

*then*

$$Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\Big\{\exists f\in\bar{F}_t,\ \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \geq \frac{nt}{4}\Big\} \leq Ke^{-cnt}.$$

**Proof:** The proof contains two parts, each based on a concentration result which uses the fact that $R_n\left(\bar{F}_t\right) \leq cnt.$

Part 1: Control of the expectation $\mathbb{E}_{\boldsymbol{\varepsilon}} Z_{\mathbf{X},\mathbf{X}'}(\boldsymbol{\varepsilon})$ :

We show first, that $Pr_{\mathbf{X},\mathbf{X}'}\left\{\widehat{R}_n\left(\bar{F}_t,\mathbf{X}\right) > nt/8\right\} \leq e^{-cnt}$ by proving a concentration of empirical averages of $\bar{F}_t$ around their expectation $R_n\left(\bar{F}_t\right)$.

Recall that the concentration result for self-bounding functions (Theorem 3.13, page 46), implies, together with a small expectation of the order of the deviation, a rate $e^{-t}$ rather than $e^{-t^2}$. Since the variables $\widehat{R}_n\left(\bar{F}_t,\mathbf{X}\right)$ satisfy the self-bounding property, by applying Theorem 3.13, if $t$ is such that $\mathbb{E}\widehat{R}_n\left(\bar{F}_t,\mathbf{X}\right) = R_n\left(\bar{F}_t\right) \leq nt/16$ then indeed

$$Pr_{\mathbf{X}}\Big\{\mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{f\in\bar{F}_t}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| > \frac{nt}{8}\Big\} \leq e^{-cnt}, \tag{4.33}$$

where $c$ is an absolute constant, which proves the first claim.

2. Control of the $\ell_2^n$ radius of the coordinate projections $\bar{F}_t/\mathbf{X}$:

In order to apply the corollary of Talagrand's convex distance inequality (Theorem 3.11, page 43), we have to show that the coordinate projections $\bar{F}_t/\mathbf{X}$ are contained in a small ball in $\ell_2^n$ of diameter $\sqrt{nt}$. The following theorem due to Bartlett et al.

(2004a) allows us to estimate, with high probability, the $\ell_2^n$ diameter of projections of functions with bounded variances if the Rademacher averages of these sets are small.

**Theorem 4.14** *(Bartlett et al. 2004a, Corollary 2.2) Let $F$ be a class of functions which map $\Omega$ into $[-b,b]$. For every $x,t>0$ which satisfy that*

$$\mathbb{E}\sup_{f\in\bar{F}_t}\Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \le \frac{nt}{10b}-\frac{11bx}{10}\,,$$

*it holds that with probability at least $1-e^{-x}$,*

$$\Big\{f\in F:\ \mathbb{E}_\mu f^2\le t\Big\}\subseteq\Big\{f\in F:\ \sum_{i=1}^n f^2(X_i)\le 2tn\Big\}\,.$$

In particular, for the case $b=1$ and by setting $x$ such that $nt/10-11x/10=nt/16$, with probability at least $1-e^{-x}=1-e^{-cnt}$ (for an appropriate constant $c$), the radius of $\bar{F}_t/\mathbf{X}$ in $\ell_2^n$ is smaller than $\sqrt{2nt}$, and thus claim 2 follows.

Combining the two results 1. and 2., we can state the following Corollary:

**Corollary 4.15** *There are absolute constants $c$ and $c_1$ for which the following holds. For every $t\ge c_1/n$ such that $R_n\left(\bar{F}_t\right)\le nt/16$, there is a set $A_t'$ of samples $(\mathbf{X},\mathbf{X}')$ which has probability larger than $1-2e^{-cnt}$, on which the set $V=V(\mathbf{X},\mathbf{X}')=\bar{F}_t/\mathbf{X}\subset\ell_2^n$ is such that*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{\mathbf{v}\in V}\Big|\sum_{i=1}^n \varepsilon_i v_i\Big|\le nt/8$$

*and*

$$\sup_{\mathbf{v}\in V}\|\mathbf{v}\|_{\ell_2^n}\le\sqrt{2nt}\,.$$

From Theorem 3.11, for every such set $V$, there is a constant $c>0$ such that

$$Pr_{\boldsymbol{\varepsilon}}\Big\{\sup_{\mathbf{v}\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big|\ge\frac{nt}{4}\Big\}\le Pr_{\boldsymbol{\varepsilon}}\Big\{\sup_{\mathbf{v}\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big|\ge\mathbb{E}_{\boldsymbol{\varepsilon}}\sup_{\mathbf{v}\in V}\Big|\sum_{i=1}^n\varepsilon_i v_i\Big|+\frac{nt}{8}\Big\}\le ce^{-c'nt}\,.$$

Hence, there are absolute constants $c$ and $K$ such that

$$Pr_{\mathbf{X},\mathbf{X}',\boldsymbol{\epsilon}}\Big\{\exists f\in\bar{F}_t,\ \Big|\sum_{i=1}^n\varepsilon_i f(X_i)\Big|\ge\frac{nt}{4}\Big\}\le Ke^{-cnt}\tag{4.34}$$

which proves the theorem. ■

Thus, we have shown that a small complexity condition together with a strong control of the variance (leading to a stronger degree of concentration) are the reason why one could obtain in Mendelson (2003) asymptotically better error probability bounds for ERM.

## 4.4 Conclusion

We have presented a new and very general framework for deriving generalization bounds for random subclasses of functions. Based on a symmetrization technique, we showed that the analysis of deviations of empirical averages from expectations of random functions can be reduced to the analysis of the behaviour of the supremum of a Rademacher process indexed by certain random coordinate projections. This allowed us to state the following two principles as *sufficient* to guarantee and quantify the generalization ability of algorithms producing functions based on a random sample:

1. For generalization error of the order of $t$, small coordinate projections, in the sense that the Rademacher averages indexed by these random coordinate projections are smaller than $nt/8$;

2. For “high confidence”, that is, a small probability of error, concentration of the suprema of Rademacher processes indexed by these random coordinate projections.

We showed that the first condition is already sufficient to ensure learnability, whereas the degree of concentration in the second determines the confidence. We then presented conditions which lead to different degrees of concentration and thus to different probabilities of error and confidence intervals.

We demonstrated the generality of our approach by presenting a range of examples of frameworks which fall into our random subclass framework. Since we showed that the Glivenko-Cantelli conditions, and the compression, sparsity, and luckiness assumptions are different ways of ensuring small coordinate projections, we proved that their underlying mechanism is the same, and we were able to relate a number of data-dependent complexities to a priori complexities. However, the derivation of the new and more general framework has not led to structurally new learning results or new insights into the *design* of learning algorithms.

Andonova Jaeger (2004) has recently derived a *relative* deviation inequality for random classes of binary functions using similar techniques. The symmetrization step is carried out similarly to the one presented here. It uses a symmetric extension of the random set, and the proof follows closely the proof for uniform relative deviations from Anthony and Shawe-Taylor (1993). The final result uses Hoeffding’s inequality combined with a union bound argument, and provides a bound in terms of the expected shattering coefficient of the symmetric extension. It states that, for any binary-valued class of functions, any $t > 0$ and any $n \geq 2/t^2$,

$$Pr_{\mathbf{X}} \left\{ \sup_{f \in F_n(\mathbf{X})} \frac{\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)}{\sqrt{\mathbb{E}_\mu f}} \geq t \right\} \leq 4\mathbb{E}_{\mathbf{X},\mathbf{X}'} |F_{2n}^{\mathrm{sym}}\left((\mathbf{X},\mathbf{X}')\right)/(\mathbf{X},\mathbf{X}')|\, e^{\frac{-nt^2}{4}} . \tag{4.35}$$

Such a result reflects better convergence rates for functions with a small expectation taken from a random subclass of functions. An interesting further step would be to investigate, through the viewpoint presented in this chapter, the potential to obtain these fast rates for data-dependent classes without using the union bound and by making use of an explicit control of the variance, possibly in a data-dependent way.

Chapter 5

# Direct Data-Dependent Bounds for Empirical Risk Minimization

## 5.1 Introduction and Overview

In this chapter we present results on data-dependent generalization bounds for a specific algorithm, namely the Empirical Risk Minimization algorithm (ERM) (see Section 2.5). We have shown in the previous chapter (Section 4.3.5) that one can obtain faster convergence rates for the function produced by the ERM algorithm in terms of the complexity of a local subset of hypothesis functions with small variance. These localized bounds, which can be recovered within our random subclass framework, are based on the analysis of uniform relative deviations of expectations and empirical averages of functions. However, recent results of Bartlett and Mendelson (2005) show that one can give performance guarantees for the ERM algorithm by *directly* bounding the expectation of an empirical minimizer, without taking the detour of the analysis of these uniform deviations. The new results are based on a notion of complexity of local subsets of hypothesis functions with fixed expectations. They were demonstrated to yield significantly better convergence rate estimates compared to previous localized approaches. It is not clear how to recover them in the random subclass framework, which is based on uniform deviation analysis. Here, we analyze the extent to which one can derive *empirical* estimates for the generalization performance for the ERM algorithm based on the new results in Bartlett and Mendelson (2005).

Bartlett and Mendelson (2005) showed the improvement of their new results by comparing them to a performance bound for empirical risk minimizers based on uniform relative deviations. This "comparison" result [1] is also in terms of the same complexity notion of local subsets of hypothesis functions with a fixed expectation, and is itself

[1] Bartlett and Mendelson (2005) call it a "comparison" result due to the fact that it is based on the comparison of expectations and empirical averages as given by the uniform relative deviations over subclasses.

new and an improvement of previous localized results. Both the "optimal" direct result and the improved result based on uniform relative deviations presented in Bartlett and Mendelson (2005) depend on the *unknown* underlying probability measure and can therefore not be computed directly.

In order to investigate *empirical* estimates of the new localized results we follow the path presented in Bartlett and Mendelson (2005). We first present an empirical estimate for the "comparison" result, which leads to a new and improved localized *data-dependent* notion of complexity which determines the generalization ability of empirical risk minimizers. We then analyze the optimality of this data-dependent estimate. We show that, although the convergence rates obtained through the "comparison" result can be significantly outperformed by these obtained through the "optimal" result, it is *in general* impossible to derive a better *empirical* estimate than one based on the "comparison" result.

Recall that, given a bounded loss function, the empirical minimization algorithm produces the function $\boldsymbol{\mathcal{A}}_{\mathrm{ERM}}(\mathbf{Z}) = \mathrm{argmin}_{h \in H} \widehat{\boldsymbol{\mathcal{R}}}(h, \mathbf{Z})$ which has the smallest *empirical* error among all hypotheses in a given hypothesis class $H$ (see Section 2.5). We assume here that such a minimizer exists. Set $F$ to be the corresponding loss or excess loss class, and denote the loss function corresponding to the empirical minimizer by $\hat{f}$, that is,

$$\mathbb{E}_n \hat{f} = \min \{\mathbb{E}_n f : f \in F\} \ .$$

The question we wish to address is how to get a high probability bound on the conditional expectation — the generalization ability — of this empirical minimizer

$$\mathbb{E}\hat{f} = \mathbb{E}(\hat{f}(X) | X_1, \ldots, X_n) \,,$$

which is solely based on and therefore computable from empirical data. Our goal is to derive, based on this bound, the fastest possible convergence rates for the ERM algorithm.

Formally, let $F$ be a class of real-valued functions defined on the probability space $(\Omega, \mu)$. We will assume in the following that $F$ is a class of functions bounded by $b$, which contains the constant function 0, and all elements of $F$ have nonnegative expectations. These assumptions are justified given that we are looking at the loss or excess loss class of a function class containing a minimizer (see also Section 2.1).

Recall that *uniform* generalization bounds which hold for *any* algorithm which picks its hypothesis from $H$ and thus can have as associated loss function any function in $F$ are based on the analysis of the supremum of the empirical process $\sup_{f \in F} (\mathbb{E}f - \mathbb{E}_n f)$. This quantity is related, via concentration results (Section 3.1), to its expectation $\mathbb{E} \sup_{f \in F} (\mathbb{E}f - \mathbb{E}_n f)$, and via symmetrization techniques (Section 3.2) to the *global* complexity of the class $F$ as measured by the Rademacher averages of $F$, $R_n(F)$.

Specifically for the ERM algorithm, one can show that the convergence rates obtained through these bounds can be improved significantly by considering only deviations of expectations and empirical averages of functions contained in small subsets of $F$. The complexities of these small subsets are called *local* complexities. The derivation of improved error bounds for the ERM algorithm based on localization goes back to Massart (2000b), Koltchinskii and Panchenko (2000), and Mendelson (2002b). The bounds in Massart (2000b), Koltchinskii and Panchenko (2000), and Mendelson (2002b) were further generalized and improved in Bousquet (2002b); Bousquet et al. (2002); Koltchinskii (2003); Lugosi and Wegkamp (2004); Bartlett et al. (2004a). The driving idea for localized results is that the risk of the empirical minimizer depends on the Rademacher complexity of *local subsets* of $F$ centered around the true minimizer (which is 0 if $F$ is the excess loss class) rather than the whole class $F$. If the function class is not "too complex" around 0 these bounds can lead to significantly faster convergence rates. The key property on which this type of derivation is based is that one has control of the variance of functions in $F$ in terms of a polynomial function of their expectation. Classes satisfying this property, called *Bernstein classes of functions*, occur naturally in machine learning settings, for example in classification problems with 0-1 loss or regression with squared-loss.

In empirical processes theory it is common to study local subsets of functions which are balls of a given radius with respect to a chosen metric. The complexity of these local subsets as a function of the radius of the balls is called the *modulus of continuity*. It is known that the fixed point of the modulus of continuity with respect to the $L_2(\mu)$ metric can be used to bound the generalization error of functions (see, e.g., Birgé and Massart 1997; van de Geer 2000). This idea has led to the results involving the generalization error of empirical minimizers in terms of Rademacher complexities of functions with small variances (balls in $L_2(\mu)$ centered around 0) or small expectations when the loss class is Bernstein with $\beta = 1$ (Koltchinskii and Panchenko 2000; Bousquet 2002b; Bousquet et al. 2002; Koltchinskii 2003; Bartlett et al. 2004a). One arrives at complexity terms which control the suprema of the empirical processes $\sup \left\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \, \mathbb{E}f^2 \leq r\right\}$, where the radius $r$ is given by a fixed-point equation and can thus be determined, or, for classes of functions with $\operatorname{Var}(f) \leq c\mathbb{E}f$, at complexity terms which control the suprema of the empirical processes $\sup \{\mathbb{E}f - \mathbb{E}_n f : f \in F, \, \mathbb{E}f \leq r\}$. Such complexities reflect the fact that, in order to estimate the expectation of the empirical minimizer $\mathbb{E}\hat{f}$, one can ignore functions with large variance or expectation since the ERM algorithm is unlikely to select them. Since the local subsets are always smaller than the whole class $F$, bounds in terms of local complexities are formally sharper than the uniform bounds.

The derivation of error bounds using localized Rademacher complexities is based (besides concentration and symmetrization) on reweighting techniques which emphasize the functions with small variance or expectation combined with a technique called *peeling* (see, e.g., van de Geer 2000, page 69), which consists of splitting the original class according to the variance of the functions. *Reweighting* the functions is equivalent to the derivation of *relative* tail inequalities, which allow one to compare the deviation of empirical averages and expectations of functions by taking into account the value of their variance (or their expectation, if one can relate it to the variance). Such relative inequalities were previously employed in statistical learning theory by Vapnik and Chervonenkis (1971); Haussler (1992); Anthony and Shawe-Taylor (1993); Lee et al. (1996); Bartlett and Lugosi (1999); Anthony and Bartlett (1999) for error bounds in terms of combinatorial or metric complexities. However, since these results involve a union bound argument, the obtained probability bounds are potentially looser than the ones obtained through the localized results.

Unfortunately, local Rademacher complexities in terms of balls of small variances or expectations depend on the underlying probability distribution and are not practically useful because this distribution is unknown. Koltchinskii and Panchenko (2000); Koltchinskii (2003); Lugosi and Wegkamp (2004); Bartlett et al. (2004a) present ways to empirically approximate the unknown distribution with the empirical distribution and provide bounds in terms of *data-dependent* local complexities, which are entirely computable from data. The local subsets in these approaches are balls of functions centered around the *empirical* minimizers.

Recent results due to Bartlett and Mendelson (2005) improve such localized estimates. They show that one can derive upper and lower bounds on the expectation of the empirical minimizer $\mathbb{E}\hat{f}$ in terms of localized subsets which are "shells" of a given expectation. The complexity notion which governs the generalization ability of functions with small empirical error was shown to be essentially determined by the behaviour of the function $\mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f = r\}$. This function measures the complexity of the local subsets of $F$ with a fixed expectation $r$, denoted here by

$$F_r = \{f \in F \,:\, \mathbb{E}f = r\}\,.$$

Clearly, the shells $F_r$ are smaller than balls $\{f \in F : \mathbb{E}f \leq r\}$. The Bernstein condition implies directly that the balls of expectation smaller than $r$ are (if $r \leq 1$) contained in $L_2(\mu)$ balls. Thus, the complexity in terms of expectations are, up to constants, better than these in terms of $L_2(\mu)$ balls, and these results improve the previous localized notions of complexity. Hence, the derived bounds for the generalization ability of the ERM algorithm are (up to constants) as good or better than previous bounds. In this chapter, we will investigate the possibility and limitations when approximating these results entirely from empirical data.

In order to present our results, following Bartlett and Mendelson (2005) we define,

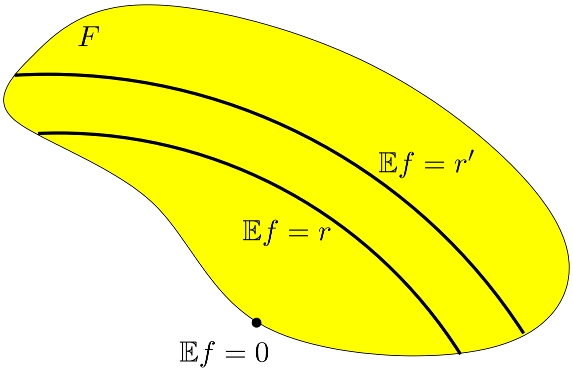


**Figure 5.1:** Graphical illustration of the local subsets $F_r = \{f \in F \ : \ \mathbb{E}f = r\} \subseteq F$ for functions with nonnegative expectations (for example excess loss classes).

for every $n$ and $\mu$, the following two functions which are measures for the complexity of the sets $F_r$:

$$\xi_{n,F,\mu}(r) = \mathbb{E}\sup\{|\mathbb{E}f - \mathbb{E}_n f| : f \in F_r\},$$
$$\xi'_{n,F,\mu}(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F_r\}.$$

These two functions control the generalization ability of all the functions in $F_r$ whenever one has a strong degree of concentration for the empirical processes $\sup_{f\in F_r}|\mathbb{E}f - \mathbb{E}_n f|$ and $\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ around their expectation. It is easy to see that for all $r$, $\xi'_{n,F,\mu}(r) \le \xi_{n,F,\mu}(r)$, and note that, in order to solely determine the generalization ability, one would need only $\xi'_{n,F,\mu}(r)$. However, for the data-dependent estimation we require the stronger control of the similarity of expectations and empirical averages through the "two-sided" complexity function $\xi_{n,F,\mu}(r)$. In the following, in cases where the underlying probability measure and the class $F$ are clear, we will refer to these functions by $\xi_n$ and $\xi'_n$.

As we will see, the behaviour of the functions $\xi_n$ and $\xi'_n$ as functions of $r$ determines the performance bounds obtained in Bartlett and Mendelson (2005). For classes of functions whose variances are bounded by their expectations, the "comparison" result shows that $\mathbb{E}\hat{f}$ is essentially upper bounded by the largest fixed points of $\xi_n$ or $\xi'_n$. More precisely, the error bounds are in terms of the quantities

$$r_n^* = \inf\{r : \ \xi_n(r) < r/4\},$$

and

$$r_n'^* = \inf\left\{r : \ \xi'_n(r) < r/4\right\},$$

which are upper bounds on the largest fixed points of the functions $4\xi_n(r)$ and $4\xi'_n(r)$. Since $\xi'_n(r) \le \xi_n(r)$, it follows immediately that $r_n'^* \le r_n^*$. The "optimal" result in

Bartlett and Mendelson (2005) improves these bounds through a direct analysis of the empirical minimizer. The analysis is based on the full strength of the very sharp concentration in Talagrand's concentration inequality (Theorem 3.12, page 44) for empirical processes indexed by each of the local subsets $F_r$. The "optimal" result shows that the expectation of the empirical minimizer is essentially determined by the largest maximizer of $\xi'_n(r) - r$. One can show that one can even derive upper and lower bounds for $\mathbb{E}\hat{f}$ in terms of upper and lower approximations of the quantity

$$s^*_n = \sup \left\{ r : \text{ argmax} \left\{ \xi'_n(r) - r \right\} \right\} .$$

If $\xi'_n(r) - r$ is peaked around $s^*_n$ and the class $F$ is not too complex around 0, then one can derive matching upper and lower bounds for $\mathbb{E}\hat{f}$ which will be of the order of $s^*_n$. One can also show that $s^*_n \leq r'^*_n$ (see Section 5.5).

The structure of this chapter is as follows: In Section 5.2 we will first present the structural assumptions which we require in order to derive the results. In Section 5.3 we then present the "comparison" estimates from Bartlett and Mendelson (2005) on which we base our data-dependent estimates for the expectation of empirical minimizers. We present our data-dependent estimate for $r^*_n$ in Section 5.4. In order to investigate the optimality of this estimate (Section 5.5), we first present the improved "optimal" result from Bartlett and Mendelson (2005) in terms of $r^*_n$ (or approximations thereof) in Section 5.5.1. This will be followed by examples showing that an estimate based on $s^*_n$ is potentially asymptotically tighter than an estimate based on $r^*_n$. However, in Section 5.5.2 we show through a counter-example that, *in general*, it is impossible to compute a *data-dependent* estimate of $s^*_n$ which is better than the empirical estimate on $r^*_n$, and thus, based solely on empirical data, our empirical estimate is optimal.

## 5.2 Structural Assumptions

In order to derive estimates for $\mathbb{E}\hat{f}$ we have to make two additional mild structural assumptions on the class $F$, namely, that $F$ is star-shaped around 0 and satisfies a Bernstein condition. Whereas the first condition imposes some "regularity" on $\xi_n$ and $\xi'_n$, the Bernstein condition allows one to control the variance of the functions in the class. Recall that Talagrand's concentration inequality ensures a degree of concentration for the supremum of the deviations of expectations and empirical averages which depends on the maximal variance of the functions in the indexing class. As we will see, the specific control of the variance in terms of expectations leads to a control of the degree of concentration in the subsets $F_r$ which depends on $r$.

**Definition 5.1** *We say that $F$ is a $(\beta, B)$-Bernstein class with respect to the probability* *measure $\mu$ (where $0 < \beta \leq 1$ and $B \geq 1$), if every $f \in F$ satisfies*

$$\mathbb{E}f^2 \leq B(\mathbb{E}f)^\beta \,.$$

*We say that $F$ has Bernstein type $\beta$ with respect to $\mu$ if there is some constant $B$ for which $F$ is a $(\beta, B)$-Bernstein class.*

Thus, for Bernstein classes of functions, the variance for every function is bounded by a power of its expectation uniformly over the class. This condition is crucial for relating a strong degree of concentration of suprema of empirical processes indexed by $F_r$ to the expectation $r$.

Note that the Bernstein condition implies, for excess loss classes, uniqueness of the minimizer. If $f^*$ is any minimizer, assume that there exists a different minimizer $f'^* \neq f^*$. Then, for $f = f'^* - f^*$ (which is a function from the excess loss class associated to $f^*$), $\mathbb{E}f = 0$ whereas $\mathbb{E}f^2 > 0$, which contradicts the Bernstein condition.

Although this might seem as a strong assumption, the Bernstein condition is indeed satisfied for a range of loss classes arising naturally in the learning setting. For example, it is satisfied for all distributions for classes of nonnegative bounded functions with $\beta = 1$, and therefore for all loss classes induced by nonnegative bounded loss functions, as for example the 0-1 loss and the square-loss. As was shown in Lee et al. (1996); Mendelson (2002b), it is also satisfied for all distributions for excess loss classes associated with learning problems where the hypothesis class is a convex class of functions bounded by 1, and the loss function is a power-type function. In particular, for the regression with square-loss, $\beta = 1$, and one can take $B = 16$ (Lee et al. 1996; Mendelson 2002b).

The Bernstein property is also satisfied for classification problems in which the data is labelled consistently, that is, in cases in which the decision for a class is not random. Such a condition was quantified in Massart and Nédélec (2004); Mammen and Tsybakov (1999) and Tsybakov (2004), by imposing that the conditional expectation $p(x) = \mathbb{E}[Y|X = x]$ is, with high probability, not "too close" to 1/2 (where $X$ denotes the input and $Y$ the label). It was shown in Mammen and Tsybakov (1999); Tsybakov (2004); Massart and Nédélec (2004) that when these "low noise" conditions are satisfied one can get faster rates of convergence to the Bayes classifier. It is easy to see that these "low noise conditions" imply that the variance is bounded by a polynomial function of the expectation (see, e.g., Boucheron et al. 2004b) and therefore that the function class is Bernstein. Thus, low noise conditions in classification are in fact Bernstein conditions, and the better results obtained are due to a stronger concentration of suprema of empirical processes.

**Definition 5.2** *$F$ is called star-shaped around 0 if for every $f \in F$ and $0 \leq \alpha \leq 1$, $\alpha f \in F$.*

(a) $F$.

(b) $\mathrm{star}(F, 0)$.

**Figure 5.2**: Enlarging the function class $F$ (its elements are depicted as dots) to $\mathrm{star}(F, 0)$.

Observe that if $F$ is an excess loss class, then any empirical minimizer over $F$ is also an empirical minimizer over its star-shaped hull

$$\mathrm{star}(F, 0) := \{\alpha f : f \in F,\, 0 \leq \alpha \leq 1\}\,.$$

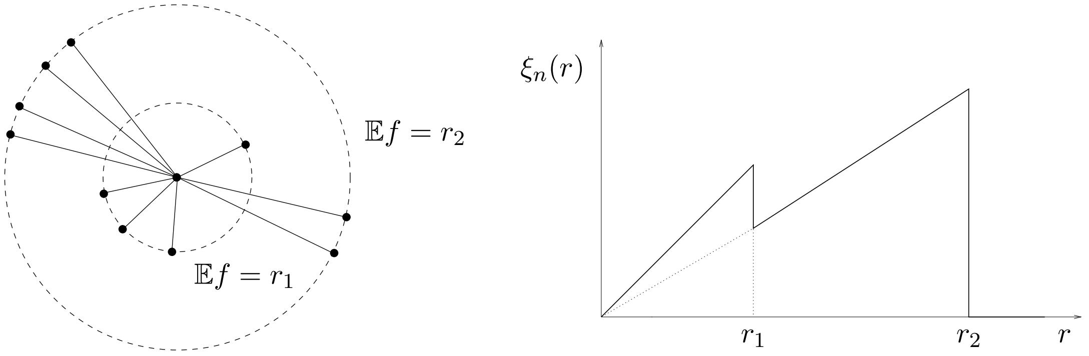
Therefore, one can replace $F$ with the larger class $\mathrm{star}(F, 0)$ (see Figure 5.2). If $F$ is a Bernstein class, note that $\mathrm{star}(F, 0)$ is also Bernstein with the same constants.

Although $F \subseteq \mathrm{star}(F, 0)$, since $\mathbb{E}$ and $\mathbb{E}_n$ are linear, the "complexity" of $\mathrm{star}(F, 0)$ is not much larger than that of $F$. One can show that the functions $\xi_n$ and $\xi'_n$ do not increase too much and thus $r^*_n$ and $r'^*_n$ essentially remain unchanged. For example, if $F$ is star-shaped and the original function class contains only functions with a given expectation $r_0$, than $\xi_n(r)$ for its star-shaped hull is a linear function of $r$ for $r \leq r_0$ and $0$ for $r > r_0$. Another example illustrating the behaviour of $\xi_n(r)$ is given in Figure 5.3 for the star-shaped hull of a class which contains solely functions whose expectations can only take on values of $r_1$ or $r_2$.

In general, the advantage one gains by replacing $F$ with its star-shaped hull is that it imposes some regularity on the complexity functions $\xi_n$ and $\xi'_n$ which allows one to analyze the complexity of the set of functions with a "large" expectation through their rescaled versions with a smaller fixed expectation. For each expectation level $r$, $F_r$ contains, rescaled, all functions from $F_{\geq r} := \{f \in F \,:\, \mathbb{E}f \geq r\}$. New functions which can lead to a considerable increase of complexity can thus only be encountered at a smaller expectation level $r' < r$ (see Figure 5.4).

**Lemma 5.3** *If $F$ is star-shaped around $0$, then for any $0 < r_1 < r_2$,*

$$\frac{\xi_n(r_1)}{r_1} \geq \frac{\xi_n(r_2)}{r_2}.$$



(a) $\mathrm{star}(F, 0)$.

(b) Graph of function $\xi_n$ corresponding to $\mathrm{star}(F, 0)$.

**Figure 5.3:** An example of a graph of a function $\xi_n$ for the class $\mathrm{star}(F, 0)$, where $F$ contains only functions with expectations $r_1$ and $r_2$.

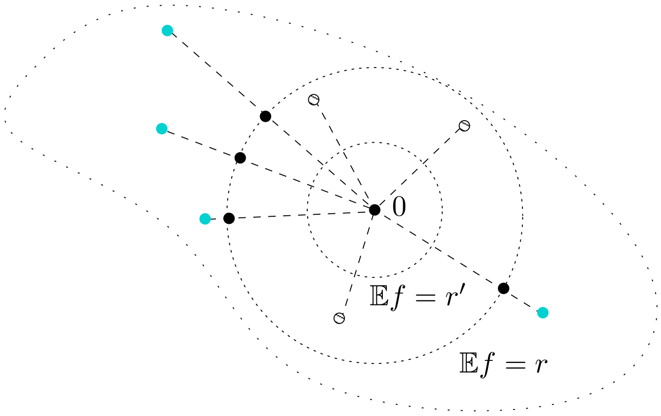


**Figure 5.4:** At each level $r$, $F_r$ (black dots) "sees" rescaled versions of all functions from $F_{\geq r}$ (gray dots). New atoms can appear only at smaller levels $r' < r$ (white dots).

*In particular, if for some $\alpha$, $\xi_n(r) \geq \alpha r$ then for all $0 < r' \leq r$, $\xi_n(r') \geq \alpha r'$. The same holds for $\xi'_n$.*

A proof for $\xi_n$ can be found in Bartlett and Mendelson (2005). It holds analogously for $\xi'_n$.

We call $\xi_n$ and $\xi'_n$ in this case "sub-linear" because, in each interval $[r, b]$, the graph of $\xi'_n$ is below the ray connecting $(0, 0)$ with $(r, \xi_n(r))$, and thus $\xi'_n$ grows slower than linearly in $r$. This is due to the easy to see fact that the functions $\xi_n(r)/r$ and $\xi'_n(r)/r$ are non-increasing, which is exactly the property which will allow us to estimate $r^*_n$ and $r'^*_n$. Figure 5.5 illustrates the graph of a typical function with this "sub-linear" property.

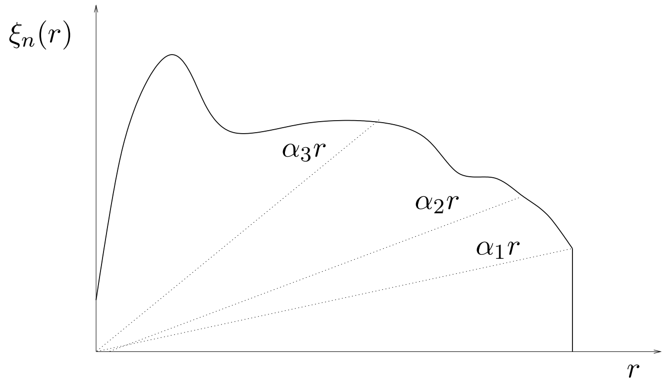


**Figure 5.5:** The graph of a function $\xi_n$ which is "sub-linear" (cf. Lemma 5.3). Observe that this implies that in the interval $[0,r]$, the graph of $\xi_n(r)$ is above or on the line connecting $(0,0)$ with $(r,\xi_n(r))$, whereas in each interval $[r,b]$, the graph of $\xi'_n$ is below or on the line connecting $(0,0)$ with $(r,\xi_n(r))$.

## 5.3 Localization for ERM

In this section we present the localized "comparison" result from Bartlett and Mendelson (2005) on which we will base our empirical estimate. In order to obtain the estimate for the empirical minimizer in terms of the complexity functions $\xi_n$ and $\xi'_n$ we will consider uniform relative deviations over $F_r$ for the random variables $|\mathbb{E}f-\mathbb{E}_n f|/\mathbb{E}f$ and $(\mathbb{E}f-\mathbb{E}_n f)/\mathbb{E}f$. Because the class $F$ is star-shaped, one can show that the projection of the function class $F_r$ and that of the class $F_{\geq r}$ onto any sample behave similarly in the following sense: for a given sample, uniform control of the deviation of $|\mathbb{E}f-\mathbb{E}_n f|/\mathbb{E}f$ over $F_r$ is equivalent to a uniform control for the same deviation over the larger set $F_{\geq r}$. We can state thus the following lemma:

**Lemma 5.4** *Let $F$ be star-shaped around $0$ and let $\mathbf{X}\in\Omega^n$ be a sample distributed according to $\mu^n$. Then, for any $r,t>0$, it holds that $\mathbb{E}f-\mathbb{E}_n f\leq t\,\mathbb{E}f$ for every $f\in F_r$ if and only if $\mathbb{E}f-\mathbb{E}_n f\leq t\,\mathbb{E}f$ for every $f\in F_{\geq r}$.*

The proof can be found in Bartlett and Mendelson (2005).

Figure 5.6 shows graphically the statement of this Theorem: for star-shaped classes, the complexity of $f\in F_{\geq r}$ is "transferred" to the set $F_r$ if we are interested in analyzing relative deviations of the form $|\mathbb{E}f-\mathbb{E}_n f|/\mathbb{E}f$ uniformly on these sets.

Since all functions in $F_r$ have the same expectation (equal to $r$), the quantities we need to control are $\sup_{f\in F_r}|\mathbb{E}f-\mathbb{E}_n f|$. Already for bounded functions, for each $F_r$, the suprema $\sup_{f\in F_r}|\mathbb{E}f-\mathbb{E}_n f|$ are highly concentrated around their mean $\xi_n(r)=\mathbb{E}\sup_{f\in F_r}|\mathbb{E}f-\mathbb{E}_n f|$ (see Section 3.1, for example Corollary 3.8, page 42). Hence, $\xi_n(r)$ is a measure for the deviation $\sup_{f\in F_r}|\mathbb{E}f-\mathbb{E}_n f|$ for most of the samples, and thus we can relate the control of $|\mathbb{E}f-\mathbb{E}_n f|/\mathbb{E}f$ in $F_r$ to that of $\xi_n(r)$. If the variance

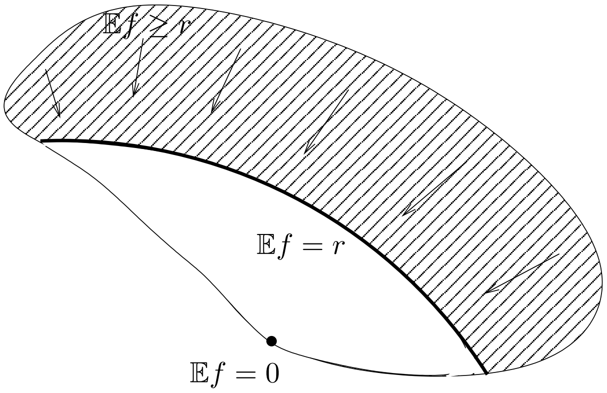


**Figure 5.6:** For star-shaped classes, uniform control of $(\mathbb{E}f - \mathbb{E}_n f)/\mathbb{E}f$ in $F_r$ is equivalent to uniform control of the the same relative deviation in $F_{\geq r}$, since each element in $F_{\geq r}$ is "seen" rescaled in $F_r$.

of this process is "small", an even stronger concentration result can be obtained from Talagrand's concentration inequality and thus, a higher degree of confidence in the error bound. Here, the fact that the class $F$ is Bernstein is crucial; for Bernstein classes, where the variance can be bounded in terms of expectations, the empirical process indexed by $F_r$ can be controlled in terms of the expectation $r$ with same high degree of concentration, allowing thus localization using $\mathbb{E}f$ instead of $\mathbb{E}f^2$. By analyzing the degree of concentration which one can obtain for the processes indexed by $F_r$ dependent on $r$, one observes an interesting phenomenon which is displayed by the following theorem. It states that, for Bernstein classes, there is a phase transition around the point where $\xi_n(r) \sim r$ . Above this point, the local subsets $F_r$ are small and the expectation and most empirical means are close. Below this point, the sets $F_r$ are too rich to allow uniform statistical control.

**Theorem 5.5** *There is an absolute constant $c > 0$ for which the following holds. Let $F$ be a class of functions defined on a probability space $(\Omega, \mu)$, such that for every $f \in F$, $\|f\|_\infty \leq b$. Assume that $F$ is a $(\beta, B)$-Bernstein class. Suppose $r \geq 0$, and $0 < \alpha, \lambda < 1$ satisfy*

$$r \geq c \max\left\{\frac{bx}{n\alpha^2\lambda}, \left(\frac{Bx}{n\alpha^2\lambda^2}\right)^{1/(2-\beta)}\right\}. \tag{5.1}$$

*1. If $\xi_n(r) \geq (1+\alpha)r\lambda$ , then with probability at least $1 - e^{-x}$,*

$$\sup_{f \in F_r} |\mathbb{E}f - \mathbb{E}_n f| \geq \lambda \mathbb{E}f\,.$$

*2. If* $\xi_n(r) \leq (1-\alpha)r\lambda$ *, then with probability at least* $1-e^{-x}$*,*

$$\sup_{f\in F_r} |\mathbb{E}f - \mathbb{E}_n f| \leq \lambda \mathbb{E}f\,.$$

The proof of this theorem can be found in Bartlett and Mendelson (2005) and is based on Talagrand's concentration inequality in Massart's version (Theorem 3.12, page 44) for the empirical process $Z = n\sup_{f\in F_r}|\mathbb{E}f - \mathbb{E}_n f|$ .

The conditions on $r$ arise from the additional terms in Talagrand's concentration inequality. Massart's version of this inequality (Theorem 3.12, page 44) resembles the form one gets for Bernstein's inequality for one function. Besides the term involving the expectation of the supremum of the empirical process, in Bernstein's inequality the degree of concentration depends on the variance term and on a bound on the $L_\infty$ norm of the function. In Massart's formulation for the concentration of suprema of empirical processes one refinds these additional terms as the maximal variance and the $L_\infty$ bound of the indexing class. The first term on the right-hand side of equation (5.1), essentially $bx/n$, arises from the $L_\infty$ term, whereas the second term, $(Bx/n)^{1/(2-\beta)}$, arises from a bound on the variance through the expectation given by the Bernstein condition.

By employing Talagrand's concentration inequality for the one-sided empirical process $Z = n\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ (Theorem 3.12, page 44), an analogous proof leads to the following theorem involving $\xi'_n$.

**Theorem 5.6** *There is an absolute constant* $c > 0$ *for which the following holds. Let* $F$ *be a class of functions defined on a probability space* $(\Omega,\mu)$*, such that for every* $f\in F$*,* $\|f\|_\infty \leq b$*. Assume that* $F$ *is a* $(\beta, B)$*-Bernstein class. Suppose* $r\geq 0$ *and* $0<\alpha$*,* $\lambda<1$ *satisfy*

$$r \geq c\max\left\{\frac{bx}{n\alpha^2\lambda}, \left(\frac{Bx}{n\alpha^2\lambda^2}\right)^{1/(2-\beta)}\right\}.$$

*1. If* $\xi'_n(r) \geq (1+\alpha)r\lambda$ *, then with probability at least* $1-e^{-x}$*,*

$$\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) \geq \lambda\mathbb{E}f.$$

*2. If* $\xi'_n(r) \leq (1-\alpha)r\lambda$ *, then with probability at least* $1-e^{-x}$*,*

$$\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) \leq \lambda\mathbb{E}f.$$

Set, for example, $\lambda=\alpha=1/2$. Then Theorem 5.5 implies that, if $r\geq c\max\{bx/n, (Bx/n)^{1/(2-\beta)}\}$ , and $\xi_n(r)\leq r/4$ , then with probability at least $1-e^{-x}$, for every

$f \in F_r$, $|\mathbb{E}f - \mathbb{E}_n f| \leq \mathbb{E}f/2$. Recall that

$$r_n^* = \inf \left\{ r : \xi_n(r) \leq r/4 \right\} .$$

Thus, if $F$ is star-shaped, by Lemma 5.3, it follows that for every $r \geq r_n^*$, if $f \in F_{\geq r}$ then $\mathbb{E}f \leq 2\mathbb{E}_n f$. The same holds for

$$r_n'^* = \inf \left\{ r : \xi_n'(r) \leq r/4 \right\} ,$$

by Theorem 5.6. Since for any $r$, $\xi_n'(r) \leq \xi_n(r)$, always $r_n'^* \leq r_n^*$.

Thus, for any function $f$ with expectation of the order $\Omega(1/n^{1/(2-\beta)})$ , with high probability either $\mathbb{E}f \leq r_n'^*$ or $\mathbb{E}f \leq 2\mathbb{E}_n f$ . In particular, for the empirical minimizer, we obtain the following estimates:

**Theorem 5.7** *Let $F$ be a class of functions defined on a probability space $(\Omega, \mu)$ which is $(\beta, B)$-Bernstein, bounded by $b$, and star-shaped around $0$. Then there is an absolute constant $c > 0$ such that if $r_n^*$ is defined as above and*

$$\tilde{r}_n := \max \left\{ r_n^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)} \right\},$$

*then with probability at least $1 - e^{-x}$, any empirical minimizer $\hat{f} \in F$ satisfies*

$$\mathbb{E}\hat{f} \leq \tilde{r}_n .$$

*Also, if $r_n'^*$ is defined as above and*

$$\tilde{r}_n' := \max \left\{ r_n'^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)} \right\} ,$$

*then with probability at least $1 - e^{-x}$, any empirical minimizer $\hat{f} \in F$ satisfies*

$$\mathbb{E}\hat{f} \leq \tilde{r}_n' .$$

Thus, with high probability, $r_n^*$ and $r_n'^*$ respectively are upper bounds for $\mathbb{E}\hat{f}$, as long as $r_n^*$ and $r_n'^*$ are larger than $c/n^{1/(2-\beta)}$ . Figure 5.7 shows graphically $r_n^*$ and $r_n'^*$. Note that $r_n'^*$ can be much smaller than $r_n^*$, and so the convergence rates obtained through $r_n'^*$ are potentially better.

For $\beta = 1$, the estimates based on $r_n'^*$ and $r_n^*$ are at best $1/n$, and in general at best $1/n^{1/(2-\beta)}$. Thus, the degree of control of the variance through the expectation, as measured by the Bernstein condition, is the parameter which influences the rate of convergence which can be obtained through the bound in terms of $r_n'^*$ and $r_n^*$ whenever

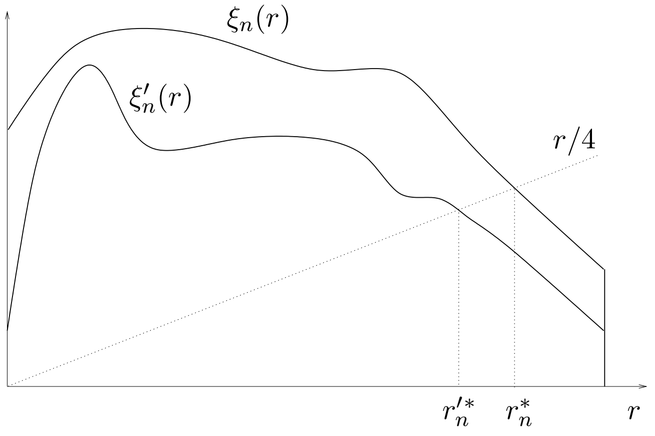


**Figure 5.7:** The graph of the functions $\xi_n$ and $\xi'_n$ , and the corresponding values for $r^*_n$ and $r'^*_n$. For any $r$, $\xi'_n(r) \leq \xi_n(r)$ , and thus always $r'^*_n \leq r^*_n$.

one requires a confidence which is arbitrarily close to 1. In particular, this approach recovers the better learning rates for convex function classes from Lee et al. (1996, 1998) and for low noise classification from Tsybakov (2004); Massart and Nédélec (2004), as both convexity of $F$ for squared-loss and low noise conditions imply a better control of the variance of functions in the class.

Thus, the quantities which upper bound the relative difference of expectations and empirical averages in the class $F$ and the error rate of the empirical minimizer are (up to constants) the fixed points of the functions $\xi_n(r)$ and $\xi'_n(r)$ respectively. Observe that these functions measure the expectation of the empirical processes $\sup_{f \in F_r} |\mathbb{E}f - \mathbb{E}_n f|$ respectively $\sup_{f \in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ indexed by the "local" subset $F_r$. Recall that in the classical results, involving a global complexity measure, the resulting bounds are given in terms of the Rademacher averages of the class, which correspond to the processes $\sup_{f \in F} |\mathbb{E}f - \mathbb{E}_n f|$ indexed by the whole set $F$. This set is larger and much more complex than $F_r$. This is the reason why one only gets rates of convergence of at best $O(1/\sqrt{n})$ through global estimates, since functions with "large" expectations contribute to $\sup_{f \in F} |\mathbb{E}f - \mathbb{E}_n f|$ , whereas with the localized approaches the rates can be as good as $O(1/n^{1/(2-\beta)})$. Also, in the previous bounds in terms of local complexity measures established in Koltchinskii and Panchenko (2000); Massart (2000b); Lugosi and Wegkamp (2004); Bartlett et al. (2004a); Koltchinskii (2003), the fixed point of the supremum of this process is indexed by the subsets $\{f \in F : \mathbb{E}f \leq r\}$ , which are all larger sets than $F_r$, or by subsets $\{f \in F : \mathbb{E}f^2 \leq r\}$ , which are also larger due to the Bernstein condition. If the class is very "complex" around 0, the difference can lead to asymptotically different estimates for the convergence rates.

## 5.4 Data-Dependent Estimation

So far, we have presented how the localization functions $\xi_n(r)$ and $\xi_n'(r)$ give an upper bound, via $r_n^*$ and ${r_n'}^*$, for the expectation of the empirical minimizer, and that these results improve previous results using local complexity measures. However, the functions $\xi_n(r)$ and $\xi_n'(r)$ depend on the unknown distribution $\mu$, and are thus unknown.

In the following, we will show that it is possible to empirically estimate $r_n^*$. We will present an algorithm, previously presented in Bartlett et al. (2004b), which enables us to estimate $r_n^*$ completely from empirical data. The estimation is similar in spirit to data-dependent estimates for localized measures proposed in Koltchinskii and Panchenko (2000); Lugosi and Wegkamp (2004); Bartlett et al. (2004a). It is based on the idea of approximating the function shells $F_r$ with "empirical error shells", an idea which was pursued in many different machine learning contexts (see, e.g., Haussler et al. 1994; Kowalczyk et al. 1995).

Although always $\xi_n'(r) \leq \xi_n(r)$, and thus $r_n'^* \leq r_n^*$, the best data-dependent estimate which we can present is one of the order of $r_n^*$. The reason for doing so is that the one-sided control of uniform deviations (as given by $\xi_n'(r)$ and $r_n'^*$), although sufficient for deriving upper bounds on the generalization error, only allows one to relate the shells $F_r$ (for large values of $r$) to empirical sets $\{f \in F \,:\, \mathbb{E}_n f \leq (1-\lambda)r\}$ . By using the two-sided localization function $\xi_n(r)$, however, we will see that one can approximate the shells $F_r$ with empirical shells

$$\hat{F}_{r_1,r_2} := \{f \in F \,:\, r_1 \leq \mathbb{E}_n f \leq r_2\}$$

in such a way that, for $n$ going to infinity, $\hat{F}_{r_1,r_2} \longrightarrow F_r$ .

In the next sections we will discuss the optimality of this data-dependent estimate, and we will show that, *in general*, a data-dependent estimate for the expectation of the empirical minimizer which can be computed solely from empirical data is not distinguishable from an estimate for $r_n^*$ (up to constants). However, in order to show that, we will also make use of the function $\xi_n'(r)$ and the estimate $r_n'^*$.

The overall idea we pursue is to estimate $r_n^*$ from an empirically computable function $\widehat{\xi}_n(r)$ which is, with high probability, an upper bound for the function $\xi_n(r)$ and therefore, its fixed point $\hat{r}_n^* = \inf\left\{r : \widehat{\xi}_n(r) \leq \frac{r}{4}\right\}$ is an upper bound for $r_n^*$. We will construct the function $\widehat{\xi}_n(r)$ such that $\widehat{\xi}_n(r)/r$ is non-increasing, and this will enable us to determine $\hat{r}_n^*$ using a binary search algorithm.

We will make use of the following direct lemma of Theorem 5.5 applied to the case $\alpha = 1/2$, $\lambda = 1/2$.

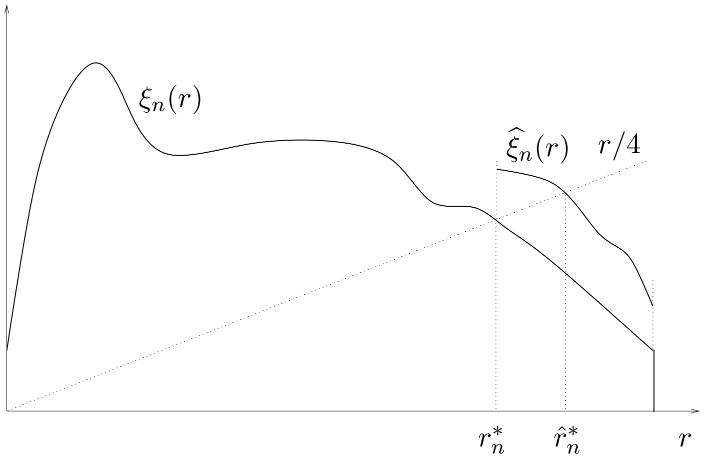


**Figure 5.8:** An empirical sub-linear function $\xi_n$ which is, for most samples, an upper bound on $\xi_n$. In this case, the empirical quantity $\hat{r}_n^* = \inf\left\{r : \widehat{\xi}_n(r) \leq r/4\right\}$ is, with high probability over a random draw of samples, an upper bound for $r_n^*$.

**Lemma 5.8** *If $F$ is $(\beta, B)$-Bernstein, and $r \geq 0$ such that*

$$r \geq c \max\left\{\frac{bx}{n}, \left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$$

*and $\xi_n(r) \leq \frac{r}{4}$ , then with probability larger than $1 - e^{-x}$ , every $f \in F_r$ satisfies that $r/2 \leq \mathbb{E}_n f \leq 3r/2$ .*

Hence, with high probability, $F_r$ is contained in an "empirical shell"

$$F_r \subseteq \{f \in F \,:\, r/2 \leq \mathbb{E}_n f \leq 3r/2\} = \hat{F}_{\frac{r}{2},\frac{3r}{2}} \,.$$

Since $F$ is star-shaped, then by Lemma 5.3, $\xi_n(r) \leq \frac{r}{4}$ if and only if $r \geq r_n^*$ . Therefore, if $r \geq \max\left\{r_n^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$, then with probability larger than $1 - e^{-x}$, $F_r \subseteq \hat{F}_{\frac{r}{2},\frac{3r}{2}}$ . We will additionally assume that $r \geq 1/n$. Thus, $\xi_n(r)$ is upper bounded by the "complexity" of (the expectation of deviations indexed by) the set $\hat{F}_{\frac{r}{2},\frac{3r}{2}}$ , a set of functions which we can determine entirely by looking at the projections of $F$ on empirical data. This complexity can be approximated (by applying symmetrization and concentration techniques) through the empirical Rademacher averages of the function class $\hat{F}_{\frac{r}{2},\frac{3r}{2}}$ .

The empirical Rademacher averages evaluated on a sample $\mathbf{X}$ of size $n$ satisfy

$$\widehat{R}_n\left(F_r, \mathbf{X}\right) \leq \widehat{R}_n\left(\hat{F}_{\frac{r}{2},\frac{3r}{2}}, \mathbf{X}\right) .$$

By symmetrization (Corollary 3.20, page 50) and concentration of Rademacher averages around their mean (Theorem 3.16, page 48), it follows thus that with probability at least $1-2e^{-x}$ over the random draw of samples,

$$\xi_n(r) \leq 2R_n\left(F_r\right) \leq 4\widehat{R}_n\left(F_r, \mathbf{X}\right) + \frac{bx}{n} \leq 4\widehat{R}_n\left(\hat{F}_{\frac{r}{2},\frac{3r}{2}}, \mathbf{X}\right) + \frac{r}{c},$$

where we used the fact that $r \geq \frac{cbx}{n}$ (and clearly we can assume that $c > 8$, a fact which will be used later).

We have thus obtained for every such $r$ an upper bound for $\xi_n(r)$ which holds with high probability and is computable from the sample. To make the upper bound hold *uniformly for all* $r$, we will divide the range of $r$ into intervals of length $1/n$. In each such interval, since $F$ is star-shaped, $\xi_n(r)$ cannot grow faster than linear and thus can be bounded. A union bound over the $O(n)$ intervals leads to an upper bound which holds for the whole range of $r$. More formally, set

$$r' = \max\left\{r_n^*, \frac{1}{n}, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$$

and

$$R = \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{\lceil bn\rceil}{n}\right\} \cap \left[\frac{\lfloor r'n\rfloor}{n}, \frac{\lceil bn\rceil}{n}\right].$$

By the union bound, and since $|R| \leq bn+1$, with probability at least $1-2(bn+1)e^{-x}$, $\xi_n(r) \leq 4\widehat{R}_n\left(\hat{F}_{\frac{r}{2},\frac{3r}{2}}, \mathbf{X}\right) + \frac{r}{c}$ for every $r \in R$. By Lemma 5.3, if $r \in \left[\frac{k}{n}, \frac{k+1}{n}\right]$, then $\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k}$. Thus, with probability at least $1-2(bn+1)e^{-x}$, every $r \in [r', b]$ satisfies

$$\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k} \leq \left(4\widehat{R}_n\left(\hat{F}_{\frac{k}{2n},\frac{3k}{2n}}, \mathbf{X}\right) + \frac{k}{cn}\right)\frac{nr}{k} \leq 8\widehat{R}_n\left(\hat{F}_{c_1r,c_2r}, \mathbf{X}\right) + \frac{r}{c},$$

where $c_1$, $c_2$ are positive absolute constants. Hence we can define

$$\widehat{\xi}_n(r) = 8\widehat{R}_n\left(\hat{F}_{c_1r,c_2r}, \mathbf{X}\right) + \frac{r}{c},$$

and with probability at least $1-2(bn+1)e^{-x}$, for every $r \in [r', b]$, it holds that $\xi(r) \leq \widehat{\xi}_n(r)$. It is easy to check that $\widehat{\xi}_n(r)/r$ is non-increasing and thus $\widehat{\xi}_n$ is also sub-linear.

Let

$$\hat{r}_n^* = \inf\left\{r : \widehat{\xi}_n(r) \leq \frac{r}{4}\right\}.$$

Then, with probability at least $1-2(bn+1)e^{-x}$, it is true that $\hat{r}_n^* \geq r_n^*$. Moreover, since $\widehat{\xi}_n(r)$ is sub-linear, it follows that $r \geq \hat{r}_n^*$ if and only if $\widehat{\xi}_n(r) \leq \frac{r}{4}$ (see Figure 5.8).

With this, the algorithm from Figure 5.9, performing a binary search for the fixed point of $4\widehat{\xi}_n(r)$, computes in $O(\log n)$ steps an upper bound on $\hat{r}_n^*$ based on the data.

```
Algorithm RSTAR($F$, $X_1, \ldots, X_n$)
    $r_L := 1/n$, $r_R := b$.
    if $\widehat{\xi}_n(r_R) \le r_R/4$ then
        for $l = 0$ to $\lceil \log_2 bn \rceil$
            $r := \frac{r_R - r_L}{2}$;
            if $\widehat{\xi}_n(r) > r/4$ then $r_L := r$,
                else $r_R := r$.
    Output $\bar{r} := r_R$.
```

**Figure 5.9:** Binary search algorithm computing $\hat{r}_n^*$ and thus an upper bound for $r_n^*$ in $O(\log n)$ steps.

By the construction, $\bar{r} - \frac{1}{n} \le \hat{r}_n^* \le \bar{r}$ . Since $r_n^*$ and thus $\hat{r}_n^*$ are always at least $cbx/n = O(1/n)$ , and $\bar{r} \in [\hat{r}_n^*, \hat{r}_n^* + 1/n]$ , it follows that $\bar{r}$ and $\hat{r}_n^*$ are of the same order and thus will lead to the same rates of convergence. Therefore, for every $n$, with probability larger than $1 - 2(bn+1)e^{-x}$ over the random draw of samples, it holds that $r_n^* \le \max\{\bar{r}, cbx/n\}$ . We are thus ready to state the following theorem:

**Theorem 5.9** *Let $F$ be a class of functions defined on a probability space $(\Omega, \mu)$ which is $(\beta, B)$-Bernstein, bounded by $b$, and star-shaped around $0$. For any $x > 0$ and any $n$, with probability at least $1 - (2bn+3)e^{-x}$ , a $\rho$-approximate empirical minimizer $\hat{f} \in F$ satisfies*

$$\mathbb{E}\hat{f} \le \max\{2\rho, r''\},$$

*where*

$$r'' = \max\left\{ \bar{r}, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)} \right\},$$

*and $\bar{r} = RSTAR(F, \mathbf{X})$.*

The value RSTAR$(F, \mathbf{X})$ is essentially the fixed point of the empirical averages $\widehat{R}_n\left(\hat{F}_{c_1 r, c_2 r}, \mathbf{X}\right)$ as a function of $r$. Thus, the quantity which matters is the (empirical) complexity of the function class $\hat{F}_{c_1 r, c_2 r}$ . Since this class can be determined empirically by looking at "empirical shells" of $F$ containing functions whose empirical averages fall in an interval of length proportional to $r$, we are thus able to determine the complexity entirely from data.

We can tighten the localization further by narrowing the size of the "shells" and replacing the empirical set $\hat{F}_{\frac{r}{2}, \frac{3r}{2}}$ with $\hat{F}_{r - \frac{r}{\log n}, r + \frac{r}{\log n}}$ . These empirical shells have the advantage that, with growing sample size, they become closer to $F_r$. The price we pay for this advantage is an extra $\log n$ factor in the final estimate, since in this case Talagrand's inequality will allow us to estimate the expectation only down to the order of $O(\log n/n)$.

With the same reasoning as before, by Theorem 5.5 for $\alpha = 1/2, \lambda = 1/\log n$ , and since $F$ is star-shaped, then, if $r \geq \max\left\{r_n^*, \frac{cbx \log n}{n}, c\left(\frac{Bx\log^2 n}{n}\right)^{1/(2-\beta)}\right\}$ , with probability larger than $1 - e^{-x}$ , $F_r \subset \hat{F}_{r-\frac{r}{\log n}, r+\frac{r}{\log n}}$ . We define

$$\widehat{\xi}_n(r) = \left(4\mathbb{E}_{\boldsymbol{\varepsilon}} R_n\left(\hat{F}_{\frac{k}{n}-\frac{k}{n\log n}, \frac{k}{n}+\frac{k}{n\log n}}\right) + \frac{k}{cn\log n}\right)\frac{nr}{k}, \quad \text{if } r \in \left[\frac{k}{n}, \frac{k+1}{n}\right].$$

Again, with probability at least $1 - 2(bn+1)e^{-x}$ , for every $r \in [r', b]$ , it holds that $\xi_n(r) \leq \widehat{\xi}_n(r)$ , where

$$r' = \max\left\{r_n^*, \frac{cbx\log n}{n}, c\left(\frac{Bx\log^2 n}{n}\right)^{1/(2-\beta)}\right\}.$$

Since $\widehat{\xi}_n(r)/r$ is non-increasing, it is possible to compute

$$\hat{r}_n^* = \inf\left\{r : \widehat{\xi}_n(r) \leq \frac{r}{2\log n}\right\}$$

with a slight modification of RSTAR (by replacing the test in the if-clause, $\widehat{\xi}_n(r) > r/4$, with $\widehat{\xi}_n(r) > r/2\log n$ ). Thus, for every $n$, with probability larger than $1 - 2bne^{-x}$ over samples of size $n$, it holds that $r_n^* \leq \bar{r}$ .

## 5.5 Optimality

In this section, we will analyze the optimality of the data-dependent estimate $\hat{r}_n^*$. Since our empirical estimate is based on the quantity $r_n^*$, we will first present a result from Bartlett and Mendelson (2005) which proves that one can obtain an estimate for the expectation of the empirical minimizer which is always better than the one based on $r_n^*$. We will also present an example (based on an example from Bartlett and Mendelson (2005)) showing that this better estimate can lead to significantly better rates for the convergence of the empirical minimizer.

However, as we will show, this improved rate of convergence cannot, in general, be recovered in a data-dependent fashion since it is *in general* impossible to distinguish (up to constants) between $r_n^*$ and $s_n^*$ based solely on empirical data. These results are work in progress contained in an unpublished manuscript Bartlett et al. (2005).

### 5.5.1 Optimal Data-Independent Result

By looking at the localized sets $F_r$, is is easy to see that, for any empirical minimizer, the quantity

$$\sup_{f \in F_r} (\mathbb{E}f - \mathbb{E}_n f) - r = - \inf_{f \in F_r} \mathbb{E}_n f$$

is maximized for the value $r = \mathbb{E}\hat{f}$ . Assume that one would have a very strong concentration of empirical processes indexed by $F_r$, that is, it holds that with high probability over the random draw of samples,

$$\sup_{f \in F_r} (\mathbb{E}f - \mathbb{E}_n f) \;\approx\; \mathbb{E} \sup_{f \in F_r} (\mathbb{E}f - \mathbb{E}_n f) \;=\; \xi'_n(r)\,.$$

In this case, it makes sense to expect that, with high probability, $\mathbb{E}\hat{f} \approx s^*_n$ , where

$$s^*_n = \operatorname{argmax}\{\xi'_n(r) - r\}\,.$$

A result formalizing this idea was proved in Bartlett and Mendelson (2005). It is based on Talagrand's concentration inequality together with the assumption of a Bernstein class, which leads to a very strong concentration of $\sup_{f \in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ around its expectation. From Talagrand's inequality (Theorem 3.12, page 44), $\sup_{f \in F_r}(\mathbb{E}f - \mathbb{E}_n f) \sim \mathbb{E}\sup_{f \in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ , where $\sim$ represents equivalence up to a multiplicative constant $(1+\rho)$. Of course, Talagrand's inequality contains additional terms besides the one involving $\xi'_n(r)$, which blow up as the multiplicative constant $(1+\rho)$ represented by $\sim$ tends to one. Hence, the claim is not so simple and one has to consider an interval containing $s^*_n$ rather than $s^*_n$ itself. Thus, for $\varepsilon > 0$, define

$$r_{n,\varepsilon,+} = \sup\left\{0 \le r \le b : \xi'_n(r) - r \ge \sup_s \left(\xi'_n(s) - s\right) - \varepsilon\right\},$$
$$r_{n,\varepsilon,-} = \inf\left\{0 \le r \le b : \xi'_n(r) - r \ge \sup_s \left(\xi'_n(s) - s\right) - \varepsilon\right\}.$$

The values $r_{n,\varepsilon,+}$ and $r_{n,\varepsilon,-}$ are upper respectively lower approximates for $s^*_n$ . They are close to $s^*_n$ if the function $\xi'_n(r) - r$ is peaked around its maximum.

Now we are ready to state the theorem from Bartlett and Mendelson (2005) which shows that one can directly bound $\mathbb{E}\hat{f}$ for the empirical minimizer. It shows that $\mathbb{E}\hat{f}$ is concentrated around $s^*_n$ and therefore, with high probability, for $\varepsilon$ of the order $\max\left\{\sup_s \left(\xi'_{n,F,\mu}(s) - s\right), r'^{\beta}\right\} \sqrt{\log n/n}$ (and thus in the most conservative case $\varepsilon \sim \sqrt{\log n/n}$ ), the expectation $\mathbb{E}\hat{f} \le r_{n,\varepsilon,+}$ . In addition, if the class is not too "rich" around 0, then with high probability, $\mathbb{E}\hat{f} \ge r_{n,\varepsilon,-}$ .

**Theorem 5.10** *For any $c_1 > 0$, there is a constant $c > 0$ (depending only on $c_1$) such* *that the following holds: Let $F$ be a class of functions defined on a probability space $(\Omega, \mu)$ which is $(\beta, B)$-Bernstein, bounded by $b$, and star-shaped around $0$. For any $x > 0$ and any $n$, define $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$ as above, and set*

$$r' = \max\left\{ r_n'^*, \frac{cb(x+\log n)}{n}, c\left(\frac{B(x+\log n)}{n}\right)^{1/(2-\beta)} \right\}.$$

*Let $\hat{f}$ denote an empirical risk minimizer. If*

$$\varepsilon \geq c\left( \max\left\{ \sup_s \left( \xi'_{n,F,\mu}(s) - s \right), r'^{\beta} \right\} \frac{(B+b)(x+\log n)}{n} \right)^{1/2},$$

*then*

1. *With probability at least $1 - e^{-x}$,*

$$\mathbb{E}\hat{f} \leq \max\left\{ \frac{1}{n}, r_{n,\varepsilon,+} \right\}.$$

2. *If*

$$\mathbb{E}\sup\left\{ \mathbb{E}f - \mathbb{E}_n f : f \in F,\, \mathbb{E}f \leq c_1/n \right\} < \sup_s \left( \xi'_{n,F,\mu}(s) - s \right) - \varepsilon,$$

*then with probability at least $1 - e^{-x}$,*

$$\mathbb{E}\hat{f} \geq r_{n,\varepsilon,-}.$$

The proof is based on a peeling technique (one peels out functions with given expectations) and uses the fact that, in small intervals, $\xi'_n(r)$ does not grow too much since $F$ is star-shaped.

Note that the upper bound $r_{n,\varepsilon,+}$ is an improvement in comparison to the bound resulting from Theorem 5.7, as long as the function $\xi'_n(r) - r$ is not "flat" around its maximizer. (A "flat" $\xi'_n(r) - r$ corresponds to no "significant atoms" appearing at a scale below some $r_0$, and thus, for $r < r_0$, $F_r$ is essentially a scaled down version of $F_{r_0}$; in this case, the two bounds will be of the same order of magnitude.) Figure 5.10 illustrates graphically such a case.

By Lemma 5.3, since $\xi'_n(r)/r$ is non-increasing,

$$\inf\left\{ r : \xi'_n(r) \leq r \right\} \leq \inf\left\{ r : \xi'_n(r) \leq \frac{r}{4} \right\}.$$

Clearly, $\xi'_n(r) \geq 0$, since $\xi'_n(r) \geq \mathbb{E}(\mathbb{E}f - \mathbb{E}_n f) = 0$ for any fixed function $f$, and thus $0 \leq s_n^* \leq \inf\left\{ r : \xi'_n(r) \leq r \right\} \leq r_n'^* \leq r_n^*$. Now, for $\beta = 1$, $\varepsilon \sim \sqrt{s_n^*/n} \ll s_n^*$ and if $\xi'_n(r)$ is sufficiently "peaked" around $s_n^*$, then the upper and lower bounds of $s_n^*$, $r_{n,\varepsilon,+}$

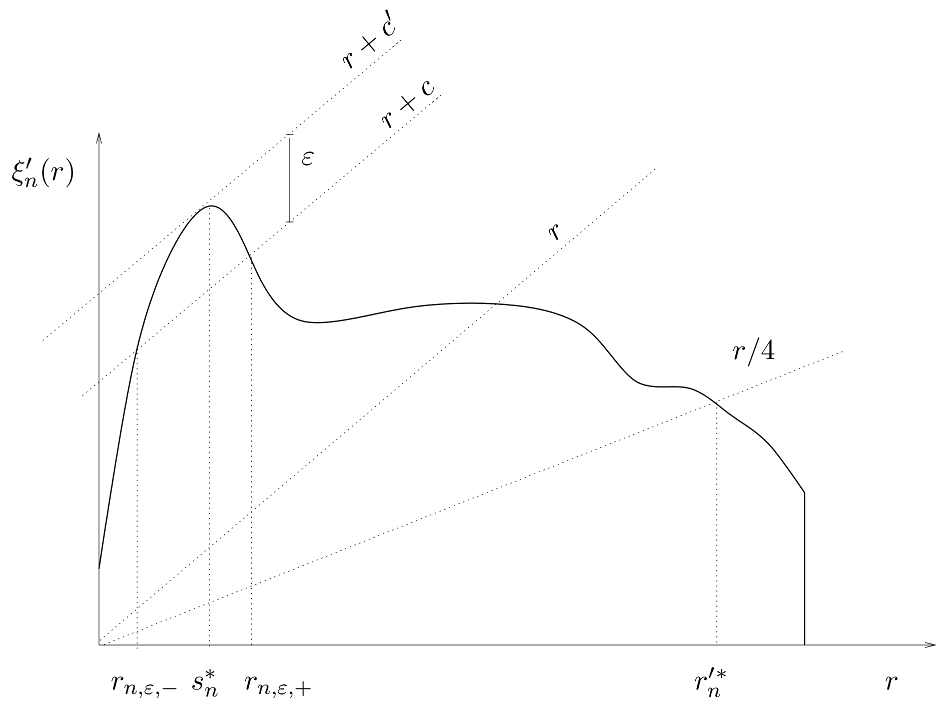


**Figure 5.10:** The graph of a function $\xi'_n$, and the corresponding values for $r'^*_n$, $s^*_n$, $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$. If $s^*_n \ll r'^*_n$ and $\xi'_n(r) - r$ is peaked around $s^*_n$, then $r_{n,\varepsilon,+}$ is smaller than $r'^*_n$.

and $r_{n,\varepsilon,-}$, will be of the order of $s^*_n$.

As we will see in the following, the bound obtained in terms of $s^*_n$ can be a significant improvement over the estimate in terms of $r^*_n$: we will present an example where $s^*_n$ is asymptotically smaller than $r^*_n$.

**Comparison**

In this section we construct a class of functions for which there is a clear gap between the result of Theorem 5.6 and the expectation of the empirical minimizer, and thus between $r'^*_n$ and $s^*_n$. The idea behind the construction is that one can have a complete freedom to choose the expectation of a function, while forcing it to have certain values on a given sample. We can therefore construct a class for which $r'^*_n$ is of the order of a constant (and thus $r^*_n$ is of the order of a constant), but which becomes very rich in subsets which have expectations close to 0 (we choose expectations close to $1/n$, since Talagrand's concentration inequality does not allow an estimate below $1/n$).

The construction is based on the idea developed in Bartlett and Mendelson (2005) of two Bernstein classes of functions satisfying the following for any fixed $n$: The first class contains all functions which vanish on a set of cardinality $n$, but have expectations equal to a given constant. The second class consists of functions which each take its minimal value on a set of cardinality $n$ but have expectations equal to $1/n$. By appropriately choosing the values of the function, one can show that the star-shaped hull of the union of these two classes has $r_n'^* \sim c$, whereas $s_n^* \sim r_{n,\varepsilon,+} \sim 1/n$. Thus, the estimate given by Theorem 5.10 is considerably better than the one resulting from Theorem 5.7. Here, we will show that one can construct a class which satisfies the same property *uniformly for all large sample sizes*, showing that, for any large sample size $n$, an estimate for the empirical minimizer based on $r_n'^*$ is asymptotically not optimal.

We will first present an almost identical construction to the one in Bartlett and Mendelson (2005) of function classes dependent on $n$. The slight modification we make is to define the classes on the interval $(0,1]$ rather than on finite sets, and such that they are constant on intervals of equal length. This modification is necessary in order to take the union over $n$ of all these classes for the final construction.

The following lemma states that, for any given sample size $n$, and for $1/n \leq \lambda \leq 1/2$, we can construct function classes $G_\lambda^n$ and $H_\lambda^n$ defined on $(0,1]$ which are both bounded and Bernstein with respect to $\mu$, and for which $\xi'_{n,H_\lambda^n,\mu}(\lambda) = \lambda$ , $\xi'_{n,G_\lambda^n,\mu}(\lambda) = \lambda + 1$ .

**Lemma 5.11** *Let $\mu$ be the uniform probability measure on $(0,1]$. Then for every $n$ and $1/n \leq \lambda \leq 1/2$ , there exists a function class $G_\lambda^n$ such that*

1. *For every $g \in G_\lambda^n$ , $-1 \leq g(x) \leq 1$ , $\mathbb{E}g = \lambda$ , and $\mathbb{E}g^2 \leq 2\mathbb{E}g$ .*
2. *For every set $\tau \subset (0,1]$ with $|\tau| \leq n$ , there is some $g \in G_\lambda^n$ such that for every $s \in \tau$, $g(s) = -1$.*

*Also, there exists a function class $H_\lambda^n$ such that*

1. *For every $h \in H_\lambda^n$ , $0 \leq h(x) \leq 1$ , $\mathbb{E}h = \lambda$ , and $\mathbb{E}h^2 \leq \mathbb{E}h$ .*
2. *For every set $\tau \subset (0,1]$ with $|\tau| \leq n$ , there is some $h \in H_\lambda^n$ such that for every $s \in \tau$, $h(s) = 0$.*

The proof of this lemma is almost identical to the proof in Bartlett and Mendelson (2005), and can be found in Appendix B.2.

Using the classes defined in Lemma 5.11, set

$$H = \bigcup_{k=5}^{\infty} H_{1/4}^k , \qquad F_k = G_{1/k}^k , \qquad G = \bigcup_{k=5}^{\infty} F_k ,$$

and

$$F = \operatorname{star}(G \cup H, 0) . \tag{5.2}$$

For all $h \in H$, $h : (0,1] \longrightarrow [0,1]$, $\mathbb{E}h = 1/4$, and $H$ is a (1,1)-Bernstein class w.r.t. $\mu$, since all functions are positive and bounded by 1. $G$ is a (1,2)-Bernstein class w.r.t. $\mu$, since all functions satisfy the Bernstein condition. Thus, $F$ is star-shaped and $(1,2)$-Bernstein with respect to $\mu$ and therefore satisfies the assumptions of Theorems

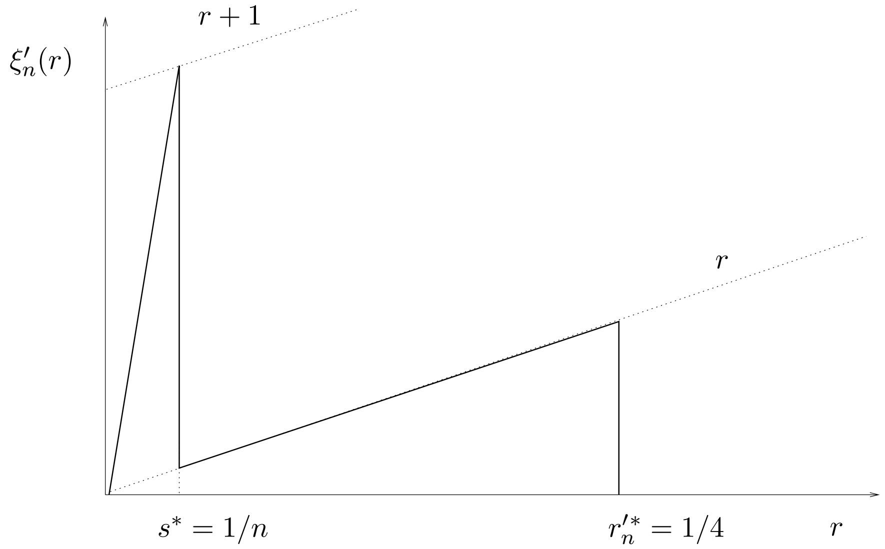


**Figure 5.11**: $\xi'_{n,\mathrm{star}(F_n \cup H^n_{1/4}),\mu}$ (as in the proof of Theorem 5.12).

5.7 and 5.10. We are now ready to show that, for $F$, the estimate given by Theorem 5.10 is asymptotically better than the one resulting from Theorem 5.7 *uniformly* for every $n \geq n_0$.

First observe that, for every $n \geq 5$ and any sample $(X_1, \ldots, X_n)$ drawn i.i.d. according to $\mu$, there is a function $f \in F$ with $\mathbb{E}f = 1/4$ and $\mathbb{E}_n f = 0$, and a function $g \in F$ with $\mathbb{E}g = 1/n$ and $\mathbb{E}_n g = -1$. Indeed, one can choose the function $f$ from $H^n_{1/4}$ and the function $g$ from $F_n = G^n_{1/n}$. Thus, for any sample of size $n \geq 5$, the localized complexity function $\xi'_n$ for the class $\mathrm{star}(F_n \cup H^n_{1/4})$ has its graph as in Figure 5.11, with $r'^*_n = 1/4$ and $s^*_n = r_{n,\varepsilon,+} = 1/n$ .

This also implies that $\xi'_{n,F_n,\mu}(1/n) = 1 + 1/n$ , since this is the maximal possible. We now show that $\xi'_{n,F_k,\mu}(1/k)$ decays rapidly in $k$, ensuring that for any $n \geq n_0$ , and every $k \leq cn$ , it holds that $\xi'_{n,F_n,\mu}(1/n) - 1/n \gg \xi'_{n,F_k,\mu}(1/k) - 1/k$ . Thus, the empirical minimizer in $\mathrm{star}(\cup_k F_k)$ is likely to be around $1/n$. Therefore, for the class $F$, $\xi'_{n,F,\mu}(r) - r$ will still achieve its maximum at $1/n$ and will decay rapidly for $r > 1/n$, ensuring that $r_{\varepsilon_n,+} \ll r'^*_n$ . Figure 5.12 illustrates the qualitative behaviour of $\xi'_{n,F,\mu}$ .

**Theorem 5.12** *For $F$ defined as above, the following holds:*

1. *For every $n \geq 5$, the function $\xi'_{n,F,\mu}$ satisfies*

$$\xi'_{n,F,\mu}(r) = \begin{cases} r + rk & \textit{if } r \in (1/(k+1), 1/k]\,, \ \textit{where } k \leq n \\ r & \textit{if } r \in (1/5, 1/4] \\ 0 & \textit{if } r > 1/4\,. \end{cases}$$

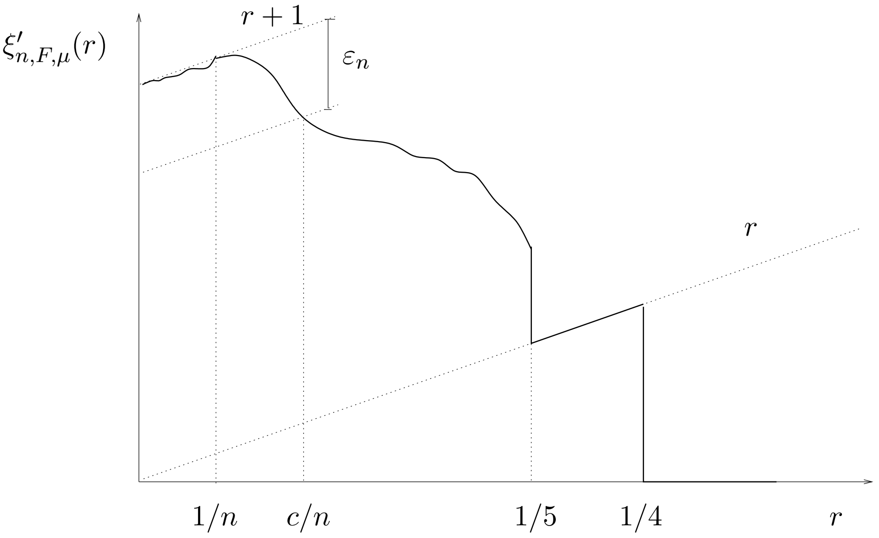


**Figure 5.12**: Qualitative behaviour of $\xi'_{n,F,\mu}$, where $F$ is defined by equation (5.2).

*In particular,* $r'^*_n = 1/4$.

*2. There exists an absolute constant* $c > 1$, *such that the following holds: for every* $x > 0$, *there exists an* $N(x)$ *such that for every* $n \geq N(x)$, *for every* $k \leq n/c$, *if* $\varepsilon_n = \sqrt{\frac{3(x+\log n)}{n}}$, *then*

$$\xi'_{n,F,\mu}(1/k) - 1/k \leq \xi'_{n,F,\mu}(1/n) - 1/n - \varepsilon_n \,.$$

*In particular,* $r_{n,\varepsilon_n,+} \leq c/n$.

Claim 1 follows directly from the linearity of the expectation and the construction of $F$. The proof of claim 2, namely that $\xi'_{n,F,\mu}(r) - r$ decays fast around $1/n$ at larger values than $r \sim c/n$, can be found in Appendix B.2.

The following direct corollary of this theorem states that, for the class $F$, for any sample size $n$, $r'^*_n = 1/4$, while the empirical minimizer is likely to be smaller than $c/n$.

**Corollary 5.13** *For* $F$ *defined as above, there is an absolute constant* $c > 0$ *for which the following holds: For any* $x > 0$ *there is an integer* $N(x)$ *such that for any* $n \geq N(x)$,

*1. With probability at least* $1 - e^{-x}$, $\mathbb{E}\hat{f} \leq c/n \sim s^*_n$ .

*2.* $r'^*_n = r^*_n = 1/4$ .

Hence, we have constructed a class $F$ which is star-shaped and Bernstein, and for which the estimate from Theorem 5.10 is asymptotically better than the one implied by Theorem 5.7. This shows that $r'^*_n$ (and even more $r^*_n$) is not an optimal estimate for $\mathbb{E}\hat{f}$, since we constructed an example of a function class for which it does not capture the correct estimate for most empirical minimizers.

However, as we will show in the next section, any *data-dependent* estimate for $\mathbb{E}\hat{f}$ based entirely on empirical data cannot be in general better than $r_n^*$.

### 5.5.2 Optimality of Data-Dependent Estimation

In this section, we will show that, in general, it is impossible to establish a data-dependent estimate of $s_n^*$ which is better than $r_n'^*$. The general idea is to construct two Bernstein classes of functions, such that the classes have asymptotically different $r_n^*$ but look identical when projected on any sample of finite size. We start by constructing two classes of functions which have identical coordinate projections on any sample of fixed size $n$. We construct them in such a way that all functions in the first class have expectations equal to some absolute constant, whereas all functions in the second class have an expectation equal to $c/n$ for an absolute constant $c$.

In the following, for a given function class $F$ and a sample $\mathbf{x} = (x_1, \ldots, x_n)$, recall that we denote the set of all coordinate projections of $F$ on $\mathbf{x}$ by

$$F/\mathbf{x} = \{(f(x_1), ..., f(x_n)) : f \in F\} .$$

In order to illustrate the idea of the construction, we start with the following theorem, which shows that we can construct, for each sample size $n$, two classes of functions as follows:

**Theorem 5.14** *There exists an integer $n_0$ and an absolute constant $c > 0$ such that, for any $n \geq n_0$, there exist two function classes $G_1^n$ and $G_2^n$ defined on a probability space $(\Omega, \mu_n)$ satisfying the following properties:*

1. *for every $f \in G_1^n$ , $\mathbb{E}f \geq 1/4$ and $\mathbb{E}f^2 \leq 2\mathbb{E}f$ ,*
2. *for every $f \in G_2^n$ , $\mathbb{E}f \leq c/n$ and $\mathbb{E}f^2 \leq 2\mathbb{E}f$ ,*
3. *for any sample $\mathbf{x} \subset \Omega^n$ with $|\{x_1, \ldots, x_n\}| = n$ , $G_1^n/\mathbf{x} = G_2^n/\mathbf{x}$ .*

**Proof:** Let $n$ be a fixed integer. Consider the following two function classes $G_1^n$ and $G_2^n$ consisting of functions defined on $\{1, ..., 4n^2\}$ taking values in $\{-1, 1, 1/n\}$ . Any function in $G_1^n$ takes the value $-1$ on exactly $n$ points, the value 1 on exactly $n^2$ points, and the value $1/n$ on $3n^2 - n$ points. Any function in $G_2^n$ takes the value $-1$ also on exactly $n$ points; however it takes the value 1 on only exactly $2n$ points, and thus the value $1/n$ on $4n^2 - 3n$ points. Hence, for any two disjoint subsets $J, I \subset \{1, ..., 4n^2\}$, $|J| = n$, $|I| = n^2$ let $f = f_{I,J} \in G_1^n$ be such that

$$f(i) = \begin{cases} -1, & \text{if } i \in J, \\ 1, & \text{if } i \in I, \\ 1/n, & \text{otherwise,} \end{cases}$$

and for any two disjoint subsets $J, I \subset \{1, ..., 4n^2\}$, $|J| = n$, $|I| = 2n$ let $f = f_{I,J} \in G_2^n$ be such that

$$f(i) = \begin{cases} -1, & \text{if } i \in J, \\ 1, & \text{if } i \in I, \\ 1/n, & \text{otherwise.} \end{cases}$$

Let $\mu_n$ be the uniform distribution on $\{1, ..., 4n^2\}$. Clearly, with respect to $\mu_n$, if $n \geq 3$, then for every $f \in G_1^n$,

$$\mathbb{E}f = \frac{-n + n^2 + (3n^2 - n)/n}{4n^2} = \frac{n^2 + 2n - 1}{4n^2} > \frac{1}{4},$$
$$\mathbb{E}f^2 = \frac{n + n^2 + (3n^2 - n)/n^2}{4n^2} = \frac{n^3 + n^2 + 3n - 1}{4n^3} \leq \frac{1}{2}.$$

Thus $\mathbb{E}f^2 \leq 2\mathbb{E}f$, which proves claim 1. On the other hand, for every $f \in G_2^n$, if $n \geq 20$,

$$\mathbb{E}f = \frac{-n + 2n + (4n^2 - 3n)/n}{4n^2} = \frac{5n - 3}{4n^2} < \frac{5}{4n},$$
$$\mathbb{E}f^2 = \frac{n + 2n + (4n^2 - 3n)/n^2}{4n^2} = \frac{3n^2 + 4n - 3}{4n^3} \leq \frac{4}{5n}.$$

It is easy to check that $G_2^n$ is a $(1, 2)$-Bernstein class of functions, since there is an $n_0$ such that if $n \geq n_0$, then for any $f \in G_2^n$, $\mathbb{E}f \geq 2/5n$ and hence $\mathbb{E}f^2 \leq 2\mathbb{E}f$. Claim 2 of the theorem is thus true for $c = 5/4$.

By construction, for any sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of size $n$ the projections

$$G_1^n/\mathbf{x} = G_2^n/\mathbf{x} = \left\{-1, 1, \frac{1}{n}\right\}^n$$

and claim 3 of the theorem follows. ■

By using the same idea with a more complicated construction, we will define two classes which have asymptotically different expectations, but have identical projections uniformly on samples *of any size.* We will take two unions of classes constructed similarly to the ones constructed in Theorem 5.14. For each sample size $n$, we will construct two function classes such that functions in the first class will have expectations of the order of a constant, while the ones in the second class will only have expectations of the order $c/n$. Since we will take the union over all such functions, we will construct the classes in a way that, for each sample size $n$, *the union* of classes with small expectations (the ones in the second class) is only "rich" at values $\mathbb{E}f \leq c/n$, implying that $r_{n,\varepsilon,+} \leq c/n$ similarly to Theorem 5.12. This can be done by ensuring that the complexity (i.e. the fat-shattering dimension) of these classes is "small". Since the two constructed union of classes look identical on any sample, this proves that it is not possible in general to distinguish empirically between $r_{n,\varepsilon,+}$ and $r'^*_n$ if all the information one has are the coordinate projections.

**Lemma 5.15** *Let $\mu$ be the uniform measure on $(0,1]$. Then, there exists an integer $n_0$ and absolute constants $c, c_1, c_2, c' > 0$ such that, for any $n \geq n_0$, there exist two function classes $F_1^n$ and $F_2^n$ defined on $(0,1]$ satisfying the following properties:*

1. *for every $f \in F_1^n$ , $\mathbb{E}f \geq 1/4$ and $\mathbb{E}f^2 \leq c_1 \mathbb{E}f$ ,*
2. *for every $f \in F_2^n$ , $\mathbb{E}f \leq c/n$ and $\mathbb{E}f^2 \leq c_2 \mathbb{E}f$ ,*
3. *for any set $\mathbf{x} \subset \Omega^n$ with $|\{x_1, \ldots, x_n\}| = n$ , $F_1^n/\mathbf{x} = F_2^n/\mathbf{x}$ ,*
4. *For any $\varepsilon > 0$ and $n \geq 2$, $\mathrm{fat}_\epsilon(F_2^n) \leq c'n/\varepsilon$ .*

**Proof:** Fix an integer $n$. We construct two function classes $F_1^n$ and $F_2^n$ on $(0,1]$ and taking values in the set

$$V_n = \left\{-1, \frac{1}{n^2}, \frac{2}{n^2}, \ldots, \frac{1}{n}, 1\right\} . \tag{5.3}$$

In both $F_1^n$ and $F_2^n$, each function is defined to be constant on each of the intervals $(j/m, (j+1)/m]$, where $m = 2n^2 + 3n$, $0 \leq j \leq m-1$. Any function in $F_1^n$ takes the value $-1$ on $n$ intervals, the value 1 on $n^2 + 2n$ intervals, and, for every $1 \leq i \leq n$ the value $i/n^2$ on $n$ intervals. Any function in $F_2^n$ takes the value $-1$ on $n$ intervals, the value 1 on $2n$ intervals, and, for every $1 \leq i \leq n$, the value $i/n^2$ on $2n$ intervals. Therefore, for any function $f \in F_1^n$,

$$\mathbb{E}f = \frac{-n + n^2 + 2n + n\sum_{i=1}^n \frac{i}{n^2}}{2n^2 + 3n} = \frac{2n^2 + 3n + 1}{4n^2 + 6n} \geq \frac{1}{4},$$

$$\mathbb{E}f^2 = \frac{n + n^2 + 2n + n\sum_{i=1}^n \frac{i^2}{n^4}}{2n^2 + 3n} = \frac{6n^4 + 18n^3 + 2n^2 + 3n + 1}{6n^2(2n^2 + 3n)} \leq 1,$$

and hence

$$\mathbb{E}f^2 \leq 4\mathbb{E}f.$$

While $F_1^n$, contains functions with expectation of the order of a constant, in $F_2^n$ all the functions have expectation $s_n \sim 1/n$. Indeed, for any function $f \in F_2^n$,

$$\mathbb{E}f = \frac{-n + 2n + 2n\sum_{i=1}^n \frac{i}{n^2}}{2n^2 + 3n} = \frac{2n + 1}{2n^2 + 3n} \leq \frac{1}{n},$$

$$\mathbb{E}f^2 = \frac{n + 2n + 2n\sum_{i=1}^n \frac{i^2}{n^4}}{2n^2 + 3n} = \frac{9n^3 + 2n^2 + 3n + 1}{3n^2(2n^2 + 3n)} \leq \frac{3}{2n}.$$

Moreover, since $\mathbb{E}f \geq 2/5n$, $\mathbb{E}f^2 \leq 5\mathbb{E}f$.

As before, it is easy to see that for any sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$,

$$F_1^n/\mathbf{x} = F_2^n/\mathbf{x} = \left\{-1, \frac{1}{n^2}, \frac{2}{n^2}, \ldots, \frac{1}{n}, 1\right\}^n .$$

We will now show that, indeed, the fat-shattering dimension of $F_2^n$ is "small". Clearly, for $1/n \leq \varepsilon \leq 1$, $\text{fat}_\epsilon\left(F_2^n\right) \leq n$. Fix $0 < \varepsilon < 1/n$. If a set $\{x_1, \ldots, x_k\}$ is $\epsilon$-shattered by $F_2^n$ (recall Definition 2.11, page 24), then there exists some $\mathbf{s} \in [-1,1]^k$ and $f \in F_2^n$ such that

$$\left\lfloor \frac{k}{2} \right\rfloor = \min\left\{\, |\{\, i \,:\, f(x_i) \in [-1, s_i - \varepsilon] \,\}|,\; |\{\, i \,:\, f(x_i) \in [s_i + \varepsilon, 1] \,\}| \,\right\}$$

(we choose $I$ such that half of the points are "above" and half "below" $\mathbf{s}$). Since the fat-shattering dimension of $F_2^n$ is the maximal $k$ for which there is a set $\{x_1, \ldots, x_k\}$ which is $\epsilon$-shattered by $F_2^n$, it follows that

$$\left\lfloor \frac{\text{fat}_\epsilon\left(F_2^n\right)}{2} \right\rfloor \leq \max_k \max_{\mathbf{s}\in[-1,1]^k} \min\left\{ |\{i \,:\, f(x_i) \in [-1, s_i-\varepsilon]\}|, |\{i \,:\, f(x_i) \in [s_i+\varepsilon, 1]\}| \right\}.$$

The maximum is achieved by taking the fixed level values $s_i = (n+1)/2n^2$, and since each interval of length $\varepsilon$ above or below this level contains at least $\varepsilon n^2 - 1$ points from $V_n$ (defined in (5.3)),

$$\left\lfloor \frac{\text{fat}_\epsilon\left(F_2^n\right)}{2} \right\rfloor \leq \big((n+1)/2 + 1 - (\varepsilon n^2 - 1)\big)\, n \leq (n + 3 - \varepsilon n^2) n\,.$$

To show that $n + 3 - \varepsilon n^2 \leq 1/\varepsilon$, which is equivalent to $-\varepsilon^2 n^2 + \varepsilon(n+3) - 1 \leq 0$, note that for any $n \geq 2$, the quadratic function $h(\varepsilon) = -\varepsilon^2 n^2 + \varepsilon(n+3) - 1 \leq 0$. This directly implies that $\text{fat}_\epsilon\left(F_2^n\right) \leq c' n/\varepsilon$ and the theorem is proved. ■

We are now ready to construct the desired classes. First, consider the set of integers $M = \{k \,:\, \exists \ell, k^2 = 2^\ell\}$. Note that if $k_1, k_2 \in M$ and $k_1 < k_2$, then $V_{k_1} \subset V_{k_2}$, where the sets $V_{k_i}$ are defined in (5.3) by setting $n = k_i$.

Set

$$F_1 = \text{star}\left(\bigcup_{k\in M} F_1^k, 0\right), \quad F_2 = \text{star}\left(\bigcup_{k\in M} F_2^k, 0\right). \tag{5.4}$$

For every sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, if $k \in M$ such that $k \geq n$, then $F_1^k/\mathbf{x} = F_2^k/\mathbf{x} = V_k \times \ldots \times V_k \subset \mathbb{R}^n$. Now the choice of $M$ becomes clear, as it ensures that for each $k' \in M$, $k' < n$, $V_{k'} \subset \bigcup_{k\in M, k\geq n} V_k$, and thus

$$F_1/\mathbf{x} = F_2/\mathbf{x} = \left(\bigcup_{k\in M, k\geq n} V_k\right)^n.$$

Hence, $F_1$ and $F_2$ are star-shaped, Bernstein classes of functions which have identical coordinate projections on any finite set. Hence, it is impossible to distinguish between these two classes based solely on empirical data. However, the expectation of the empirical minimizer is very different for these two classes, as stated in the following theorem, whose proof can be found in Appendix B.2.

**Theorem 5.16** *For $F_1$ and $F_2$ defined as in* (5.4)*, there is an absolute constant $c > 0$ for which the following holds: For any $x > 0$ there is an integer $N(x)$ such that for any $n \geq N(x)$,*

1. *for $F_1$, with probability at least $1 - e^{-x}$, $\mathbb{E}\hat{f} \geq 1/4 \sim r_n^*(F_1)$ ;*
2. *for $F_2$, with probability at least $1 - e^{-x}$, $\mathbb{E}\hat{f} \leq c/n \sim s_n^*(F_2)$ .*

Thus, the estimates for the convergence rate of the ERM algorithm based on $s_n^*$ are significantly better for the class $F_2$ than for $F_1$. However, the classes have identical coordinate projections on any sample, and hence are indistiguishable empirically. Thus, one can not get a better *empirical* estimate for the convergence rate for $F_2$ than by an empirical estimate for $r_n^*$.

## 5.6 Conclusion

In this section we investigated results regarding the extent to which one can derive *empirical* estimates for the generalization performance for the ERM algorithm. Our investigations are based on two recent results for the convergence rate of the ERM algorithm from Bartlett and Mendelson (2005). These two results use localization with shells $F_r$ of the excess loss function class of fixed expectations and localized complexity functions $\xi_{n,F,\mu}(r) = \mathbb{E}\sup\{|\mathbb{E}f - \mathbb{E}_n f| : f \in F_r\}$ and $\xi'_{n,F,\mu}(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F_r\}$ . The "optimal" result in Bartlett and Mendelson (2005) provides upper and lower bounds on the expectation of empirical minimizers in terms of the largest maximizer $s_n^*$ of $\xi'_n(r) - r$. Since this result is based on a direct analysis of the expectation of the empirical minimizer and does not take the detour of analyzing uniform deviations of expectations and empirical averages, it is not known if it can be recovered within the random subclass framework which we presented in Section 4. Bartlett and Mendelson (2005) also prove a "comparison" result which is based on uniform relative deviations which provides upper bounds on the expectation of empirical minimizers in terms of the fixed points $r_n^*$ and $r_n'^*$ of the functions $\xi_{n,F,\mu}(r)$ and $\xi'_{n,F,\mu}(r)$. For "well-behaved" classes, $s_n^* \leq r_n'^* \leq r_n^*$ .

We first presented a *data-dependent* upper bound on the expectation of the empirical minimizer produced by the ERM algorithm. This estimate, which is entirely computable from empirical data, is the tightest data-dependent estimate based on localized complexity notions which we are aware of. It is based on the distribution-dependent $r_n^*$ "comparison" result in Bartlett and Mendelson (2005) for which it is an upper bound.

We extended an example from Bartlett and Mendelson (2005) to show that there exist classes for which the estimates based on $s_n^*$ can lead to significantly better rates of convergence than the ones based on $r_n'^*$ and thus on $r_n^*$.

Finally, we presented an example showing that, in general, this potentially better estimate based on $s_n^*$ cannot be recovered solely based on empirical data. We presented an example of two function classes for which the $s_n^*$ estimate would lead to asymptotically different convergence rates but which look identical projected on any sample of empirical data. For these classes, one cannot get better empirical estimates than estimates of $r_n^*$ or $r_n'^*$.

Chapter 6

# Conclusion

This thesis studied the problem of bounding the performance of machine learning algorithms in a statistical setting. It presented two contributions concerned with the data-dependent analysis of the generalization error of learning algorithms.

We developed first a general framework for deriving generalization bounds for data-dependent random subclasses of functions based on the comparison of empirical averages and expectations of functions in these random classes. Our approach was motivated by the fact that one is interested in bounds for the function which a particular learning algorithm produces from the actual data, whereas the standard worst-case uniform bounds in learning theory hold simultaneously for any function in the hypothesis class. Such a function produced from the actual random data is a random function.

We showed, based on symmetrization techniques, that the analysis of deviations of empirical averages from expectations of random functions can be reduced to the analysis of the behaviour of the supremum of a Rademacher process indexed by certain random coordinate projections. We identified two separate principles which are sufficient to guarantee and quantify the generalization ability of algorithms producing functions based on a random sample:

1. *Learnability* and *the convergence rate* are ensured by small Rademacher averages of certain symmetric subsets.

2. *High-probability convergence rates* and thus *small confidence intervals* for the generalization error are ensured by the degree of concentration of the suprema of Rademacher processes indexed by these symmetric subsets.

We showed that geometric properties of the random coordinate projections of these random subsets directly influence the degree of concentration and thus the confidence estimates for the generalization error.

We then demonstrated the generality of our approach by showing that the standard uniform approaches based on complexities which characterize Glivenko-Cantelli classes, and the compression, sparsity, and luckiness frameworks all fall into our random subclass framework. It was known before that some of the uniform complexity measures and the assumptions in the above data-dependent and algorithm-dependent frameworks are intrinsically similar though the nature of this similarity was not clear. We showed here that the underlying mechanism which makes them work is the fact that these assumptions are all different ways of ensuring that a "typical" coordinate projection is small. We were able to improve the bounds given in the above mentioned frameworks by avoiding the potentially loose union bound. Our approach related the complexity notions in these frameworks directly to the Rademacher complexity measures.

In addition, we showed that the faster convergence rates for the Empirical Risk Minimization algorithm (ERM) in the learning sample complexity results due to Mendelson (2003) and in results based on local (non-random) subclasses of functions with small variance can be recovered from our framework by using the full power of concentration results established via a strong control on the variance. Although we did not present here any new bounds for the generalization error, the potential of our random subclass framework is that it opens the avenue to exploit information on the variance of functions in order to derive, in a similar fashion, faster high-probability convergence rates for *random* classes of functions.

We conclude that, in order to determine the convergence rates for the error of learning algorithms *based on comparisons of empirical and expected errors*, a key quantity is the Rademacher complexity of random coordinate projections, whereas a key property for determining confidence interval estimates for this error is the $\ell_2^n$ geometry of these coordinate projections. However, given the aim to determine the real convergence rate for learning algorithms, more refined methods of analysis might be needed which are not necessarily based on comparisons of empirical and expected errors.

We gave one example of such a more refined analysis for the estimation of the generalization performance of the ERM, which is a central algorithm in statistical learning theory. We investigated new results on the generalization performance of the ERM algorithm due to Bartlett and Mendelson (2005) which are not based on the comparison of empirical averages and expectations of functions since they directly bound the expectation of the empirical minimizer. These direct estimates based on a new localized notion of complexity of subsets of hypothesis functions with identical expected errors give, under certain circumstances, matching upper and lower bounds and thus essentially optimal estimates for the convergence rates of empirical minimizers.

We studied the extent to which one can obtain empirical versions of these direct estimates with the same "optimal" convergence rates. We first presented an algorithm which computes a data-dependent upper bound for the expected error of empirical minimizers in terms of the complexity of data-dependent local subsets. Although the computation of these complexities can be potentially very expensive, they can be determined based solely on empirical data. This approach improves previous data-dependent results based on localized estimates. We then constructed a counter-example which shows that the direct estimate in Bartlett and Mendelson (2005) can not be recovered universally from empirical data.

These results deepen our understanding of the possibilities and limitations in estimating the generalization performance of the ERM. Although, for well-behaved classes, it is now understood which parameters and structural properties completely characterize the high-probability convergence rates for the expected error of empirical minimizers, we conclude that there are inherent limits in quantifying these convergence rates from observed empirical data universally, without further assumptions. Some questions that are left unanswered are whether our counter-example is "typical" and thus relevant in practical situations, whether it is possible to characterize problems for which one has such a "gap" between the true and the empirical estimates, and whether similar results can be derived for other algorithms than ERM.

Like most results in statistical learning theory, although our results hold true for training samples of finite size and thus are non-asymptotic in this sense, they are useful only for sufficiently large sample sizes since they are based on concentration of measure which is a high-dimensional phenomenon. From the practical point of view, as long as the sample size is sufficiently large, our results are a step towards understanding some mechanisms and parameters which are essential for obtaining probabilistic error bounds for learning algorithms. However, for sample sizes which are significantly smaller than the ones required for our results, the parameters and insights can be misleading and our conclusions are thus not valid in such cases.

Appendix A

# Empirical Processes

Let $V$ be a subset of $\mathbb{R}^n$. We consider in the following stochastic processes which are collections of centered random variables $\{X_{\mathbf{v}} : \mathbf{v} \in V\}$ indexed by $V$. In particular, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ be a vector of independent Rademacher variables (i.e. $Pr\left\{\varepsilon_i = -1\right\} = Pr\left\{\varepsilon_i = 1\right\} = 1/2$ ), and let $\mathbf{g} = (g_1, \ldots, g_n)$ be a vector of independent standard Gaussian random variables. Both these vectors can be seen as models for a random noise sequence of length $n$. The stochastic processes which quantify the correlations of vectors in $V$ with the random noise sequences $\boldsymbol{\varepsilon}$ and $\mathbf{g}$,

$$\left\{ Y_{\mathbf{v}} = \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| \; : \; \mathbf{v} \in V \right\} \quad \text{and} \quad \left\{ Z_{\mathbf{v}} = \left| \sum_{i=1}^{n} g_i v_i \right| \; : \; \mathbf{v} \in V \right\}$$

are the *Rademacher (Gaussian) process indexed by the set* $V$. One of the central questions of empirical process theory is to find upper and lower bounds for the quantity $\mathbb{E} \sup_{\mathbf{v} \in V} X_{\mathbf{v}}$, where the random variable $\sup_{\mathbf{v} \in V} X_{\mathbf{v}}$ is the *supremum of the stochastic process indexed by* $V$. We call $R_n(V) = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v} \in V} Y_{\mathbf{v}}$ the *Rademacher averages* and $G_n(V) = \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{v} \in V} Z_{\mathbf{v}}$ the *Gaussian averages associated with* $V$.

It turns out that for Gaussian processes one can study the quantity $\mathbb{E} \sup_{\mathbf{v} \in V} Z_{\mathbf{v}}$ by looking at the metric space $(V, \ell_2^n)$ and vice-versa. The crucial property of Gaussian processes which allows one to take advantage of the structure of $(V, \ell_2^n)$ is the fact that

$$Pr\left\{|Z_{\mathbf{v}} - Z_{\mathbf{u}}| > t\right\} \leq 2e^{\frac{-t^2}{2\|\mathbf{v}-\mathbf{u}\|_2^2}} .$$

A stochastic process is called *sub-Gaussian with respect to a pseudo-metric* $d$ *on* $V$, if for any $\mathbf{v}, \mathbf{u} \in V$, and every $x > 0$

$$Pr\left\{|X_{\mathbf{v}} - X_{\mathbf{u}}| > t\right\} \leq 2e^{\frac{-t^2}{2d^2(\mathbf{v},\mathbf{u})}} .$$

Gaussian processes are therefore sub-Gaussian with respect to $\ell_2^n$ on $V$. By Hoeffding's inequality (Theorem 3.3, page 37), it follows immediately that Rademacher processes are also sub-Gaussian with respect to the Euclidean metric. One can extend many results which hold for Gaussian processes to the sub-Gaussian case and therefore to Rademacher processes. In particular, majorizing measures, developed by Fernique and Talagrand, which are a more general form of the chaining method due to Kolmogorov, can be employed to provide optimal upper and lower bounds for the expectation of suprema of sub-Gaussian stochastic processes (Fernique 1975; Talagrand 1996a, 2005).

In the following we will state some basic theorems from empirical process theory regarding Rademacher and Gaussian processes which will be used in this thesis. [1]

The following theorem due to Maurey and Pisier (1976) shows that Rademacher and Gaussian averages are similar. A proof can be also found, for example, in Ledoux and Talagrand (1991), on page 97.

**Theorem A.1 (Comparison of Rademacher and Gaussian averages)** *There are absolute constants $c$ and $C$ such that for every set $V \subset \mathbb{R}^n$, $n \geq 2$,*

$$c\,\mathbb{E}_{\boldsymbol{\varepsilon}} \left( \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| \right) \leq \mathbb{E}_{\mathbf{g}} \left( \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} g_i v_i \right| \right) \leq C\,\mathbb{E}_{\boldsymbol{\varepsilon}} \left( \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| \right) \log n \,.$$

A useful property of Gaussian and Rademacher averages is that a contraction applied to each coordinate of $V$ does not change the averages by much.

**Theorem A.2 (Contraction principle (Ledoux and Talagrand 1991, page 95))** *Let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ be a Lipschitz function with Lipschitz constant $L_\phi$, such that $\phi(0) = 0$ . Then,*

$$\mathbb{E}_{\mathbf{g}} \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} g_i \phi(v_i) \right| \leq 2L_\phi \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} g_i v_i \right| ,$$

*and the same holds for the Rademacher averages of $V$.*

The following theorem reflects a relationship between Rademacher averages and the metric $\ell_2^n$-entropy and was originally proved in Dudley (1967) for Gaussian processes. The most general version is due to Pisier (1989). The following version is from van der Vaart and Wellner (1996) (Corollary 2.2.8, page 101), it holds for sub-Gaussian processes and thus applies to Rademacher processes.

**Theorem A.3 (Dudley's entropy integral)** *Let $V$ be a bounded subset of $\ell_2^n$. Then there exists a constant $C$ such that*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left( \sup_{\mathbf{v}\in V} \left| \sum_{i=1}^{n} \varepsilon_i v_i \right| \right) \leq C \int_0^\infty \sqrt{\log N(u, V, \ell_2^n)}\, du \,.$$

[1] For more details on suprema of empirical processes, see, for example, Dudley (1984); Ledoux and Talagrand (1991); van der Vaart and Wellner (1996); Fernique (1997); Dudley (1999); van de Geer (2000); Ledoux (2001); Bousquet (2002b); Massart (2003); Wellner (2004).

Clearly, the upper limit of integration can be taken to be the $\ell_2^n$ radius of a ball centered at the origin which contains the set.

For finite subsets $V \subseteq B_2^n$, since $N(u, V, \ell_2^n) \leq |V|$ for any $u > 0$, one can derive the following corollary. It states that for every finite set $V \subset \ell_2^n$, the Gaussian averages (and therefore the Rademacher averages) can be upper bounded by a quantity which depends on the cardinality of $V$ and the $\ell_2^n$-diameter of $V$.

**Corollary A.4 (Comparison theorem for finite sets)** *There is an absolute constant $C$ such that for every finite set $V \subset \ell_2^n$,*

$$\mathbb{E}_{\mathbf{g}} \sup_{\mathbf{v} \in V} \left| \sum_{i=1}^{n} g_i v_i \right| \leq C \sqrt{\log |V|} \sup_{\mathbf{v} \in V} \|\mathbf{v}\|_2,$$

*where $(g_i)_{i=1}^n$ are independent standard Gaussian random variables.*

The "converse" of Dudley's entropy integral was originally proved by Sudakov, and provides a lower bound for Gaussian averages in terms of the metric $\ell_2^n$-entropy. It was also extended to Rademacher processes, a case in which the $\ell_2^n$ is replaced by an "intermediate" between the $\ell_2^n$ and $\ell_1^n$ norms (see, e.g., Ledoux and Talagrand 1991, Proposition 4.15, page 117). Here we only state the result for Gaussian processes.

**Theorem A.5 (Sudakov's minoration (Ledoux and Talagrand 1991, Theorem 3.18, page 80))** *There exists a constant $c > 0$ such that, for every bounded set $V$ of $\ell_2^n$,*

$$\sup_{u > 0} u \sqrt{\log N(u, V, \ell_2^n)} \leq c\, \mathbb{E}_{\mathbf{g}} \left( \sup_{\mathbf{v} \in V} \left| \sum_{i=1}^{n} g_i v_i \right| \right).$$

Let $F$ be a class of measurable functions defined on a space $\Omega$ with underlying probability measure $\mu$. Then

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i f(X_i) \;:\; f \in F \right\}$$

is a Gaussian process with respect to $L_2(\mu_n)$, and by Hoeffding's inequality, the process

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \;:\; f \in F \right\}$$

is sub-Gaussian with respect to $L_2(\mu_n)$ (recall that $\|f\|_{L_2(\mu_n)}^2 = 1/n \sum_{i=1}^n f^2(X_i)$ ) (see, e.g. Mendelson 2002c; Wellner 2004). This will enable us to use the theorems of this chapter for Rademacher averages of classes of functions.

Another useful fact is that the suprema of Rademacher and Gaussian processes are norms on the dual space of $\ell_2^n$. Recall that a norm in the dual space (that is, the vector space of linear functionals from $\ell_2^n$ into $\mathbb{R}$) is $\|\xi\| = \sup_{\|\mathbf{v}\|_2^n=1} |\xi(\mathbf{v})|$ . One can define a norm whose unit ball is the absolute convex hull of $F$. With that, one can show that,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i f(X_i) = \left\| \sum_{i=1}^{n} g_i e_i \right\|_{F^0}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) = \left\| \sum_{i=1}^{n} \varepsilon_i e_i \right\|_{F^0} ,$$

where $e_1, e_2, \ldots e_n$ is an orthonormal basis of the dual space of $\ell_2^n$, and $\| \cdot \|_{F^0}$ denotes the dual norm of the norm whose unit ball is the absolute convex hull of $F$, that is, the convex hull of $F \cup -F$ (see, e.g., Mendelson 2002c). The function $\mathbf{x} \mapsto \|\sum_{i=1}^n x_i e_i\|_{F^0}$ , being a norm, is convex. One can easily show by the triangle inequality and the Cauchy-Schwartz inequality that it is also Lipschitz, and its Lipschitz constant depends on the geometry of $F$. Thus, the functions $G(\mathbf{g}) = \| \sum_{i=1}^n g_i e_i \|_{F^0}$ and $R(\boldsymbol{\varepsilon}) = \| \sum_{i=1}^n \varepsilon_i e_i \|_{F^0}$ are convex Lipschitz functions with Lipschitz constants $\sup_{f \in F} \|(f(x_1), \ldots, f(x_n))\|_2 / \sqrt{n}$. Thus, we can employ concentration results for convex Lipschitz functions (like Talagrand's convex distance inequality, see for example Ledoux (2001), page 78 and 136) to show that Rademacher and Gaussian averages are concentrated around their expectation (see Theorem 3.11).

Appendix B

# Proofs

## B.1 Proofs for Chapter 4

### Proof of Lemma 4.9

Recall that $A_{n,d}(\mathbf{X}) \subseteq \mathcal{H}$ is the set of all functions with loss classes associated to lucky sets of size smaller than or equal to $2^d$,

$$A_{n,d}(\mathbf{X}) = \{f \in \mathcal{H} \,:\, M\big(\tfrac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \leq 2^d\}$$

and that

$$H_{n,d}(\mathbf{X}) := \bigcup_{f \in A_{n,d}(\mathbf{X})} H_l(f,\mathbf{X})\,.$$

1. $M\big(\frac{1}{n}, H_{n,d}(\mathbf{X}), L_1(\mu_n)\big) \leq 2^d$.

To prove the first property, assume that $M\big(\frac{1}{n}, H_{n,d}(\mathbf{X}), L_1(\mu_n)\big) > 2^d$. Then, by the definition of packing numbers, there exists a $\frac{1}{n}$ separated set $K \subseteq H_{n,d}(\mathbf{X})$ with respect to $L_1(\mu_n)$ of cardinality $|K| = 2^d + 1$. For all $k \in K$, by the definition of $H_{n,d}(\mathbf{X})$, there exists an $f_k \in A_{n,d}(\mathbf{X})$ such that $k \in H_l(f_k,\mathbf{X})$. Let $f' = \operatorname{argmin}_{f_k \in A_{n,d}(\mathbf{X})} L(f_k,\mathbf{X})$. Then $K \subseteq H_l(f',\mathbf{X})$. This means that $K$ is a $\frac{1}{n}$ separated set in $H_l(f',\mathbf{X})$ of cardinality larger than $2^d$, which contradicts $f' \in A_{n,d}(\mathbf{X})$.

2. If $f \in F$ satisfies $M\big(\frac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \leq 2^d$, then $l_f \in H_{n,d}(\mathbf{X})$.

The fact that Property 2 is satisfied follows directly from the definition of $H_{n,d}(\mathbf{X})$: if $f \in \mathcal{H}$ such that $M\big(\frac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \leq 2^d$, then $f \in A_{n,d}(\mathbf{X})$, and because $l_f \in H_l(f,\mathbf{X})$ it follows that $l_f \in H_{n,d}(\mathbf{X})$.

3. Uniqueness.

To prove uniqueness, assume that there is another set $H'_{n,d}(\mathbf{X})$ satisfying the properties 1 and 2. By property 2, $H_{n,d}(\mathbf{X}) \subseteq H'_{n,d}(\mathbf{X})$. We show that if $f$ is such that $l_f \in H'_{n,d}(\mathbf{X})$ then $M\big(\frac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \leq 2^d$. Indeed, for every $f$ such that $l_f \in H'_{n,d}(\mathbf{X})$ there exists a $g \in A_{n,d}(\mathbf{X})$ such that $l_f \in H_l(g,\mathbf{X})$, which implies

$H(f,\mathbf{X}) \subseteq H(g,\mathbf{X})$. Therefore, since $g \in A_{n,d}(\mathbf{X})$, $M\big(\frac{1}{n}, H_l(f,\mathbf{X}), L_1(\mu_n)\big) \leq 2^d$ , and thus also $H'_{n,d}(\mathbf{X}) \subseteq H_{n,d}(\mathbf{X})$ .

## Proof of Lemma 4.10

For a fixed double sample $(\mathbf{X},\mathbf{X}')$ let $\mu_{2n}$ be the empirical measure supported on $(\mathbf{X},\mathbf{X}')$. Put

$$A(\mathbf{X},\mathbf{X}') = \Big\{ l_f \,:\, f \in \mathcal{H},\, M\big(\tfrac{1}{2n}, H_l(f,(\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \leq \omega\big(L(f,\mathbf{X}), n, \delta\big) \Big\}$$

and

$$B_d(\mathbf{X},\mathbf{X}') = \Big\{ l_f \,:\, f \in \mathcal{H},\, M\big(\tfrac{1}{2n}, H_l(f,(\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \leq 2^d \Big\}.$$

Note that

$$F_{n,d}(\mathbf{X}) \cap A(\mathbf{X},\mathbf{X}') \subseteq B_d(\mathbf{X},\mathbf{X}') \subseteq F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')).$$

By the $\omega$-smallness condition (4.17),

$$Pr_{\mathbf{X},\mathbf{X}'}\left\{ \exists f \in \mathcal{H} \,:\, l_f \in \big(A(\mathbf{X},\mathbf{X}')\big)^c \right\} \leq \delta,$$

and by the union bound for disjoint sets,

$$\begin{aligned} &Pr_{\mathbf{X},\mathbf{X}'}\Big\{ \exists f \in F_{n,d}(\mathbf{X}),\, \Big| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - f(X'_i)\big) \Big| \geq t \Big\} \\ &\quad = Pr_{\mathbf{X},\mathbf{X}'}\Big\{ \exists f \in F_{n,d}(\mathbf{X}) \cap A(\mathbf{X},\mathbf{X}'),\, \Big| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - f(X'_i)\big) \Big| \geq t \Big\} \\ &\qquad + Pr_{\mathbf{X},\mathbf{X}'}\Big\{ \exists f \in F_{n,d}(\mathbf{X}) \cap (A(\mathbf{X},\mathbf{X}'))^c,\, \Big| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - f(X'_i)\big) \Big| \geq t \Big\}, \end{aligned}$$

and our claim follows.

## Proof of Lemma 4.11

For every double sample $(\mathbf{X},\mathbf{X}')$, let $\mu_{2n}$ be the empirical measure supported on $(\mathbf{X},\mathbf{X}')$ and define two random sets in the following manner. Let

$$A(\mathbf{X},\mathbf{X}') = \begin{cases} \{l_{\mathcal{A}(\mathbf{X})}\}, & \text{if } M\big(\frac{1}{n}, G_l((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) < \omega\big(L(\mathcal{A}(\mathbf{X})), n, \delta\big) \\ \emptyset, \quad \text{otherwise}\,, & \end{cases}$$

and put

$$B(\mathbf{X},\mathbf{X}') = \begin{cases} \{l_{\mathcal{A}(\mathbf{X})}\}, & \text{if } M\big(\frac{1}{n}, G_l((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \le 2^d \\ \emptyset, \ \text{otherwise}. \end{cases}$$

Note that for every $(\mathbf{X},\mathbf{X}')$,

$$F_{n,d}(\mathbf{X}) \cap A(\mathbf{X},\mathbf{X}') \subset B(\mathbf{X},\mathbf{X}') \subset F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')).$$

Moreover, if $F_{n,d}(\mathbf{X}) \cap \big(A(\mathbf{X},\mathbf{X}')\big)^c \neq \emptyset$, then $F_{n,d}(\mathbf{X}) = \{l_{\mathcal{A}(\mathbf{X})}\}$ and $A(\mathbf{X},\mathbf{X}') = \emptyset$. Thus, by the $\omega$-smallness condition,

$$Pr_{\mathbf{X},\mathbf{X}'}\Big\{F_{n,d}(\mathbf{X}) \cap \big(A(\mathbf{X},\mathbf{X}')\big)^c \neq \emptyset\Big\} \le Pr_{\mathbf{X},\mathbf{X}'}\Big\{A(\mathbf{X},\mathbf{X}') = \emptyset\Big\} < \delta.$$

The claim of the lemma follows now directly from the union bound for the disjoint sets $F_{n,d}(\mathbf{X}) \cap A(\mathbf{X},\mathbf{X}')$ and $F_{n,d}(\mathbf{X}) \cap \big(A(\mathbf{X},\mathbf{X}')\big)^c$.

### Proof of Theorem 4.12

Let $F_{n,d}$ and $F^{\mathrm{sym}}_{2n,d}$ be defined as above, and recall that (equation (4.27))

$$M\big(\tfrac{1}{n}, F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}')), L_1(\mu_{2n})\big) \le 2^d$$

for every $(\mathbf{X},\mathbf{X}')$. Let $V := F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))/\mathbf{X} \subset \ell_2^n$, put $\mu_{2n}$ to be the empirical measure supported on $\mathbf{X} = (\mathbf{X},\mathbf{X}')$, and set $\nu_n$ to be the empirical measure supported on $\mathbf{X}$. Note that for every $f,g$, $\mathbb{E}_{\mu_{2n}}|f-g| \ge \mathbb{E}_{\nu_n}|f-g|/2$. Thus, every $1/n$-cover of $F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ in $L_1(\mu_{2n})$ is a $2/n$-cover of the same set in $L_1(\nu_n)$. In particular, if $A$ is a maximal $1/n$-packing of $F^{\mathrm{sym}}_{2n,d}((\mathbf{X},\mathbf{X}'))$ in $L_1(\mu_{2n})$, it is a $2/n$ cover of that set in $L_1(\nu_n)$. It is easy to verify that, up to isomorphism, $B\big(L_1(\nu_n)\big) = nB_1^n$, and in particular,

$$V \subseteq A + \frac{2}{n} \cdot nB_1^n = A + 2B_1^n,$$

and by the triangle inequality,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v}\in V} \Big|\sum_{i=1}^n \varepsilon_i v_i\Big| \le \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{a}\in A} \Big|\sum_{i=1}^n \varepsilon_i a_i\Big| + 2\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{b}\in B_1^n} \Big|\sum_{i=1}^n \varepsilon_i b_i\Big|.$$

By equation (4.3.2) and because $|A| \le 2^d$ by (4.20), the first term can be bounded by $\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{a}\in A} \Big|\sum_{i=1}^n \varepsilon_i a_i\Big| \le C\sqrt{\log|A|}\sqrt{n} \le C\sqrt{nd}$. For the second term, one can apply the triangle inequality to show that $\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{b} \in B_1^n} \left| \sum_{i=1}^n \varepsilon_i b_i \right| \leq 1$. In conclusion,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in F_{2n,d}^{\mathrm{sym}}((\mathbf{X},\mathbf{X}'))} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \leq C\sqrt{nd}.$$

To complete the proof, apply Corollary 4.6 for $t = C\sqrt{\frac{d}{n}\log(1/\delta)}$.

## B.2 Proofs for Chapter 5

### Proof of Lemma 5.11

Let $m = 2(n^2+n)$. We will define functions which are constant on the intervals $\left(\frac{j}{m}, \frac{j+1}{m}\right]$ , $0 \leq j \leq m-1$ . Define $G_\lambda^n$ to be the function class containing all functions taking the value $-1$ on exactly $n$ such intervals, that is, constructed as follows: Let $J \subset \{1,...,m\}$ , $|J| = n$, and define $g_J \in G_\lambda^n$ as:

$$g_J(x) = \begin{cases} -1, & \text{if } x \in (\frac{j-1}{m}, \frac{j}{m}] \text{ and } j \in J, \\ t_\lambda, & \text{otherwise,} \end{cases}$$

where

$$t_\lambda = \frac{\lambda m + n}{m-n} = \frac{\lambda(2n^2+2n)+n}{2n^2+n}.$$

Since $0 \leq \lambda \leq 1/2$, then $0 < t < 1$.

By construction, all functions $g \in G_\lambda^n$ have expectation $\lambda$ with respect to the uniform measure on (0,1], and they are (1,2)-Bernstein, since

$$\mathbb{E}g^2 = \frac{1}{m}\left(n + t_\lambda^2(m-n)\right) \leq \mathbb{E}g + 2\frac{n}{m} < \mathbb{E}g + \frac{1}{n} \leq 2\mathbb{E}g$$

and $\lambda \geq 1/n$.

The construction of $H_\lambda^n$ is similar, and one can enforce the desired behaviour by defining functions in $H_\lambda^n$ to take the values $\{0, t'_\lambda\}$, for an appropriately chosen $t'_\lambda$.

### Proof of Theorem 5.12

Claim 1 follows directly from the linearity of the expectation and the construction of $H$.

Claim 2: Fix $n$. In order to estimate the value $\xi'_{n,F,\mu}(1/k)$ for $k < n$, we will first estimate the quantity $\sup_{f \in F_k}(\mathbb{E}f - \mathbb{E}_n f)$ for a given sample $\mathbf{X} = (X_1, \ldots, X_n)$ drawn i.i.d. according to $\mu$. Fix $k$ and denote by $m = 2(k^2+k)$. By the construction of $F_k$, every function $g \in F_k$ is of the form $g_J$ for some $J \subset \{1,...,m\}$, $|J| = k$.

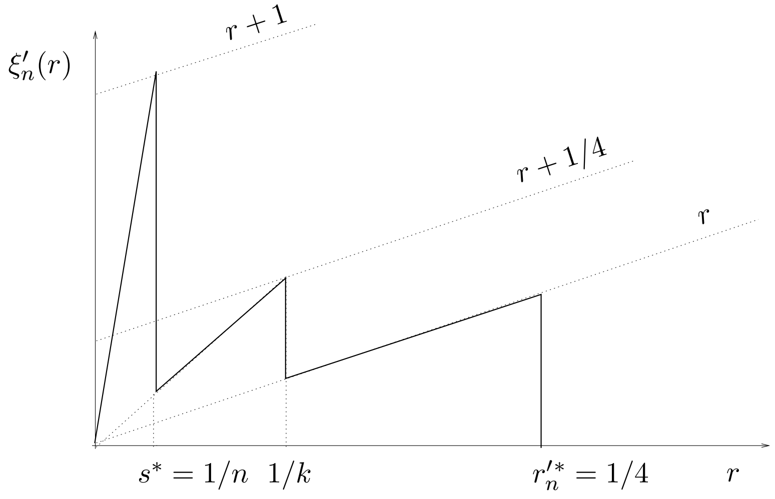


**Figure B.1:** The graph of $\xi'_n$ for the class $\text{star}(F_n \cup H^n_{1/4} \cup F_k)$ in the case in which $\mathbb{E}\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) - 1/k = 1/4$.

Next, we show that $\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f)$ is with high probability significantly smaller than $1/n + 1$. Indeed, we will show that, with high probability, $\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) \le 1/k + 1/4$. To illustrate the idea, assume that this would happen with probability 1, and thus the expectation of this quantity would be at most $1/k + 1/4$. Then, $\xi'_n$ for $\text{star}(F_n \cup H^n_{1/4} \cup F_k)$ would behave like in Figure B.1, and not, say as in Figure B.2.

To show that indeed, for most samples, $\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) \le 1/k + 1/4$, set $\Phi_k$ to be the class of indicator functions:

$$\Phi_k = \{\mathbf{1}_J : (0,1] \longrightarrow \{0,1\} |\ J \subset \{1,...,m\}, |J| = k\},$$

where $\mathbf{1}_J(x) = 1$ if there exists $j \in J$ such that $x \in (\frac{j-1}{m}, \frac{j}{m}]$ and 0 otherwise. For every $\phi \in \Phi_k$, $\mathbb{E}\phi = k/m$. Note that $\text{VC}(\Phi_k) = k$, since no set of $k+1$ distinct points in $(0,1]$ can be shattered by $\Phi_k$, but there is a set of cardinality $k$, namely $\{1/k, 1/(k-1), \ldots, 1\}$, which is shattered by $\Phi_k$.

For each $J$, set

$$\ell_J(\mathbf{X}) = \sum_{i=1}^{n} \mathbf{1}_J(X_i).$$

The random variable $\ell_J(\mathbf{X})$ counts the number of points $X_i$ which fall in the intervals $(\frac{j-1}{m}, \frac{j}{m}]$, where $j \in J$, which is precisely the number of points from the sample on which $g_J$ takes the value $-1$. Hence, by the definition of $g_J$,

$$\mathbb{E}_n g_J = \frac{-2\ell_J(\mathbf{X})(k+1)^2 + 3kn + 2n}{kn(2k+1)},$$

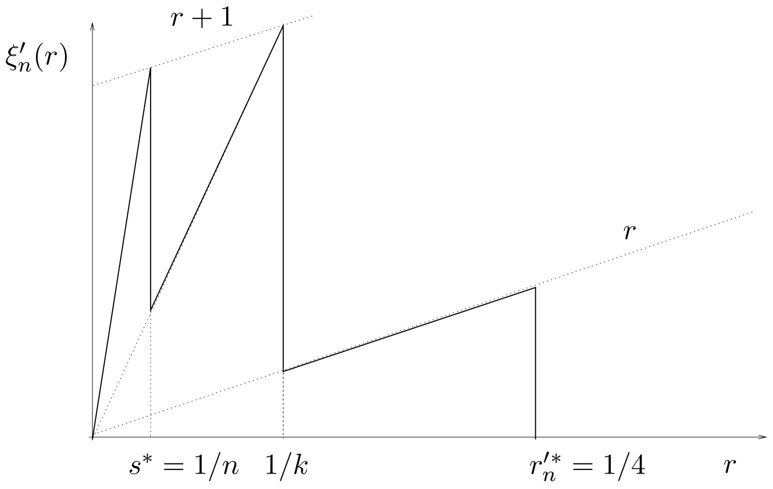


**Figure B.2:** The graph of $\xi'_n$ for the class $\text{star}(F_n \cup H^n_{1/4} \cup F_k)$ in the case in which $\mathbb{E}\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) - 1/k = 1$ .

and

$$\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) = \frac{1}{k} + \frac{2(k+1)^2 \sup_J \ell_J(\mathbf{X}) - 3kn - 2n}{kn(2k+1)},$$

where the supremum for $\ell_J(\mathbf{X})$ is taken over all sets $J \subset \{1, ..., m\}$, such that $|J| = k$.

Next, we show that with high probability $\sup_J \ell_J(\mathbf{X}) \le n/4$, which then implies that for $k \ge 5$,

$$\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) \le \frac{1}{k} + \frac{(k+1)^2/2 - 3k - 2}{k(2k+1)} \le \frac{1}{k} + \frac{1}{4}. \tag{B.1}$$

Indeed,

$$\sup_J \ell_J(\mathbf{X}) = \sup_{f\in\Phi_k} \sum_{i=1}^n f(X_i),$$

and the latter quantity can be controlled through the complexity of $\Phi_k$. From Talagrand's concentration inequality (Theorem 3.12, page 44), for the random variable $Z = \sup_{f\in\Phi_k}(1/n \sum_{i=1}^n f(X_i) - \mathbb{E}f)$, for the case $\rho = 1$, $b = 1$ (which implies $\sigma \le \sqrt{n}$), it follows that there exist absolute constants $c_1, c_2 > 0$ and an integer $n_0$ such that for $n \ge n_0$, with probability larger than $1 - e^{-c_1 n t^2}$,

$$\sup_{f\in\Phi_k} \sum_{i=1}^n f(X_i) \le \frac{kn}{m} + 2nR_n(\Phi_k) + n\max\{t, t^2\}t \le \frac{kn}{m} + 2c_2\sqrt{kn} + nt,$$

where we have used the fact that for any $f \in \Phi_k$ , $\mathbb{E}f = k/m$ , and $t < 1$. The last inequality holds since $\text{VC}(\Phi_k) = k$ , and therefore, by Theorem 2.30, the Rademacher averages can be bounded by $R_n(\Phi_k) \leq c_2\sqrt{\mathrm{VC}(\Phi_k)/n}$ for some absolute constant $c_2$. Setting t=1/10, and since $kn/m \leq n/10$ for any $k \geq 5$, it follows that there exists an absolute constant $c > 0$ such that for any $k \leq n/c$, with probability at least $1-e^{-c'n}$,

$$\sup_J \ell_J(\mathbf{X}) \leq \frac{n}{5} + 2c_2\sqrt{kn} \leq \frac{n}{4}\,,$$

and thus equation (B.1) is true.

Now, we are ready to show that equation (B.1) is sufficient to ensure that the function $\xi'_{n,\mathrm{star}(\cup_{k=5}^{\infty} F_k),\mu}(r)-r$ attains its maximum at $r < c/n$, and decays sharply for larger values of $r$. In particular, for $\varepsilon \sim \sqrt{\log n/n}$ it holds that $r_{n,\varepsilon,+} \sim c/n$. For this, let $A_k$ be the set of all functions in $\mathrm{star}(\cup_{k=5}^{\infty} F_k)$ of expectation $1/k$ (i.e., containing both functions from $F_k$ and the rescaled versions of functions from $F'_k$, where $k' < k$). Hence,

$$A_k = \bigcup_{k'=5}^{k} \frac{k'}{k} F_{k'}\,.$$

By the union bound, with probability at least $1 - ke^{-c'n} \geq 1 - ne^{-c'n}$,

$$\sup_{f\in A_k} (\mathbb{E}f - \mathbb{E}_n f) \leq \frac{1}{k} + \frac{1}{4},$$

and therefore (as the rescaled functions from $H$ do not contribute to the supremum of $\sup_{f\in F,\mathbb{E}f=1/k} (\mathbb{E}f - \mathbb{E}_n f)$ )

$$\begin{aligned}\xi'_{n,F,\mu}(1/k) = \mathbb{E}\sup_{f\in A_k} (\mathbb{E}f - \mathbb{E}_n f) &\leq \frac{1}{k} + (1-ne^{-c'n})\frac{1}{4} + (1-(1-ne^{-c'n})) \\ &\leq \frac{1}{k} + \frac{1}{4} + c''ne^{-c'n}\,.\end{aligned}$$

Hence $\xi'_{n,F,\mu}(1/k) - 1/k \leq 1/4 + c''ne^{-c'n}$, and there exists an $n_0$ such that if $n \geq n_0$ and $k \leq n/c$, then $\xi'_{n,F,\mu}(1/k) - 1/k \leq 1/2$. However, $\xi'_{n,F,\mu}(1/n) - 1/n - \varepsilon_n = 1-\varepsilon_n$, and setting $\varepsilon_n = \sqrt{3(x+\log n)/n}$, it follows that there is a $n(x)$ such that for all $n \geq n(x)$, $1-\varepsilon_n \geq 1/2$. Thus, setting $N(x) = \max\{n_0, n(x)\}$, then for all $n \geq N(x)$ and $k \leq n/c$, $\xi'_{n,F,\mu}(1/k) - 1/k \leq \xi'_{n,F,\mu}(1/n) - 1/n - \varepsilon_n$, which completes the proof.

### Proof of Theorem 5.16

Fix $n \geq N(x)$, where $N(x)$ will be specified later.

It is easy to see that for $F_1$ it holds that $r_n^* \geq r_n'^* \geq 1/4$, and thus $\mathbb{E}\hat{f} \geq 1/4$ and claim 1 follows.

We will now show that for $F_2$, for any sample of size $n$, the empirical minimizer is likely to have expectation smaller than $c/n$ which is thus asymptotically smaller than that of any minimizer in $F_1$. The idea and the main steps of the proof are similar to the proof of Theorem 5.12.

First, set $\varepsilon_n = \sqrt{3(x+\log n)/n}$.

Let $k_n$ be the smallest element in $M$, such that $n \leq k_n$. Thus, there exists an $\ell$ such that $k_n = 2^{2l}$. Hence, $k_n/4 = 2^{2l-2} \in M$ and $k_n/4 < n \leq k_n$ . Note that since $n \leq k_n$, $\inf_{f \in F_2^{k_n}} \mathbb{E}_n f = -1$ , and therefore $\xi'_{n,F_2,\mu}(s_{k_n}) - s_{k_n} = 1$ , where $s_{k_n} \sim 1/k_n$ is the expectation of functions in $F_2^{k_n}$ (cf. proof of Lemma 5.15). This means that a maximal value of $\xi'_{n,F_2,\mu}(r) - r$ is attained at $s_{k_n} \sim 1/k_n$ , and since $k_n/4 < n \leq k_n$ , $s_{k_n} \sim c/n$ for an absolute constant $c$.

The main part of the proof is to show that $\xi'_{n,F_2,\mu}(r)$ is peaked enough around $s_{k_n} \sim c/n$, such that $r_{n,\varepsilon_n,+} \leq c/n$. For this, we will show that there exists an absolute constant $c > 0$, such that for large enough values of $n$, and for $r \geq c/n$, $\xi'_{n,F_2,\mu}(r) - r \leq 1/2$ and thus $\xi'_{n,F_2,\mu}(r) - r$ is significantly smaller than $\xi'_{n,F,\mu}(s_{k_n}) - s_{k_n}$. This follows from the fact that for $k \leq n/c$ , with $k \in M$ , the functions $F_2^k$, are not "complex" enough when projected onto samples of size $n$.

The fat-shattering dimension of $F_2^k$ at any scale $\varepsilon > 0$ is smaller than $c'k/\varepsilon$ (cf. Lemma 5.15), and therefore, by Theorem 2.31, there is a constant $c_2 > 0$ such that $\mathbb{E}R_n(F_2^k) \leq c_2\sqrt{k/n}$ . By Talagrand's concentration inequality (Theorem 3.12, page 44) for the empirical process $Z = \sup_{f \in F_2^k} (\mathbb{E}f - \mathbb{E}_n f)$, since all functions are bounded by 1, it follows that for any $1 > t > 0$, with probability at least $1 - e^{-c_1 n t^2}$ ,

$$\sup_{f \in F_2^k} (\mathbb{E}f - \mathbb{E}_n f) \leq 2R_n\left(F_2^k\right) + t \leq 2c_2\sqrt{\frac{k}{n}} + t.$$

By setting $t = 1/4$, it follows that there exists a constant $c > 0$, such that with probability at least $1 - e^{-c_1' n}$ , for any $k \leq n/c$ ,

$$\sup_{f \in F_2^k} (\mathbb{E}f - \mathbb{E}_n f) \leq 1/2.$$

Let $A_k$ be the set of all functions of expectation $s_k$, containing both functions from $F_2^k$ and the rescaled versions of functions from $F_2^{k'}$, where $k' \in M$, $k' < k$, that is,

$$A_k = \bigcup_{k' \in M, k' \leq k} \frac{s_k}{s_{k'}} F_2^{k'} .$$

Therefore, for any $k \leq n/c$, by the union bound, it follows that with probability at least $1 - \log n e^{-c_1' n}$ (in fact, the union bound is taken over $\log k$ sets, but $\log k \leq \log n$)

$$\sup_{f \in A_k} (\mathbb{E}f - \mathbb{E}_n f) \leq \frac{1}{2}.$$

Estimating the expectation,

$$\begin{aligned}\xi'_{n,F_2,\mu}(s_k) &\leq (1-\log n\, e^{-c'_1 n})\frac{1}{2} + \log n\, e^{-c'_1 n}(s_k+1)\\ &\leq (1-\log n\, e^{-c'_1 n})\frac{1}{2} + 2\log n\, e^{-c'_1 n}\\ &\leq \frac{1}{2} + c' \log n\, e^{-c'_1 n}\,.\end{aligned}$$

Thus, since $\xi'_{n,F,\mu}(s_{k_n}) - s_{k_n} - \varepsilon_n = 1 - \varepsilon_n \longrightarrow 1$ as $n$ tends to $\infty$, there exists $n_0$ such that for all $n \geq n_0$, and for $k \leq n/c$, it holds that $\xi'_{n,F,\mu}(s_k) - s_k \leq \xi'_{n,F,\mu}(s_{k_n}) - s_{k_n} - \varepsilon_n$. Thus, $r_{n,\varepsilon_n,+} \leq c/n$ .

# Bibliography

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the 1993 IEEE Sympositτm on Foundations of Computer Science. IEEE Press*, pages 292–301, 1993.

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

S. Andonova Jaeger. *Theoretical and Experimental Analysis of the Generalization Ability of Some Statistical Learning Algorithms.* PhD thesis, Boston University, 2004.

M. Anthony. Uniform Glivenko-Cantelli theorems and concentration of measure in the mathematical modelling of learning. In *CDAM Research Report Series Centre for Discrete and Applicable Mathematics, LSE CDAM-LSE-2002-07*, 2002.

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

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.

K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J.*, 19(2):357–367, 1967.

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 2004a. To appear.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

P. L. Bartlett and G. Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, 44:55–62, 1999.

P. L. Bartlett and S. Mendelson. Empirical risk minimization. *Probability Theory and Related Fields*, 2005. To appear.

P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 270–284. Springer, 2004b.

P. L. Bartlett, S. Mendelson, and P. Philips. Work in progress, 2005.

P. L. Bartlett and A. Tewari. Sparseness versus estimating conditional probabilities: Some asymptotic results. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 564–578. Springer, 2004.

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.

G. Bennett. Probability inequalities for sums of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

J. Berger. The frequentist viewpoint and conditioning. In *Proccedings of the Berkley Symposium*, pages 15–44, 1985.

S. N. Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

P. Billingsley. *Probability and Measure.* John Wiley & Sons, New York, second edition, 1986.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *A Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

S. Boucheron, O. Bousquet, and G. Lugosi. Concentration inequalities. In O. Bousquet, U.v. Luxburg, and G. Rätsch, editors, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004a.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. Available at http://www.econ.upf.es/∼lugosi/surveys.html, 2004b.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.

S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16(3):277–292, 2000.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31:1583–1614, 2003.

O. Bousquet. A Bennett concentration inequality and its application to empirical processes. *CR. Acad. Sci. Paris*, I(334):495–500, 2002a.

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms.* PhD thesis, École Polytechnique, 2002b.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to Statistical Learning Theory. In O. Bousquet, U.v. Luxburg, and G. Rätsch, editors, *Advanced Lectures in Machine Learning*, pages 169–207. Springer, 2004.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In J. Kivinen and R.H. Sloan, editors, *Proceedings of the 15th Annual Conference on Computational Learning Theory, COLT 2002*, pages 59–73, Sydney, Australia, July 8-10, 2002. Springer.

A. H. Cannon, J. M. Ettinger, D. R. Hush, and J. C. Scovel. Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, 2:335–358, 2002.

G. Casella. Conditionally acceptable frequentist solutions. *Statistical Decision Theory*, 1:73–84, 1988.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other Kernel-Based Learning Methods).* Cambridge University Press, 2001.

A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: On the bias-variance problem. *Foundation of Computational Mathematics*, 2:413–428, 2002.

S. Dasgupta and P. M. Long. Boosting with diverse base classifiers. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory, COLT 2003*, pages 273–287. Springer, 2003.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Number 31 in Applications of Mathematics. Springer, New York, 1996.

L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7):1011–1018, 1995.

L. Devroye and T. Wagner. Distribution-free probability inequalities for the deleted and holdout estimates. *IEEE Transactions on Information Theory*, 25:202–207, 1979.

R. O. Duda, P. E. Hart, and D. G. Sto. *Pattern Classification.* John Wiley & Sons, second edition, 2000.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097: 2–142, 1984.

R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 14(4):1306–1326, 1987.

R. M. Dudley. *Real Analysis and Probability.* Mathematics Series. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

R. M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

R. M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

R. Durrett. *Probability: Theory and Examples.* Duxbury Press, second edition, 1996.

W. Feller. *An Introduction to Probability Theory and its Applications.* John Wiley & Sons, New York, third edition, 1971.

X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes (French). In *Lecture Notes in Mathematics, Vol. 480, École d'Été de Probabilités de Saint-Flour, IV-1974*, pages 1–96. Springer, Berlin, 1975.

X. Fernique. *Fonctions aléatoires gaussiennes, vecteurs aléatoires gaussiens (French).* Universite de Montreal, Centre de Recherches Mathematiques , Montreal, CMP 98:02, 1997.

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Y. Freund. Self bounding learning algorithms. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 247–258. ACM Press, New York, 1998.

Y. Freund and R. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

M. Fromont. Model selection by bootstrap penalization for classification. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 285–299. Springer, 2004.

D. Gamarnik. Extension of the PAC framework to finite and countable Markov chains. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pages 308–317. ACM Press, 1999.

Y. Gat. A bound concerning the generalization ability of a certain class of learning algorithms. Technical Report 548, University of California, Berkeley, March 1999.

E. Giné, V. Koltchinskii, and J. A. Wellner. Ratio limit theorems for empirical processes. In E. Giné, editor, *Stochastic Inequalities and Applications*, pages 249–278. Springer, 2004.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag, Berlin, 2002.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory (A)*, 69(2):217–232, 1995.

D. Haussler, H. S. Seung, M. Kearns, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *COLT '94: Proceedings of the seventh annual conference on Computational learning theory*, pages 76–87, New York, NY, USA, 1994. ACM Press.

R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms.* The MIT Press, 2002.

R. Herbrich and R. C. Williamson. Learning and generalization: Theoretical bounds. In M. A. Arbib, editor, *Handbook of Brain Theory and Neural Networks(2nd edition)*, pages 3140–3150. The MIT Press, 2002.

R. Herbrich and R. C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3(2):175–212, 2003.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. Hoffmann-Jørgensen. Probability in Banach space. In *Lecture Notes in Mathematics, Vol. 598, École d'Été de Probabilités de Saint Flour VI-1976*, pages 1–186, 1977.

W. Jiang. Process consistency for AdaBoost. *The Annals of Statistics*, 32(1):13–30, 2004.

T. Joachims. *Learning to Classify Text using Support Vector Machines.* PhD thesis, Cornell University, 2002.

M. Kääriäinen, T. Malinen, and T. Elomaa. Selective Rademacher penalization and reduced error pruning of decision trees. *Journal of Machine Learning Research*, 5: 1107–1126, 2004.

J.-P. Kahane. *Some Random Series of Functions.* Cambridge University Press, 1968.

M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1):7–50, 1997.

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.

J. Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72:789–807, 1977.

T. Klein. Une inégalité de concentration gauche pour les processus empiriques. *CR. Acad. Sci. Paris*, 334(6):501–504, 2002.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. Technical report, University of New Mexico, August 2003.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459, 2000.

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.

V. Koltchinskii and D. Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *The Annals of Statistics*, 2005. To appear.

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.

A. Kowalczyk, J. Szymanski, and R.C. Williamson. Learning curves from a modified VC-formalism: a case study. In *Proceedings of IEEE International Conference on Neural Networks (ICNN'95)*, volume 6, pages 2939–2943, 1995.

S. R. Kulkarni, G. Lugosi, and S. S. Venkatesh. Learning pattern classification–A survey. *IEEE Transactions on Information Theory*, 44(6):2178–2206, 1998.

J. Langford and A. Blum. Microchoice bounds and self bounding learning algorithms. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pages 209–214. ACM Press, New York, 1999.

M. Ledoux. Isoperimetry and Gaussian analysis. In *Lecture Notes in Mathematics 1648, École d'Été de Probabilités de Saint Flour 1994*, 1994.

M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer, 1991.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6): 2118–2132, 1996.

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.

N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms.* PhD thesis, University of California Santa Cruz, 1989.

N. Littlestone and M. Warmuth. Relating data compression and learnability. University of California Santa Cruz, 1986. Unpublished manuscript.

Philip M. Long. Minimum majority classification and boosting. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 181–186. American Association for Artificial Intelligence, 2002.

G. Lugosi. Concentration-of-measure inequalities, 2003. Available at http://www.econ.upf.es/∼lugosi/anu.ps.

G. Lugosi and A. B. Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864, 1999.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30–55, 2004.

G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the 15th Annual Conference on Computational Learning Theory, COLT 2002*, pages 319–333. Springer-Verlag, 2002.

M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3:723–746, 2002.

P. Massart. About the constants in Talagrand's concentration inequality for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000a.

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000b.

P. Massart. St. Flour lecture notes, July 2003. Available at http://www.math.u-psud.fr/∼massart/flour.pdf.

P. Massart and E. Nédélec. Risk bounds for statistical learning, 2004. Available at http://www.math.u-psud.fr/∼massart/page5.html.

B. Maurey and G. Pisier. Séries de variables aléatoires vectorielles indépendantes et géométrie des espaces de Banach (French). *Studia Mathematica*, 58(1):45–90, 1976.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, London Math. Soc. Lect. Note Series 141*, pages 148–188, 1989.

C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.

R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 119–184. Springer, 2003.

S. Mendelson. Geometric parameters of kernel machines. In J. Kivinen and R.H. Sloan, editors, *Proceedings of the 15th Annual Conference on Computational Learning Theory, COLT 2002*, pages 29–43, Sydney, Australia, July 8-10, 2002a. Springer.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002b.

S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002c.

S. Mendelson. A few notes on Statistical Learning Theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

S. Mendelson. Geometric parameters in learning theory. In *GAFA lecture notes*, 2005. To appear.

S. Mendelson and P. Philips. Random subclass bounds. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory, COLT 2003*, pages 329–343. Springer, 2003.

S. Mendelson and P. Philips. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152(1):37–55, 2003.

M. Molloy. The probabilistic method. In M. Habib, C. McDiarmid, Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–35. Springer, 1998.

W. S. Noble. Support vector machine applications in computational biology. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 71–92. The MIT Press, 2004.

A. Pajor. *Sous-espaces $l_1^n$ des espaces de Banach. (French) [$l_1^n$-subspaces of Banach spaces]*. Hermann, Paris, 1985.

D. Panchenko. A note on Talagrand's concentration inequality. *Electronic Communications in Probability*, 6:55–65, 2001.

D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability*, 31(4):2068–2081, 2003.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics, 94. Cambridge University Press, 1989.

D. Pollard. *Convergence of Stochastic Processes*. Springer, 1984.

M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, Inc., 1980. revised and enlarged edition.

E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119(2):163–175, 2001.

G. K. Robinson. Conditional properties of statistical procedures. *The Annals of Statistics*, 7:742–755, 1979.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 2005. To appear.

R. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

R. E. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5): 1651–1686, 1998.

B. Schölkopf. *Support vector learning*. Oldenbourg Verlag, Munich, 1997.

B. Schölkopf and A. Smola. *Learning with Kernels*. The MIT Press, 2002.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

Shiryayev. *Probability.* Springer-Verlag, Berlin, 1984.

S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

M. J. Steele. *Probability Theory and Combinatorial Optimization.* Number 69 in Cbms-Nsf Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1997.

I. Steinwart. On the influence of the kernel on the consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

I. Steinwart. Sparseness of Support Vector Machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

I. Steinwart. Consistency of Support Vector Machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 51:128–142, 2005.

A. Stuart, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics, Volume 2, Classical Inference and the Linear Model.* Oxfor University Press Inc., sixth edition, 1999.

M. Talagrand. The Glivenko-Cantelli problem. *The Annals of Probability*, 15:837–870, 1987.

M. Talagrand. Type, infratype and the Elton-Pajor theorem. *Inventiones Mathematicae*, 107:41–59, 1992.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22:20–76, 1994.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l' I.H.E.S.*, 81:73–205, 1995.

M. Talagrand. Majorizing measures: The generic chaining. *The Annals of Probability*, 24:1049–1103, 1996a.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996b.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24:1–34, 1996c.

M. Talagrand. *The Generic Chaining.* Springer, 2005.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics.* Springer, New York, 1996.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer, New York, 1982.

V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.

V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 1999.

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.

M. Vidyasagar. *A Theory of Learning and Generalization.* Springer, New York, 1997.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundation of Computational Mathematics*, 5 (1):59–85, 2005.

U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In . K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17.* MIT Press, Cambridge, MA, 2005.

G. Wahba. Estimating derivatives from outer space. Technical Report TSR 989, Mathematics Research Center, U.S. Army, The University of Wisconsin, May 1969.

J. Wellner. Empirical processes: Theory and applications, 2004. Available at http://www.stat.washington.edu/jaw/RESEARCH/TALKS/BocconiSS/emp-prc-bk-big2.pdf.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004a.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004b.

# Glossary of Symbols

| | |
|---|---|
| $\boldsymbol{\mathcal{A}}$ | algorithm |
| $B_p^n$ | unit ball of $\ell_p^n$ |
| $B\big(L_\infty(\Omega)\big)$ | unit ball of $L_\infty(\Omega)$ |
| $\mathbb{E}$ | expectation with respect to all random variables |
| $\mathbb{E}_X$ | expectation with respect to the distribution of $X$ |
| $\mathbb{E}f$ | expectation of the random variable $f(X)$ |
| $\mathbb{E}_\mu$ | expectation with respect to the measure $\mu$ |
| $\mathbb{E}_\mu f$ | expectation of the random variable $f(X)$, where $X$ is distributed according to $\mu$ |
| $\mathbb{E}_n$ | random variable, defined as the expectation with respect to the empirical measure $\mu_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ supported on the random variables $(X_1, \ldots, X_n)$ |
| $F$ | class of functions (often loss class) |
| $F/X$ | coordinate projection of the set $F$ onto the set of coordinates $X$ |
| $\mathrm{fat}_\epsilon\,(F)$ | fat-shattering dimension of $F$ |
| $\widehat{\mathrm{fat}}_\epsilon\,(F, \mathbf{X})$ | empirical fat-shattering dimension of $F$ on sample $\mathbf{X}$ |
| $G_n(F)$ | Gaussian averages of $F$ |
| $G_n(V)$ | Gaussian averages associated with $V$ |
| $H$ | class of hypothesis functions |
| $H_{VC}\,(F)$ | VC-entropy of $F$ |
| $\widehat{H}_{VC}\,(F, \mathbf{X})$ | empirical VC-entropy of $F$ on sample $\mathbf{X}$ |
| $I_A$ | indicator of an event $A$ |
| $I_S(x)$ | indicator function of a set $S$, equal to $I_{x \in S}$ |
| $L_\infty(\Omega)$ | the set of bounded functions on $\Omega$ with respect to the norm $\lVert f \rVert_\infty := \sup_{\omega\in\Omega} \lvert f(\omega)\rvert$ |
| $L_p(\mu)$ | the space of measurable functions on $\Omega$ with a finite norm $\lVert f \rVert_{L_p(\mu)} := (\int \lvert f\rvert^p d\mu)^{1/p}$ |
| $\ell_\infty^n$ | $\mathbb{R}^n$ with the norm $\lVert x \rVert_\infty := \sup_{1\le i\le n} x_i$ |
| $\ell_p^n$ | $\mathbb{R}^n$ with the norm $\lVert x \rVert_p := \left(\sum_{i=1}^n \lvert x_i\rvert^p\right)^{1/p}$ |
| $l$ | loss function |
| $l_2$ | square-loss function |
| $l_h$ | loss function associated to hypothesis $h$ and loss function $l$ |
| $l_{0-1}$ | 0-1 loss function |

| | |
|---|---|
| $M(\varepsilon, F, d)$ | packing number |
| $N(\varepsilon, F, d)$ | covering number |
| $\mathbb{N}$ | the set of natural numbers |
| $n$ | sample size |
| $\Omega$ | measurable space, usually $\Omega = \mathcal{X} \times \mathcal{Y}$ |
| $\mathcal{P}(S)$ | the power set of a set $S$, that is the set of all subsets of $S$ |
| $Pr$ | probability with respect to all random variables |
| $Pr_X$ | probability with respect to the distribution of $X$ |
| $Pr_\mu$ | probability with respect to the measure $\mu$ |
| $R_n(V)$ | Rademacher averages associated with $V$ |
| $\boldsymbol{\mathcal{R}}\,(h)$ | expected loss for hypothesis function $h$ |
| $\widehat{\boldsymbol{\mathcal{R}}}\,(h, \mathbf{z})$ | empirical loss for hypothesis function $h$ on sample $\mathbf{z}$ |
| $\mathbb{R}$ | the set of real numbers |
| $R_n\,(F)$ | Rademacher averages of $F$ |
| $R_n\,(f, \mathbf{X}, \boldsymbol{\varepsilon})$ | Rademacher sum, often denoted by $R_n f$ |
| $\widehat{R}_n\,(F, \mathbf{X})$ | empirical Rademacher averages of $F$ |
| $\overline{R}_n\,(F)$ | uniform Rademacher averages of $F$ |
| $S^c$ | the complement of a set $S$ |
| $VC(F)$ | VC-dimension of $F$ |
| $\widehat{VC}\,(F, \mathbf{X})$ | empirical VC-dimension of $F$ on sample $\mathbf{X}$ |
| $\mathrm{Var}\,(X)$ | variance of the random variable $X$ |
| $\mathrm{Var}\,(f)$ | variance of the random variable $f(X)$ |
| $(X_1', \ldots, X_n')$ | ghost sample, iid copy of $(X_1, \ldots, X_n)$ |
| $X, Y, Z, \ldots$ | random variables |
| $\mathcal{X}$ | input space |
| $\mathbf{X}$ | the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ drawn according to $\mu^n$ |
| $\mathbf{X}\vert_{i=j}^{m}$ | $(X_j, ..., X_m)$ for $\mathbf{X} = (X_1, ..., X_n)$ |
| $\mathbf{x}$ | the vector $\mathbf{x} = (x_1, \ldots, x_n)$, a particular instance of the random vector $\mathbf{X}$ |
| $x, y, z, \ldots$ | particular instances of the random variables $X, Y, Z, \ldots$ |
| $\mathcal{Y}^{\mathcal{X}}$ | set of all functions mapping $\mathcal{X}$ to $\mathcal{Y}$ |
| $\mathcal{Y}$ | output (label) space |
| $\mathcal{Z}$ | input-output space $\mathcal{X} \times \mathcal{Y}$ |
| $\mu$ | unknown probability distribution on $\Omega = \mathcal{X} \times \mathcal{Y}$ |
| $\mu_n$ | empirical measure |

# Index