University of Amsterdam

MSc Mathematics

Master Thesis

Fast Rate Conditions in Statistical Learning

Author:
Muriel Felipe Pérez Ortiz

Supervisor:
Prof. dr. P.D. Grünwald
Prof. dr. J.H. van Zanten

Examination date:
September 20, 2018

Korteweg-de Vries Institute for Mathematics

Centrum voor Wiskunde en Informatica

Abstract

We study conditions under which the order of convergence of algorithms in statistical learning can be improved from $O_{\mathbb{P}}(n^{-1/2})$ to $O_{\mathbb{P}}(n^{-1})$ (up to logarithmic factors) in the number of data points. If excess losses are bounded, it is known that two conditions, called Bernstein's and strong central condition, are equivalent and lead to fast rates both for Empirical Risk Minimization and for randomized algorithms. If the excess losses are unbounded, they are no longer equivalent and are known to lead to faster rates either under additional assumptions or for specific randomized algorithms. We investigate their relation in the unbounded case and show weak, realistic assumptions under which they become equivalent. Furthermore, in this regime we show tighter bounds than those presented in the literature.

Title: Fast Rate Conditions in Statistical Learning
Author: Muriel Felipe Pérez Ortiz, muriel.perezortiz@student.uva.nl, 11391758
Supervisor: Prof. dr. P.D. Grünwald, Prof. dr. J.H. van Zanten
Second Examiner: Prof. dr. B.J.K. Kleijn
Examination date: September 20, 2018

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 105-107, 1098 XG Amsterdam
http://kdvi.uva.nl

Centrum voor Wiskunde en Informatica
Science Park 123, 1098 XG Amsterdam
http://www.cwi.nl

Contents

1. Introduction

In machine learning, statistics and pattern recogition the goal is often to make optimal decisions based on data. This work is about conditions under which the quality of the conclusions of such procedures increases *fast* as the amount of data increases.

In order to illustrate the kind of problem that concerns us, let us first consider the standard pattern recognition problem of *supervised classification* [Gyorfi et al., 1996]. In this situation, we want to be able to classify objects in one of two classes, say positive $(+)$ or negative $(-)$. Suppose that we can observe n objects and that for each of them we record a set of m features which we believe are informative of the class to which they belong. Thus, suppose that we number the objects from 1 to n and that for the i-th one we encode its features in an ordered vector $X_i = (x_i^1, \ldots, x_i^m)$, which can consist of categorical, numeric or other type of entries. The vector X_i is an element of the set $\mathcal{X}$ that contains all the possible values that the features might take. Suppose that we can also observe the class to which each one belongs, so that for the i-th one we know that it belongs to class Y_i, which is either equal to $+$ or $-$. Thus, the task of finding a classifier based on our observations can be seen as that of finding a correspondence g_n that outputs a class ($+$ or $-$) given the features of an object, that is, $g_n : \mathcal{X} \to \mathcal{Y} := \{+, -\}$. An example of this problem is that of predicting if the land use of some area will change or not in a given period of time. In this particular case $+$ would correspond to land use change occurring, while $-$ would correspond to the opposite. Thus, data $(X_1, Y_1), \ldots, (X_n, Y_n)$ would consist of records of variables that are believed to be predictive of land use change and whether it occurred or not in n areas. A classifier would take the values of the predictive variables for an area where it is unknown if land use will change or not and output our best guess based on data.

Naturally there are many ways in which such a classifier might be found, based on the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. We consider a special model for analyzing these situations, that of statistical learning, which has proven to be very useful [Vapnik, 1999]. This model includes three ingredients. First, a probabilistic model for how data is generated. Second, a criterion or algorithm for choosing a classifier based on data. Third, a restricted set of classifiers from which to choose. We assume that each data point (X_i, Y_i) is generated independently of the others and follows a probability distribution $\mathbb{P}$ (which is typically unknown) on all possible pairs of features and classes $\mathcal{X} \times \mathcal{Y}$. This means that $\mathbb{P}\{x, y\}$ is the probability of observing an object with the features x (out of all possible features $\mathcal{X}$) that belongs to the class y (an element of $\mathcal{Y} = \{+, -\}$). With this probabilistic model in mind, we can already set as a goal to minimize the probability of error, that is, we set ourselves to find a classifier g^* that minimizes $\mathbb{P}\{g(X) \neq Y\}$. Often the relationship between the features that we observe and the class to which objects belong is not one to one, that is, it might occur that two objects that have the same

features belong to different classes. This implies that in general it can occur that the probability of error is never zero, that is $\mathbb{P}\{g(X) \neq Y\} > 0$ for each classifier g. Additionally, the distribution $\mathbb{P}$ is often unknown and one must do the best one can to find a classifier using the data available. Thus, perhaps the most intuitive idea is that of choosing the classifier that does best according to the data available, that is, a classifier g_n^* that minimizes the estimated probability of error $\mathbb{P}_n\{g(X) \neq Y\}$ given by

$$\mathbb{P}_n\{g(X) \neq Y\} = \frac{\text{Number of data points where } g(X_i) \neq Y_i}{n}.$$

Unfortunately, this idea, called *Empirical Risk Minimization* (ERM), might fail as one can always choose a classifier $\tilde{g}$ that does not make any mistakes on the data. For instance, we can choose it to satisfy $\tilde{g}(X_i) = Y_i$ for $i = 1, \ldots, n$ and $\tilde{g}(x) = +$ in any other case. This classifier achieves zero error in the data, but intuitively, it is not what we want. Thus, the set of classifiers that we chose g from should not contain all possible classifiers and we adopt the point of view that one chooses from a set of classifiers $\mathcal{G}$ that has been fixed in advance and is not too big (in a sense discussed below).

Given this setup, we are ready to illustrate the first paragraph of this introduction and the scope of this work. By *decision* we mean the act of choosing one of the classifiers from $\mathcal{G}$ to make predictions. The speed at which our best guess g_n^* (that minimizes the error in the sample) becomes optimal is the rate at which its probability of error converges to the lowest possible in $\mathcal{G}$ as more data is gathered, that is, the rate at which

$$\mathbb{P}\{g_n^*(X) \neq Y\} \xrightarrow{\mathbb{P}} \inf_{g \in \mathcal{G}} \mathbb{P}\{g(X) \neq Y\}$$

as $n \to \infty$ (see Section 2.3 for a precise definition of $\xrightarrow{\mathbb{P}}$). We will call rates of order $n^{-1/2}$ in the number of data points *slow*, because they can be attained under very weak conditions on $\mathbb{P}$ and $\mathcal{G}$. On the other hand, we refer to rates of order n^{-1} as *fast*, because even faster rates are usually not achievable. Rates in-between $n^{-1/2}$ and n^{-1} are achievable under *reasonable* conditions that hold in many but by no means all situations. A *fast rate* roughly means that in order to to increase the relative performance by a factor of 100, one needs to gather ~ 10000 times more data points when rates are slow, while one needs ~ 100 times more data points when the rates are fast. This means that finding conditions under which rates are fast is important for practical purposes. This work is about such conditions.

Much has been said about classification in particular [see Gyorfi et al., 1996, Boucheron et al., 2005], but in this work we will focus on the more general abstract learning problem. Consider a set of hypotheses $\mathcal{H}$ and suppose that we observed data $Z_1, \ldots, Z_n$, which is assumed to be identically and independently distributed according to a distribution $\mathbb{P}$ on a space $\mathcal{Z}$. For instance, in the case of classification $\mathcal{H}$ can be identified with a set of classifiers $\mathcal{G}$ and $\mathcal{Z}$ with the set of all possible pairs of features and classes $\mathcal{X} \times \mathcal{Y}$. Thus, given a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ that quantifies how bad choosing h is when we observe Z, we judge a decision h as good if its expected loss over the distribution of

the data is small. We denote the expected loss of $h \in \mathcal{H}$ by $L(h)$ and write the expected value as

$$L(h) = \mathbb{P}[\ell(h, Z)] = \int \ell(h, Z) \mathrm{d}\mathbb{P}.$$

In the classification case, one has that $\ell(g, x, y) = \mathbb{1}\{g(x) \neq y\}$, so that for the classifier g the expected loss $L(g)$ is nothing more than the probability of error $\mathbb{P}\{g(X) \neq X\}$. Thus, minimizing the expected loss is equivalent to minimizing the probability of error in the classification case. Since it is typically not known what the optimal expected loss might be, we are happy to consider the performance relative to the best possible achievable performance on the class $\mathcal{H}$, which is measured by the expected excess loss $E(h)$ given by

$$E(h) = L(h) - \inf_{h \in \mathcal{H}} L(h).$$

The study of the expected excess loss, also called excess risk, has been central to the theory of statistical learning [Vapnik, 1998]. We will focus on the case in which there is a $h^* \in \mathcal{H}$ for which the infimum on the right is achieved. Consequently, we can define the excess loss $\varepsilon(h, Z)$ as

$$\varepsilon(h, Z) = \ell(h, Z) - \ell(h^*, Z)$$

and note that $E(h) = \mathbb{P}[\varepsilon(h, Z)]$. This work deals with conditions on the excess losses and their distributions in the case that they might be unbounded but are still constrained to have exponentially small left tails. This means that we concern ourselves with situations in which the probability that the hypotheses in the class perform better than the optimal is exponentially small. Obviously, this makes the task of empirically identifying elements $h \in \mathcal{H}$ with small expected excess loss easier.

As we noted before, there are many valid ways in which one can choose a hypothesis $h \in \mathcal{H}$ based on the available data $Z_1, \ldots, Z_n$. Here we will consider two. The first method is ERM, which consists of choosing a hypothesis h_n^* that minimizes the *empirical risk*, that is

$$h_n^* \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$$

from a restricted class of hypotheses $\mathcal{H}$. This is the hypothesis that looks best according to our data. Second, we consider the case of *randomized prediction*. In this case, predictions are made according to a random hypothesis $\hat{h}_n$ sampled from a data dependent distribution Π_n on $\mathcal{H}$. Usually one establishes a *prior* distribution Π_0 and bases the distribution Π_n both on the data and Π_0 in analogy to Bayesian procedures. More generally, the distribution Π_n is built strategically in such a way that hypotheses with lower empirical risk are sampled with higher probability. This approach also includes ERM and other deterministic predictors as a special case, as they can be viewed as sampling one hypothesis with probability one for prediction.

As we will recall in Chapter 2, deviation bounds for both ERM and randomized prediction have been proven. Under mild conditions it has been shown that

$$E(h_n^*) = O_{\mathbb{P}}(n^{-1/2} \mathrm{Comp}^{1/2}(n)) \tag{1.1}$$

[see Vapnik, 1998] where $\mathrm{Comp}(n)$ is a specific measure of the complexity of the hypotheses class $\mathcal{H}$ for which $\mathrm{Comp}(n) \leq \log|\mathcal{H}|$ if $\mathcal{H}$ has finite size (see also Appendix A for the exact meaning of the $O_{\mathbb{P}}$ notation). For the randomized case, through the so-called PAC-Bayesian inequalities, it has been established that

$$\Pi_n[E(h)] \leq \Pi_n[E_n(h)] + O_{\mathbb{P}}(n^{-1/2}\mathrm{Comp}(n)),$$

where Π_n is any suitable distribution on $\mathcal{H}$ and and E_n is the empirical excess loss

$$E_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i)$$

[see McAllester, 1998]. In this work we deal with conditions under which fast rates can be obtained, that is, situations in which $O_{\mathbb{P}}(n^{-1}\mathrm{Comp}(n))$ can replace its counterparts from the previous equations. We next introduce the conditions with which we will deal.

Consider first the case in which excess losses are bounded, that is, when

$$\sup_{h,z} \varepsilon(h, z) < \infty.$$

This case has been extensively studied. It it has been shown that if $\mathcal{H}$ has finite size

$$\sup_{h\in\mathcal{H}} \frac{\mathbb{P}[\varepsilon^2(h, Z)]}{\mathbb{P}[\varepsilon(h, Z)]} \leq B,$$

where $\mathbb{P}[\varepsilon^2(h, Z)]$ is the second moment of the excess loss, faster rates can be achieved. This condition is the strongest version of what has been called *Bernstein's condition* (see Condition 1 on page 26) in association with the use of Bernstein's inequality to obtain faster concentration (see Section 3.2.1). Bartlett and Mendelson [2006] proved that also in the case that the class $\mathcal{H}$ is infinite and the excess losses are uniformly bounded, Bernstein's condition implies faster concentration. Van Erven et al. [2015] found that it is equivalent to the cumulant generating function of $-\varepsilon(h, Z)$ being non positive for all h at some point, that is,

$$\sup_{h\in\mathcal{H}} \log \mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h,Z)}] \leq 0$$

for some $\eta > 0$, where $\mathbb{P}$ is the data generating distribution. This condition is the strong version of what the authors called *Central Condition* (see Condition 2 on page 28). They noted that it had been used in the past, often implicitly, in order to obtain fast rates. For instance, it can be related to having a well specified model in the case of density estimation (see Example 3.2.4). Consequently, when viewed as a statistical learning problem, the central condition also allows for fast rates (see Example 2.2.2) for density estimation. Furthermore, Grünwald and Mehta [2016] showed that from the strong central condition condition fast rates could also be established using a PAC-Bayesian-style analysis, where the study of cumulant generating functions is essential.

Nevertheless, the landscape changes when the excess losses might be unbounded, which is the case in many practical situations. In the unbounded case, Audibert [2009]

showed a nonstandard randomized learning algorithm that achieved fast rates if Bernstein's condition is satisfied. To the best of our knowledge, it is still not known if satisfying Bernstein's condition also leads to fast rates for ERM in the unbounded case. One might hope that this question can be answered using the relation of Bernstein's condition to the strong central condition. This is not the case because on the one hand Grünwald and Mehta [2016] showed that the strong central condition was not enough to obtain fast rates and on the other, as we will show in Chapter 4, the strong central condition and Bernstein's condition are not equivalent in the unbounded case even under additional conditions. Despite this, Grünwald and Mehta [2016] showed that under an additional condition on the tails of the excess losses and the strong central condition, fast rates could be obtained for randomized algorithms and for ERM. The condition, called the *witness condition*, is the existence of some $u > 0$ such that

$$0 < \sup_{h \in \mathcal{H}} \frac{\mathbb{P}[\varepsilon(h, Z) \mathbb{1}\{\varepsilon(h, Z) > u\}]}{\mathbb{P}[\varepsilon(h, Z)]} \leq 1,$$

that is, it is a mild condition on the relative weight of the right tail of the excess losses. It is easy to see that it is implied by Bernstein's condition (see Section 3.2.1). Grünwald and Mehta [2016] related this condition to conditions for the convergence of likelihood ratios that had been previously studied by Wong and Shen [1995], who use lower truncations of log likelihood ratios in their analysis. Recall that $\varepsilon(h, Z) > u$ implies for the losses that $\ell(h, Z) > u + \ell(h^*, Z)$, thus we are dealing with restrictions on how much worse h can be compared to the optimal h^* on average.

Apart from the conditions that we previously mentioned, little is known about conditions on the excess losses that lead to fast rates. In this work we make a first step in understanding the relation between these conditions, and the uses and limitations of the tools at hand. The main contributions of this work are contained in Chapter 4 and are the following:

1. As we mentioned before, we establish by way of counterexamples (Counterexamples 4.1.1 and 4.1.4) that even under stringent tail restrictions, it is not true that Bernstein's condition and the combination of the central condition and the witness condition are equivalent. Since the combination of the strong central condition and the witness condition lead to fast rates (see Section 3.3), this means that the question of whether Bernstein's condition leads to fast rates for ERM in the unbounded case remains open as Bernstein's condition does not imply them.

2. We establish that if there exists $r > 1$ such hat

$$\sup_{h \in \mathcal{H}} \frac{(\mathbb{P}[|\varepsilon(h)|^{2r}])^{1/r}}{\mathbb{P}[\varepsilon^2(h)]} < \infty,$$

(see Condition 4 on page 39) Bernstein's condition and the strong central condition are equivalent as long as the cumulant generating function of $-\varepsilon(h)$ satisfies

$$\sup_{h \in \mathcal{H}} \log \mathbb{P}[\mathrm{e}^{-\eta \varepsilon(h)}] < \infty \tag{1.2}$$

for some $\eta > 0$ (Theorem 4.2.3). The requirement in (1.2) means that the left tail probabilities of the excess loss can be uniformly bounded by an exponential function, a natural relaxation of the bounded excess loss condition. The condition on the moments of random variables had been considered before, for instance by Mendelson [2014, Lemma 6.1]. It is satisfied in the case that for each h the excess loss $\varepsilon(h, Z)$ has a Gaussian, Laplacian or uniform distribution and many other distributions with well-behaved tails also satisfy it (see Example 4.2.2). This means that this condition can be used in place of the witness condition (because Bernstein's condition implies it) in addition to the central condition to achieve fast rates. In this case, we proved a tighter excess risk bound (Corollary 4.2.6) as our next contribution.

3. We notice that in the case that the previous condition and the central condition hold, a condition (see Condition 5 on page 41) also considered by Vapnik [1998] holds. This condition, which is also on the relative size of a moment of $\varepsilon(h, Z)$, is the existence of some $r > 1$ such that

$$\sup_{h \in \mathcal{H}} \frac{(\mathbb{P}[|\varepsilon(h, Z)|^{2r}])^{1/2r}}{\mathbb{P}[\varepsilon(h, Z)]} < \infty.$$

   Vapnik [1998] argued that this condition characterizes light tails and proved that it implied slow rates for ERM in the case that it was imposed on the loss function $\ell(h, Z)$. We prove (Theorem 4.2.4) that this condition and (1.2) are enough to obtain sharper bounds than those obtained by Grünwald and Mehta [2016]. We also prove that if (1.2) holds, Vapnik's condition implies the strong central (Lemma 4.2.8) and Bernstein's condition (Lemma 4.2.10), and thus also the witness condition. Since this condition holds in many practical cases, the resulting bounds are relevant.

4. We prove that if (1.2) holds for some $\eta > 0$, then either the witness condition alone or an alternative second moment condition implies that slow rates can be achieved (Theorem 4.3.1).

With this in mind, we now describe how this thesis is organized.

Chapter 2. In Section 2.1 we describe formally the basic theory of statistical learning and the set-up for the remainder of this thesis. In Section 2.2 we introduce empirical risk minimization. Since before worrying about rates of convergence one needs to establish consistency first, for the sake of completeness we recall in Section 2.3 the classic theorem on consistency of ERM due to Vapnik and Chervonenkis [1991]. In Section 2.4 we make some considerations about excess losses, the main object of study. In Section 2.5 we show how concentration inequalities can be used to obtain rates of convergence for ERM for finite classes. In Section 2.6 we introduce the PAC-Bayesian setting, and recall and prove the PAC-Bayesian inequalities that are the backbone of PAC-Bayesian bounds in Section 2.6.1, which are also valid in the case of infinite classes.

Chapter 3. In Section 3.1 we describe the inequality obtained by Zhang [2006b] on which the bounds obtained by Grünwald and Mehta [2016] are based. We do this because we later use Zhang's inequality to obtain tighter bounds using Vapnik's condition in Chapter 4. We then describe the bounds on the expected excess loss obtained by Grünwald and Mehta [2016] in Section 3.3 after having explained how Bernstein's, Central and the Witness condition come into play in Section 3.2.

Chapter 4 It is in this chapter where the main contributions of this work can be found. In Section 4.1 we establish counterexamples that relate the strong central, the witness and Bernstein's condition. In Section 4.2 we establish our new condition, recall Vapnik's condition and prove tighter excess risk bounds than those of Grünwald and Mehta [2016]. In Section 4.3 we prove that under the witness condition or an alternative mild condition, slow rates can be obtained as long as (1.2) holds.

Conclusion and Appendices We finish with concluding remarks in Chapter 5, where we also present some open questions that arise as a result of this thesis. Furthermore, we present four appendices. In Appendix A we describe what rates of convergence in probability mean. In Appendix B we explain the Cramér-Chernoff method for proving deviation inequalities and we show why it yields $n^{-1/2}$ rates. In Appendix C we discuss subgamma tails for random variables and present *Bernstein's moment condition*. In Appendix D we point at results on how the results derived in Section 2.4 can be extended to infinite classes using a celebrated and now classical technique called *chaining*. Although there are other methods for obtaining rates of convergence in the case of infinite classes, chaining is known to yield optimal rates.

2. Basic Theory

There is a vast family of problems in statistics, pattern recognition and machine learning in which one wants to make an optimal decision based on data. We will take the point of view that data are independent realizations of a random process. In this chapter we describe a useful high-level model for such situations, that of statistical learning theory [Vapnik, 1998], which is framed in the language of probability theory. We introduce the notation that we will use throughout the rest of the document. In Section 2.1 we describe the set up of an abstract learning problem and give some examples of how this model applies in concrete statistical applications. We will treat two situations: Empirical Risk Minimization (ERM from now on), and randomized prediction. In Section 2.2 we explain the method of ERM and we describe what its consistency means. Since before wondering about rates of convergence one needs to worry about whether consistency happens at all, in Section 2.3 we recall necessary and sufficient conditions for it to happen, which are classical. In Section 2.4 we introduce excess losses, the main object of analysis. We focus on them because their study is useful to infer nonasymptotic bounds on the performance of learning algorithms. In Section 2.5 we introduce excess losses with subgamma and subgaussian tails and explain how rates of convergence can be obtained for ERM. We focus on the case in which the hypotheses class is finite and point at how to generalize this analysis to infinite classes in Appendix D. In Section 2.6 we introduce the PAC-Bayesian model for situations in which our prediction might not depend deterministically on the observed data and we prove the main PAC-Bayesian inequality, which is central in this type of analysis.

2.1. Statistical Learning Theory

The statistical learning model consists of three parts. First, a probabilistic model for the generation of the data that we observe. Second, a set of hypothes (or hypotheses class) from which we want to choose. Third, a notion of optimality for choosing a hypothesis. We assume that data take values in a set $\mathcal{Z}$ and we endow it with a probability space structure. Thus, let $(\mathcal{Z}, \mathcal{F}, \mathbb{P})$ be a probability space. We model data $Z_1, \dots, Z_n$ as independent and identically distributed random variables taking values on $\mathcal{Z}$ with distribution $\mathbb{P}$. For any random variable Z we write as $\mathbb{P}Z$ or $\mathbb{P}[Z]$ its expectation with respect to $\mathbb{P}$. We follow the same notation when taking expectations with respect to other distributions.

We adopt the point of view that set of hypotheses $\mathcal{H}$ from which we want to choose has been fixed in advance, that is, it does not depend on the number of data points n. Since we will also consider situations in which our predictions do not depend deter-

ministically on the data that is observed, we endow $\mathcal{H}$ with a sigma-algebra $\mathcal{G}$. In this situation, we also need to endow $\mathcal{Z} \times \mathcal{H}$ with a measurable structure, for which we choose the product sigma-algebra. We consider $\mathbb{R}$ as a measurable space with the Borel sigma-algebra generated by its usual topology.

We encode our notion of optimality in a measurable function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. We interpret $\ell : (h, Z) \mapsto \ell(h, Z)$ as the loss associated to using h to make predictions when observing the data point Z and refer to it as the loss function. When it is clear from context, we will drop the Z-dependence of ℓ and just write $\ell(h)$ instead of $\ell(h, Z)$, and $\ell_i(h)$ instead of $\ell(h, Z_i)$. This notation also stresses the fact that we might also view ℓ as a random function (or a stochastic process) taking values in some subset of $\mathbb{R}^{\mathcal{H}}$ and $\ell_1, \ldots, \ell_n$ as a random sample of such functions. For a fixed $h \in \mathcal{H}$ let the expected loss be

$$L(h) = \mathbb{P}[\ell(h)],$$

also called the risk of h.

With this in mind, we will concern ourselves with an *abstract learning problem*, which is that of finding elements $h \in \mathcal{H}$ as close to the infimum over $\mathcal{H}$ of the expected excess loss as possible based on data. This notion is encoded in the expected excess loss, also called excess risk functional. We denote it by $E(h)$ and define it as

$$E(h) = L(h) - \inf_{h \in \mathcal{H}} L(h).$$

2.2. Empirical Risk Minimization (ERM)

Empirical Risk Minimization (ERM) is the idea of choosing the hypotheses that looks best on the available data. First, we define the empirical risk, which is nothing more than the average of the loss on the sample. As we pointed out already in the introduction, ERM is not always consistent and problems may arise if, roughly speaking, the hypothesis class at hand is too big. We will cite a condition that is necessary and sufficient for the consistency of ERM in the next section.

For a random sample $Z_1, \ldots, Z_n$ consider the empirical measure $\mathbb{P}_n$ given by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i},$$

where δ_z is the probability measure that concentrates all of its mass on the point z. Consequently, $\mathbb{P}_n$ is a (random) measure that puts mass $1/n$ at each of the (random) points $Z_1, \ldots, Z_n$. Call empirical risk, or empirical loss $L_n(h)$ the empirical mean of the loss function, that is,

$$L_n(h) = \mathbb{P}_n[\ell(h)] = \frac{1}{n} \sum_{i=1}^{n} \ell_i(h).$$

Empirical Risk Minimization (ERM) is the idea of picking from $\mathcal{H}$ an element h_n^* that minimizes the empirical risk, that is,

$$h_n^* \in \arg\min_{h \in \mathcal{H}} L_n(h).$$

Thus, the ERM is the best fitting hypothesis according to the data that is available. Nevertheless, the size of $\mathcal{H}$ is a compromise. For instance in the case of classification, if the hypotheses class $\mathcal{H}$ includes all possible functions $\mathcal{X} \to \mathcal{Y}$, then the empirical loss will always be zero, while the expected loss can remain positive. Thus, two questions arise. The first question is about consistency, that is, finding necessary and sufficient conditions for the convergence of the excess risk to zero. It turns out that the expected excess risk can be bounded in a distribution-free manner, and necessary and sufficient conditions for its convergence are known. While this is remarkable, in practice we are also interested in non asymptotic results about the speed at which this convergence occurs, as it might be arbitrarily slow [see Gyorfi et al., 1996, Chapter 7].

In the introduction we already highlighted how the problem of classification can be cast in these terms. We now give two more examples, that of regression and that of density estimation. This means that our analysis also has implications for these problems.

Example 2.2.1 (Least Squares Regression with Random Design)**.** Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Identify $\mathcal{H}$ with a set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ and let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an iid sample from $\mathbb{P}$. Thus, we say that the design is random because the points $X_1, \ldots, X_n$ not fixed by us. In this case, $\mathcal{Y} = \mathbb{R}$ and for $f \in \mathcal{F}$, the loss function is given by

$$\ell(f(X), Y) = \frac{1}{2}(f(X) - Y)^2.$$

One is then interested in finding a function f that minimizes

$$L(f) = \mathbb{P}[\ell(f(X), Y)] = \frac{1}{2}\mathbb{P}[f(X) - Y]^2,$$

often called the mean quadratic error. It is known that under certain regularity conditions the estimator f_n^* that minimizes the error in the sample, that is,

$$f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

is a consistent estimator of the best possible f in the class of functions $\mathcal{F}$.

Example 2.2.2 (Density Estimation)**.** Say that given a random sample $Z_1, \ldots, Z_n$ from a probability distribution $\mathbb{P}$, one is interested in estimating its density $\mathbb{P}$ by choosing one from a set of distributions $\mathcal{Q}$, called *statistical model*. Suppose that there is a dominating measure ν such that each element $\mathbb{Q} \in \mathcal{Q}$ has a density q with respect to ν. Identify $\mathcal{H}$ with the set $\mathcal{Q}$ of probability distributions over $\mathcal{Z}$. If we choose for $\mathbb{Q} \in \mathcal{Q}$ the loss function $\ell(\mathbb{Q}, z) = -\log q(z)$, then the expected loss

$$L(\mathbb{Q}) = -\mathbb{P}\left[\log q(Z)\right]$$

is the negative likelihood. Thus, minimizing L in this case amounts to maximizing the likelihood.

It is known that under certain regularity conditions, the maximum likelihood estimator $\mathbb{Q}_n^*$ with density q_n^*

$$q_n^* = \arg\max_{\mathbb{Q}\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} \log q(Z_i)$$

is unique and is a consistent estimator of $\mathbb{P}$ in the case that $\mathbb{P} \in \mathcal{Q}$, that is, in the case that the model is well specified.

In the next section we a classical result on the necessary and sufficient conditions for the consistency of ERM which is due to Vapnik and Chervonenkis [1991] and was later compiled by Vapnik [1998]. The result gives necessary and sufficient conditions for the consistency of ERM and provides a link to the study of uniform deviations of empirical averages from their meas.

2.3. Consistency of Empirical Risk Minimization

We say that ERM is consistent if both its risk and its empirical counterpart become as small as possible as $n \to \infty$, that is, if

$$L(h_n^*) \xrightarrow{\mathbb{P}} \inf_{h\in\mathcal{H}} L(h)$$

and

$$L_n(h_n^*) \xrightarrow{\mathbb{P}} \inf_{h\in\mathcal{H}} L(h).$$

where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability (see Appendix A). Nevertheless, Vapnik [2000] argued that this concept of consistency allows for trivial cases in which it fails to capture the intrinsic capacity of the hypotheses class $\mathcal{H}$ at hand. Their argument goes as follows. If it has been established that ERM is not consistent for a specific learning problem, an easy trick can turn it into a consistent one. It is enough to add a hypothesis and to redefine the loss function so that the newly added hypothesis minorizes the loss function. This means that consistency in this sense would not necessarily depend on the intrinsic *capacity* of the hypotheses class at hand, but it may depend on the existence of a hypothesis that minorizes the loss function, that is, the existence of some h^* such that $\ell(h^*) \leq \ell(h)$ almost surely for all $h \in \mathcal{H}$. Thus, procedures should be consistent even if hypotheses with small losses are removed. Consequently, they proposed a stronger concept, that of *nontrivial consistency*, In order to define it, we need to introduce the set of hypotheses $\mathcal{H}_c \subseteq \mathcal{H}$ with loss bigger than the threshold $c > 0$, that is,

$$\mathcal{H}_c = \{h \in \mathcal{H} : L(h) \geq c\}.$$

Definition 2.3.1 (Nontrivial Consistency)**.** We say that ERM is *nontrivially consistent* if for each nonempty $\mathcal{H}_c$ (defined above) it holds that

$$\inf_{h\in\mathcal{H}_c} L_n(h) \xrightarrow{\mathbb{P}} \inf_{h\in\mathcal{H}_c} L(h).$$

Nontrivial consistency implies consistency and Vapnik and Chervonenkis [1991] gave necessary and sufficient conditions for it to happen. They call this result The Key Theorem, which reads as follows.

Theorem 2.3.2 (The Key Theorem, Vapnik and Chervonenkis [1991]). *Let $\mathcal{H}$ be a hypotheses class, let $\ell(h)$ be a loss function with expected loss $L(h)$ such that*

$$\sup_{h\in\mathcal{H}} |L(h)| < \infty$$

Then the two following statements are equivalent

- *The nontrivial consistency of ERM*
- *The existence of one-sided uniform convergence over $\mathcal{H}$ of the empirical losses to their expectations, that is,*

  $$\sup_{h\in\mathcal{H}} L(h) - L_n(h) \xrightarrow{\mathbb{P}} 0$$

  as $n \to \infty$

This result has as important consequence that the analysis of the consistency of ERM has a worst case flavor. On the other hand it establishes a link with the study of the uniform convergence of empirical averages $L_n(h)$ to their expectations $L(h)$, to which much attention has been paid [see Van der Vaart and Wellner, 1996]. Nevertheless, the theorem does not provide tools for deriving nonasymptotic results on speeds of convergence.

The random variable

$$\sup_{h\in\mathcal{H}} L(h) - L_n(h)$$

has been called one-sided empirical process. Note that its measurability can be a concern as the supremum is taken over the set $\mathcal{H}$, which might be uncountable. In order to avoid this issue, we will assume that the empirical processes in question are *separable*, that is, that there exists some countable subset of $\mathcal{H}$ such that the supremum of the empirical processes over $\mathcal{H}$ equals the supremum taken over the countable subset. Note that this is the case in which $\mathcal{H}$ is a complete, separable, metric space (i.e. a Polish space) and $\ell : h \mapsto \ell(h)$ is almost surely continuous.

2.4. Excess Losses

In the remainder of this thesis, we will focus on the case in which there exists a unique element $h^* \in \mathcal{H}$ that minimizes the expected loss, that is,

$$h^* = \arg\min_{h\in\mathcal{H}} L(h).$$

In this case, one can define the excess loss $\varepsilon : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ as

$$\varepsilon(h) = \ell(h) - \ell(h^*),$$

so that the expected excess loss $E(h)$ can be written as

$$E(h) = \mathbb{P}[\varepsilon(h)] = \mathbb{P}[\ell(h) - \ell(h^*)] = L(h) - L(h^*),$$

which is non negative. Analogously, we define its empirical counterpart $E_n(h) = \mathbb{P}[\varepsilon(h)]$.

The reason for studying excess losses lies in the fact that if for some estimator $\hat{h}_n$ dependent on the data it happens that

$$E(\hat{h}_n) - E_n(\hat{h}_n) \xrightarrow{\mathbb{P}} 0,$$

then the estimator is (non-trivially) consistent, as discussed in the previous section. In the special case of the ERM h_n^*, this implies that

$$E(h_n^*) = E_n(h_n^*) + o_{\mathbb{P}}(1).$$

But, because of the definition of the excess loss, the empirical excess loss at the ERM is negative while the expected excess loss itself is positive. This means that the last equation would imply that

$$E(h_n^*) = o_{\mathbb{P}}(1).$$

Since

$$E(h_n^*) - E_n(h_n^*) \leq \sup_{h \in \mathcal{H}} E(h) - E_n(h), \tag{2.1}$$

much attention has been paid to the study of the right hand side, that is, to the study of uniform deviations random variables from their averages. This has resulted in a fruitful approach.

2.5. Rates from Concentration Inequalities

Concentration inequalities quantify the size of the deviations of a random variable from its mean, in other words, to which extent they *concentrate* around it. The phenomenon of concentration holds in many situations [see Boucheron et al., 2013]. We will focus on excess losses for which a uniform subgaussian or subgamma bound bound holds. We will aslo focus on finite classes $\mathcal{H}$ and explain the methods that can be used to extend these results to infinite classes in Appendix D. The reason for considering the following types of inequality is that it is exactly inequalities of this type that one obtains through methods such as the Cramér-Chernoff method (see Appendix B).

Uniformly subgaussian excess losses include those that are bounded (via Hoefding's inequality) and uniformly subgamma excess losses include those that satisfy Bernstein's moment condition for fixed constants (see Appendix C). Said excess losses satisfy the deviation inequality

$$\mathbb{P}\left\{n^{1/2}(E(h) - E_n(h)) \geq t\right\} \leq \mathrm{e}^{-\phi(t/\tau)} \tag{2.2}$$

for all $h \in \mathcal{H}$, some $\tau > 0$, and $\phi(t) = \phi_2(t) := t^2$ in the subgaussian case. In the subgamma case

$$\phi(t) = \phi_{\text{Bern}}^L(t) = \left(\frac{\sqrt{1 + 2Lt/\sqrt{n}} - 1}{L/\sqrt{n}} \right)^2$$

with some $L > 0$. These two last inequalities can be written equivalently in the more enlightening form

$$\mathbb{P}\left\{ n^{1/2}(E(h) - E_n(h)) \geq \tau \left(\sqrt{t} + \frac{L}{2} \frac{t}{\sqrt{n}} \right) \right\} \leq \mathrm{e}^{-t},$$

for the subgamma case by inverting ϕ_{Bern}^L and as

$$\mathbb{P}\left\{ n^{1/2}(E(h) - E_n(h)) \geq \tau\sqrt{t} \right\} \leq \mathrm{e}^{-t},$$

in the subgaussian case. This means that for small values of t, subgamma random variables behave essentially as subgaussian, while they have exponential tails for larger values of t. Note also that subgaussian excess losses are also subgamma.

In order to illustrate how rates of convergence are obtained from concentration inequalities, let us first consider finite classes. In Appendix D we point at how these results can be extended to infinite classes.

Let $\mathcal{H}$ be a finite hypotheses class of size $|\mathcal{H}| = N < \infty$. If the excess losses satisfy a concentration inequality such as those discussed earlier, with a simple union bound one can obtain that

$$\mathbb{P}\left\{ n^{1/2} \sup_{h \in \mathcal{H}} E(h) - E_n(h) \geq t \right\} \leq \mathrm{e}^{-\phi(t/\tau)}$$

which means that with probability higher than $1 - \delta$ it holds simultaneously for all $h \in \mathcal{H}$ that

$$E(h) \leq E_n(h) + \frac{\tau}{n^{1/2}} \phi^{-1}\left(\log \frac{N}{\delta} \right).$$

Note that at an empirical risk minimizer h_n^*, the empirical excess risk $E_n(h_n^*)$ is negative while the excess risk $E(h_n^*)$ itself is positive. This implies that for the ERM rule h_n^* with probability higher than $1 - \delta$

$$E(h_n^*) \leq \frac{\tau}{n^{1/2}} \phi^{-1}\left(\log \frac{N}{\delta} \right).$$

In the case that excess losses are subgamma,

$$E(h_n^*) \leq \tau \left(\sqrt{\frac{\log \frac{N}{\delta}}{n}} + \frac{L}{2} \frac{\log \frac{N}{\delta}}{n} \right)$$

which implies rates of order $n^{-1/2}$ for excess losses. The ubiquity of this type of rate is inherent to Cramer-Chernoff Method (see Appendix B).

Example 2.5.1 (Simple Normal Means Model)**.** Consider the task of estimating the mean of a sample $Z_1, \ldots, Z_n$ which is known to be generated from a normal distribution with variance 1 and unknown mean μ^*. Identify $\mathcal{H} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$ and use the maximum likelihood estimator μ_n^*

$$\mu_n^* = \arg\max_{\mu \in \mathbb{R}} \mathbb{P}_n[\ell(\mu, Z_i)] = \arg\min_{\mu \in \mathbb{R}} \frac{1}{2}\mathbb{P}_n[\mu - Z_i]^2.$$

Differentiation leads to

$$\mu_n^* = \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

which is normally distributed with variance $1/n$ and mean μ^*. Thus, applying the Cramér-Chernoff method to both tails leads to

$$\mathbb{P}\left\{|\mu_n^* - \mu^*| \geq \epsilon\right\} \leq 2\mathrm{e}^{-n\epsilon^2/2},$$

that is, μ_n^* tends to μ^* at a $n^{-1/2}$ rate. Since the expected excess loss can be written in this case as

$$E(\mu_n^*) = L(\mu_n^*) - L(\mu^*) = \frac{1}{2}(\mu_n^* - \mu^*)^2,$$

the previous bound translates into

$$\mathbb{P}\left\{L(h_n^*) - L(h^*) \geq \epsilon\right\} \leq 2\mathrm{e}^{-n\epsilon},$$

Thus the expected excess loss converges at a n^{-1} rate to zero. Note that even though $\mu_n^* \xrightarrow{\mathbb{P}} \mu^*$ at a $n^{-1/2}$ rate, the loss converges at a faster rate to zero.

2.6. The PAC-Bayesian Model

There are situations in which it is either desirable or natural for a learning algorithm to provide as an output a probability distribution over the class of hypotheses. Predictions are thus made according to a randomized hypothesis drawn from the output distribution so that the predictions no longer depend deterministically on the data. An example of such procedures are Gibbs learning algorithms, which make predictions according to a randomized hypothesis drawn from a data dependent distribution constructed in a pseudo-Bayesian fashion (see Remark 3.1.5). In addition to Gibbs algorithms, PAC-Bayesian bounds have been proven to lead to faster rates of convergence in situations in which little is known for ERM [Audibert, 2009]. Randomized algorithms and their PAC-Bayesian analysis have proven to be necessary to obtain nontrivial rates in situations in which data are not assumed to be iid and might even be adversarial [Cesa-Bianchi and Lugosi, 2006].

One model for such situation is the so-called *PAC-Bayesian* model, where PAC stands for *Probably Approximately Correct*. In this model, it is assumed that there exists an initial or *prior* distribution Π_0 over the hypotheses class $\mathcal{H}$. Given the data $Z_1, \ldots, Z_n$ an

algorithm outputs a *posterior* distribution Π_n, in analogy to the Bayesian terminology. Note that these distributions need not be related to each other through Bayes formula (see Remark 3.1.5) and can even be concentrated on a single element, in which case we recover the deterministic case (and consequently ERM). In this setting we are interested in studying the Π_n-expected excess loss

$$\Pi_n[E(h)] = \Pi_n \mathbb{P}[\varepsilon(h)]$$

We stress the fact that the distribution Π_n is allowed to depend on the sample $Z_1, \dots, Z_n$, turning $\Pi_n[E(h)]$ into a random variable.

We now recall the fundamental inequalities that have been used in PAC-Bayesian analysis.

2.6.1. PAC-Bayesian Inequalities

The first PAC-Bayesian bounds have been attributed to McAllester [1998] and overviews were given by McAllester [2013], Audibert [2004] and van Erven [2014]. We will prove a PAC-Bayesian result in Theorem 2.6.2. PAC-Bayesian theorems can be thought of as refined union bounds and are based in the careful study of the cumulant generating function of random variables. Remarkably, they hold in great generality, including for instance infinite hypotheses classes $\mathcal{H}$. The main tool is Donsker-Varadhan's variational formula, which we now recall [see Boucheron et al., 2013, Corollary 4.14].

Theorem 2.6.1 (Donsker-Varadhan Variational Formula)**.** *Let X be a real valued integrable random variable. Then for every $\eta \in \mathbb{R}$ and every distribution $\mathbb{P}$*

$$\log \mathbb{P} \mathrm{e}^{\eta X} = \sup_{\mathbb{Q}} \eta \mathbb{Q} X - \mathrm{KL}(\mathbb{Q}, \mathbb{P}) \tag{2.3}$$

where

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}) = \mathbb{Q}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] \tag{2.4}$$

is the Kullback-Leibler divergence and the supremum is taken with respect to all probability measures $\mathbb{Q}$ that are absolutely continuous[1] with respect to $\mathbb{P}$. Furthermore, the supremum is achieved at $\mathbb{Q} = \mathbb{P}_\eta$, where $\mathbb{P}_\eta$ is the probability measure with density

$$\frac{\mathrm{d}\mathbb{P}_\eta}{\mathrm{d}\mathbb{P}} = \frac{\mathrm{e}^{\eta X}}{\mathbb{P}[\mathrm{e}^{\eta X}]}$$

As a consequence, we have that for each distribution Π_n and Π_0 (with Π_n absolutely continuous with respect to Π_0) it holds that

$$\Pi_n X \leq \inf_{\eta > 0} \frac{1}{\eta} (\log \Pi_0[\mathrm{e}^{\eta X}] + \mathrm{KL}(\Pi_n, \Pi_0)). \tag{2.5}$$

[1] Remember that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ if and only if for any measurable set S, the fact that $\mathbb{P}(S) = 0$ implies that $\mathbb{Q}(S) = 0$.

This means that if we choose X carefully to be a suitable function of $E(h) - E_n(h)$, roughly speaking, the problem of bounding $\Pi_n[E(h) - E_n(h)]$ can be translated into the problem of bounding a cumulant generating function defined by

$$\psi_X(\eta) = \log \mathbb{P}[\mathrm{e}^{\eta X}], \tag{2.6}$$

a well-known provider of deviation inequalities (see Appendix B). Since we focus on excess losses that satisfy inequalities of the form (2.2), a natural choice for X is

$$X_h = \phi\left(\frac{n^{1/2}(E(h) - E_n(h))_+}{\tau}\right). \tag{2.7}$$

With this in mind, the following is the anticipated PAC-Bayesian inequality and its proof.

Theorem 2.6.2. *Let ε an excess loss function over the hypothesis class $\mathcal{H}$ that satisfies a concentration inequality of the form*

$$\mathbb{P}\left\{n^{1/2}(E(h) - E_n(h)) \geq t\right\} \leq \mathrm{e}^{-\phi(t/\tau)}$$

for $\phi = \phi_2$ or $\phi = \phi_{Bern^L}$. Then it holds that with $\mathbb{P}$-probability higher than $1 - \delta$

$$\Pi_n E(h) \leq \Pi_n E_n(h) + \frac{\tau}{n^{1/2}}\phi^{-1}\left(2\log\frac{2}{\delta} + 2\mathrm{KL}(\Pi_n, \Pi_0)\right)$$

Remark 2.6.3. In the subgaussian case we recover the result originally obtained by McAllester [1998]. Our proof is a reworking of the exposition of Boucheron et al. [2005, Section 6]. Their proof is given in the context of classification. We note that the only property that is used in their proof is the subgaussianity of excess losses, which is a consequence of the boundedness of the excess loss and Hoefding's inequality.

Proof. Take X_h as in (2.7). We will use (2.5), so we need to bound the quantity

$$\log \Pi_0[\mathrm{e}^{\eta X_h}].$$

Using Markov's inequality and Fubinni's theorem we obtain that for any $\epsilon > 0$ it holds that

$$\mathbb{P}\left\{\Pi_0[\mathrm{e}^{\eta X_h}] \geq \epsilon\right\} \leq \frac{\Pi_0 \mathbb{P}[\mathrm{e}^{\eta X_h}]}{\epsilon}.$$

Note that

$$\begin{aligned}
\mathbb{P}\mathrm{e}^{\eta X_h} &= 1 + \int_0^\infty \mathbb{P}\left\{\mathrm{e}^{\eta X_h} \geq t\right\} \mathrm{d}t \\
&= 1 + \int_0^\infty \mathbb{P}\left\{X_h \geq t/\eta\right\} \mathrm{e}^t \mathrm{d}t \\
&= 1 + \int_0^\infty \mathbb{P}\left\{n^{1/p}(E(h) - E_n(h)) \geq \tau\phi^{-1}(t/\eta)\right\} \mathrm{e}^t \mathrm{d}t \\
&\leq 1 + \int_0^\infty \mathrm{e}^{-t(1/\eta - 1)} \mathrm{d}t \\
&= 1 + \frac{\eta}{1-\eta}
\end{aligned}$$

so that picking

$$\eta = \frac{1}{2}$$

and

$$\delta = \frac{2}{\epsilon}$$

one obtains that with probability higher than $1 - \delta$

$$\Pi_0[\mathrm{e}^{\eta X_h}] \leq \frac{2}{\delta}.$$

Using Jensen's inequality, (2.5) and the inequality that we just obtained one obtains that with probability higher than $1 - \delta$

$$\phi\left(\Pi_n\left[\frac{n^{1/2}(E(h) - E_n(h))_+}{\tau}\right]\right) \leq 2\left(\log\frac{2}{\delta} + \mathrm{KL}(\Pi_n, \Pi_0)\right).$$

Inversion, rearrangement and the fact that $x \leq x_+$ lead to the result. □

Remark 2.6.4. Note that this bound holds for general hypotheses classes and for arbitrary distributions Π_n. Thus, in the case that $n^{1/2}(E(h) - E_n(h))$ has a subgaussian right tail with scale parameter τ, then the bound reads

$$\Pi_n E(h) \leq \Pi_n E_n(h) + \tau\sqrt{\frac{2\log\frac{2}{\delta} + 2\mathrm{KL}(\Pi_n, \Pi_0)}{n}}$$

while in the case that $n^{1/2}(E(h) - E_n(h))$ is sub gamma then

$$\Pi_n E(h) \leq \Pi_n E_n(h) + \tau\left(\sqrt{\frac{2\log\frac{2}{\delta} + 2\mathrm{KL}(\Pi_n, \Pi_0)}{n}} + L\frac{\log\frac{2}{\delta} + \mathrm{KL}(\Pi_n, \Pi_0)}{n}\right),$$

that is, a $n^{-1/2}$ rate in both cases.

Remark 2.6.5. This inequality can be applied also for the analysis of ERM in the case that $\mathcal{H}$ is finite. Indeed, suppose that Π_0 is the uniform distribution on $\mathcal{H}$, then if Π_n is the distribution concentrated at the ERM h_n^*, Theorem 2.6.2 gives that

$$E(h_n^*) \leq \frac{\tau}{n^{1/2}}\phi^{-1}\left(2\log\frac{2|\mathcal{H}|}{\delta}\right)$$

in the case that the excess losses are either subgaussian ($\phi = \phi_2$) or subgamma ($\phi = \phi_{\mathrm{Bern}}^L$).

As we have seen, this PAC-Bayesian inequality makes use of the tail behavior of the excess losses. In the first section of the next chapter we will use a different approach to obtaining such inequalities.

3. Toward Fast Rates

In this section we present a second risk bound derived by Grünwald and Mehta [2016] based on an inequality previously obtained by Zhang [2006b], itself a variation of the PAC-Bayesian inequalities studied in Section 2.6. These excess risk bounds hold in the case that two conditions are met and they lead to potentially fast rates. The first of these two conditions (Condition 3) is on the relative weight of the right tail of the excess losses and is called *witness condition*. The second condition is called *(strong) central condition* and is a uniform condition on the rate of decrease of the cumulant generating function of the negative excess losses close to the origin.

In Section 3.1 we introduce Zhang's inequality. We explain how this inequality can be used either as a bound on the expected loss, from which the results obtained in Section 2.6 can be recovered, and how it was used both by Zhang and by Grünwald and Mehta as a bound on a quantity closely related to the cumulant generating function of the excess loss, its *annealed expectation* (which we define below). In Section 3.2 we introduce the conditions that Grünwald and Mehta used to obtain fast bounds using Zhang's inequality, and that will be the main topic of analysis of Chapter 4. In Section 3.3 we present Grünwald and Mehta's bound, which leads to fast concentration and explain in Example 3.3.5 how it can be used to obtain fast rates for ERM in the case that losses are Lipschitz continuous under the conditions introduced in Section 3.2.

3.1. Second PAC-Bayesian Inequality

In this section we consider another type of PAC-Bayesian Inequalities. They have been attributed to Zhang [2006a] and Zhang [2006b]. These inequalities are are also based on the use of Donsker-Varadhan's variational formula from Theorem 2.6.1. Later, Grünwald and Mehta [2016] used these formulas to derive rates of convergence for unbounded loss functions under conditions on the tails of the excess losses and its cumulant generating function. Just as it was the case in the previous section, the inequalities that we explain here relate the cumulant generating function of $E(h) - E_n(h)$ and the Kulback-Leibler divergence and they can be used to bound the expected excess loss directly. But now, these inequalities will be used as a bound on the *annealed expectation*, which at the same time and under additional conditions can be used to bound the expected excess loss. We start by explaining some consequences of Zhang's inequality, the main element of this discussion, which we then prove it in Theorem 3.1.1.

Zhang's inequality implies that for each $\delta \in (0,1)$ with probability higher than $1-\delta$ the relation

$$\Pi_n\left[-E_n(h) - \frac{1}{\eta}\psi_{-\varepsilon(h)}(\eta)\right] - \frac{\mathrm{KL}(\Pi_n, \Pi_0)}{\eta n} \leq \frac{\log\frac{1}{\delta}}{\eta n} \tag{3.1}$$

holds for every choice of prior Π_0 and posterior Π_n, where $\psi_{-\varepsilon(h)}(\eta)$ is the cumulant generating function of $-\varepsilon(h)$ as defined in (2.6) on page 20. One could rewrite this inequality as

$$\Pi_n\left[E(h)\right] \leq \Pi_n\left[E_n(h) + \frac{1}{\eta}\psi_{E(h)-\varepsilon(h)}(\eta)\right] + \frac{\mathrm{KL}(\Pi_n, \Pi_0) + \log\frac{1}{\delta}}{\eta n}$$

and use it as a bound on the expected excess risk (see Remark 3.1.2). Instead, Grünwald and Mehta [2016] (and Zhang himself in his original article) use it as a bound on

$$A_\eta(h) = -\frac{1}{\eta}\psi_{-\varepsilon(h)}(\eta) = -\frac{1}{\eta}\log\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h)}]. \tag{3.2}$$

which is called the *annealed expectation* of the excess loss, or *free energy* in the statistical mechanics community. Additionally define the *Information Complexity* $\mathrm{IC}(\eta, n)$ as

$$\mathrm{IC}(\eta, n) = \Pi_n E_n(h) + \frac{\mathrm{KL}(\Pi_n, \Pi_0)}{\eta n}. \tag{3.3}$$

which is a combination of the empirical excess loss and the Kullback Leibler divergence, as defined in (2.4). This rearrangement leads to the conclusion that

$$\Pi_n\left[A_\eta(h)\right] \leq \mathrm{IC}(\eta, n) + \frac{\log\frac{1}{\delta}}{\eta n}.$$

with probability higher than $1-\delta$. Grünwald and Mehta [2016] and others have observed that rates of convergence for $\mathrm{IC}(\eta, n)$ can be obtained as $n \to \infty$. Indeed, if concentration inequalities are available for the excess losses $\varepsilon(h)$, rates for $\Pi_n E_n(h)$ can be obtained. At the same time, randomized algorithms can be designed with appropriate choice of prior Π_0 and posteriors Π_n which lead to a fast decrease of the KL term. We show exactly this in the Example 3.3.5.

Nevertheless, even if it is possible to obtain rates of convergence for the annealed expectation, this does not imply the same for the expected excess loss, the main quantity of interest. Recall that by Jensen's inequality $A_\eta(h) \leq E(h)$, while in light of our previous observations we are interested in bounds on the opposite direction, that is, bounds of the form $E(h) \lesssim A_\eta(h)$. Accordingly, Grünwald and Mehta [2016] found two additional mild conditions under which such an inequality can be derived (see Theorem 3.3.1). Those conditions will be the focus of Section 3.2.

With this in mind, we give the Theorem due to Zhang [2006b, Lemma 2.1], which implies (3.1) at the beginning of this discussion.

Theorem 3.1.1 (Information Exponential Inequality)**.** *Let ε an excess loss function over the hypothesis class $\mathcal{H}$. Then for any prior distribution Π_0 and any posterior Π_n it holds that*

$$\mathbb{P}\left[\exp\left(\eta n \Pi_n\left[A_\eta(h) - \mathrm{IC}(\eta, n)\right]\right)\right] \leq 1.$$

Proof. The starting point is again Donsker-Varadhan's variational formula as described in Theorem 2.6.1 and its consequence from (2.5), which we rewrite with ηn instead of η and rearrange as

$$\eta n \Pi_n[X] - \mathrm{KL}(\Pi_n, \Pi_0) \leq \log \Pi_0[\mathrm{e}^{\eta n X}].$$

for any $\eta > 0$ and any random variable X. Taking exponentials and taking $\mathbb{P}$-expectations we obtain that

$$\mathbb{P}[\mathrm{e}^{\eta n \Pi_n X - \mathrm{KL}(\Pi_n, \Pi_0)}] \leq \mathbb{P}\Pi_0[\mathrm{e}^{\eta n X}]. \tag{3.4}$$

In (3.4) take

$$\begin{aligned} X_h &= -E_n(h) - \frac{1}{\eta n}\psi_{-E_n(h)}(\eta n) \\ &= -E_n(h) - \frac{1}{\eta}\psi_{-\varepsilon(h)}(\eta) \\ &= A_\eta(h) - E_n(h). \end{aligned}$$

This choice and Fubini's theorem makes the right hand side of the inequality in (3.4) smaller than one and thus the result follows. □

The main consequence of Theorem 3.1.1 is that it implies inequalities both in probability and in expectation for $\Pi_n[A_\eta(h) - \mathrm{IC}(\eta, n)]$. Indeed, by Markov's inequality

$$\begin{aligned} \mathbb{P}\left\{\Pi_n\left[A_\eta(h) - \mathrm{IC}(\eta, n)\right] \geq t\right\} &= \mathbb{P}\left\{\exp\left(\eta n \Pi_n\left[A_\eta(h) - \mathrm{IC}(\eta, n)\right]\right) \geq \mathrm{e}^{\eta n t}\right\} \\ &\leq \mathrm{e}^{-\eta t}. \end{aligned}$$

This means that with probability higher than $1 - \delta$ it holds that

$$\Pi_n\left[A_\eta(h) - \mathrm{IC}(\eta, n)\right] \leq \frac{\log \frac{1}{\delta}}{\eta n},$$

which is just a rewriting of (3.1). Additionally, by Jensen's inequality

$$\mathbb{P}\left[\Pi_n\left[A_\eta(h) - \mathrm{IC}(\eta, n)\right]\right] \leq 0.$$

Remark 3.1.2. The inequality obtained in Theorem 3.1.1 can alternatively be written as

$$\mathbb{P}\left[\exp\left(\eta n \Pi_n\left[E(h) - E_n(h) - \frac{1}{\eta}\psi_{E(h)-\varepsilon(h)}(\eta) - \frac{\mathrm{KL}(\Pi_n, \Pi_0)}{\eta n}\right]\right)\right] \leq 1 \tag{3.5}$$

By our previous remark, this implies that with probability higher than $1-\delta$ the inequality

$$\Pi_n[E(h)] \leq \Pi_n\left[E_n(h) + \frac{1}{\eta}\psi_{E(h)-\varepsilon(h)}(\eta)\right] - \frac{\mathrm{KL}(\Pi_n, \Pi_0) + \log \frac{1}{\delta}}{\eta n}$$

holds. This implies two things. First, that a bound on the cumulant generating function of the centered excess loss $E(h) - \varepsilon(h)$ can be turned into bound on the excess loss. Second, that since we have the freedom to choose η, this bound can be optimized. Even though it is not novel, we think it is instructive to also show how such bounds can be derived in this way. We do this in the following corollary, which is the same result as the obtained in Theorem 2.6.2 up to constants.

Corollary 3.1.3. *Let ε such that for $\varepsilon(h) - \mathbb{P}[\varepsilon(h)]$ a uniform subgamma bound holds (see Section 2.5). Then, with probability higher than $1-\delta$,*

$$\Pi_n\left[E(h)\right] \leq \Pi_n\left[E_n(h)\right] + \sqrt{2v\frac{\mathrm{KL}(\Pi_n,\Pi_0)+\log\frac{1}{\delta}}{n}} + c\frac{\mathrm{KL}(\Pi_n,\Pi_0)+\log\frac{1}{\delta}}{n}.$$

for positive constants v and c and for all prior distributions Π_0 and posterior distributions Π_n.

Remark 3.1.4. Notice that with $\tau = \sqrt{2v}$ and $L/2 = c/\sqrt{2v}$ we obtain the same bound as we would have obtained in Theorem 2.6.2 up to constants.

Proof. Recall that if an excess loss function has a subgamma right tail, then there exists v and c such that (see Appendix C)

$$\psi_{E(h)-\varepsilon(h)}(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta} \tag{3.6}$$

for each $h \in \mathcal{H}$ and for $0 < c\eta < 1$. Because of Bernstein's moment condition, the bound from (3.6) holds for every $\eta \in [0, 1/c)$. By the previous remark and and (3.6) one obtains that with probability higher than $1-\delta$

$$\Pi_n\left[E(h)\right] \leq \Pi_n\left[E_n(h)\right] + \frac{1}{2}\frac{v\eta}{1-c\eta} + \frac{\mathrm{KL}(\Pi_n,\Pi_0)+\log\frac{1}{\delta}}{\eta n}.$$

as long as $0 < c\eta < 1$. Since this inequality holds for all η, we can take η so as to minimize the right hand. Since the function $x \mapsto \frac{x}{1-Cx} + \frac{K}{x}$ is minimized at $x^* = \frac{\sqrt{K}}{1+C\sqrt{K}}$ and has maximum equal to $2\sqrt{K} + CK$ we obtain the result. □

Remark 3.1.5 (Information Risk Minimization). Zhang [2006b] used his inequality to in order to produce randomized algorithms according to what he called *Information Risk Minimization*. Note three things. First, if we write (3.1) explicitly in terms of the losses ℓ and their expected values L, we obtain that with probability higher than $1-\delta$ and for every choice of prior Π_0 and posterior Π_n it holds that

$$\Pi_n\left[-\frac{1}{\eta}\psi_{-\ell(h)}(\eta)\right] \leq \Pi_n\left[L_n(h)\right] + \frac{\mathrm{KL}(\Pi_n,\Pi_0)}{\eta n} + L(h^*) - L_n(h^*) + \frac{\log\frac{1}{\delta}}{\eta n}. \tag{3.7}$$

Second, by Jensen's inequality the left hand side of this equation is smaller than the expected loss $\Pi_n L(h)$. Third, the right hand consists of two parts: one which does

not depend on Π_n, and another one that can be interpreted as a penalized empirical risk. Indeed, it is the sum of the empirical risk and the Kullback Leibler divergence, a measure of the inefficiency of using the distribution Π_0 on $\mathcal{H}$ when the true distribution is Π_n [see Cover and Thomas, 2006, Section 2.3]. Consequently, if one is to minimize the right hand side over the choices of Π_n, from Donsker Varadhan's variational formula it follows that the minimum is attained at the distribution Π_n^* with density with respect to Π_0 given by

$$\frac{\mathrm{d}\Pi_n^*}{\mathrm{d}\Pi_0} = \frac{\mathrm{e}^{-\eta \sum_{i=1}^n \ell_i(h)}}{\Pi_0[\mathrm{e}^{-\eta \sum_{i=1}^n \ell_i(h)}]}.$$

Information Risk Minimization is exactly the procedure of choosing the posterior Π_n^* in this way. Notice that in the case of density estimation (see Example 2.2.2) the loss is the negative log-likelihood, which turns the posterior distribution Π_n^* into the Bayesian posterior by taking $\eta = 1$.

3.2. Conditions for Faster Concentration

In the previous section we proved bounds on the annealed expectation $A_\eta(h)$ of the excess loss (see its definition in (3.2) on page 23) and noticed that in general it is smaller than the expected excess loss $E(h)$ by Jensen's inequality. We anticipated that, under suitable conditions, Grünwald and Mehta [2016] obtained an estimate of the form $E(h) \lesssim A_\eta(h)$. This section is about those conditions, and an additional condition, Bernstein's condition, which has been proven to lead to fast bounds.

We start by describing Bernstein's condition. We show how it leads to fast rates in the case that excess losses are bounded and hypothesis classes are finite and cite results that extend this to infinite classes. After that, we describe two conditions used by Grünwald and Mehta [2016] to obtain a bound on the expected excess loss through this method. These two conditions are the *Strong Central Condition* (Condition 2 on page 28), and the *Witness Condition* (Condition 3 on page 29). We recall Grünwald and Mehta's exact result in Theorem 3.3.1 in the next section, where we also discuss its implications.

3.2.1. Bernstein's Condition

This condition leads to advantages when using Bernstein's inequality (see Appendix C), hence the name. The condition reads

Condition 1. The excess losses $\varepsilon(h) = \ell(h) - \ell(h^*)$ on a hypotheses class $\mathcal{H}$ with expected values $E(h) = \mathbb{P}[\varepsilon(h)]$ satisfy for all $h \in \mathcal{H}$

$$\mathbb{P}[\varepsilon^2(h)] \leq B(\mathbb{P}[\varepsilon(h)])^\beta$$

for some $B > 0$ and some $0 \leq \beta \leq 1$.

Remark 3.2.1. We refer to the condition with $\beta = 1$, which is its strongest version, simply as Bernstein's condition. This is because we are mainly interested in its implications in

this case. We will mention explicitly when we use other values of β. In order to avoid confusion we refer to it simply as Bernstein's condition while we refer to condition in Theorem C.0.1 on page 53, which is also commonly associated with Bernstein's name, as Bernstein's *moment* condition.

In order to motivate why Condition 1 leads to faster rates, let $\mathcal{H}$ be a finite hypotheses class of size $|\mathcal{H}| = N$ and suppose that the excess loss ε is uniformly bounded from above by some positive $b < \infty$, that is $\varepsilon(h) \leq b$ almost surely for all $h \in \mathcal{H}$. Then, Bernstein's inequality and the union bound gives that with probability higher than $1-\delta$ it holds for each $h \in \mathcal{H}$ that

$$E(h) \leq E_n(h) + \sqrt{2\mathbb{P}[\varepsilon^2(h)]\frac{\log\frac{N}{\delta}}{n}} + \frac{b}{3}\frac{\log\frac{N}{\delta}}{n}.$$

If as in Condition 1 it holds for each $h \in \mathcal{H}$

$$\mathbb{P}[\varepsilon^2(h)] \leq BE(h)^\beta$$

for fixed $\beta > 0$ and $B > 0$, then also using the fact that at the empirical risk minimizer h_n^* empirical risk $E_n(h_n^*)$ is negative one obtains that

$$E(h_n^*) \leq \sqrt{2BE(h_n^*)^\beta\frac{\log\frac{N}{\delta}}{n}} + \frac{b}{3}\frac{\log\frac{N}{\delta}}{n}.$$

Rearranging this one obtains that

$$E(h_n^*) \leq \left(2B\frac{\log\frac{N}{\delta}}{n}\right)^{\frac{1}{2-\beta}},$$

which implies faster rates than $n^{-1/2}$ in the case that $\beta > 0$. This result was extended to infinite classes by Bartlett and Mendelson [2006]. Their argument is more involved and uses a generalization of Bernstein's inequality due to Talagrand [1994]. In the case that losses are unbounded, to our best knowledge, it is still an open question whether Condition Bernstein's condition leads to fast rates for ERM. Nevertheless, Audibert [2009] found a nonstandard randomized algorithm that achieves fast rates if it holds.

In the special case of classification, Bernstein's condition has been known to also lead to faster rates. It can be shown that it is implied by a well studied condition found by Mammen and Tsybakov [1999], which also leads to fast rates [see also Boucheron et al., 2005, Section 5].

Remark 3.2.2. For finite classes Bernstein's condition holds every time that the expected loss minimizer $h^* = \arg\min_h L(h)$ is unique and the excess losses are bounded from above. Indeed, in that case for each $h \neq h^*$ it holds that for some $m > 0$ the expected loss is bigger than m, that is, $\mathbb{P}[\varepsilon(h)] > m$. On the other hand, by the boundedness of the excess losses, there is some b such that $\mathbb{P}[\varepsilon^2(h)] \leq b$. This means that if $h \neq h^*$, then

$$\mathbb{P}[\varepsilon^2(h)] \leq \frac{b^2}{m}\mathbb{P}[\varepsilon(h)],$$

that is, Bernstein's condition holds for $B = b^2/m$.

Still, in practical situations, with finite classes one might not always expect a fast rate even if the expected loss minimizer h^* is unique (see Remark 4.0.1). In those cases Bernstein's condition becomes harder to satisfy.

3.2.2. The Central Condition

The central condition is a uniform condition on the rate of decrease near the origin of the cumulant generating function of $-\varepsilon(h)$. We will show that it leads to fast rates for the empirical risk of ERM and even though we will mainly focus on its strong version, we cite also its weaker version as proposed by Grünwald and Mehta [2016], which leads to intermediate rates.

Condition 2 (The Central Condition)**.** The excess loss $\varepsilon(h)$ on the hypotheses class $\mathcal{H}$ satisfies the central condition if there exists a bounded non-decreasing function $\eta : [0, \infty) \to [0, \infty)$ and an element $h^* \in \mathcal{H}$ so that for all $\epsilon > 0$, for $\eta = \eta(\epsilon)$

$$\sup_{h \in \mathcal{H}} \psi_{-\varepsilon(h)}(\eta) \leq \eta\epsilon. \tag{3.8}$$

If it happens that η can be chosen not to depend on ϵ, then

$$\sup_{h \in \mathcal{H}} \psi_{-\varepsilon(h)}(\eta) \leq 0.$$

for some $\eta > 0$ and we say that the *strong central condition* holds.

Remark 3.2.3. In both its strong and weaker version after using Jensen's inequality, the central condition implies that for any h

$$L(h) \geq L(h^*),$$

that is, h^* minimizes the expected loss.

Example 3.2.4 (Maximum Likelihood Estimation and the Central Condition)**.** In the case of well specified maximum likelihood estimation the strong central condition is satisfied. Indeed, if $\mathcal{P}$ is a set of probability densities and for $p \in \mathcal{P}$ the loss function $\ell(p) = -\log p$ and $\mathbb{P}$ has density p^*, then the the cumulant generating function of the excess loss $\varepsilon(p) = \ell(p) - \ell(p^*)$ has cumulant generating function $\psi_p(\eta)$ given by

$$\psi_{-\varepsilon(p)}(\eta) = \log \mathbb{P}\mathrm{e}^{\eta(\log(p) - \log(p^*))} = \log \mathbb{P}\left[\frac{p}{p^*}\right]^{\eta}$$

which equals zero for $\eta = 1$.

Remark 3.2.5. From the central condition it is possible to derive rates at which the empirical excess loss tends to zero. The rate at which $\eta(\epsilon)$ decreases as $\epsilon \downarrow 0$ determines its

rate of convergence. Consider a hypotheses class of size $\mathcal{H} = N$. If the central condition holds we can write for any $\epsilon > 0$ and for each $h \in \mathcal{H}$

$$\mathbb{P}\left\{E_n(h)_- \geq 2\epsilon\right\} \leq \mathrm{e}^{-n(2\eta\epsilon - \psi_{-\varepsilon(h)}(\eta))} \leq \mathrm{e}^{-n\eta\epsilon}$$

with $\eta = \eta(\epsilon)$. The union bound implies that

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} E_n(h)_- \geq 2\epsilon\right\} \leq N\mathrm{e}^{-n\eta\epsilon}$$

Thus, if $\eta(\epsilon) = O(\epsilon^\beta)$ as $\epsilon \downarrow 0$ for some $\beta \in [0, 1]$, we can conclude that for n big enough with probability higher than $1 - \delta$ the inequality

$$|E_n(h_n^*)| \leq \left(\frac{1}{n}\log\frac{N}{\delta}\right)^{\frac{1}{1+\beta}}$$

holds, that is, a $n^{-\frac{1}{1+\beta}}$ rate. In the case that the strong version of the central condition holds, a similar analysis and a change of variable leads to the conclusion that the identity

$$\mathbb{P}\left\{nE_n(h)_- \geq t\right\} \leq \mathrm{e}^{-\eta t/2}$$

holds *for all* values of t. Corollary D.0.5 can be used to obtain n^{-1} rates of convergence for $E_n(h_n^*)$ in the case that the excess loss is Lipschitz.

As we showed in the previous remark, rates for the expected loss of ERM can be obtained from the central condition. Nevertheless the main quantity of interest is the expected excess loss. We show in the next subsection an additional condition that allows us to do just that.

3.2.3. The Witness Condition

The second condition is implied by Condition 1 with $\beta = 1$ and is an uniform condition on the relative weight of the upper tail of the excess loss. It was named *Witness Condition* by Grünwald and Mehta [2016] and it reads:

Condition 3 (Witness Condition)**.** The excess loss ε on the hypotheses class $\mathcal{H}$ satisfies the witness condition if for all $h \in \mathcal{H}$ it is true that

$$\sup_{h\in\mathcal{H}} \frac{\mathbb{P}[\varepsilon(h)\mathbb{1}\left\{\varepsilon(h) > u\right\}]}{\mathbb{P}[\varepsilon(h)]} \leq c$$

for some $u > 0$ and $c \in (0, 1]$.

Remark 3.2.6. Condition 3 holds for bounded excess losses trivially and proving that Condition 1 with $\beta = 1$ implies Condition 3 (that Bernstein's condition implies the

witness condition) is not hard. Indeed, using Hölder's and Markov's inequalities one obtains

$$\begin{aligned}\mathbb{P}[\varepsilon(h)]\mathbb{1}\left\{\varepsilon(h) > u\right\} &\leq \mathbb{P}[\varepsilon^2(h)]^{1/2}(\mathbb{P}\left\{\varepsilon(h) > u\right\})^{1/2} \\ &\leq \frac{1}{u}\mathbb{P}[\varepsilon^2(h)] \\ &\leq \frac{B}{u}\mathbb{P}[\varepsilon(h)]\end{aligned}$$

so that it is enough to choose $u > B$.

3.3. Expected Excess Loss Bound

Under Conditions 3 and the strong version of Condition 2, Grünwald and Mehta [2016, Lemma 16] proved the following, which, chained with the inequality from Theorem 3.1.1, leads to potentially faster rates of convergence as we show in Corollary 3.3.3. We show how these ideas can be applied to ERM in Example 3.3.5.

Theorem 3.3.1. *Let $\bar{\eta}$ such the strong central condition (see Condition 2) holds. Let $u > 0$ and $c \in (0, 1)$ such that the witness condition (Condition 3) holds. Then*

$$E(h) \leq \mu(\eta) A_\eta(h)$$

with

$$\mu(\eta) = \frac{1}{c}\frac{\eta u + 1}{1 - \eta/\bar{\eta}} \tag{3.9}$$

for $\eta \in (0, \bar{\eta})$.

Remark 3.3.2. Note that the two assumptions of this theorem are very different in nature. On the one hand the strong central condition implies the existence of all moments of $\varepsilon(h)_-$ while the witness condition can hold even if the second moment of $\varepsilon(h)_-$ does not exist.

It is possible to chain both the preceding theorem and Theorem 3.1.1 to obtain an expected excess risk bound:

Corollary 3.3.3. *Under the same conditions as those of of Theorem 3.3.1 it holds that with probability higher than $1 - \delta$*

$$\Pi_n[E(h)] \leq \mu(\eta)\left(\mathrm{IC}(\eta, n) + \frac{1}{\eta n}\log\frac{1}{\delta}\right)$$

where μ is as defined in (3.9) and Π_n that is absolutely continuous with respect to Π_0.

Remark 3.3.4. In the case that the excess losses are bounded the witness condition is trivially satisfied. This means that the strong central condition is sufficient for obtaining fast rates. Indeed, Van Erven et al. [2015] proved that Bernstein's and the strong central condition are equivalent for bounded losses.

Example 3.3.5 (A Bound for ERM Using Corollary 3.3.3)**.** Consider a loss function ℓ on a hypotheses class $\mathcal{H}$ equipped with a semi-norm d, a sigma algebra and a finite measure μ. Suppose that ℓ is Lipschitz continuous with Lipschitz constant L, that is,

$$|\ell(h_1) - \ell(h_2)| \leq L d(h_1, h_2)$$

almost surely for each $h_1, h_2 \in \mathcal{H}$. Suppose that for each $\epsilon > 0$ the covering number $N = N(\mathcal{H}, d, \epsilon)$ (see Appendix D for the definition of covering number) is finite and satisfies

$$\log N(\mathcal{H}, d, \epsilon) = O(\log \epsilon^{-1})$$

as $\epsilon \downarrow 0$. Lastly, Suppose that the conditions of Theorem 3.3.1 (and consequently those of Corollary 3.3.3) are satisfied.

We obtain rates for ERM by constructing an algorithm by specifying a prior probability Π_0 and a posterior probability Π_n. In some cases, the same bound without a $\log n$ factor can be obtained using *chaining*. This technique is considered in Appendix D but we do not consider its use for obtaining fast rates explicitly. Choose Π_0 to be uniform over $\mathcal{H}$ so that $\mathrm{d}\Pi_0 = \mu(\mathcal{H})^{-1}\mathrm{d}\mu$. Choose Π_n in to be the uniform distribution on a ball of radius ϵ around the ERM h_n^*. Notice that Π_n makes predictions in the set of elements that are at distance smaller than ϵ to $\mathcal{H}$. Call the ϵ-ball around h_n^* as $B_\epsilon(h_n^*)$, then

$$\mathrm{d}\Pi_n = \mathbb{1}\left\{B_\epsilon(h_n^*)\right\} \frac{d\mu}{\mu(B_\epsilon(h_n^*))}.$$

Consider a random sample $Z_1, \ldots, Z_n$. Because ℓ is Lipschitz

$$L(h_n^*) \leq \Pi_n[L(h)] + L\epsilon,$$

so that

$$E(h_n^*) \leq \Pi_n[E(h)] + L\epsilon$$

and we can use Corollary 3.3.3. In order to do so, we need to bound the Kullback Leibler divergence $\mathrm{KL}(\Pi_n, \Pi_0)$ and the empirical risk $\Pi_n[E_n(h)]$. First, note that

$$\frac{\mathrm{d}\Pi_n}{\mathrm{d}\Pi_0} = \frac{\mu(\mathcal{H})}{\mu(B_\epsilon(\hat{h}_n^*))} \mathbb{1}\left\{B_\epsilon(\hat{h}_n^*)\right\}$$

which implies that the Kulback Leibler divergence satisfies

$$\mathrm{KL}(\Pi_n, \Pi_0) = \log \frac{\mu(\mathcal{H})}{\mu(B_\epsilon(\hat{h}_n^*))} \leq \log N(\mathcal{H}, d, \epsilon).$$

On the other hand the empirical excess risk $\Pi_n[E_n(h)]$ also satisfies

$$\Pi_n E_n(h) \leq E_n(h_n^*) + L\epsilon \leq L\epsilon$$

because at the empirical risk minimizer $E_n(h_n^*) \leq 0$. This means that, using Corollary 3.3.3, the excess risk satisfies with probability higher than $1 - \delta$

$$E(h_n^*) \leq \mu(\eta) \left(L\epsilon + \frac{\log N(\mathcal{H}, d, \epsilon) + \log \frac{1}{\delta}}{\eta n} \right)$$

for each $\epsilon > 0$, each η sufficiently small and μ as defined in (3.9). Thus, in the last equation one can choose ϵ to depend on n as $\epsilon = 1/n$ and the assumption on the metric entropy leads to the conclusion that

$$E(h_n^*) = O_{\mathbb{P}}(n^{-1} \log(n)).$$

4. Relation Among Conditions

This chapter contains the main contributions of this work. We start by highlighting the panorama so far, and how our contribution fits into it. We first describe what has been established so far in the case of unbounded and unbounded excess losses. Initially, whenever we refer to the possibility of obtaining a fast (or slow) rate, we refer to the case when $\mathcal{H}$ is finite, for ease of exposition. We will point at an extension of the situation of finite classes and comment on infinite classes on Remark 4.0.1.

Bounded Excess Losses The case that the excess losses are bounded, that is,

$$\sup_{h \in \mathcal{H}} |\varepsilon(h)| < \infty$$

almost surely has been thoroughly studied and the following has been established.

1. Bernstein's condition (Condition 1 on page 26) leads to fast rates of convergence for ERM (see Section 3.2.1).
2. Bernstein's condition is equivalent to the strong central condition (Condition 2 on page 28) [Van Erven et al., 2015] and it leads to fast rates (see Remark 3.3.4 after Corollary 3.3.3)
3. Neither the strong central condition nor Bernstein's condition hold automatically.
4. Slow rates hold without any additional condition and are provable via Bernstein's or Hoeffding's inequality.

Unbounded Excess Losses Little is known about sufficient conditions for fast rates in the case that excess losses might be unbounded but

$$\sup_{h \in \mathcal{H}} \mathbb{P}[\varepsilon(h)] < \infty.$$

1. Audibert [2009] showed a nonstandard randomized algorithm that achieves fast rates if Bernstein's condition holds. The same result is, to our best knowledge, unknown for ERM.
2. If both the strong central condition (Condition 2 on page 28) and the witness condition (Condition 3 on page 29) hold simultaneously, fast rates are possible in the PAC-Bayesian setting (see Corollary 3.3.3) and for ERM.

Our Contribution We make a number of contributions in the unbounded case, that is, when excess losses might be unbounded but

$$\sup_{h \in \mathcal{H}} \mathbb{P}[\varepsilon(h)] < \infty$$

1. We prove by way of counterexample that Bernstein's condition does not imply the strong central condition even if excess losses are forced to have exponential tails (Counterexample 4.1.1).
2. We show an excess loss function for which both the strong central condition and the witness condition hold but Bernstein's condition does not (Counterexample 4.1.4).

We considered the situation in which

$$\sup_{h \in \mathcal{H}} \psi_{-\varepsilon(h)}(\eta) < \infty \tag{4.1}$$

for some $\eta > 0$. This is the situation in which the probability of any fixed hypothesis being better than the best hypothesis is exponentially small, a natural weakening of the bounded case. The following results are proven under this assumption.

3. We find a new condition (Condition 4 on page 39) on the moments of the excess loss under which Bernstein's condition and the strong central condition become equivalent, just as in the bounded case. In consequence, this new condition can replace the witness condition in Corollary 3.3.3 in order to obtain fast rates.
4. We note that if both Bernstein's condition and the new condition from the last item (Condition 4) hold, then a condition (Condition 5 on page 41), under which Vapnik [1998] proved expected loss bounds for ERM, holds.
5. Using Vapnik's condition (Condition 5) we obtain fast rates by proving a tighter analogue (Theorem 4.2.4) of Corollary 3.3.3, improving the result of Grünwald and Mehta [2016], although under a stronger assumption.
6. We prove that the witness condition or the right tail of the excess loss having bounded second moment is sufficient to obtain slow rates (see Subsection 4.3).

Remark 4.0.1. The case in which $\mathcal{H}$ is finite also includes the situation in which one might allow the size of $\mathcal{H}$ to depend on the number of data points. More formally, we have at hand a sequence of hypotheses classes $\mathcal{H}_1, \mathcal{H}_2, \ldots$ and, for each sample of size n we learn $\hat{h}_n$ from the n-th hypotheses class $\mathcal{H}_n$, that is, $\hat{h}_n \in \mathcal{H}_n$. If the conditions that we discuss here hold on $\mathcal{H} = \cup_n \mathcal{H}_n$ and the size of each $\mathcal{H}_i$ remains bounded by some constant N, then the analysis and the rates that we present also apply predictions made according to $\hat{h}_n \in \mathcal{H}_n$.

All results mentioned for fast (or slow) rates can be extended up to $\log n$ factors for ERM in the case where $\mathcal{H}$ is infinite and has logarithmic metric entropy (see Definition D.0.1) with respect to a suitable semimetric and the losses satisfy a Lipschitz condition (see Example 3.3.4).

4.1. Strong Central, Witness and Bernstein's Conditions

We have seen that in cases when both the witness condition (Condition 3 on page 29) and the central condition (Condition 2 on page 28) hold fast rates are possible. As shown by Audibert [2009] and Bartlett and Mendelson [2006], the same occurs when Bernstein's condition (Condition 1 on page 26) holds. Thus the question of their equivalence arises. In the case that the losses are totally bounded, that is, in the case that $\mathbb{P}$-almost surely for some $M < \infty$

$$\sup_{h \in \mathcal{H}} |\varepsilon(h)| < M,$$

Van Erven et al. [2015] showed that Bernstein's and the central condition are equivalent.

In the case in which excess losses might be unbounded, these two conditions cannot be equivalent without further restrictions. First, Bernstein's condition cannot imply the strong central condition as the latter implies the existence of all moments of $\varepsilon(h)_- = \max\{-\varepsilon(h), 0\}$ and Bernstein's condition is only a relationship between the first and the second moments of $\varepsilon(h)$. It is then enough to consider an excess loss for which $\operatorname{Var} \varepsilon(h) < \infty$ but for which the third moment of its left tail is infinite, that is, $\mathbb{P}[(-\varepsilon(h)_-)^3]$ is infinite. In the next counterexample we show that there even exists an excess loss for which $\varepsilon(h)$ has finite support for each $h \in \mathcal{H}$ for which Bernstein's condition holds but the strong central condition does not.

Counterexample 4.1.1. There exists an excess loss $\varepsilon(h)$ such that for each $h \in \mathcal{H}$ it has exponential and Bernstein's condition holds (and thus $\varepsilon(h)$ has bounded variance and mean) but the strong central condition does not. Indeed, let $\mathcal{H} = \mathbb{N}$ and let the excess loss ε be such that

$$\varepsilon(i) = \begin{cases} -\log^2 i & \text{with probability } \frac{1}{i} \\ 1 & \text{with probability } 1 - \frac{1}{i}. \end{cases}$$

Its mean is

$$\mathbb{P}[\varepsilon(i)] = 1 - \frac{1 + \log^2 i}{i}$$

and its second moment

$$\mathbb{P}[\varepsilon^2(i)] = 1 + \frac{\log^4 i - 1}{i}.$$

Note two things. First, that $\varepsilon(1) = 0$ and for any other $i \in \mathcal{H}$, $\mathbb{P}[\varepsilon(i)] > 0$ so that in this case $h^* = 0$. Second, Bernstein's condition holds for $B = 11$ but the strong central

condition does not hold. In order to see why, rewrite the definition of $\varepsilon(i)$ as

$$\varepsilon(i) = \begin{cases} -a_i & \text{with probability } \mathrm{e}^{-\eta_i a_i} \\ 1 & \text{with probability } 1 - \mathrm{e}^{-\eta_i a_i} \end{cases}$$

with $\eta_i = 1/\log i$ and $a_i = \log^2 i$. Then, for any $\eta > 0$

$$\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(i)}] > \mathrm{e}^{(\eta-\eta_i)a_i}.$$

Thus, we can chose i big enough so that $\eta - \eta_i > 0$ because $\eta_i \to 0$ as $i \to \infty$. Consequently

$$\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(i)}] > 1$$

for big enough i, which means that the strong central condition does not hold.

On the other hand Bernstein's condition implies that both the mean and the variance of the excess losses are bounded as shown in the following lemma. Since there are excess loss functions without an uniformly bounded variance or mean that satisfy the strong central condition, it cannot imply Bernstein's condition.

Lemma 4.1.2. Let ε and excess loss function on a hypotheses class $\mathcal{H}$. If ε satisfies Bernstein's condition (Condition 1) then both its mean and is variance are uniformly bounded over $\mathcal{H}$.

Proof. This is the case for the variance since

$$\mathbb{P}[\varepsilon(h)^2] = \mathrm{Var}[\varepsilon(h)] + (\mathbb{P}[\varepsilon(h)])^2 \leq B(\mathbb{P}[\varepsilon(h)])^\beta$$

it holds that

$$\mathrm{Var}[\varepsilon(h)] \leq B(\mathbb{P}[\varepsilon(h)])^\beta - (\mathbb{P}[\varepsilon(h)])^2. \tag{4.2}$$

Since $\mathbb{P}[\varepsilon(h)] \geq 0$ and the function $x \mapsto Bx^\beta - x^2$ is bounded by its value at $x^* = \left(\frac{B\beta}{2}\right)^{\frac{1}{2-\beta}}$, we can conclude that

$$\mathrm{Var}[\varepsilon(h)] \leq B\left(\frac{B\beta}{2}\right)^{\frac{\beta}{2-\beta}} - \left(\frac{B\beta}{2}\right)^{\frac{2}{2-\beta}}.$$

which is a uniform bound over all h. In turn (4.2) implies that

$$B(\mathbb{P}[\varepsilon(h)])^\beta - (\mathbb{P}[\varepsilon(h)])^2 \geq 0$$

which means that

$$\mathbb{P}[\varepsilon(h)] \leq B^{\frac{1}{2-\beta}},$$

again a uniform bound over $\mathcal{H}$. □

Lemma 4.1.2 indicates that there are cases in which the strong central condition holds but Bernstein's conditions does not, because the latter implies uniformly bounded means and variances of the excess losses. In the next counterexample we exhibit an excess loss for which its means and variances are uniformly bounded and still the strong central and witness condition hold, but Bernstein's does not. In order to show this we reasoned as follows. We first showed that the combination of witness and central condition implies Bernstein's condition for a family of trimmed excess losses. We then looked for an excess loss whose trimmed version satisfied Bernstein's condition but the untrimmed did not. Even though the counterexample does not use the following lemma, we show it to present the need to create an excess loss that behaves very differently from their truncated versions.

Lemma 4.1.3. Let ε be an excess loss function over a hypotheses class $\mathcal{H}$ that satisfies both the witness condition (Condition 3) for some $u > 0$ and $c \in (0, 1)$, and the strong central condition (Condition 2). Then the trimmed excess loss $\varepsilon'(h) = \varepsilon(h)\mathbb{1}\{\varepsilon(h) \leq u\}$ satisfies Bernstein's condition.

Proof. Let η such that for all $h \in \mathcal{H}$ it happens that $\psi_h^-(\eta) \leq 0$. This means that $\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h)}] \leq 1$. Fix $h \in \mathcal{H}$. By Taylor's theorem there exists $\eta^* \in (0, \eta)$ depending such that

$$\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h)}] = 1 - \eta\mathbb{P}[\varepsilon(h)] + \frac{\eta^2}{2}\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon^2(h)],$$

which in addition to the central condition means that

$$\frac{\eta}{2}\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon^2(h)] \leq \mathbb{P}[\varepsilon(h)]$$

By the witness condition the expected value of $\varepsilon(h)$ can be bounded by that of its trimmed version $\varepsilon'(h)$ so that

$$\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon^2(h)] \leq C\mathbb{P}[\varepsilon'(h)]$$

for some constant C independent of h. Thus it is enough to prove that $\mathbb{P}\varepsilon'^2(h)$ can be bounded by the term in the left hand of the previous inequality. Write

$$\begin{aligned}\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon^2(h)] &= \mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon'(h)}\varepsilon'^2(h)] + \mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon^2(h)\mathbb{1}\{\varepsilon(h)] \geq u\} \\ &\geq \mathbb{P}[\mathrm{e}^{-\eta_h\varepsilon'(h)}\varepsilon'^2(h)] \\ &\geq \mathrm{e}^{-\eta u}\mathbb{P}[\varepsilon'^2(h)].\end{aligned}$$

Conclude that

$$\mathbb{P}[\varepsilon'^2(h)] \leq B\mathbb{P}[\varepsilon'(h)]$$

for some constant B independent of h. □

Counterexample 4.1.4. There exists an excess loss function $\varepsilon(h)$ with uniformly bounded mean and variance that satisfies both the witness condition for some $u > 0$ and $c \in (0, 1)$, and the central condition for which Bernstein's condition does not hold.

Indeed, let $\mathcal{H} = \mathbb{N}$ and let the the excess loss $\varepsilon(i)$ be such that

$$\varepsilon(i) = \begin{cases} i-1 & \text{with probability } \frac{1}{i^2} \\ \frac{1}{i} & \text{with probability } 1 - \frac{1}{i^2}. \end{cases}$$

Note that $\varepsilon(1) = 0$ so that in this case $h^* = 1$ and that since the excess losses are non-negative, the strong Central Condition is satisfied. It is clear that ε has uniformly bounded mean and variance. For this choice, the excess loss $\varepsilon(i)$ satisfies the witness condition for $u = 1$ and $c = 1/2$. This is because the excess loss $\varepsilon(i)$ and its trimmed version $\varepsilon'(h) = \varepsilon(h)\mathbb{1}\{\varepsilon(h) \leq u\}$ satisfy

$$\mathbb{P}[\varepsilon(i)] = \frac{2}{i} - \frac{1}{i^2} - \frac{1}{i^3} \tag{4.3}$$

$$\mathbb{P}[\varepsilon'(i)] = \frac{1}{i} - \frac{1}{i^3} \tag{4.4}$$

so that for $i \geq 2$

$$\frac{\mathbb{P}[\varepsilon(i)]}{\mathbb{P}[\varepsilon'(i)]} = \frac{2i^2 - i - 1}{i^2 - 1} \leq 2,$$

which means that

$$c\mathbb{P}[\varepsilon(i)] \leq \mathbb{P}[\varepsilon'(i)].$$

The second moment of the excess loss and its trimmed version satisfy

$$\mathbb{P}[\varepsilon^2(i)] = 1 - \frac{2}{i} + \frac{2}{i^2} - \frac{1}{i^4} \tag{4.5}$$

$$\mathbb{P}[\varepsilon'^2(i)] = \frac{1}{i^2} - \frac{1}{i^4} \tag{4.6}$$

so that on the one hand

$$\frac{\mathbb{P}[\varepsilon'^2(i)]}{\mathbb{P}[\varepsilon'(i)]} = \frac{i^2 - 1}{i^3 - i} \leq 1$$

which implies that the Bernstein condition holds for the trimmed excess loss with $B = 1$. On the other hand for the excess loss

$$\frac{\mathbb{P}[\varepsilon^2(i)]}{\mathbb{P}[\varepsilon(i)]} \asymp \frac{i}{2}$$

as $i \to \infty$, that is, Bernstein's condition does not hold.

4.2. A New Condition and Vapnik's Condition

In this section we show how a moment condition on the excess losses makes Bernstein's condition and the strong Central Condition equivalent. It was used by Mendelson [see 2014, Lemma 6.1] as a sufficient condition to obtain rates of convergence without using concentration inequalities. The condition reads:

Condition 4. $\varepsilon(h)$ is an excess loss function on a hypotheses class $\mathcal{H} \ni h$ such that for some $r > 1$ it holds that

$$\sup_{h \in \mathcal{H}} \frac{(\mathbb{P}[|\varepsilon(h)|^{2r}])^{1/r}}{\mathbb{P}[\varepsilon^2(h)]} < \infty.$$

and

$$\sup_{h \in \mathcal{H}} \mathbb{P}[\varepsilon^2(h)] < \infty$$

Remark 4.2.1. Note that this condition implies that $\sup_h \mathbb{P}[\varepsilon(h)] < \infty$.

The following example is adapted in part from Vapnik [1998, p. 208] and it shows how this condition holds in many practical settings.

Example 4.2.2. Consider again the simple normal means model from Example 2.5.1 where $Z \sim \mathcal{N}(\mu^*, 1)$ and $\ell(\mu, Z) = \frac{1}{2}(Z - \mu)^2$. A quick calculation shows that

$$\varepsilon(\mu, Z) = \frac{\mu^2 - \mu^{*2}}{2} + (\mu^* - \mu)Z$$

so that under $\mathbb{P}$ the excess loss $\varepsilon(h)$ has a normal distribution. There are also other common situations in which the excess loss is normally distributed, such as minimum squares regression with normal residuals. Thus, consider in general normally distributed excess losses $\varepsilon(h)$ with mean μ_h and variance σ_h^2 which are allowed to vary with h. Since for the normal distribution it holds that[1] $\mathbb{P}[(\varepsilon(h) - \mu_h)^n] = \sigma^n(n-1)!!$ for even n, then, defining $m_4 = \mathbb{P}[(\varepsilon(h) - \mu_h)^4]$ we obtain that

$$\frac{m_4}{\sigma_h^2} = 3.$$

With this in mind,

$$\begin{aligned}
\frac{(\mathbb{P}[\varepsilon^4(h)])^{1/2}}{\mathbb{P}[\varepsilon^2(h)]} &= \frac{(\mathbb{P}[((\varepsilon(h) - \mu_h) + \mu_h)^4])^{1/2}}{\mathbb{P}[((\varepsilon(h) - \mu_h) + \mu_h)^2]} \\
&= \frac{\sqrt{m_4 + 6\mu^2\sigma^2 + 3\sigma^4}}{\sigma^2 + \mu^2} \\
&= \frac{\sqrt{3\sigma^2 + 6\mu^2\sigma^2 + 3\sigma^4}}{\sigma^2 + \mu^2} \\
&\leq \sqrt{3}.
\end{aligned}$$

This means that if the excess losses are Gaussian under $\mathbb{P}$, then the Condition 4 holds for $r = 2$ and $\tau = \sqrt{3}$. The same analysis holds for this choice of ℓ and for distributions of Z (and consequently of ε) that are symmetric scale-location families. In that case, we recognize τ to be the square root of the kurtosis. This includes Laplacian, uniform, logistic and many other distributions.

[1] !! makes reference to the double factorial which is defined as $n!! = n(n-2)(n-4)\ldots 2$

In the next Theorem we show how condition 4 makes the strong central condition and Bernstein's condition equivalent.

Theorem 4.2.3. *Let ε be an excess loss on a hypotheses class such that (4.1) holds for some $\eta > 0$ and Condition 4 holds. Then Bernstein's condition and the strong central condition are equivalent.*

Proof. This proof will rest on the fact for each η and a fixed $h \in \mathcal{H}$ where

$$\sup_{h\in\mathcal{H}} \psi_{-\varepsilon(h)}(\eta) < \infty$$

there exists an $\eta^* \in (0,1)$ (depending on h) such that

$$\mathbb{P}\mathrm{e}^{-\eta\varepsilon(h)} = 1 - \eta\mathbb{P}\varepsilon(h) + \frac{1}{2}\eta^2\mathbb{P}\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2. \tag{4.7}$$

Fix $h \in \mathcal{H}$ and suppose that Bernstein's Condition Holds. Define $r^* = (1-1/r)^{-1}$ and choose η such that

$$\sup_{h\in\mathcal{H}} \psi(2r^*\eta) < \infty$$

By (4.7) it is enough to prove that for each h one has

$$\frac{1}{2}\eta^2\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2] \leq \eta\mathbb{P}[\varepsilon(h)].$$

By Hölder's Inequality and our assumptions

$$\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2] \leq \mathbb{P}[\mathrm{e}^{-r^*\eta^*\varepsilon(h)}]^{1/r^*}\mathbb{P}[|\varepsilon(h)|^{2r}]^{1/r} \leq C\mathbb{P}[\varepsilon(h)^2].$$

for some constant $C > 0$ independent from h. The results follows from Bernstein's condition and taking η sufficiently small.

Fix $h \in \mathcal{H}$ and suppose that the central condition holds for some $\eta > 0$. Then by Equation (4.7) for each h there exists a η^* such that

$$\frac{1}{2}\eta^2\mathbb{P}[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2] \leq \eta\mathbb{P}[\varepsilon(h)].$$

Let $K > 0$. Note that

$$\begin{aligned}\mathbb{P}\left[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2\right] &= \mathbb{P}\left[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2\mathbb{1}\left\{\varepsilon(h) > K\right\}\right] + \mathbb{P}\left[\mathrm{e}^{-\eta^*\varepsilon(h)}\varepsilon(h)^2\mathbb{1}\left\{\varepsilon(h) \leq K\right\}\right] \\ &\geq \mathrm{e}^{-\eta K}\mathbb{P}\left[\varepsilon(h)^2\mathbb{1}\left\{\varepsilon(h) \leq K\right\}\right]\end{aligned}$$

so it is enough to provide a constant $c \in (0,1)$ such that

$$\mathbb{P}\left[\varepsilon(h)^2\mathbb{1}\left\{\varepsilon(h) \geq K\right\}\right] \leq c\mathbb{P}\left[\varepsilon(h)^2\right].$$

We focus on proving this last relation. Use Hölder's inequality to obtain that

$$\mathbb{P}\left[\varepsilon(h)^2\mathbb{1}\left\{\varepsilon(h) \geq K\right\}\right] \leq \left(\mathbb{P}\left\{\varepsilon(h) \geq K\right\}\right)^{1/r*}\mathbb{P}\left[|\varepsilon(h)|^{2r}\right]^{1/r}$$

We can use Markov's inequality to obtain that

$$\mathbb{P}\left\{\varepsilon(h) \geq K\right\} \leq \mathbb{P}\left\{|\varepsilon(h)|^{2r} \geq K^{2r}\right\} \leq \frac{\mathbb{P}[|\varepsilon(h)|^{2r}]}{K^{2r}}$$

and that consequently

$$\mathbb{P}\left[\varepsilon(h)^2 \mathbb{1}\left\{\varepsilon(h) \geq K\right\}\right] \leq \frac{\mathbb{P}\left[|\varepsilon(h)|^{2r}\right]}{K^{2r/r^*}} \leq \left(\frac{\mathbb{P}\left[\varepsilon(h)^2\right]}{K^{2/r^*}}\right)^r$$

because $1/r + 1/r^* = 1$ and the assumption. Choose K so big so that

$$K > \max\left\{\sup_{h\in\mathcal{H}} \mathbb{P}\left[\varepsilon(h)^2\right]^{r^*/2}, 1\right\}.$$

and define $c = 1/K^{r^*/2}$. Together with the fact that $x^r \leq x$ for $x \in [0,1]$, this choice of c leads to

$$\mathbb{P}\left[\varepsilon(h)^2 \mathbb{1}\left\{\varepsilon(h) \geq K\right\}\right] \leq c\mathbb{P}\left[\varepsilon(h)^2\right],$$

as it was to be shown. □

A close inspection of the last proof, shows that if Condition 4 and Bernstein's condition (Condition 1) hold, then the following condition holds

Condition 5. An excess loss function $\varepsilon(h)$ on a hypotheses class $\mathcal{H}$ satisfies

$$\sup_{h\in\mathcal{H}} \frac{(\mathbb{P}[|\varepsilon(h)|^{2r}])^{1/2r}}{\mathbb{P}[\varepsilon(h)]} \leq \tau$$

for some $r > 1$ and

$$\sup_{h\in\mathcal{H}} \mathbb{P}\left[\varepsilon(h)\right] < \infty.$$

Vapnik [1998, Section 5.7] called the distributions for which this condition holds *light tailed* since the random variables that satisfy it tend to have less mass on the tails. Furthermore, Vapnik [1998] found that this condition implied bounds for the expected loss when the loss function is required to satisfy it.

We now prove that Condition 5 is enough to obtain an inequality of the form of Theorem 3.3.1, that is, of the form

$$A_\eta(h) \lesssim E(h).$$

Recall that $A_\eta(h)$ is the annealed expectation defined in (3.2). This means that, just as in Corollary 3.3.3, one can use the inequality from Theorem 3.1.1 to obtain expected excess loss bounds, but with a better leading constant factor. We further show that it implies the witness condition (Condition 3), the Strong central condition (Condition 2) and Bernstein's condition (Condition 1).

Theorem 4.2.4. *Let ε be an excess loss function on a hypotheses class $\mathcal{H}$ such that Condition 5 holds for some finite non zero τ and $r > 1$. Assume also that (4.1) holds for some $\eta > 0$. Then there exists a positive constant C such that if*

$$\bar{\eta} = \frac{1}{C \sup_{h \in \mathcal{H}} E(h)}$$

and

$$\mu(\eta) = \frac{1}{1 - \frac{\eta}{\bar{\eta}}}$$

then

$$E(h) \leq \mu(\eta) A_\eta(h).$$

Proof. Consider a fixed $h \in \mathcal{H}$. We know that for each $\eta > 0$ where $\sup_h \psi_{-\varepsilon(h)}$ it holds that

$$\log \mathbb{P}[\mathrm{e}^{-\eta \varepsilon(h)}] \leq -\eta \mathbb{P}[\varepsilon(h)] + \frac{1}{2} \eta^2 \mathbb{P}[\mathrm{e}^{-\eta^* \varepsilon(h)} \varepsilon^2(h)]$$

for some $\eta^* \in (0, \eta)$. Define $r^* = (1 - 1/r)^{-1}$. By Hölder's inequality, the assumption

$$\begin{aligned} \mathbb{P}[\mathrm{e}^{-\eta^* \varepsilon(h)} \varepsilon^2(h)] &\leq \left(\mathbb{P}[\mathrm{e}^{-r^* \eta^* \varepsilon(h)}] \right)^{1/r^*} \left(\mathbb{P}[\varepsilon^{2r}(h)] \right)^{1/r} \\ &\leq C \left(\mathbb{P}[\varepsilon(h)] \right)^2 \end{aligned}$$

for some finite constant C independent of h. Bringing together these two inequalities we obtain that

$$\begin{aligned} \log \mathbb{P}[\mathrm{e}^{-\eta \varepsilon(h)}] &\leq -\eta \mathbb{P}[\varepsilon(h)] + C \eta^2 \mathbb{P}[\varepsilon(h)]^2 \\ &= \eta \mathbb{P}[\varepsilon(h)] (C \eta \mathbb{P}[\varepsilon(h)] - 1) \\ &\leq \eta \mathbb{P}[\varepsilon(h)] (C \eta \sup_{h \in \mathcal{H}} \mathbb{P}[\varepsilon(h)] - 1) \end{aligned} \tag{4.8}$$

Define $\bar{\eta}$ as

$$\bar{\eta} = \frac{1}{C \sup_{h \in \mathcal{H}} \mathbb{P}[\varepsilon(h)]}$$

and take $\eta < \bar{\eta}$. Rearrange to obtain that

$$E(h) \leq \mu(\eta) A_\eta(h)$$

with

$$\mu(\eta) = \frac{1}{1 - \frac{\eta}{\bar{\eta}}}.$$

□

Remark 4.2.5. Our claim that the bound in Theorem 4.2.4 is tighter than that of Grünwald and Mehta [2016] (see Theorem 3.3.1) is based on the fact the function μ is in this case smaller than that of Theorem 3.3.1 and that as $\eta \downarrow 0$ the inequality becomes an equality because the leading constant is exactly 1. In the same fashion as in Corollary 3.3.3, we can obtain an excess risk bound chaining Zhang's inequality and the previous.

Corollary 4.2.6. *Under the conditions of Theorem 4.2.4 it holds for every prior distribution Π_0 and posterior distribution Π_n that with probability higher than $1-\delta$*

$$\Pi_n[E(h)] \leq \mu(\eta)\left(\mathrm{IC}(\eta, n) + \frac{\log \frac{1}{\delta}}{\eta n}\right)$$

is satisfied for $\eta < \bar{\eta}$ where $\bar{\eta}$ and μ are as defined in Theorem 4.2.4 and $\mathrm{IC}(\eta, n)$ is as defined in (3.3).

We now present an example to show that this is indeed the case in situations in which fast rates are possible, as in Example 4.2.2.

Example 4.2.7. As we saw in Example 4.2.2, there are many distributions with well behaved tails and hypotheses classes that satisfy Condition 4. In the case that those distributions either satisfy the strong central or Bernstein's condition, and thus are in the fast rate regime, then they satisfy Condition 5. For instance, if $\varepsilon(h)$ is distributed according to a normal distribution with mean μ_h and variance σ_h^2 possible dependent on h, then

$$\psi_{-\varepsilon(h)}(\eta) = -\mu_h \eta + \frac{1}{2}\sigma_h^2 \eta^2$$

so that the central condition holds if and only if there is some η such that for all h

$$\sigma_h^2 \leq 2\eta\mu_h,$$

which can be seen to be equivalent to Bernstein's condition. In the case that this is satisfied, Condition 5 holds and thus the tighter bound from Corollary 4.2.6 holds. This means that this result can be used in many practical applications.

Additionally, Condition 5 implies the witness condition, the strong central condition and Bernstein's condition.

Lemma 4.2.8. Condition 5 implies the strong central condition (Condition 2) if (4.1) holds for some $\eta > 0$.

Proof. A careful reading of the proof of Theorem 4.2.4 shows that from Equation (4.8) it follows that the strong central condition holds for $\eta < \bar{\eta}$, where $\bar{\eta}$ is as defined in the same theorem. □

Lemma 4.2.9. Condition 5 implies the witness condition (Condition 3)

Proof. Suppose that ε satisfies Condition 5. Let $u > 0$ and define $r^* = (1 - 1/2r)^{-1}$. Use Hölder's and Markov's inequality to obtain that

$$\begin{aligned}
\mathbb{P}\left[\varepsilon(h)\mathbb{1}\left\{\varepsilon(h) > u\right\}\right] &\leq \mathbb{P}\left[\varepsilon(h)^{2r}\right]^{1/2r} \mathbb{P}\left\{\varepsilon(h) > u\right\}^{1/r^*} \\
&\leq \frac{\mathbb{P}\left[\varepsilon(h)^{2r}\right]}{u^{2r/r^*}} \\
&\leq \tau \frac{\mathbb{P}\left[\varepsilon(h)\right]^{2r}}{u^{2r/r^*}}.
\end{aligned}$$

Choosing u sufficiently large, specifically

$$u > \max\left\{\tau^{r^*/2r} \sup_{h\in\mathcal{H}} \mathbb{P}\left[\varepsilon(h)\right]^{r^*}, 1\right\}$$

implies that

$$\mathbb{P}\left[\varepsilon(h)\mathbb{1}\left\{\varepsilon(h) > u\right\}\right] \leq c\mathbb{P}\left[\varepsilon(h)\right]$$

with $c \in (0,1)$ because $x^{2r} \leq x$ for $x \in [0,1]$. Thus, the witness condition holds. □

Lemma 4.2.10. Condition 5 implies Bernstein's condition (Condition 1).

Proof. From Hölder's inequality and Condition 5 we obtain that

$$\mathbb{P}\left[\varepsilon^2(h)\right] \leq \mathbb{P}\left[\varepsilon^{2r}(h)\right]^{1/r} \leq \tau\mathbb{P}\left[\varepsilon(h)\right]^2 \leq \tau \sup_{h'} \mathbb{P}\left[\varepsilon(h')\right]\mathbb{P}\left[\varepsilon(h)\right],$$

that is, Bernstein's condition holds with $B = \tau \sup_h \mathbb{P}\left[\varepsilon(h)\right]$. □

4.3. Slow Rates from the Witness Condition

In this section we prove risk bounds for excess losses that satisfy either the witness condition (Condition 3 on page 29) or an alternative second moment condition under the assumption that (4.1) holds for some $\eta > 0$. We do this by using Bernstein's moment condition (see Theorem C.0.1) in order to prove that uniform subgamma bounds hold for certain random variables related to $E(h) - \varepsilon(h)$. Recall that we say that a random variable has a subgamma right tail if for some constants $v, c > 0$ it holds that

$$\psi_X(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta}$$

for $\eta \in (0, 1/c)$. Recall also that we are interested in the situation in which subgamma bounds hold uniformly over $\mathcal{H}$, that is, if $\{X_h\}_{h\in\mathcal{H}}$, then

$$\sup_{h\in\mathcal{H}} \psi_{X_h}(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta},$$

in which case we say that X_h has a uniformly subgamma right tail.

Recall also that using Corollary 3.1.3, one can use bounds on the cumulant generating functions to obtain of order $n^{-1/2}$. With this in mind the result reads:

Theorem 4.3.1. *Let ε an excess loss over a hypotheses class $\mathcal{H}$ such that*

$$\sup_{h\in\mathcal{H}} \psi_{-\varepsilon(h)}(\eta) < \infty$$

for some $\eta > 0$.

1. *If the witness condition holds, that is, there exist $u > 0$ and $c' \in [0, 1]$ so that for every h*

$$c'\mathbb{P}[\varepsilon(h)] \leq \mathbb{P}[\varepsilon(h)\mathbb{1}\{\varepsilon(h) \leq u\}]$$

then $c'E(h) - \varepsilon(h)$ has a uniformly subgamma right tail.

2. *If the second moment of the positive part $\varepsilon_+(h) = \max\{0, \varepsilon(h)\}$ of the excess loss is uniformly bounded, that is, if*

$$\sup_h \mathbb{P}[\varepsilon(h)_+^2] < \infty.$$

Then $E(h) - \varepsilon(h)$ has a uniformly subgamma right tail.

Proof. Suppose that 1 holds. Let ε' be the trimmed excess loss $\varepsilon'(h) = \varepsilon(h)\mathbb{1}\{\varepsilon(h) \leq u\}$. Let $\eta > 0$ be so that $\sup_h \psi_{E(h)}(\eta) < \infty$. Note that for each h

$$\mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h)}] \leq \mathbb{P}[\mathrm{e}^{-\eta\varepsilon'(h)}]$$

so that $\psi_{-\varepsilon(h)}(\eta) \leq \psi_{-\varepsilon'(h)}(\eta)$. At the same time

$$\mathbb{P}[\mathrm{e}^{-\eta\varepsilon'(h)}] = \mathbb{P}[\mathrm{e}^{-\eta\varepsilon'(h)}(\mathbb{1}\{\varepsilon(h) \leq u\} + \mathbb{1}\{\varepsilon(h) > u\})] \leq 1 + \mathbb{P}[\mathrm{e}^{-\eta\varepsilon(h)}].$$

These two facts mean that a bound on $\psi_{\varepsilon'(h)}(\eta)$ implies a bound on $\psi_{\varepsilon(h)}(\eta)$ and that the $\varepsilon'(h)$ also have bounded cumulant generating function at η for each $h \in \mathcal{H}$. Consequently assume that there is a finite M such that

$$\sup_{h\in\mathcal{H}} \mathbb{P}[\mathrm{e}^{-\eta\varepsilon'(h)}] \leq M$$

Since the $\varepsilon'(h)$ are bounded by u from the right, $\mathbb{P}\varepsilon'_+(h)^2 \leq u^2$ and one can write

$$\begin{aligned}
\mathbb{P}[\varepsilon'(h)_-^n] &= \mathbb{P}[\varepsilon(h)_-^n] \\
&= \mathbb{P}\int_0^{\varepsilon_-(h)} nt^{n-1}\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\{\varepsilon_-(h) > t\}\, nt^{n-1}\mathrm{d}t \\
&\leq Mn\int_0^\infty \mathrm{e}^{-\eta t}t^{n-1}\mathrm{d}t \\
&= M\frac{1}{\eta^n}n\Gamma(n) \\
&= M\frac{1}{\eta^n}n!
\end{aligned}$$

For each $n \geq 2$. Thus

$$\mathbb{P}[\varepsilon'(h)^2] = \mathbb{P}[\varepsilon'(h)_-^2] + \mathbb{P}[\varepsilon'(h)_+]^2 \leq M\frac{2}{\eta^2} + u^2. \tag{4.9}$$

Consequently define $v = M\frac{2}{\eta^2} + u^2$ and $c' = \frac{1}{\eta}$ and to obtain that

$$\mathbb{P}\varepsilon'_-(h)^n \leq \frac{1}{2}n!vc'^{n-2}, \tag{4.10}$$

that is, ε' satisfies Bernstein's moment condition and, by Theorem C.0.1 the result follows.

In the second case, the analogue of Equation (4.10) can be obtained for $\varepsilon(h)$ directly by choosing $v = M\frac{2}{\eta^2} + \sup_h \mathbb{P}[\varepsilon(h)_+^2]$ and $c = \frac{1}{\eta}$, so the result also follows in the same manner. □

Consequently, the result from Theorem 4.3.1 can be used directly in conjunction with Corollary 3.1.3 to obtain slow rates.

Remark 4.3.2. The two conditions for the previous theorem are not not comparable, that is, one does not imply the other. Take $\mathcal{H} = \{h\}$ a hypotheses class with one element. Choose $\varepsilon(h)$ to have a Gaussian distribution on $\mathbb{R}$ with mean zero and unit variance. Then for every $u > 0$ it happens that $\mathbb{P}E\mathbb{1}\{E \leq u\} < 0$. This means that the witness condition does not hold but $\varepsilon(h)$ has bounded second moment. On the other hand take again a class of hypotheses $\mathcal{H}$ with a single element h and let $\varepsilon(h)$ distributed on the positive integers with distribution function so that $\mathbb{P}\{E = n\} \propto n^{-3}$. It follows that E has no second moment but $\mathbb{P}E\mathbb{1}\{E \geq u\} > 0$ for each $u > 0$, that is, it satisfies the witness condition.

Remark 4.3.3. We proved the last theorem using Bernstein's moment condition. This means under both the assumptions of Theorem 4.3.1 that

$$\sup_{h\in\mathcal{H}} \psi_{-\varepsilon(h)}(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta}$$

for the constants given in the proof. This is a consequence of the fact that $\mathbb{P}[\varepsilon(h)] \geq 0$. This means in the terms of the of the (weaker) central condition (Condition 2, on page 28) that it is satisfied with a function $\eta(\epsilon)$ which is linear asymptotically as $\epsilon \downarrow 0$. More specifically, inversion shows that $\eta(\epsilon) = \frac{\epsilon}{\frac{v}{2}+c\epsilon} = O(\epsilon)$ as $\epsilon \downarrow 0$. Grünwald and Mehta [2016] also proved that slow rates were attainable using this point of view.

5. Conclusion

We considered the abstract learning problem. Using a $\mathcal{Z}$-valued iid sample $Z_1, \ldots, Z_n$, the problem is choosing a hypothesis h from a hypotheses class $\mathcal{H}$ that minimizes the expected loss $L(h) = \mathbb{P}[\ell(h, Z)]$ according to the loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

In this work we considered conditions for finding bounds on the expected excess loss

$$E(h) = L(h) - \inf_{h \in \mathcal{H}}(h)$$

in the case in which there exists a $h^* \in \mathcal{H}$ at which the expected loss $L(h)$ is minimized, when one can write

$$E(h) = \mathbb{P}[\varepsilon(h, Z)]$$

where $\varepsilon(h, Z) = \ell(h, Z) - \ell(h^*, Z)$ is the excess loss. We consider conditions under which the order of the bounds can be improved from $n^{-1/2}$ to n^{-1} up to logarithmic factors, which is significant for practice. The conditions that we considered are on the excess losses themselves and the bounds that we reviewed are valid both for the method of Empirical Risk Minimization and in the PAC-Bayesian setting.

In the case that the excess losses are bounded it is known that a condition, called Bernstein's condition (Condition 1 on page 1) gives fast rates and that it is equivalent to the strong central condition (Condition 2 on page 28). However, when excess losses might be unbounded, this landscape changes. Under an additional condition on the right tail of the distribution of the excess loss (the witness condition, Condition 3 on page 29) it is known that the central condition leads to fast rates both for ERM and in the PAC-Bayesian setting [Grünwald and Mehta, 2016] while the same is only known under Bernstein's condition for a nonstandard randomized algorithm [Audibert, 2009]. Both conditions, however, are no longer equivalent in the unbounded case, even under stringent tail and moment conditions, which we showed in Section 4.1.

We considered a natural weakening of the case in which excess losses are bounded, the situation in which the probability of any nonoptimal hypotheses being better than the optimal is exponentially small. In this regime, in Section 4.2, nevertheless, a mild condition on the moments of the excess loss (Condition 4 on page 39) made both the central and Bernstein's conditions equivalent. Also in this regime, we showed that when the strong central condition holds (and thus fast rates are possible) tighter bounds than those of Grünwald and Mehta [2016] (Section 4.2), are possible, although under a stronger assumption. We showed that this is indeed the regime of many practical situations, which means that our new bound is relevant for applications. Finally, in Section 4.3 we showed that under our weakest assumptions, slow bounds could also be proven.

The ultimate goal of this line of research would be to completely characterize when fast rates are attainable in the regime that we studied. With this in mind, we present some open questions.

Open Questions As we showed, in the regime in which excess losses are negative with exponentially small tails (Condition in (4.1)), under a moment condition both Bernstein's and the strong central condition are equivalent and lead to fast rates. This leads to the first question:

1. Are there other conditions that lead to fast rates in this regime?
2. In relation to the last question, do central and Bernstein's condition characterize completely fast rates in this case, that is, is it possible to prove that if fast rates are achieved, then these conditions must hold?

We also proved in Theorem 4.3.1 that in the case that the witness or a second moment condition holds then slow rates are also possible.

3. Is this second moment condition also sufficient to achieve fast rates in the case that the strong central condition holds? This is relevant as this condition might be easier to check.

A. Rates of Convergence in Probability

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and let X be a random variable. We say that X_n tends to X in probability if for any $\epsilon > 0$

$$\mathbb{P}\left\{|X_n - X| > \epsilon\right\} \to 0$$

as $n \to 0$. We denote this as $X_n \xrightarrow{\mathbb{P}} X$ for short. For some real function $r(n)$ we say that X_n is of order $r(n)$ in probability or $X_n = O_{\mathbb{P}}(r(n))$ for short, if for each $\delta > 0$ there is a finite $M > 0$ such that with probability higher than $1 - \delta$

$$|X_n - X| \leq M|r(n)|$$

for n big enough. In the special case that $r(n) \to 0$ as $n \to \infty$ we say that X_n tends to X at a a rate $r(n)$ (in probability).

We are interested in bounds of the form

$$\mathbb{P}\left\{|X_n - X| > \epsilon\right\} \leq a\mathrm{e}^{-nf(\epsilon)}$$

for some constant a and positive function $f(\epsilon)$ tending to 0 as $\epsilon \downarrow 0$. Set $\delta = a\mathrm{e}^{-nf(\epsilon)}$. If f is one-to-one, inversion shows that

$$\epsilon = f^{-1}\left(\frac{1}{n}\log\frac{a}{\delta}\right),$$

and that consequently with probability higher than $1 - \delta$

$$|X_n - X| \leq f^{-1}\left(\frac{1}{n}\log\frac{a}{\delta}\right).$$

In the case that for some $\alpha > 0$ the function f^{-1} satisfies $f^{-1}(x) = O(x^\alpha)$ as $x \to 0$, the last equation implies that X_n tends to X at a $n^{-\alpha}$ rate. Thus, the behavior of the inverse and thus of $f(x)$ itself as $x \downarrow 0$ determines the rate of convergence in probability of the sequence X_n. In particular if $f(x) = O(x^{1+\beta})$ as $x \downarrow 0$, for some $\beta \in [0, 1]$, then the convergence happens at a $n^{-1/(1+\beta)}$ rate, which leads to rates between n^{-1} for $\beta = 0$, and $n^{-1/2}$ for $\beta = 1$.

B. The Cramér-Chernoff Method

B.1. Cumulant Generating Functions

The Cramér-Chernoff method is behind many basic (and advanced) methods for obtaining concentration inequalities. It is based on the use of the cumulant generating function to obtain tail bounds. Let X be a random variable with mean $\mathbb{P}X$ and let $\psi_X(\eta)$ defined by

$$\psi_X(\eta) = \log \mathbb{P}[\mathrm{e}^{\eta X}]$$

be its cumulant generating function.

If there is some $\eta^* > 0$ such that $\psi_X(\eta^*) < \infty$, then $\psi_X(\eta)$ exists for all η in the interval $[0, \eta^*]$. This is a consequence of the exponential function's convexity. Indeed, for any $\eta \in [0, \eta^*]$ it is true that $\eta = \theta\eta^*$ for some $\theta \in [0, 1]$ and consequently $\mathbb{P}\mathrm{e}^{\eta X} \leq \theta\mathbb{P}\mathrm{e}^{\eta^* X} < \infty$. Furthermore, $\psi_X(0) = 0$ and it is smooth.

Define the probability measure $\mathbb{P}_\eta$ by its density with respect to $\mathbb{P}$ given by

$$\frac{\mathrm{d}\mathbb{P}_\eta}{\mathrm{d}\mathbb{P}} = \frac{\mathrm{e}^{\eta X}}{\mathbb{P}[\mathrm{e}^{\eta X}]}.$$

Direct calculation shows that

$$\psi'_X(\eta) = \mathbb{P}_\eta X$$

and that

$$\psi''_X(\eta) = \mathrm{Var}_\eta\, X,$$

where Var_η makes reference to the variance taken with respect to the probability distribution $\mathbb{P}_\eta$. This means that the function $\eta \mapsto \psi_X(\eta)$ is convex.

B.2. Cramér - Chernoff Tail Bounds

The Cramér-Chernoff method is based on a smart use of Markov's inequality. If $\psi(\eta) < \infty$ for some $\eta > 0$, for $t > 0$ Markov's inequality gives

$$\mathbb{P}\left\{X \geq t\right\} = \mathbb{P}\left\{\mathrm{e}^{\eta X} \geq \mathrm{e}^{\eta t}\right\} \tag{B.1}$$

$$\leq \mathrm{e}^{\psi_X(\eta) - \eta t} \tag{B.2}$$

In any case, one can optimize the upper bound and obtain that

$$\mathbb{P}\left\{X \geq t\right\} \leq \mathrm{e}^{-\psi_X^*(t)}$$

where

$$\psi_X^*(t) = \sup_{\eta>0} \eta t - \psi_X(\eta).$$

By Jensen's inequality $\psi_X(\eta) \geq \eta \mathbb{P}X$, so that the bound becomes trivial for $t \leq \mathbb{P}X$ because in that case $\eta\epsilon - \psi_X(\eta) \leq 0$. Thus, one might extend the supremum in the definition of ψ^* to

$$\psi_X^*(t) = \sup_{\eta\in\mathbb{R}} \eta\epsilon - \psi_X(\eta).$$

This means that for $t \geq \mathbb{P}X$ the function $\psi^*(t)$ coincides with the convex conjugate of $\psi(\eta)$, which is also called the Fenchel-Legendre dual function.

B.3. Possible Rates Via The Cramér-Chernoff Method

The Cramér-Chernoff method is behind many basic (and advanced) concentration inequalities for random variables. Here we show that under no additional assumptions, it is not possible to obtain faster rates than $n^{-1/2}$ from it. Even though not necessarily in the context of the search for faster rates, this fact has been previously noted. We follow part of the development of McAllester and Ortiz [2003] in this section. A more detailed description of the Cramér-Chernoff method can be found in Boucheron et al. [2013, Chapter 2].

With the observations of Appendix A and of the previous section, in many situations the the rate of convergence of a succession of random variables depends on the behavior of the function $\psi^*_{X-\mathbb{P}X}(\epsilon)$ for small ϵ. In the following we show for small ϵ, the function $\psi^*_{X-\mathbb{P}X}$ behaves as a quadratic function. Given the discussion of Appendix A, this implies rates of order $n^{-1/2}$. We show this by noticing that the zeroth and first order terms in its Taylor approximation of $\psi^*_{X-\mathbb{P}X}$ are zero.

First we write a more explicit expression for $\psi^*_{X-\mathbb{P}X}(\eta)$ from which we can obtain more information. Since $\psi_{X-\mathbb{P}X}$ is smooth, the infimum that defines its dual $\psi^*_{X-\mathbb{P}}$ can be found by differentiation with respect to η. The infimum is attained at the η such that $\epsilon = \mathbb{P}_\eta X - \mathbb{P}X$. Such η exists because the function $\eta \mapsto \mathbb{P}_\eta\varepsilon$ is monotonously increasing. By a slight abuse of notation, refer to that special η as $\eta(\epsilon)$. With this in mind $\psi^*_{E-\varepsilon}$ can be written as

$$\psi^*_{X-\mathbb{P}X}(\epsilon) = \eta(\epsilon)\epsilon - \psi_{X-\mathbb{P}X}(\eta(\epsilon)).$$

We can now proceed to study the behavior of $\psi^*_{E-\varepsilon}(\epsilon)$ for small ϵ through its Taylor approximation. Note that for a sufficiently smooth function f, we can write

$$f(x) = f(0) + xf'(0) + x^2 \int_0^1 \int_0^t f''(xs)\mathrm{d}s\mathrm{d}t$$

and that by Taylor's theorem the last term is approximately quadratic for small x, more precisely,

$$x^2 \int_0^1 \int_0^t f''(xs)\mathrm{d}s\mathrm{d}t = \frac{f''(0)}{2}x^2 + o(x^2)$$

as $x \downarrow 0$. Note also that because $\eta(0) = 0$ and $\psi_{E-\varepsilon}(0) = 0$, we have that $\psi^*_{E-\varepsilon}(0) = 0$ and that $(\psi^*_{E-\varepsilon})'(0) = \eta(0) = 0$. Additionally, direct computation reveals that

$$(\psi^*_{X-\mathbb{P}X})''(\epsilon) = \frac{1}{\operatorname{Var}_{\eta(\epsilon)} X}.$$

Together these observations imply that

$$\psi^*_{X-\mathbb{P}X}(\epsilon) = \epsilon^2 \int_0^1 \int_0^t \frac{1}{\operatorname{Var}_{\eta(\epsilon s)} X} \mathrm{d}s\mathrm{d}t.$$

and that for small ϵ

$$\psi^*_{X-\mathbb{P}X}(\epsilon) = \frac{1}{2 \operatorname{Var} \varepsilon} \epsilon^2 + o(\epsilon^2)$$

as $\epsilon \downarrow 0$, a quadratic function.

C. Subgamma Random Variables and Bernstein's Inequality

Call ψ_X the cumulant generating function of a random variable X, that is,

$$\psi_X(\eta) = \log \mathbb{P}[\mathrm{e}^{\eta X}].$$

Following Boucheron et al. [2013], we say that a random variable X has a subgamma right tail with variance factor and scale parameter c if its cumulant generating function satisfies

$$\psi_X(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta} \tag{C.1}$$

for $0 < c\eta < 1$.

The following is Theorem 2.10 in Boucheron et al. [2013], and it proves that a certain condition, which we call *Bernstein's moment condition*, implies that the centered version $X - \mathbb{P}[X]$ of a (possibly unbounded) random variable X has a subgamma right tail.

Theorem C.0.1 (Bernstein's Inequality)**.** *Let X be a random variable. If there exists $v > 0$ so that $\mathbb{P}X^2 \leq v$ and a positive c so that for each $n \geq 3$*

$$\mathbb{P}X_+^n \leq \frac{1}{2}n!vc^{n-2} \tag{C.2}$$

then the centered random variable $X - \mathbb{P}[X]$ has a subgamma right tail, that is,

$$\psi_{X-\mathbb{P}[X]}(\eta) \leq \frac{1}{2}\frac{v\eta^2}{1-c\eta} \tag{C.3}$$

for $0 < c\eta < 1$.

Using Cramér-Chernoff's method from Appendix B, it is possible to show that if X has a subgamma right tail (it satisfies (C.1)), then for each $t > 0$ it holds that

$$\mathbb{P}\left\{X > \sqrt{2vt} + ct\right\} \leq \mathrm{e}^{-t},$$

which can also be written in the form

$$\mathbb{P}\left\{X > \tau\left(\sqrt{t} + \frac{L}{2}t\right)\right\} \leq \mathrm{e}^{-t}$$

for $\tau = \sqrt{2v}$ and $L = 2c/\sqrt{2v}$. Equivalently, by inverting the function $x \mapsto \sqrt{t} + \frac{L}{2}t$

$$\mathbb{P}\{X > t\} \leq \mathrm{e}^{-\phi^L_{\text{Bern}}(t/\tau)}$$

with

$$\phi^L_{\mathrm{Bern}}(t/\tau)(t) = \left(\frac{\sqrt{1+2Lt}-1}{L}\right)^2.$$

Additionally, using simple properties of cumulant generating functions it follows that if $X_1, \ldots, X_n$ are iid copies of a random variable with subgamma right tail, then it holds that

$$\mathbb{P}\left\{n^{1/2}\mathbb{P}_n[X] > t\right\} \leq \mathrm{e}^{-\phi^{L/\sqrt{n}}_{\mathrm{Bern}}(t/\tau)}$$

where

$$\mathbb{P}_n[X] = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

D. Rates for Infinite Classes: Chaining

Families of random variables that satisfy concentration inequalities as discussed in Section 2.5 can be characterized in terms of a certain type of norms called Orlicz norms. These norms are a generalization of p-norms that arises naturally in the study of uniform integrability [see de la Vallé-Poussin's theorem in Rao and Ren, 1991, Chapter 1]. We first define these norms, then show how they have been used to study suprema of stochastic processes for infinite families through a technique called chaining. We only point at the main results, which are classical. The first ideas in these direction are conventionally attributed to the work of Kolmogorov and we point at the expositions of Van der Vaart and Wellner [1996, Chapter 2], Pollard [1984, Section 7.2], and the generalization called generic chaining, by Talagrand [2014, Chapter 2]. We view the excess loss as a stochastic process $h \mapsto \varepsilon(h)$ and we present the results in the general setting.

The main idea is to approximate the suprema of stochastic processes using finite sets in a sequential manner. For this covering numbers and metric entropies are used.

Definition D.0.1 (Covering Number and Metric Entropy)**.** Let (S, d) be a semimetric[1] space. For $\epsilon > 0$ the *covering number* $N(d, S, \epsilon)$ is the minimum amount of balls of radius ϵ with centers in S whose union contains S. The logarithm of the covering number $\log N(d, S, \epsilon)$ is called *metric entropy*.

Let $\Phi : [0, \infty) \to [0, \infty)$ a function that is positive, convex, increasing, and satisfies $\Phi(0) = 0$. Then we define the Φ-Orlicz semi norm $\|\cdot\|_\Phi$ as

$$\|X\|_\Phi = \inf \left\{ t : \mathbb{P}\Phi(|X|/t) \leq 1 \right\}. \tag{D.1}$$

By dominated convergence a random variable X has finite Φ-norm if and only if $\mathbb{P}\Phi(|X|) < \infty$. Just as in the case of $\mathcal{L}^p$ spaces of p-integrable random variables, the space $\mathcal{L}^\Phi$ of random variables for which the norm $\|\cdot\|_\Phi$ is finite is a (Banach) normed space after identification of random variables according to almost sure equality.

A bound on the Orlicz norm of a random variable implies rates of decay for its tails. Indeed, if $\|X\|_\Phi \leq \tau$, then by Markov's inequality

$$\mathbb{P}\left\{|X| \geq t\right\} = \mathbb{P}\left\{\Phi(|X|/\tau) \geq \Phi(t/\tau)\right\} \leq \frac{1}{\Phi(t/\tau)}.$$

This means that if Φ is of the form

$$\Phi(t) = \mathrm{e}^{\phi(t)} - 1 \tag{D.2}$$

[1] Recall that a semimetric d over a set S is a symmetric function $d : S \times S \to \mathbb{R}^+$ that satisfies the triangle inequality and $d(s, s) = 0$ for each $s \in S$. Note that the fact that $d(s_1, s_2) = 0$ does not imply that $s_1 = s_2$, which is why semimetrics fail in general to be metrics.

for a increasing non negative convex function ϕ that satisfies $\phi(0) = 0$ (see Section 2.5), then the previous implies bounds of the form

$$\mathbb{P}\left\{|X| \geq t\right\} \leq \mathrm{e}^{-\phi(t/\tau)}.$$

It is also true that an inequality of this form implies a bound on the Orlicz norm induced by a function Φ of the form of Equation (D.2) in certain situations. In those situations, Orlicz norms characterize tail behavior of random variables. We explore two cases. Since we are interested in studying successions of random variables such as $E_n(h)$, we introduce the parameter n in the following definitions. With this in mind, define

$$\Phi_p(t) = \mathrm{e}^{t^p} - 1 = \mathrm{e}^{\phi_p(t)} - 1$$

and

$$\Phi^L_{\text{Bern}}(t) = \mathrm{e}^{\left(\frac{\sqrt{1+2Lt}-1}{L}\right)^2} - 1 = \mathrm{e}^{\phi^L_{\text{Bern}}(t)} - 1$$

The converse holds true for ϕ_p [Van der Vaart and Wellner, 1996, Lemma 2.2.1] and for ϕ^L_{Bern} [Van de Geer and Lederer, 2013, Lemma 2]. We write down the results in the following lemma

Lemma D.0.2. Let $\{X_n\}$ be a succession of random variables. The following holds.

- If $\mathbb{P}\left\{n^{1/p}|X_n| \geq t\right\} \leq C\mathrm{e}^{-\phi_p(t/\tau)}$, then $\left\|n^{1/p}X_n\right\|_{\Phi_p} \leq (1+C)^{1/p}\tau$
- $\mathbb{P}\left\{n^{1/2}|X_n| \geq t\right\} \leq 2\mathrm{e}^{-\phi^{L/\sqrt{n}}_{\text{Bern}}(t/\tau)}$, then $\left\|n^{1/2}X_n\right\|_{\Phi^{\sqrt{3}L/\sqrt{n}}_{\text{Bern}}} \leq \sqrt{3}\tau$

If $X_1, \ldots, X_N$ is a collection of random variables for which $\|X_i\|_\Phi \leq \tau$. It is easy to bound the expectation of $\max_i X_i$ through Jensen's inequality

$$\Phi(\mathbb{P}[\max_i |X_i|/\tau]) \leq \mathbb{P}[\max_i \Phi(|X_i|/\tau)] \leq \sum_i \mathbb{P}[\Phi(|X_i|/\tau)] \leq N$$

so that

$$\mathbb{P}[\max_i |X_i|] \leq \tau\Phi^{-1}(N).$$

On the other hand, with a little more work, it is possible to prove that if there is a constant c such that $\Phi(x)\Phi(y)\Phi(cxy)$ remains bounded as $x, y \to \infty$ (which is the case for Φ_p and Φ^L_{Bern}) then

$$\left\|\max_i \Phi(|X_i|/\tau)\right\|_\Phi \lesssim \tau\Phi^{-1}(N).$$

Consider now an uncountable family of random variables $\{X_t\}_{t\in T}$. Through a series successive approximations in finite subsets of T the last equations can be used to obtain an bound for $\|\sup_{t\in T}|X_t|\|_\Phi$.

Theorem D.0.3. *Let Φ be a convex, non decreasing, nonzero function with $\Phi(0)$ and*

$$\limsup_{x,y\to\infty} \Phi(x)\Phi(y)/\Phi(cxy) < \infty$$

for some constant c. Let $\{X_t\}_{t\in T}$ be a separable stochastic process on the semi metric space (T,d) with

$$\|X_s - X_t\|_\Phi \leq Cd(s,t)$$

for every $s,t \in T$ and a constant C. Then there exists a constant K depending only on Φ such that for $t_0 \in T$

$$\left\|\sup_t |X_t|\right\|_\Phi \leq \|X_{t_0}\|_\Phi + K\int_0^{\mathrm{diam}T} \Phi^{-1}(N(d,T,\epsilon))\mathrm{d}\epsilon$$

where $\mathrm{diam}T$ is the diameter of T.

The following lemma will be useful.

Lemma D.0.4. let (T,d) be a semi metric space and $\{X_t\}_{t\in T}$ a stochastic process with almost surely Lipschitz continuous paths, that is,

$$|X_s - X_t| \leq Ld(s,t)$$

almost surely for every $s,t \in T$. If $\|X_t\|_\Phi < \infty$ for each t, then

$$\|X_s - X_t\|_\Phi \leq Cd(s,t).$$

for some constant $C > 0$.

Proof. It is easy to check from the definition of the Orlicz norm $\|\cdot\|_\Phi$ (see Equation (D.1)) that

$$\begin{aligned}\|X_s - X_t\|_\Phi &= \inf\{K : \mathbb{P}[\Phi(|X_s - X_t|/K)] \leq 1\}\\ &\leq \inf\{K : \Phi(Ld(s,t)/K) \leq 1\}\\ &= \frac{L}{\Phi^{-1}(1)}d(s,t)\end{aligned}$$

so that taking $C = L/\Phi^{-1}(1)$ we obtain the result. □

Corollary D.0.5. *Let $\{X\}_{t\in T}$ be a stochastic process. Assume that there is a semi metric d on T and that X has almost surely Lipschitz continuous paths. Assume that X satisfies an inequality of the form*

$$\mathbb{P}\left\{n^{1/p}|X| \geq t\right\} \lesssim \mathrm{e}^{-n\phi(t/\tau)} \tag{D.3}$$

as in Lemma D.0.2. Assume that there is a t_o such that $X_{t_0} = 0$ almost surely and define $\Phi(x) = \mathrm{e}^{\phi(x)} - 1$. In the case that $\phi = \phi_{Bern}^L$ then

$$\left\|n^{1/2}\sup_{t\in T}|X_t|\right\|_{\Phi_{Bern}^{L/\sqrt{n}}} \lesssim \int_0^{\mathrm{diam}\mathcal{H}} \left(\phi_{Bern}^{L/\sqrt{n}}\right)^{-1}(\log(N(d,\mathcal{H},\epsilon)+1))\,\mathrm{d}\epsilon.$$

In the case that $\phi = \phi_p$, then

$$\left\| n^{1/p} \sup_{t \in T} |X_t| \right\|_{\Phi_p} \lesssim \int_0^{\operatorname{diam} \mathcal{H}} (\phi_p)^{-1} \left(\log(N(d, \mathcal{H}, \epsilon) + 1) \right) \mathrm{d}\epsilon.$$

Proof. We prove the case for ϕ_p, the other case is similar. From Lemma D.0.2 it follows that

$$\left\| n^{1/p} |X| \right\|_{\Phi_p} \lesssim \tau$$

for each $h \in \mathcal{H}$. With this in mind, from Lemma D.0.4 it follows that

$$\left\| n^{1/p} (X_t - X_s) \right\|_{\Phi_n} \leq C d(h_1, h_2)$$

for each $s, t \in T$ and some constant $C > 0$. Thus, we can apply Theorem D.0.3 by noticing that that the Lipschitz continuity of $t \mapsto X_t$ makes it separable. □

This implies rates of order $n^{-1/p}$ for ϕ_p and of order $n^{-1/2}$ for ϕ^L_{Bern} as long as the integral converges. This is the case for ϕ_p because, if all the conditions in the previous theorem are satisfied, the Orlicz norm

$$\left\| n^{1/p} \sup_{t \in T} |X_t| \right\|_{\Phi_p} \leq \tau$$

for some finite $\tau > 0$ involving the integral of the metric entropy given that it converges. Then, by our previous considerations

$$\mathbb{P} \left\{ n^{1/p} \sup_{t \in T} |X_t| \geq t \right\} \leq \mathrm{e}^{-\phi_p(t/\tau)} = \mathrm{e}^{-(t/\tau)^p},$$

which means that with probability higher than $1 - \delta$ the inequality

$$\sup_{t \in T} |X_t| \leq \frac{\tau}{n^{1/p}} \left(\log \frac{1}{\delta} \right)^{1/p}$$

holds, a $n^{-1/p}$ rate. Similarly, a $n^{-1/2}$ rate holds for ϕ^L_{Bern}.

Popular Summary

Many problems in statistics, pattern recognition, and machine learning can be formulated in the following way. We are given a set of data $Z_1, \ldots, Z_n$ in which each data point is produced by a unique mechanism independently from the rest. We have a set of hypotheses $\mathcal{H}$ from which we want to choose one. And we have a notion of what it means for a decision to be bad given a data point, encoded on a function ℓ, called *loss function*. This function depends on which data point we observe and which hypothesis we choose, that is, $\ell = \ell(h, Z)$. The goal is then to choose a hypotheses such that the expected loss $L(h)$ (over the data) is small. Examples of these problems include prototypical machine learning and statistical problems such as classification, regression and density estimation. In these problems, hypotheses with low expected losses are better at making predictions on future data.

The expected loss of each hypothesis is, however, usually unknown. Perhaps the most intuitive idea is to minimize an empirical estimate of it. In order to estimate the expected loss using the data, one can use the empirical average $\hat{L}_n(h)$ of the loss function (over the data)

$$\hat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i).$$

We then choose an element h_n^* that minimizes $\hat{L}_n(h)$ and use it as our best guess. This technique is called Empirical Risk Minimization and, remarkably, it has proven to be very useful. Since also the lowest expected loss is unknown, for any estimate $\hat{h}_n$ the main quantity of interest is how far it is from optimal, that is,

$$L(\hat{h}_n) - \inf_{h' \in \mathcal{H}} L(h'),$$

called expected excess loss. Notably, and in agreement with our intuition, choosing elements $\hat{h}_n$ based on data that have a small $\hat{L}_n(\hat{h}_n)$ can and does (in many situations) lead to small values of the expected excess loss.

In this work we investigated how the expected excess loss of data-based estimates $\hat{h}_n$ decreases to zero for different choices of $\hat{h}_n$ as the number of data points n increases. In that case, it is known that under weak conditions the expected excess loss decreases at a rate of $n^{-1/2}$, which we called *slow rate*. We considered situations in which this rate can be improved to be n^{-1}, which we called *fast rate*. These conditions are well known when losses are bounded; we investigated them in the unbounded case. Achieving fast rates means that for decreasing the expected excess loss by a factor of 100, ~ 100 times more data points are needed, while ~ 10000 times more are needed in the slow case.

Bibliography

Jean-Yves Audibert. *Théorie statistique de l'apprentissage : une approche PAC-Bayésienne*. PhD thesis, Paris 6, January 2004. URL http://www.theses.fr/2004PA066003.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, August 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS623. URL https://projecteuclid.org/euclid.aos/1245332827.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, July 2006. ISSN 0178-8051, 1432-2064. doi: 10.1007/s00440-005-0462-3. URL https://link.springer.com/article/10.1007/s00440-005-0462-3.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, November 2005. ISSN 1292-8100, 1262-3318. doi: 10.1051/ps:2005018. URL https://www.cambridge.org/core/journals/esaim-probability-and-statistics/article/theory-of-classification-a-survey-of-some-recent-advances/42A9912D17169A650AB06244820464BC.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, February 2013. ISBN 978-0-19-953525-5. Google-Books-ID: koNqWRluhP0C.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, NY, USA, 2006. ISBN 978-0-471-24195-9.

Peter D. Grünwald and Nishant A. Mehta. Fast Rates for General Unbounded Loss Functions: from ERM to Generalized Bayes. *arXiv:1605.00252 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1605.00252. arXiv: 1605.00252.

L. Devroye L. Gyorfi, Gabor Lugosi, and L. Devroye. *A probabilistic theory of pattern recognition*. Springer-Verlag, 1996.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, December 1999. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1017939240. URL https://projecteuclid.org/euclid.aos/1017939240.

David McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. *arXiv:1307.2118 [cs]*, July 2013. URL http://arxiv.org/abs/1307.2118. arXiv: 1307.2118.

David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003.

David A. McAllester. Some PAC-Bayesian Theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 230–234, New York, NY, USA, 1998. ACM. ISBN 978-1-58113-057-7. doi: 10.1145/279943.279989. URL http://doi.acm.org/10.1145/279943.279989.

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

D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984. ISBN 978-1-4612-9758-1. URL //www.springer.com/us/book/9781461297581.

Malempati Madhusudana Rao and Zhong Dao Ren. *Theory of Orlicz spaces*. M. Dekker New York, 1991.

M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28–76, January 1994. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176988847. URL https://projecteuclid.org/euclid.aop/1176988847.

Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer-Verlag, Berlin Heidelberg, 2014. ISBN 978-3-642-54074-5. URL //www.springer.com/la/book/9783642540745.

Sara Van de Geer and Johannes Lederer. The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1-2):225–250, October 2013. ISSN 0178-8051, 1432-2064. doi: 10.1007/s00440-012-0455-y. URL https://link.springer.com/article/10.1007/s00440-012-0455-y.

A. W. Van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 978-0-387-94640-5. URL //www.springer.com/la/book/9780387946405.

Tim van Erven. PAC-Bayes Mini-tutorial: A Continuous Union Bound. *arXiv:1405.1580 [stat]*, May 2014. URL http://arxiv.org/abs/1405.1580. arXiv: 1405.1580.

Tim Van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Vladimir Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer-Verlag, New York, 2 edition, 2000. ISBN 978-0-387-98780-4. URL //www.springer.com/br/book/9780387987804.

Vladimir N. Vapnik and A. Ja Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk. *Pattern Recognition and Image Analysis*, 1(3):284–305, 1991.

Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Wing Hung Wong and Xiaotong Shen. Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339–362, April 1995. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176324524. URL https://projecteuclid.org/euclid.aos/1176324524.

Tong Zhang. From -entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, October 2006a. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053606000000704. URL https://projecteuclid.org/euclid.aos/1169571794.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, April 2006b. ISSN 0018-9448. doi: 10.1109/TIT.2005.864439.