Diss. ETH No. 18908

Oracle inequalities using truncation and cross-validation

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
CHARLES MOVERLY MITCHELL
Dipl. Math. ETH
born April 1st, 1980
citizen of Allschwil BL and British citizen

accepted on the recommendation of
Prof. Dr. Sara van de Geer, examiner
Prof. Dr. Peter Bühlmann, co-examiner

2010

First and foremost, I'd like to thank my advisor, Sara van de Geer, for all the opportunities and collaboration she has given me over these last few years. She has enabled this algebraicist to branch out into statistics and learn a new trade. Her attention to detail has been of immeasurable use to me, especially in fixing flawed proofs in a hurry. Despite being an impeccable theoretician, she has also displayed wise pragmatism in assessing what is important, and what is not. Sara has probably given me more advice than I managed to take in, though I've tried to heed the important parts. Moreover, she enabled – and encouraged – me to attend important conferences and workshops at a key time during my work. Oberwolfach certainly was impressive.

I'd also like to thank my co-examiner, Peter Bühlmann. I think it is safe to say that an important contribution he makes to everybody's theses at the Seminar for Statistics – not just those of his own students – is in creating a wonderfully friendly and co-operative environment at this institute. The Seminar for Statistics is a very sociable and active place, not just academically, but for instance also in sports, and in this Peter is very much a leader by example.

Academic work is never an entirely solitary pursuit, even if it may sometimes seem so. In my work on cross-validation, Guillaume Lecué has been a very helpful collaborator. His quick thinking and sharp theoretical mind were sometimes too fast for me to follow, but have proven invaluable to me in an entire section of my work. For that, I am very thankful.

Further thanks extends to Bernadetta Tarigan, with whom I shared my first office at the Seminar for Statistics, and who has been and remained both helpful and a good source of conversation. Thank you especially, Bernadetta, for ploughing through my manuscript and pointing out how it appears to the "ordinary reader" (if there is such a person).

In summary, I'd like to thank the entire Seminar for Statistics and its studious yet cheery denizens for a wonderful and blessed 3 1/2 years here. Where there's always something new to learn, where people have broad interests beyond the bare academic necessities, and where even each tea break becomes a fascinating window on the world. Thank you all!

Contents

Data-driven model selection is becoming ever more important. With the advent of strong computing power and the ever-increasing number of estimating methods available for a given problem, it has become more possible – and more crucial – to compute a large number of models and then choose between them using empirical criteria. Understanding how the properties of such a selection-based estimator derive from those of the candidate estimators is then key.

A typical setup we are interested in involves a continuous, infinite class of models and a *finite* set of estimation procedures which take an arbitrary amount of training data and yield estimators in the full class. The procedures may be completely different, or they may differ only by the tuning parameters they use. A family of loss functions – indexed in the full model class – describes the "regret" or loss incurred when a specific model is chosen and a particular sample observed. Each loss gives rise to a true expected loss, or risk, over the true distribution of the data – which is then approximated empirically for the purpose of model selection. For a given set of (finitely many) training data, the true model (by which we merely mean the model associated with the globally smallest true risk) may not actually be produced by one of the estimation procedures. Instead, there is some "best estimator", the so-called *oracle*, one of the finitely many estimators trained by the data. Ideally, we would like to use the oracle for estimation, but to find it in every situation does not constitute an estimator computable from the finitely many data available. However, it is the performance of this oracle that provides the theoretical benchmark to which the performance of the selected estimator is compared.

In this thesis, we investigate to what extent single-split and cross-validation procedures have risks close to that of the oracle. Such proximity of the selected model to the best candidate model is formulated in terms of so-called *oracle inequalities*, the type of result we show here. To obtain such results, we use a line of argument that takes tail (large-loss) conditions, and crucially also margin (small-loss) conditions controlling the noise close to the optimum, and use empirical process theory with them. The results are quite general, permitting application to different shapes of margin conditions and to loss functions that only have power tails, rather than exponential ones. Thus examples from the three main statistical problems, regression, classification and density estimation, can be treated simultaneously with this theory.

Datenbasierte Modellwahl wird immer wichtiger. Mit der Verfügbarkeit starker Rechenleistung, sowie einer zunehmenden Anzahl von Schätzmethoden in jeder beliebigen Situation, ist es nötig – und unabdingbar – geworden, viele potentielle Modelle anzupassen und dann mittels empirischer Kriterien unter ihnen auszuwählen. Wichtig ist dann zu verstehen, wie die Eigenschaften eines solchen durch empirische Modellwahl produzierten Schätzers sich aus den Eigenschaften der kandidierenden Schätzer zusammensetzen.

In einer typischen Situation, die uns interessiert, haben wir eine kontinuierliche, unendliche Familie von Modellen sowie eine *endliche* Klasse von Schätzverfahren, die aus einer beliebigen Menge von Trainingsdaten Schätzer in der grossen Klasse produzieren. Die Schätzverfahren können ganz verschiedener Natur sein, oder sich auch einfach durch einzelne Parameter unterscheiden. Eine Familie von Verlustfunktionen mit Index in der gesamten Modellklasse beschreibt den Verlust (oder die Inkompatibilität) eines Modells bei bestimmten potentiellen Daten. Jede Verlustfunktion führt zu einem erwarteten Verlust, d.h. ein Risiko, für die wahre Verteilung der Daten, und es ist dieses Risiko, dass zwecks Modellwahl empirisch genähert wird. Für eine gegebene (endliche) Menge von Trainingsdaten wird das wahre Modell (was hier lediglich das Modell mit dem global kleinsten Risiko bedeutet) nicht unbedingt von einer der kandidierenden Schätzverfahren produziert. Stattdessen gibt es einen "besten Schätzer", das sogenannte *Orakel*, das eines der endlich vielen Schätzer ist. Wir würden sehr gerne dieses Orakel verwenden, aber es ist in der Praxis unmöglich, diesen mit endlich vielen Daten zu finden: es ist kein Schätzer, das nur die Daten verwendet. Das Orakel gibt uns jedoch den theoretischen Vergleichswert, an dem wir die Güte eines auf Modellwahl beruhenden Schätzers messen können.

In dieser Doktorarbeit untersuchen wir, inwiefern Modellwahlverfahren mit einfacher Datenspaltung, sowie das mehrfach spaltende Verfahren der Kreuzvalidierung, Risiken erlangen, die ähnlich tief liegen wie das des Orakels. Diese Nähe des Schätzers zum besten Kandidaten wird als sogenannte *Orakelungleichung* ausgedrückt, einer Form, der wir uns hier bedienen. Um zu unseren Resultaten zu gelangen, betrachten wir Bedingungen auf den Wahrscheinlichkeiten hoher Verluste, und ganz zentral auch Bedingungen auf dem Verhalten kleiner Verluste (Margin-Bedingungen), damit wir nahe bei der Wahrheit noch die Varianz unter Kontrolle haben. Wir verwenden empirische Prozesse, um unter diesen

Veraussetzungen zu allgemeinen Resultaten zu gelangen, welche für unterschiedlich strenge Margin-Bedingungen und unterschiedlich schnell abklingende Wahrscheinlichkeiten hoher Verluste (nicht nur exponentiell abklingende, sondern auch solche, die nur in der Ordnung einer Potenz abklingen) ihre Gültigkeit beibehalten. So lassen sich mit dieser Theorie simultan Beispiele aus den drei klassischen Statistikproblemen der Regression, Klassifikation und Dichteschätzung behandeln.

Chapter 1

Empirical Processes and M-Estimators

1.1 Introduction

This thesis is about model selection, the data-driven choice of models and their tuning parameters when we have no other way of knowing which choice is best. The procedures we use are neither new nor complex: it is on simple risk minimization and the popular technique of cross-validation that our particular focus lies here. Difficulties arise primarily when we try to make statements on the performance of such supposedly simple procedures – all the more so when these statements are to be wide-ranging in their nature. Thus while this work is theoretical in nature, it is our endeavour that it address issues that are relevant in realistic examples. As hinted at in the dissertation title, the two key "realistic" issues that are addressed are power tails—tackled by truncation arguments in Chapter 2—and cross-validation, which is the subject of Chapter 3.

1.2 The empirical distribution

At its core, statistics concerns itself with the extraction of knowledge on random variables and their connections using only a finite amount of data. There generally is some unknown distribution P involved in computing the quantities of interest (by probabilities and expectations). The simplest, most general—and indeed most common—means of working with this is to at some stage replace the unknown distribution P by a distribution that approximates it; if n independent and identically distributed data $X_1, \ldots, X_n \sim P$ are present, such an approximation is to be found in the shape of the empirical distribution

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

the average of the Dirac distributions $\delta_{X_i} : A \mapsto 1_A(X_i)$ for the given data. This succinct approximation enables the estimation of any function $f(X)$ of a random variable X following the probability law P, where f takes values in a vector space; the estimate here is $\widehat{f(X)} := P_n f(X) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$. Here by a common abuse of notation, P_n denotes the expectation with respect to the empirical distribution. Furthermore, moments $E[|X|^p]$ can be estimated by empirical moments $P_n|X|^p = \frac{1}{n} \sum_{i=1}^{n} |X_i|^p$, enabling the construction of moment estimators that estimate parameters of a distribution by taking those for which the empirical moments match the true ones, and so forth.

1.3 M-estimators

A frequent aim in the construction of estimators is that some functional of them be minimized or maximized. Most often this takes the shape of minimizing an expectation of the form $R(f) := P\ell(f; X)$ over some class $\mathcal{F} \ni f$ of potential parameters. This can be interpreted as minimizing a *risk* R, which in turn is computed from a loss $\ell(f, x)$ that quantifies the negative consequences (cost, distance, regret or similar) when a new datum x is observed after f has been chosen as the (estimated) parameter.

The empirical distribution from the available data, then, can be used

to estimate the true risk $R(f)$ incurred by a parameter choice by way of its empirical risk $R_n(f) := P_n\ell(f; X)$. Thus the true risk minimizer

$$f_* := \min_{f\in\mathcal{F}} R(f)$$

can be estimated by the empirical risk minimizer

$$\hat{f} := \min_{f\in\mathcal{F}} R_n(f) \ .$$

As this estimator minimizes (or in other situations, maximizes) a criterion, it is termed an *M-estimator*. If the target criterion is the additive inverse of a log-likelihood function, we precisely obtain the maximum likelihood estimator of the parameter of interest. Estimating the location of a univariate real-valued distribution by the mean or median of sample data constitutes M-estimation, and the loss functions are the L^2-loss $\ell(\mu, x) = (\mu - x)^2$ and the L_1-loss $\ell(\mu, x) = |\mu - x|$, respectively. Naturally, the choice of the underlying class $\mathcal{F}$ is crucial for the process of M-estimation, too: in regression, for example, where the regression function f can be the parameter of interest, smaller classes (such as that of linear functions) lead to more regularized but inflexible solutions, whereas larger classes (such as that of all integrable functions) lead to more variable and closer-fit parameter choices. This is the typical bias-variance tradeoff. The set $\mathcal{F}$ of target parameters is frequently not a fixed set, but increases with the sample size n; the resulting parameter sets $(\mathcal{F}_n)_{n\in\mathbb{N}}$ should ideally constitute a sieve that is dense in some larger parameter class $\mathcal{F}_0$ containing the "true" parameter, the global optimum f_0 of the risk R.

1.4 Empirical processes

The asymptotic properties of what is perhaps the simplest and most common M-estimator, the arithmetic mean, are very easy to derive: the law of large numbers gives consistency under simple conditions, and the central limit theorem subsequently supplies the rate and limiting distributions for the same. An arbitrary M-estimator, however, has neither an explicit expression, nor is it so easy to treat. However, the quantity being minimized—the empirical risk $P_n\ell(f, \cdot)$—is an arithmetic mean; thus for each fixed parameter f, we generally know the behaviour

of $P_n\ell(f,\cdot)$ under simple conditions. In particular, we know how the "empirical increment" $P_n\ell(f,\cdot)-P\ell(f,\cdot)$ concentrates around zero when the sample size is made arbitrarily large, and we know the asymptotic behaviour of the normalized increment $\sqrt{n}\cdot(P_n\ell(f,\cdot)-P\ell(f,\cdot))$. The corresponding M-estimator, however, depends on the whole *class* of these increments indexed in $f\in\mathcal{F}$. Such a class is a stochastic process termed an *empirical process*; typically, it is the normalized process

$$\mathbb{G}_n\ell(f,\cdot) := \sqrt{n}\cdot(P_n\ell(f,\cdot)-P\ell(f,\cdot))\ , f\in\mathcal{F}$$

which is referred to as the empirical process. See also [50], p.42, and [52], p. 80ff., for this definition, and especially the latter reference for the most of the discussion of the empirical process contained in this chapter. The example-based brief introduction to empirical processes given in [39] formulates the empirical process quite nicely as "an operator that acts on a function ... to produce a properly standardized sample average". This is perhaps also a good way to think of it here: an empirical process merely transforms a class of functions we are studying (such as loss functions) to yield approximation bounds for the (estimation) procedure we are performing. Any bounds or geometric properties of this function class that may be known to us can be used to handle the empirical process.

1.5 Uniform asymptotics

With the simultaneous description of a class of loss functions in an empirical process also comes the need to examine the properties of these losses simultaneously, or *uniformly*. If we perform estimation in the class $\mathcal{F}$ and thus obtain the loss class $\mathcal{G} := \{\ell(f,\cdot)|f\in\mathcal{F}\}$, the empirical process $\mathbb{G}_n$ only tells us something about the quality of this estimation if we know the worst-case behaviour of this process, i.e. if we can show some properties of its supremum

$$\sup_{g\in\mathcal{G}}|\mathbb{G}_n g| =: \|\mathbb{G}_n\|_{\mathcal{G}} = \sqrt{n}\cdot\|P_n-P\|_{\mathcal{G}}\ .$$

Empirical process theory essentially is the study of empirical processes and their supremum bounds. Thus the central question of empirical process theory is as follows: Given a distribution P on some space

S, a class of functions $g : S \to \mathbb{R}$ and a convex increasing function Φ on $\mathbb{R}^+$, find bounds for the expression

$$P^{\otimes n}\Phi\left(\|P_n - P\|_{\mathcal{G}}\right) \ .$$

See Pollard's introductory volume [40] for a succinct presentation and treatment of this problem.

Some results on the suprema of empirical processes are in fact uniform versions of point results. The most basic property that should be required of an empirical process is the uniform convergence to 0 of its supremum, constituting a uniform law of large numbers. This is the *Glivenko-Cantelli property*

$$\|P_n - P\|_{\mathcal{G}} \xrightarrow{P} 0 \ (n \to \infty),$$

best known from the Glivenko-Cantelli theorem on uniform convergence of the empirical distribution function $F_n(x) := P_n(X \leq x)$ to the true distribution function, a situation which corresponds to the family $\mathcal{G} = (1\{(-\infty, c]\})_{c\in\mathbb{R}}$ of functions on the real line.

Theorem 1.5.1 (Glivenko-Cantelli; Theorem 19.1 of [50])**.** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then $\|F_n - F\|_\infty \xrightarrow{\text{as}} 0$.*

The Glivenko-Cantelli property is at heart a geometric one, and can be shown using bracketing numbers (or, similarly, using covering numbers):

Theorem 1.5.2 (Theorem 2.4.1 of [52])**.** *Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is Glivenko-Cantelli.*

Here a bracket $[l, u]$ bounded by two functions l and u—themselves not necessarily in $\mathcal{F}$—is the set of all functions $f \in \mathcal{F}$ with $l \leq f \leq u$, and the ε-bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of brackets with $\|u - l\| < \varepsilon$ needed to cover $\mathcal{F}$ (Definition 2.1.6 of [52]). A classic geometrical quantity that is used for bounding the bracketing numbers is the *Vapnik-Chervonenkis dimension* (or Vapnik-Chervonenkis index), which for a class of sets denotes the maximum size of a set of which the class can distinguish all subsets (by intersection), and for a function f is just the Vapnik-Chervonenkis dimension

of its subgraph $\{(x,t) : t < f(x)\}$. Function classes with finite Vapnik-Chervonenkis dimensions generally have covering numbers that only grow polynomially in $1/\varepsilon$; this gives them a low complexity and is the basis for good bounds on their empirical processes. See [52], Chapter 2.6, for these definitions and the theory associated with them, and [53], Chapter XI, for background on shatter coefficients, which are closely related. Examples of the Vapnik-Chervonenkis dimension of classes include:

- Any finite class $\mathcal{A}$ of subsets of some $\mathbb{R}^d$ has a Vapnik-Chervonenkis dimension which is at most $\log_2 |\mathcal{A}|$. ([18], Theorem 13.6)
- The class of lower cells $(-\infty, c] := \{x \in \mathbb{R}^d : x_j \leq c_j \forall j\}$ in d-dimensional real space has Vapnik-Chervonenkis dimension $d+1$. ([52], Example 2.6.1)
- The class of all closed balls $\{x \in \mathbb{R}^d : \|x-a\|_2 \leq b\}$ (where $a \in \mathbb{R}^d$ and $b > 0$) in $\mathbb{R}^d$ has Vapnik-Chervonenkis dimension $d+2$. ([18], Corollary 13.2)
- The class of all convex polygons in $\mathbb{R}^2$ has infinite Vapnik-Chervonenkis dimension. ([18], Theorem 13.10)

Once such a uniform law of large numbers has been established, the next property of interest is a uniform central limit theorem. In analogy to the normal distribution of the limit of $\sqrt{n}(P_n - P)(g)$ in the univariate central limit theorem, a uniform central limit theorem should contain the convergence of the whole stochastic process $(\sqrt{n}(P_n - P)(g))_{g \in \mathcal{G}}$ to a Brownian bridge on $\mathcal{G}$, with the covariance at any pair $g_1, g_2 \in \mathcal{G}$ being $P[(g_1 - Pg_1)(g_2 - Pg_2)])$ for a random variable X with distribution P. Such an empirical process $\mathbb{G}_n$ with weak convergence $\mathcal{G}_n \rightsquigarrow \mathcal{G}$ in $\ell^\infty(\mathcal{F})$ to a P-Brownian bridge is a called a P-*Donsker class* ([52], p.81). The set of all monotone functions $f : \mathbb{R} \to [0,1]$ is P-Donsker for every probability measure on $[0,1]$ ([52], Example 2.6.21). In general, we usually require both an entropy condition and some envelope (moment) bound to obtain the Donsker property. In this thesis, however, we perform finite model selection and can thus dispense with entropy conditions, using the cardinality of the function class (i.e. the number of candidate models) instead.

1.6 Orlicz norms

Beside the structure of the empirical process—or the cardinality of its index set—the random variables that give rise to it must also be controlled. Naturally this could be done by absolute bounds, but they are too rigid, and in many examples they do not hold. Moment bounds or exponential moment bounds are good alternatives. At the most general level, such bounds can be expressed using Orlicz norms (see [40], p.3):

Definition 1.6.1. Let X be a real-valued random variable, and let $\Phi : \mathbb{R} \to \mathbb{R}^+$ be a convex, increasing function on $\mathbb{R}^+$ with $\Phi(0) \in [0, 1)$. Then the Φ-*Orlicz norm* of X is defined as

$$\|X\|_\Phi := \inf \left\{ C > 0 : P\Phi\left(\frac{|X|}{C}\right) \leq 1 \right\} .$$

This is a norm, and through this it defines a complete normed space $\mathcal{L}^\Phi$ of those real-valued random variables (up to almost sure equality) with finite $\|\cdot\|_\Phi$, the Φ-Orlicz space.

Any random variable with a finite Orlicz norm has a corresponding tail bound by a simple application of Markov's inequality (see [52], p. 96):

$$P(|X| > z) \leq \frac{1}{\Phi\left(\frac{z}{\|X\|_\Phi}\right)} .$$

The functions $\Phi = \phi_p : x \mapsto x^p$ yield the standard L^p-norms as their Orlicz norms (when $p \geq 1$). The other important class generating Orlicz norms consists of the functions

$$\psi_p : x \mapsto \exp(x^p) - 1$$

for $p \geq 1$. The tail bounds these give rise to are roughly exponential. For instance, bounds on the ψ_1-Orlicz norm of X, which we shall use here, imply the exponential moment bound

$$E\left[\exp\left(\frac{|X|}{M}\right) - 1 - \frac{|X|}{M}\right] \cdot M^2 \leq \frac{v}{2} \tag{1.1}$$

for the parameters $M := \|X\|_{\psi_1}$ and $v := 2 \cdot \|X\|_{\psi_1}^2$. Up to scaling, Inequality 1.1 is equivalent to the family of moment bounds,

$$E\left[|X|^m\right] \leq m! M^{m-2} \cdot \frac{v}{2} \quad \forall m \geq 2 . \tag{1.2}$$

For the ψ_1-Orlicz norm to be finite, it suffices that there exist some finite value of $E\psi_1(|X|/C)$. The following lemma quantifies this:

Lemma 1.6.2 (Orlicz norm)**.** *Let X be a real-valued random variable, and assume that constants $C > 0, K > 0$ are known for which $\mathbb{E}\exp(|X|/C) \leq K$. Then for any $C' > C$,*

$$\mathbb{E}\exp(|X|/C') \leq 1 + K \cdot \frac{C}{C' - C} \ ,$$

and thus the ψ_1-Orlicz norm of X has the upper bound

$$\|X\|_{\psi_1} \leq (K+1)C \ .$$

As the functions ψ_p increase much more rapidly than the power functions ϕ_p, boundedness conditions on their Orlicz norms are stronger, and harder to obtain, than the corresponding conditions on the L^p-norms, and the resulting tail bounds decay more quickly. For instance, we have

$$\|X\|_p \leq p! \ \|X\|_{\psi_1}$$

(see [52], p. 95).

1.7 Concentration of empirical measure

An important input into the analysis of empirical processes takes the form of moment bounds or absolute bounds on the random variables involved. To draw conclusions about the behaviour of even just one empirical increment $(P_n - P)g$ from such conditions, we need inequalities that tell us how the empirical mean of a random variable concentrates around its true mean. Many useful concentration inequalities can be derived from the most basic one, Chebyshev's inequality:

Theorem 1.7.1 (Chebyshev's inequality)**.** *(Reference: [10], A.15.4) Consider a random variable $X \in \mathbb{R}$ with distribution P, and an increasing function $\phi : \mathbb{R} \to [0, \infty)$. Then for all a with $\phi(a) > 0$, we have*

$$P\left[X \geq a\right] \leq \frac{E\left[\phi(X)\right]}{\phi(a)} \ .$$

Two of the standard types of concentration inequalities are Hoeffding's and Bernstein's. The original references for these are given by [49] as being [24] and [9], respectively, though a plethora of versions of these can be found in the literature. Hoeffding's inequality uses hard bounds on the random variable involved, while Bernstein's inequality presupposes exponential moments:

Theorem 1.7.2 (Hoeffding's inequality)**.** *(Reference:[52], Prop. A.6.1) Let $X_1, \ldots, X_n$ be independent random variables taking values in $[0,1]$. Let $\mu = E\left[\bar{X}_n\right]$. Then for $0 < t < 1 - \mu$,*

$$P\left[\bar{X}_n - \mu \geq t\right] \leq e^{-2nt^2}.$$

Theorem 1.7.3 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with zero mean such that $E\left[|X_i|^m\right] \leq m! M^{m-2} v/2$, for every $m \geq 2$ (and all i) and some constants M and v. Then*

$$P\Big(\frac{1}{n}\sum_{i=1}^{n} X_i > x\Big) \leq e^{-\frac{1}{2}\frac{nx^2}{v+Mx}} \ .$$

This form of Bernstein's inequality is one of the classical types, as e.g. [52], Lemma 2.2.11 is too. Its classical proof is also reprised at the end of this section. In Chapter 2 we shall prove variants of it that are tailored to the derivations of oracle inequalities we use it for. Other variants that it gives rise to include a version for the ψ_1-Orlicz norm that we shall use here:

Lemma 1.7.4. *Let $X, X_1, ..., X_m$ be i.i.d random variables such that $\|X\|_{\psi_1} < \infty$. Then, for any $u > 0$,*

$$\mathbb{P}\Big(\mathbb{E}X - \frac{1}{m}\sum_{i=1}^{m} X_i > 4u\|X\|_{\psi_1}\Big) \leq \exp\left(-\frac{m}{2} \cdot (u^2 \wedge u)\right) \ .$$

1.8 Literature

Contemporary books that provide good references for empirical process theory are van der Vaart and Wellner [52] and van de Geer [48], which provide a careful treatment and examples. The notation we use here mostly follows these books, and the former is the source of many remarks

in this chapter. A now more classical and succinct introduction to empirical process theory is given by Pollard [40], which provides the essentials, including Orlicz norms, symmetrization, chaining, maximal inequalities, uniform laws of large numbers and functional central limit theorems, in a series of concise chapters. That book is preceded by Pollard [39], a paper provides a more simply sketched introduction to the topic on the basis of location estimation. Looking back to the early days of empirical process theory, Giné and Zinn [20] and Dudley [19] give central limit theorems for empirical processes, and Vapnik and Chervonenkis' [53] seminal book on pattern recognition provides the starting point for Vapnik-Chervonenkis classes.

The exponential moment bound 1.1 is used by van der Vaart et al. [51] to define the concept of *Bernstein numbers* (M, v) of Z, in a setup where $Z = f(X)$ for a random variable X and a class of measurable, real-valued functions $\mathcal{F} \ni f$.

1.9 Proofs

Proof of Lemma 1.6.2 By Chebyshev's inequality, the known bound $\mathbb{E}\exp(|X|/C) \leq K$ gives rise to a tail bound $P[|X| \geq t] \leq K\exp(-t/C)$ for $t > 0$. Then for any $C' > C$,

$$\begin{aligned}
\mathbb{E}\exp(|X|/C') &= \int_0^\infty P[\exp(|X|/C') > t]dt \\
&= \int_0^\infty P[|X| > C'\ln(t)]dt \\
&\leq 1 + \int_1^\infty Kt^{-C'/C}dt \\
&= 1 + K \cdot \frac{C}{C'-C} \ .
\end{aligned}$$

This is ≤ 2 iff $C' \geq (K+1)C$, from which

$$\mathbb{E}\left[\exp\left(\frac{|X|}{(K+1)C}\right) - 1\right] \leq 1$$

follows. ■

Proof of Lemma 1.7.3 First we derive a bound for the expectation of $\exp(X_i/L)$, where L is an arbitrary scaling factor strictly greater than M:

$$\begin{aligned}\mathbb{E}\exp\left(\frac{X_i}{L}\right) &\leq 1+\sum_{m=2}^{\infty}\frac{m!M^{m-2}v/2}{m!L^m} \leq 1+\frac{v}{2L^2}\sum_{m=0}^{\infty}\left(\frac{M}{L}\right)^m \\ &= 1+\frac{v}{2L(L-M)} \leq \exp\left(\frac{v}{2L(L-M)}\right) .\end{aligned}$$

Now the independence of the X_i combines with this bound to give us the inequality

$$P\left[\frac{1}{n}\sum_{i=1}^{n}X_i > x\right] \leq \frac{\prod_{i=1}^{n}\mathbb{E}\exp\left(\frac{X_i}{L}\right)}{\exp(nx/L)} = \exp\left(\frac{n}{L}\left(\frac{v}{2(L-M)}-x\right)\right) .$$

Choosing $L = M + v/x$ then gives us the desired upper bound. ■

Proof of Lemma 1.7.4 Define the random variable $X' := E[X]-X$ and i.i.d. copies thereof, $X_i' := E[X] - X_i$. By the convexity and monotonicity of ψ_1, we have

$$\psi_1\left(\frac{|E[X]|}{\|X\|_{\psi_1}}\right) \leq \psi_1\left(\frac{E[|X|]}{\|X\|_{\psi_1}}\right) \leq E\left[\psi_1\left(\frac{|X|}{\|X\|_{\psi_1}}\right)\right] \leq 1 ,$$

and therefore $\|E[X]\|_{\psi_1} \leq \|X\|_{\psi_1}$. Thus $\|X'\|_{\psi_1} \leq 2\cdot\|X\|_{\psi_1}$ by the triangle inequality, and in particular $\|X'\|_{\psi_1} < \infty$. Because of this, the moment bound $E[|X'|^m] \leq m!M^{m-2}\cdot v/2$ holds for $M = \|X'\|_{\psi_1}$, $v = 2\cdot\|X'\|_{\psi_1}^2$ and for every $m \geq 2$. Now we can produce the following upper bound:

$$\begin{aligned}P\left(E[X]-\frac{1}{m}\sum_{i=1}^{m}X_i > 4u\|X\|_{\psi_1}\right) &= P\left(\frac{1}{m}\sum_{i=1}^{m}X_i' > 4u\|X\|_{\psi_1}\right) \\ &\leq P\left(\frac{1}{m}\sum_{i=1}^{m}X_i' > 2u\|X'\|_{\psi_1}\right) .\end{aligned}$$

By Lemma 1.7.3, this possesses an upper bound in

$$2\exp\left(-\frac{4mu^2\cdot\|X'\|_{\psi_1}^2}{2(2\|X'\|_{\psi_1}^2+\|X'\|_{\psi_1}\cdot 2u\|X'\|_{\psi_1})}\right) = 2\exp\left(-\frac{mu^2}{1+u}\right) ,$$

which in turn is bounded from above by

$$2\exp\left(-\frac{m}{2}\cdot(u^2\wedge u)\right)\ .$$

■

Chapter 2

Model Selection Oracle Inequalities using Truncation

This chapter is a revised and extended version of the paper [38].

2.1 Introduction

Consider a sample $Z_1, \ldots, Z_N$ of independent random variables in some space $\mathcal{Z}$, whose distribution depends on an unknown parameter f. To estimate f, we split the sample into two parts: a test set $Z_1, \ldots, Z_n$ and a training set $Z_{n+1}, \ldots, Z_N$. Based on the training set various estimators of f are constructed, say $\hat{f}_1, \ldots, \hat{f}_p$. To decide among these estimators, we use the test set. Suppose that $\gamma_f : \mathcal{Z} \to \mathbf{R}$ is a loss function. The final estimate $\hat{f}$ is now chosen to minimize the empirical risk:

$$\hat{f} := \arg \min_{\hat{f}_j : 1 \le j \le p} \frac{1}{n} \sum_{i=1}^{n} \gamma_{\hat{f}_j}(Z_i) \ .$$

In this chapter, we examine whether this empirical risk minimization

leads to taking, among the p estimators, the "nearly best" one. Here, "nearly best" will be defined in terms of the excess risk of the estimators.

The behavior of the excess risk near f will be called the margin behavior. We not only consider the classical case, which is quadratic margin behavior, but also a more general case. For the tails of our excess loss functions, we consider both an exponential moment condition and a more general power tail condition. We prove a risk inequality under the most general combination of these conditions, and in doing so automatically obtain risk inequalities for more restricted situations. These latter situations represent examples we give from regression, classification and density estimation.

A common and succinct way of expressing the quality of an aggregated estimator is by way of an oracle inequality of the form

$$\mathbf{E}R(\hat{\gamma}) \leq A \cdot \inf_{\gamma \in \mathbf{\Gamma}} R(\gamma) + C(\mathbf{\Gamma}, n) \ .$$

Here $R(\gamma) := \mathbf{E}_Z \gamma(Z)$ is the risk of the procedure that has loss γ, and $C(\mathbf{\Gamma}, n)$ is a quantity that depends on the cardinality (when finite) or complexity (such as the metric entropy) of the class $\mathbf{\Gamma}$ of models or aggregates up for selection, as well as on the sample size n.

When the number of procedures being aggregated is a finite number $p := |\mathbf{\Gamma}|$, most of the results in the literature set $\mathcal{O}(\log(p)/n)$ to be the benchmark for the rate of the term $C(\mathbf{\Gamma}, n)$ above. For instance, Bunea et al. [14] give this rate for Gaussian regression and a linear aggregate that minimizes a penalized sum of squares. For a more general risk problem, Györfi and Wegkamp [22] obtain a similar result, and Lecué [31] achieves the same rate for the Cumulative Aggregation with Exponential Weights (CAEW) procedure in a classification setup with bounded loss. Other types of results in this vein include Bartlett and Mendelson's [8] high probability bounds for the estimator risk of empirical risk minimization, done for the estimation of functions from a class with a uniform bound.

The analysis of empirical risk minimization stands on two major pillars. The first of these is empirical process theory. In Vapnik and Chervonenkis' seminal work on pattern recognition [53], the importance of the empirical process

$$((P_n - P)(f))_{f \in \mathcal{F}}$$

of the class $\mathcal{F}$ of candidate procedures for the study of empirical risk minimizers was already recognized. More recently, van de Geer [48] also describes the use of empirical processes in understanding such estimators. The second foundation we need is the study of concentration inequalities, which describe the concentration of random variables and their empirical means around their true means. The value of such inequalities in the analysis of model selection via empirical risk minimization is recognized, and put to use, in the papers of Barron et. al. [7] and Birgé and Massart [12].

In much of the literature, the quantities to be estimated are assumed to be uniformly bounded. Another very important condition for ensuring good rates in oracle inequalities is the margin condition, which controls the "noise" between procedures that differ only very slightly in risk, and thus makes assumptions on the *small-scale* behaviour of the family of losses. Thus the margin properties of a family of losses describe how easily identifiable the optimal loss is inside the whole family. For some regression setups, a uniform bound on the target and the estimates already dispenses with the need for a margin condition, as in the results of Bunea et al. [14]. (We shall see in Example 2.6.2 that such a uniform bound implies the margin condition when using L^2-loss.) In classification, though, which is the original area for margin conditions, the situation is somewhat more complex. Here the margin conditions that hold are generally weaker than the ones known in regression or density estimation setups. Tsybakov [46] provides a good treatment of this case. Koltchinskii [27] looks at a wider range of situations, generalizing Tsybakov's results, among others; besides a margin condition, his approach also requires direct conditions on the empirical process or on the complexity of the candidate class Γ in lieu of boundedness conditions.

Generally, most of the literature deals with only one particular problem, such as regression; furthermore, the strong boundedness conditions usually imposed are not always necessary. It is well-known that some conditions must be imposed in order to obtain risk rates that are better than $\mathcal{O}(1/\sqrt{n})$. For example, Lee et al. [33] give an overview of risk rates in an agnostic learning setup and show that convexity properties on the class of candidate functions lead to risk rates around $\mathcal{O}(1/n)$ rather than $\mathcal{O}(1/\sqrt{n})$. Mendelson [36] uses a least-squares regression example to also show that $\mathcal{O}(1/\sqrt{n})$ cannot be improved upon without

assuming something like a Bernstein-type inequality. (While convexity assumptions can suffice for obtaining fast risk rates, they are not always necessary, as also shown by Mendelson [37]). Our interest lies in inequalities for a general loss function setup, with boundedness conditions replaced by suitably loose requirements on the tails, at least when conditioning on the training set. Such conditioning on the training set is common practice; to average the results over the training data then requires margin and power tail conditions to hold uniformly over all trained versions of the estimators used, if possible – or if not, then other, possibly more stringent, conditions.

Another fairly general approach is taken by Audibert [3], who looks at the general prediction problem, i.e. regression and classification, and uses a progressive mixture rule for aggregation, but with only a brief reference to averaging over the training stage, which would be part of the full sample splitting problem. On the other hand, Rigollet [41] examines sample splitting schemes with multiple splits and thus comes close to cross-validation, but does so only for the problem of density estimation. A direct treatment of a cross-validation scheme is to be found in van der Vaart et al. [51]; Chapter 3 of this thesis also focusses on cross-validation. And in the context of classification, recent inequalities are given for recursive aggregation by mirror descent by Juditsky et al. [26] and for aggregation with exponential weights by Lecué [31].

2.1.1 Notation

The results will be conditional on the training set. We use $\mathbf{P}$ to denote the distribution of the test sample, and $\mathbf{E}$ denotes the expectation of random variables depending on the test sample.

For $\gamma : \mathcal{Z} \to \mathbf{R}$, we write

$$P\gamma := \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\gamma(\tilde{Z}_i) \ ,$$

where $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ is an i.i.d. copy of $(Z_1, \ldots, Z_n)$, and

$$P_n\gamma := \frac{1}{n}\sum_{i=1}^{n} \gamma(Z_i) \ .$$

Let $\gamma_j : \mathcal{Z} \to \mathbf{R}$, $j = 1, \ldots, p$, be given loss functions in a class $\mathbf{\Gamma}$. We consider the empirical risk minimization estimator

$$\hat{\gamma} := \arg \min_{1 \le j \le p} P_n \gamma_j \ .$$

The target is

$$\gamma_0 := \arg \min_{\gamma \in \mathbf{\Gamma}} P\gamma \ ,$$

whose best approximation is

$$\gamma_* := \arg \min_{1 \le j \le p} P\gamma_j \ .$$

We will write f_* for the corresponding parameter value (or an arbitrary choice thereof, if it is not unique) at which this minimum is attained, i.e. for which $\gamma_* = \gamma_{f_*}$. We define the excess risks

$$\hat{\mathcal{E}} := P(\hat{\gamma} - \gamma_0)$$

(a random variable, as it depends on the test sample),

$$\mathcal{E}_j := P(\gamma_j - \gamma_0)$$

and

$$\mathcal{E}_* := P(\gamma_* - \gamma_0) \ .$$

Without loss of generality, we assume that $\mathbf{\Gamma}$ is of the form $\mathbf{\Gamma} := \{\gamma_f : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a subset of a semi-metric space with semi-metric d, and given the training set, we write (with some abuse of notation) γ_{f_j} as γ_j, $\{f_j\}_{j=1}^p \subset \mathcal{F}$.

2.1.2 Goal

Our goal is now to show that $\hat{\mathcal{E}}$ is close to $\mathcal{E}_*$ (with large probability or in expectation). The results are modifications of inequalities of the form

$$(1 - \delta)\mathbf{E}\hat{\mathcal{E}} \le (1 + \delta)\mathcal{E}_* + \frac{\Delta}{\delta} \ ,$$

where $\delta > 0$ is an arbitrary small constant, and with Δ of order $\log(p)/n$ and not depending on $\mathcal{E}_*$ – see for example Chapter 7 in Györfi et al. [21].

In the standard setup of Section 2.7 and under a quadratic "margin condition", for instance, we show that for $1 \leq m \leq 1 + \log p$

$$\mathbf{E}\hat{\mathcal{E}}^{\frac{m}{2}} \leq \left(\sqrt{\mathcal{E}_*} + \sqrt{\Delta}\right)^m ,$$

with Δ of order $\log(2p)/n$ and not depending on $\mathcal{E}_*$. In particular, with $m = 2$, this reads

$$\mathbf{E}\hat{\mathcal{E}} \leq \left(\sqrt{\mathcal{E}_*} + \sqrt{\Delta}\right)^2 .$$

This gives rise to a non-sharp oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \Delta , \quad \delta > 0 .$$

A sharp ($\delta = 0$) and rate-optimal (correction term $\mathcal{O}(\Delta)$) oracle inequality cannot be established in a general setup by empirical risk minimization (cf. Lecué [31]). Instead, methods such as mirror averaging could be used, as by Juditsky et al. [25]. See also Audibert ([5] and [6]) for some limitations of empirical risk minimization, and alternative approaches to overcome the limitations. We however believe empirical risk minimization remains an important topic of study because it is widely applied in practice, and is closely related to various cross-validation schemes, as we shall see in Chapter 3.

2.1.3 Convex loss

In our proofs, we only use the property

$$P_n\hat{\gamma} \leq P_n\gamma_* .$$

In the convex case, this sometimes means that conditions can be weakened. Let $\mathbf{F}$ be a convex subset of a linear vector space, and suppose that $\mathbf{\Gamma} := \{\gamma_f : \ f \in \mathbf{F}\}$, with $f \mapsto \gamma_f$ convex, $\mathbf{P}$-almost everywhere. Then for $0 \leq \alpha \leq 1$, we have the inequality

$$P_n\gamma_{\alpha\hat{f}+(1-\alpha)f_*} \leq \alpha P_n\hat{\gamma} + (1-\alpha)P_n\gamma_* \leq P_n\gamma_* .$$

This means that we can replace $\hat{\gamma}$ by $\gamma_{\alpha\hat{f}+(1-\alpha)f_*}$ throughout, leading to inequalities for the excess risk

$$\hat{\mathcal{E}}_\alpha = P\gamma_{\alpha\hat{f}+(1-\alpha)f_*} - P\gamma_0 .$$

From these, we can often deduce inequalities for the original $d(\hat{f}, f_0)$. As we shall see, this extension (with $\alpha < 1$) allows us to work with weaker conditions (than with $\alpha = 1$). In particular, the example on maximum likelihood will take a similar approach with α set to $1/2$.

2.1.4 Organization of this chapter

The next section develops the truncation idea in a gentle fashion, going through all the steps in the development of a model selection oracle inequality. The two other pivotal instruments, a Bernstein inequality and a margin condition, are included at their point of use. After obtaining this first result (and a variant of the proof involving Orlicz norms) we shall set out again and refine the truncation argument in a more powerful line of proof that encompasses more general margin conditions. Here we work with examples from the outset, and leave the details of the proofs for the end.

Section 2.5 presents Bernstein's inequality. It is stated in the form of a probability bound and a moment bound. Section 2.6 presents the margin condition and some examples where it holds. Section 2.7 gives the main results, both one for exponential moments and a very general margin condition, and one for power tails and a particular form of margin condition. Subsequently, Section 2.8 applies the main results to the examples already given, and the proofs for the latter half of the chapter are in Section 2.9. The optimality of these results in certain cases is examined in Section 2.3.

2.2 Developing oracle inequalities and using the truncation argument

We would like to find oracle inequalities to compare the true additional risk incurred by the M-estimator over the oracle. These we formulate in terms of excess risk over a further comparison loss function γ_0 that does not depend on any training data, and may or may not be in the class $(\gamma_j)_{j=1,\dots,p}$ of trained estimators in any situation. This formulation is necessary, as margin conditions, which are important in oracle inequality

proofs using empirical process theory, are generally only true in terms of excess risk of the true best estimator (eg. the true regression function $E[Y|X=x]$), which may or may not lie in the class of models under investigation.

The prototype of an oracle inequality is

$$\left(\mathbf{E}P\hat{\gamma} - P\gamma_0\right) \leq \left(P\gamma_* - P\gamma_0\right) + C_n \ , \tag{2.1}$$

where C_n is some quantity depending on n only, the **correction term** of the oracle inequality (2.1).

One established example of an oracle inequality that formed part of the motivation for this work is found in [21], Theorem 7.1. It is an important precursor to this work as it also examines a hold-out estimate and also uses the Bernstein inequality to show a concentration of empirical measure about the mean and to thus prove an oracle inequality. In this example, a regression function m stemming from an unknown distribution μ of data pairs (X,Y) is to be estimated. This estimation is done by way of training a pre-determined finite set $\mathcal{Q}_n$ of models using the first n_l data pairs to yield estimators $m_{n_l}^{(h)}$ of the true regression function m, and then selecting from these estimators that which empirically minimizes the quadratic loss for a further n_t data pairs. The resulting estimator is written as $m_{n_l}^{(H)}$ by Györfi et al. This estimator is then compared to the true best choice of a model, the oracle $m_{n_l}^{(\hat{h})}$, yielding the non-sharp oracle inequality

$$\begin{aligned} & E\left[\int |m_{n_l}^{(H)}(x) - m(x)|^2 \mu(dx)\right] \\ \leq \ & (1+\delta)E\left[|m_{n_l}^{(\hat{h})}(x) - m(x)|^2 \mu(dx)\right] \\ & + L^2(16/\delta + 35 + 19\delta)\frac{1+\log(|\mathcal{Q}_n|)}{n_t} \end{aligned} \tag{2.2}$$

for all $\delta > 0$. The proof given requires all responses and regression estimates involved to be bounded by L, and crucially uses this fact to apply Bernstein's inequality.

In this motivating example above, our models are indexed by $h \in \mathcal{Q}_n$, our data comes as a pair (X,Y), and the loss functions are $\gamma_h(x,y) =$

$|m_{nl}^{(h)}(x)-y|^2$. Thus in our notation, the oracle inequality (2.2) becomes

$$\mathbf{E}P\hat{\gamma}(X) \leq (1+\delta)P\gamma_* + L^2(16/\delta + 35 + 19\delta)\frac{1+\log(M)}{n} \ .$$

Our first step towards oracle inequalities is a fairly standard one; it is then in the analysis of empirical processes that the real hurdles in oracle inequality proofs are to be taken. This first step involved in the derivation of an oracle inequality like (2.1) uses the optimality of $\hat{\gamma}$ for the empirical risk to derive a correction term in the shape of an empirical process:

$$\begin{aligned}
\mathbf{E}P\hat{\gamma} - P\gamma_0 &= \mathbf{E}P\hat{\gamma} - P\gamma_* - \mathbf{E}P_n\hat{\gamma} + \mathbf{E}P_n\gamma_* \\
&\quad + \underbrace{\mathbf{E}P_n\hat{\gamma} - \mathbf{E}P_n\gamma_*}_{\leq 0} + P\gamma_* - P\gamma_0 \\
&\leq \mathbf{E}[-(P_n - P)(\hat{\gamma} - \gamma_*)] + (P\gamma_* - P\gamma_0) \\
&\leq \mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| + (P\gamma_* - P\gamma_0) \ .
\end{aligned}$$

2.2.1 Splitting the empirical process – naively

The empirical process

$$|(P_n - P)(\hat{\gamma} - \gamma_*)|$$

is cumbersome to deal with as is. To work with it, we first split it (as an upper bound) into two factors

$$|(P_n - P)(\hat{\gamma} - \gamma_*)| = \frac{|(P_n - P)(\hat{\gamma} - \gamma_*)|}{\|\hat{\gamma} - \gamma_*\|} \cdot \|\hat{\gamma} - \gamma_*\|$$

using the L_2-norm $\|\cdot\| = \|\cdot\|_2$, and then replace $\hat{\gamma}$ in the first factor by a supremum over γ_j to obtain

$$|(P_n - P)(\hat{\gamma} - \gamma_*)| \leq Z_n \cdot \|\hat{\gamma} - \gamma_*\|$$

for

$$Z_n := \max_{j=1,\ldots,p} Z_{n,j}, \quad Z_{n,j} := \frac{|(P_n - P)(\gamma_j - \gamma_*)|}{\|\gamma_j - \gamma_*\|} \ .$$

Now we take expectations, and further decouple the two factors in our upper bound for $|(P_n - P)(\hat{\gamma} - \gamma_*)|$ using Hölder's inequality:

$$\begin{aligned} \mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| &\leq \mathbf{E}[Z_n \cdot \|\hat{\gamma} - \gamma_*\|] \\ &\leq (\mathbf{E}[Z_n^2])^{1/2} \cdot (\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2])^{1/2} \end{aligned}$$

Our intention now would be to bound the first factor using empirical process theory, and to tackle the second factor by assuming the (typical) margin condition

$$C \cdot P(\gamma_j - \gamma_0) \geq \|\gamma_j - \gamma_0\|^2 \ \forall j \tag{2.3}$$

for the L_2-norm. The latter step is easily possible; the former must, however, cope with the possibility of arbitrarily small (or even zero) values of the denominator $\hat{\gamma} - \gamma_*$.

2.2.2 Correction by enforcing lower bounds

Arbitrarily small values of $\hat{\gamma} - \gamma_*$ can cause problems with Z_n, in whose denominator they appear. However, such small values are precisely what we are interested in for our oracle inequality. So we would like to make some kind of case distinction where small values of $\hat{\gamma} - \gamma_*$ are regarded separately from large ones. This we can do by truncating $\hat{\gamma} - \gamma_*$ at some (as yet unspecified) norm value $\tau_n > 0$. (We use notation that indicates dependence on n, as our ultimate interest lies in the rate – and constant factors – involved in the oracle inequality for increasing sample size n.) So we redefine Z_n as

$$Z_n := \max_{j=1,\ldots,p} Z_{n,j}, \quad Z_{n,j} := \frac{|(P_n - P)(\gamma_j - \gamma_*)|}{\|\gamma_j - \gamma_*\| \vee \tau_n} \ .$$

and reformulate our correction term by

$$\mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| \leq (\mathbf{E}[Z_n^2])^{1/2} \cdot (\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2 \vee \tau_n^2])^{1/2} \ .$$

2.2.3 Introducing Orlicz norms and Bernstein inequalities to the proof

We will now focus on the term $\mathbf{E}[Z_n^2]$, which is the difficult part to bound. We would like to find bounds for it using assumptions on the

$\gamma_j(X)$ or $\gamma_j(X) - \gamma_*(X)$. One straightforward type of such assumptions consists of uniform ones, such as bounds on – or other finite moments of –

$$\Gamma(X) := \max_{j=1,\ldots,p} |\gamma_j(X) - \gamma_*(X)| \ .$$

Proving things about $\mathbf{E}[Z_n^2]$ using such envelope conditions on the γ_js involves switching suprema and expectations, and this is where Orlicz norms and Bernstein inequalities come into the picture. Orlicz norms $\|\cdot\|_\psi$ can make use of [52], Problem 2.2.8, which when applied to the $Z_{n,j}$ implies

$$\mathbf{E}[Z_n^2] \leq \psi^{-1}(M) \cdot \max_{j=1,\ldots,p} \|Z_{n,j}^2\|_\psi \ .$$

For instance for $\psi = \psi_1 := \exp(x) - 1$:

$$\mathbf{E}[Z_n^2] \leq \log(1+M) \cdot \max_{j=1,\ldots,p} \|Z_{n,j}^2\|_{\psi_1} = \log(1+M) \cdot \max_{j=1,\ldots,p} \|Z_{n,j}\|_{\psi_2}^2 \ .$$

This of course is only of any use if a sensible upper bound on the $\|Z_{n,j}\|_{\psi_2}$ (or suchlike) can be found.

Bernstein inequalities for Z_n, on the other hand, look something like

$$\mathbf{E}\exp(\beta Z_n) \leq 2M \exp\big(\frac{\beta^2/n}{2(1-\beta/(n\tau_n))}\big) \quad \forall \beta \in (0, n\tau_n) \ ; \tag{2.4}$$

this particular assumption implies that

$$\mathbf{E}[Z_n^2] \leq \Big(\sqrt{\frac{2\log(6M)}{n}} + \frac{\log(6M)}{n\tau_n}\Big)^2 \ , \tag{2.5}$$

as we shall see a little later. But assuming that such a Bernstein inequality holds is also a big step.

2.2.4 Finding the convergence rate of the correction term for a given approach

At the end of the day, the quantity we want to optimize in our oracle inequality is the convergence rate of the correction term: that convergence should be as swift as possible. So in order to evaluate any attempted realisation of the approaches listed above, we would be well-advised to find this convergence rate for them at the earliest possible opportunity. This we do with the following lemma:

Lemma 2.2.1. *Assume that* $\mathbf{E}[Z_n^2] \leq B(M, n, \tau_n)$ *for some function* B*, and that the margin condition (2.3) holds. Then we have the oracle inequality*

$$\begin{aligned}(1-\delta)(\mathbf{E}P\hat{\gamma} - P\gamma_0) \quad &\leq \quad (1+\delta)(P\gamma_* - P\gamma_0) + \frac{C}{2\delta} \cdot B(M, n, \tau_n) \\ &\quad + \frac{\delta}{2C} \cdot \tau_n^2 \ .\end{aligned}$$

Proof. Under the assumptions given, we have

$$\begin{aligned}&\mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| \\ \leq \quad &(\mathbf{E}[Z_n^2])^{1/2} \cdot (\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2 \vee \tau_n^2])^{1/2} \\ \leq \quad &(B(M, n, \tau_n))^{1/2} \cdot (\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2 \vee \tau_n^2])^{1/2} \ ,\end{aligned}$$

and thus for all $\delta' > 0$,

$$\begin{aligned}&\mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| \\ \leq \quad &(\frac{B(M, n, \tau_n)}{\delta'})^{1/2} \cdot (\delta' \mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2 \vee \tau_n^2])^{1/2} \\ \leq \quad &\frac{B(M, n, \tau_n)}{2\delta'} + \frac{\delta'}{2}\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2 \vee \tau_n^2] \\ \leq \quad &\frac{B(M, n, \tau_n)}{2\delta'} + \frac{\delta'}{2}\big(\mathbf{E}[\|\hat{\gamma} - \gamma_*\|^2] + \tau_n^2\big) \ .\end{aligned}$$

Applying first the triangle inequality, and then the margin condition, to the norm term gives us

$$\begin{aligned}\mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| \quad &\leq \quad \frac{\delta'}{2}\big(\mathbf{E}[(\|\hat{\gamma} - \gamma_0\| + \|\gamma_* - \gamma_0\|)^2]\big) \\ &\quad + \frac{B(M, n, \tau_n)}{2\delta'} + \frac{\delta'}{2}\tau_n^2 \\ &\leq \quad \delta'\big(\mathbf{E}[\|\hat{\gamma} - \gamma_0\|^2] + \|\gamma_* - \gamma_0\|^2\big) \\ &\quad + \frac{B(M, n, \tau_n)}{2\delta'} + \frac{\delta'}{2}\tau_n^2 \\ &\leq \quad \delta'\big(C(\mathbf{E}P\hat{\gamma} - P\gamma_0) + CP(\gamma_* - \gamma_0)\big) \\ &\quad + \frac{B(M, n, \tau_n)}{2\delta'} + \frac{\delta'}{2}\tau_n^2 \ .\end{aligned}$$

With this, we obtain the oracle inequality

$$(1-\delta' C)(\mathbf{E}P\hat{\gamma} - P\gamma_0) \quad \leq \quad (1+\delta' C)(P\gamma_* - P\gamma_0) + \frac{B(M,n,\tau_n)}{2\delta'} + \frac{\delta'}{2}\tau_n^2 \ ,$$

which the substitution $\delta = \delta' C$ turns into the desired result. ■

Note. In this (non-sharp, but nonetheless useful) oracle inequality, it is the term

$$\frac{B(M,n,\tau_n)}{2\delta'} + \frac{\delta'}{2}\tau_n^2$$

which – when optimized over τ_n – determines the convergence rate.

2.2.5 Using the Bernstein inequality for bounded loss functions

First we need to underpin the derivation of inequality (2.5) from Bernstein conditions. This requires a lemma:

Lemma 2.2.2. *Let X be a non-negative real random variable such that for some constants $a, r > 0$ and $C \geq 1$,*

$$E\left[\exp(\beta X)\right] \leq C \cdot \exp\left(\frac{a\beta^2}{2(1-r\beta)}\right) \quad \forall \beta \in (0, 1/r) \ .$$

Then

$$\|X\|_{L^2(P)} = \left(E\left[X^2\right]\right)^{1/2} \leq \sqrt{2a\log(3.1C)} + r \cdot \log(3.1C) \ .$$

Proof. Fix $\beta \in (0, 1/r)$ and define the functions

$$h : u \mapsto \exp(\beta\sqrt{u}) - 1 \ ,$$

with domain $[0, \infty)$, and

$$g := h^{-1} : v \mapsto \left(\frac{1}{\beta}\log(v+1)\right)^2 \ ,$$

also with domain $[0, \infty)$. Then

$$h'(u) = \beta \exp(\beta\sqrt{u}) \cdot \frac{1}{2\sqrt{u}} \ ,$$

which is non-negative on the domain of h, and

$$h''(u) = \frac{\beta \exp(\beta\sqrt{u})}{4u} \cdot (\beta - \frac{1}{\sqrt{u}}) \ ,$$

which for $u > 0$ is strictly positive iff $u > 1/\beta^2$. Thus h is convex on $(\frac{1}{\beta^2}, \infty)$, and g is concave on $(h(1/\beta^2), \lim_{u\to\infty} h(u)) = (e-1, \infty)$.

Write the function g as a sum $g_1 + g_2$, where

$$g_1(v) = \begin{cases} (\frac{1}{\beta}\log(v+1))^2 & v \geq e-1 \\ \frac{1}{\beta^2} - (e-1-v)\cdot\frac{2}{\beta^2 e} & 0 \leq v \leq e-1 \end{cases}$$

$$g_2(v) = \begin{cases} 0 & v \geq e-1 \\ (\frac{1}{\beta}\log(v+1))^2 - \frac{1}{\beta^2} + (e-1-v)\cdot\frac{2}{\beta^2 e} & 0 \leq v \leq e-1 \end{cases}$$

As g_1 is constructed piecewise from two concave functions and is continuous with continuous first derivative in the point $e-1$ where these two functions are spliced together, it is concave on all of its domain $[0, \infty)$. Furthermore, g_2 is non-negative and

$$\begin{aligned} \max_{v\in[0,\infty)} g_2(v) &= \max_{v\in[0,e-1]} \big((\frac{1}{\beta}\log(v+1))^2 - \frac{2(v+1)}{\beta^2 e}\big) - \frac{1}{\beta^2} + \frac{2}{\beta^2} \\ &\leq \max_{v\in[0,e-1]} \big(\frac{1}{\beta}\log(v+1)\big)^2 + \max_{v\in[0,e-1]} (\frac{-2(v+1)}{\beta^2 e}) + \frac{1}{\beta^2} \\ &= \frac{1}{\beta^2} - \frac{2}{\beta^2 e} + \frac{1}{\beta^2} \\ &= \frac{2}{\beta^2}(1 - \frac{1}{e}) \ . \end{aligned}$$

We can now compute upper bounds for $E\left[X^2\right]$:

$$\begin{aligned} E\left[X^2\right] &= E\left[g(h(X^2))\right] \\ &= E\left[g(\exp(\beta X) - 1)\right] \\ &= E\left[g_1(\exp(\beta X) - 1)\right] + E\left[g_2(\exp(\beta X) - 1)\right] \\ &\leq E\left[g_1(\exp(\beta X) - 1)\right] + \frac{2}{\beta^2}(1 - \frac{1}{e}) \\ &\leq g_1(E\left[\exp(\beta X) - 1\right]) + \frac{2}{\beta^2}(1 - \frac{1}{e}) \\ &\leq g(E\left[\exp(\beta X) - 1\right]) + \frac{2}{\beta^2}(1 - \frac{1}{e}) \end{aligned}$$

$$
\begin{aligned}
&\leq g\big(C \cdot \exp(\frac{a\beta^2}{2(1-r\beta)}) - 1\big) + \frac{2}{\beta^2}(1 - \frac{1}{e}) \\
&= \Big(\frac{1}{\beta}\big(\log(C \cdot \exp(\frac{a\beta^2}{2(1-r\beta)}))\big)\Big)^2 + \frac{2}{\beta^2}(1 - \frac{1}{e}) \\
&\leq \frac{1}{\beta^2} \cdot \Big(\log(C \cdot \exp(\frac{a\beta^2}{2(1-r\beta)})) + \sqrt{2(1 - \frac{1}{e})}\Big)^2 \\
&\leq \frac{1}{\beta^2} \cdot \Big(\log(3.1C) + \frac{a\beta^2}{2(1-r\beta)}\Big)^2
\end{aligned}
$$

The optimal value of β to use in this bound is that which minimizes

$$
\frac{1}{\beta} \cdot \Big(\log(3.1C) + \frac{a\beta^2}{2(1-r\beta)}\Big) = \frac{1}{\beta}\log(3.1C) + \frac{a}{2(1/\beta - r)} \ ,
$$

i.e. $\beta = 1/\zeta$, where $\zeta \in (r, \infty)$ minimizes

$$
\zeta \log(3.1C) + \frac{a}{2(\zeta - r)} \ .
$$

Thus

$$
\begin{aligned}
0 = \frac{d}{d\zeta}\Big(\zeta \log(3.1C) + \frac{a}{2(\zeta - r)}\Big) \quad &= \quad \log(3.1C) - \frac{a}{2(\zeta - r)^2} \\
\Rightarrow \zeta = \sqrt{\frac{a}{2\log(3.1C)}} + r \ , &
\end{aligned}
$$

which does indeed lie in (r, ∞), and for which computation of the second derivative quickly shows that a minimum is attained there. Furthermore,

$$
\begin{aligned}
(E\left[X^2\right])^{1/2} \quad &\leq \quad \zeta \log(3.1C) + \frac{a}{2(\zeta - r)} \\
&= \quad \sqrt{\frac{a\log(3.1C)}{2}} + r \cdot \log(3.1C) + \sqrt{\frac{a\log(3.1C)}{2}} \\
&= \quad \sqrt{2a\log(3.1C)} + r \cdot \log(3.1C) \ .
\end{aligned}
$$

■

Now if we assume that the excess loss over the oracle is bounded, i.e. there is a $K \in \mathbb{R}$ such that $\|\gamma_j - \gamma_*\|_\infty \leq K$ for all $j = 1, \ldots, p$

(for instance if all the loss functions are bounded), then we can derive a Bernstein inequality as follows: (denoting $\tilde{\gamma}_j := \gamma_j - \gamma_* - P(\gamma_j - \gamma_*)$)

$$\begin{aligned}
\mathbf{E}[\exp\Big(\frac{\beta\tilde{\gamma}_j}{n(\|\gamma_j-\gamma_*\|\vee\tau_n)}\Big)] &\leq 1+\sum_{m=2}^{\infty}\frac{1}{m!}\mathbf{E}\frac{\beta^m\tilde{\gamma}_j^m}{n^m\|\gamma_j-\gamma_*\|^2\tau_n^{m-2}}\\
&\leq 1+\frac{\beta^2}{n^2}\sum_{m=2}^{\infty}\Big(\frac{2\beta K}{n\tau_n}\Big)^{m-2}\\
&\leq 1+\frac{\beta^2}{n^2(1-2\beta/n\cdot K/\tau_n)}\\
&\leq \exp\Big(\frac{\beta^2}{n^2(1-2\beta/n\cdot K/\tau_n)}\Big)
\end{aligned}$$

for all $j=1,\ldots,p$, and so

$$\begin{aligned}
&\mathbf{E}\exp(\beta Z_n)\\
\leq\ &\sum_{j=1}^{p}\Big[\mathbf{E}\exp\Big(\frac{\beta}{n\tau_n}\sum_{i=1}^{n}\tilde{\gamma}_j(X_i)\Big)+\mathbf{E}\exp\Big(-\frac{\beta}{n\tau_n}\sum_{i=1}^{n}\tilde{\gamma}_j(X_i)\Big)\Big]\\
\leq\ &\sum_{j=1}^{p}\Big[\prod_{i=1}^{n}\mathbf{E}\exp\Big(\frac{\beta}{n\tau_n}\tilde{\gamma}_j(X_i)\Big)+\prod_{i=1}^{n}\mathbf{E}\exp\Big(-\frac{\beta}{n\tau_n}\tilde{\gamma}_j(X_i)\Big)\Big]\\
\leq\ &2M\cdot\exp\Big(\frac{\beta^2/n}{(1-2\beta/n\cdot K/\tau_n)}\Big)
\end{aligned}$$

follows for all $\beta\in(0,n\tau_n/K)$. By Lemma 2.2.2, we obtain the upper bound

$$\mathbf{E}[Z_n^2]\leq\Big(\sqrt{\frac{4\log(6.2M)}{n}}+\frac{2K\cdot\log(6.2M)}{n\tau_n}\Big)^2.$$

Combined with Lemma 2.2.1, this then yields the following oracle

inequality:

$$\begin{aligned}
& (1-\delta)(\mathbf{E}P\hat{\gamma} - P\gamma_0) \\
\leq\ & (1+\delta)(P\gamma_* - P\gamma_0) + \frac{C}{2\delta}\Big(\sqrt{\frac{4\log(6.2M)}{n}} + \frac{2K\log(6.2M)}{n\tau_n}\Big)^2 \\
& + \frac{\delta}{2C}\tau_n^2 \\
\leq\ & (1+\delta)(P\gamma_* - P\gamma_0) + \frac{C}{\delta}\Big(\frac{4\log(6.2M)}{n} + \frac{4K^2\log(6.2M)^2}{n^2\tau_n^2}\Big) \\
& + \frac{\delta}{2C}\tau_n^2 \ .
\end{aligned}$$

Optimising over τ_n subsequently gives us

$$\begin{aligned}
(1-\delta)(\mathbf{E}P\hat{\gamma} - P\gamma_0) \quad \leq \quad & (1+\delta)(P\gamma_* - P\gamma_0) \\
& + \frac{2\log(6.2M)}{n} \cdot \Big(\frac{2C}{\delta} + K\sqrt{2}\Big) \ . \quad (2.6)
\end{aligned}$$

2.2.6 Extension to unbounded loss functions

As already mentioned, we would like to extend our oracle inequality to situations with unbounded loss functions. Among other things, this would then also allow for statements about regression problems with Gaussian errors. We will go still further and not assume the existence of any exponential moments on the loss functions (as would be the case for loss functions with Gaussian distributions), but only assume the existence of some finite moment. To perform this extension of the oracle inequality, we can use a truncation argument, where we split the correction term into a truncated part, which satisfies boundedness conditions that allow the result from the previous section to be applied, and an upper part, which can be dealt with using Chebyshev's inequality.

Theorem 2.2.3. *Assume that we have an envelope*

$$\Gamma \geq |\gamma_j - \gamma_*| \quad \forall j = 1, \ldots, p$$

for the excess losses over the oracle, assume that this envelope function has some finite moment of order $s > 1$:

$$\|\Gamma\|_s^s \leq \infty \ ,$$

and assume that the margin condition (2.3) holds. Then we have the following oracle inequality for all $\delta > 0$:

$$(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \le (1+\delta)(P\gamma_* - P\gamma_0) + \frac{2\log(6.2M)}{n} \cdot \frac{2C}{\delta}$$
$$+\Big(\frac{\log(6.2M)}{n}\Big)^{1-1/s} \cdot \|\Gamma\|_s \cdot 2^{3/2-1/2s} \cdot (s-1)^{1/s-1} \cdot s \ .$$

Proof. For an arbitrary constant $K > 0$, we have a correction term upper bound

$$\begin{aligned}\mathbf{E}|(P_n - P)(\hat{\gamma} - \gamma_*)| \quad &\le \quad \mathbf{E}[|(P_n - P)\big((\hat{\gamma} - \gamma_*)1\{\Gamma \le K\}\big)|] \\ &\quad +\mathbf{E}[|(P_n - P)\big((\hat{\gamma} - \gamma_*)1\{\Gamma > K\}\big)|]\end{aligned}$$

which according to the preceding section can be bounded by

$$2\frac{\log(6.2M)}{n} \cdot \Big(\frac{2C}{\delta} + K\sqrt{2}\Big) + \mathbf{E}[|(P_n - P)\big((\hat{\gamma} - \gamma_*)1\{\Gamma > K\}\big)|] \ .$$

As the envelope function Γ has a finite s-th moment, a bound for the second summand is given by

$$\begin{aligned}&\mathbf{E}[|(P_n - P)\big((\hat{\gamma} - \gamma_*)1\{\Gamma > K\}\big)|] \\ = \quad &\mathbf{E}[\Big|\frac{1}{n}\sum_{i=1}^n \big((\hat{\gamma}(X_i) - \gamma_*(X_i))1\{\Gamma(X_i) > K\} \\ &-P[(\hat{\gamma} - \gamma_*)1\{\Gamma > K\}]\big)\Big|] \\ \le \quad &\frac{1}{n}\sum_{i=1}^n \big(\mathbf{E}[|\hat{\gamma}(X_i) - \gamma_*(X_i)|1\{\Gamma(X_i) > K\}] \\ &+\mathbf{E}P[|\hat{\gamma} - \gamma_*|1\{\Gamma > K\}]\big) \\ \le \quad &2P[\sup_j |\gamma_j - \gamma_*|1\{\Gamma > K\}] \\ \le \quad &\frac{2}{s-1}\|\Gamma\|_s^s \cdot K^{1-s}\end{aligned}$$

As in Lemma 2.2.1, we thereby obtain the general oracle inequality

$$\begin{aligned}(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \quad &\le \quad (1+\delta)(P\gamma_* - P\gamma_0) \\ &\quad +\frac{2\log(6.2M)}{n} \cdot \Big(\frac{2C}{\delta} + K\sqrt{2}\Big) \\ &\quad +\frac{2}{s-1}\|\Gamma\|_s^s \cdot K^{1-s} \ .\end{aligned}$$

Optimising this over $K > 0$, we get

$$(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \leq (1+\delta)(P\gamma_* - P\gamma_0) + \frac{2\log(6.2M)}{n} \cdot \frac{2C}{\delta}$$
$$+\Big(\frac{\log(6.2M)}{n}\Big)^{1-1/s} \cdot \|\Gamma\|_s \cdot 2^{3/2-1/2s} \cdot (s-1)^{1/s-1} \cdot s \ .$$

■

2.2.7 Good oracle inequalities using Orlicz norms

So far, the best oracle inequality we have seen has come about using only Bernstein inequalities, and without utilizing Orlicz norms. We can, however, prove a good oracle inequality with an Orlicz norm argument, too – though at first we will construct a proof that relies on Bernstein inequalities.

We have the following lemma connecting Bernstein inequality and Orlicz norm conditions:

Lemma 2.2.4. *Let $Z_{n,j}$ and Z_n be as before. Then if the Bernstein-type inequality*

$$P\exp(\beta Z_{n,j}) \leq 2\exp\Big(\frac{\beta^2/n}{2(1-\beta/n\cdot K/\tau_n)}\Big)$$

holds for all $\beta \in (0, \frac{n\tau_n}{K})$, we have

$$\|Z_n\|_{\psi_1} \leq K' \cdot \Big(\frac{K}{n\tau_n}\log(1+M) + \sqrt{\frac{1}{n}} \cdot \sqrt{\log(1+M)}\Big)$$

for some universal constant K'.

Proof. Applying Chebyshev's inequality to the functions $x \mapsto \exp(\beta x)$ and combining this with the Bernstein-type inequality assumed gives us

$$\mathbf{E}[Z_{n,j} > x] \leq 2\exp\Big(\frac{\beta^2/n}{2(1-\beta/n\cdot K/\tau_n)} - \beta x\Big)$$

for all $\beta \in (0, \frac{n\tau_n}{K})$. For $\beta = \frac{n\tau_n x}{Kx+\tau_n}$, we receive the inequality

$$\mathbf{E}[Z_{n,j} > x] \leq 2\exp\Big(-\frac{n\tau_n x^2}{2(Kx+\tau_n)}\Big) = 2\exp\Big(-\frac{x^2}{2(\frac{K}{n\tau_n}\cdot x + \frac{1}{n})}\Big) \ ,$$

which by Lemma 2.2.10 of [52] gives us

$$\|Z_n\|_{\psi_1} \leq K' \cdot \Big(\frac{K}{n\tau_n} \log(1+M) + \sqrt{\frac{1}{n}} \cdot \sqrt{\log(1+M)} \Big)$$

for some universal constant K'. ∎

From this lemma we can now deduce an oracle inequality similar to Theorem 2.2.3:

Theorem 2.2.5. *Assume that we have an envelope*

$$\Gamma \geq |\gamma_j - \gamma_*| \quad \forall j = 1, \ldots, p$$

for the excess losses over the oracle, assume that this envelope function has some finite moment of order $s > 1$:

$$\|\Gamma\|_s^s \leq \infty \ ,$$

and assume that the margin condition (2.3) holds. Then we have the following oracle inequality for all $\delta > 0$ and for the same universal constant K' as in Lemma 2.2.4:

$$(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \leq (1+\delta)(P\gamma_* - P\gamma_0) + \frac{\log(1+M)}{n} \cdot \frac{4CK'^2}{\delta}$$
$$+ \Big(\frac{K' \cdot \log(1+M)}{n} \Big)^{1-1/s} \cdot 2^{3/2-1/2s} \|\Gamma\|_s \cdot (s-1)^{1/s-1} \cdot s \ .$$

Proof. As before, we first look at the case where differences in the loss functions are absolutely bounded by K, and thus the Bernstein inequality holds. By [52], Section 2.2, $\|Z_n\|_2 \leq 2 \cdot \|Z_n\|_{\psi_1}$, and thereby

$$\mathbf{E}[Z_n^2] \leq 4\|Z_n\|_{\psi_1}^2 \leq 8K'^2 \cdot \Big(\frac{1}{n} \cdot \log(1+M) + \frac{K^2}{n^2\tau_n^2} \cdot (\log(1+M))^2 \Big) \ .$$

From this, and Lemma 2.2.1, we once more obtain a non-sharp oracle inequality:

$$\begin{aligned} & (1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \\ \leq \ & (1+\delta)(P\gamma_* - P\gamma_0) \\ & + \frac{4CK'^2}{\delta} \Big(\frac{1}{n} \cdot \log(1+M) + \frac{K^2}{n^2\tau_n^2} \cdot (\log(1+M))^2 \Big) + \frac{\delta}{2C} \tau_n^2 \ . \end{aligned}$$

Optimising over τ_n then gives us

$$(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \leq (1+\delta)(P\gamma_* - P\gamma_0) + \frac{2K' \log(1+M)}{n} \cdot \Big(\frac{2CK'}{\delta} + K\sqrt{2}\Big) \; . \tag{2.7}$$

If we now consider unbounded loss functions, as in the previous section, we receive the inequality

$$\begin{aligned}(1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \leq\; & (1+\delta)(P\gamma_* - P\gamma_0) \\ & + \frac{2K' \log(1+M)}{n} \cdot \Big(\frac{2CK'}{\delta} + K\sqrt{2}\Big) \\ & + 2\|\Gamma\|_s^s \cdot K^{1-s} \; .\end{aligned}$$

After optimisation over $K > 0$, we get

$$\begin{aligned}& (1-\delta)(P\mathbf{E}\hat{\gamma} - P\gamma_0) \\ \leq\; & (1+\delta)(P\gamma_* - P\gamma_0) + \frac{\log(1+M)}{n} \cdot \frac{4CK'^2}{\delta} \\ & + \Big(\frac{2\sqrt{2}K' \cdot \log(1+M)}{n}\Big)^{1-1/s} \cdot 2^{1/s}\|\Gamma\|_s \cdot (s-1)^{1/s-1} \cdot s \\ =\; & (1+\delta)(P\gamma_* - P\gamma_0) + \frac{\log(1+M)}{n} \cdot \frac{4CK'^2}{\delta} \\ & + \Big(\frac{K' \cdot \log(1+M)}{n}\Big)^{1-1/s} \cdot 2^{3/2-1/2s}\|\Gamma\|_s \cdot (s-1)^{1/s-1} \cdot s \; .\end{aligned}$$

■

2.3 Lower bounds

In much of the literature, an oracle inequality for some estimation, selection or aggregation procedure is immediately followed by a lower bound statement, in which a concrete example shows that the bound just given cannot be improved upon. This can then be used to compare different methods, for instance model selection by M-estimation and mirror averaging. For density estimation on finite-dimensional real spaces, Chapter 3 of Rigollet's thesis [41] makes a comparison between the rates for model selection, linear aggregation and convex aggregation.

The same is done for least-squares regression in Tsybakov [45] (see also Bunea et al. [14]). The rates in these are essentially $\mathcal{O}(1/n)$ for sample size n with varying types of dependence on the number of procedures, M. One caveat that appears there, though, is that for convex aggregation, this nice rate only holds if $M \leq \sqrt{n}$. This same caveat also appears in Yang [55] (oracle inequalities in Theorem 1 and lower bounds in Theorem 2), where the oracle inequalities are for a special type of linear aggregation of regression procedures with squared loss. As for density estimation in [41], the optimal rate is $\mathcal{O}(M/n)$. Audibert [4] takes a more general approach for regression, and using a generalized version of Assouad's Lemma (for the original, see Assouad [2]), proves very general oracle inequalities about regression estimation procedures. Applied to L^q-regression, this amounts to correction term rates $\mathcal{O}((M/n)^r)$, where the rate r is a decreasing function (constructed from two pieces) of the regression order q.

For classification problems, there exists literature on (oracle) convergence rates both for empirical risk minimization (Tsybakov [46]) and exponentially weighted aggregation (Lecué [29]). See also Lecué [31] for a comparison of these two. There margin conditions are important for these rates to hold, and depending on the strength of margin condition presupposed, the convergence rate of the correction term varies between $\mathcal{O}(1/\sqrt{n})$ and $\mathcal{O}(1/n)$.

The rate of our Bernstein-derived oracle equality is in general slightly worse than the $\mathcal{O}(1/n)$ obtained by some of the authors cited above. However, unlike [45], [14] and [55], we do not demand that all regression functions involved (true and estimated alike) be uniformly bounded, but only ask for (fairly flexible) moment conditions to hold on the error term. Thus our result is more comparable to Audibert's [4], although in cases other than L^q-regression, the latter's oracle inequalities need quite some unpacking, and in any case do not cover non-regression problems such as density estimation.

Having established, then, that our oracle inequality offers a reasonable rate at the level of generality it permits, we will now show that this rate is indeed optimal, i.e. that the same oracle inequality does not hold in the same form with any strictly faster convergence rate.

Assume that $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ for some space $\mathcal{X}$ - in other words, that we can divide the data space into an input and an output space which we

can subsequently use to construct a regression example. Then for some given integer r write $\mathcal{X}$ as the disjoint union $\mathcal{X}_0 \cup \ldots \cup \mathcal{X}_r$ of non-empty sets (for this, $|\mathcal{X}|$ must be at least $r+1$), and assume that we are using L^2-loss. Let $p_1 \leq 1/r$ be a given probability, let p_2 and p_3 be arbitrary probabilities, and let $y_1 > 0$ be a fixed response level in the output space $\mathbb{R}$. Construct a hypercube H of distributions on $\mathcal{X} \times \mathbb{R}$ indexed in $H_r := 2^{\{1,\ldots,r\}}$ as follows: for $I \subset \{1,\ldots,r\}$, let $P(X \in \mathcal{X}_0) = 1 - rp_1$, $P(X \in \mathcal{X}_i) = p_1$ when $i \in \{1,\ldots,r\}$, and let Y generally take on the value 0 with the exception that

$$P_I(Y = y_1 | X \in \mathcal{X}_i) = \begin{cases} p_2 & i \in I \\ p_3 & i \notin I \end{cases}$$

for indices $i \in \{1,\ldots,r\}$.

For each distribution P_I in this hypercube we have a corresponding regression function $g_I = E_{P_I}[Y|X = x]$, which at $x \in \mathcal{X}_i$ takes on values

$$g_I(x) = \begin{cases} p_2 \cdot y_1 & i \in I \\ p_3 \cdot y_1 & i \in \{1,\ldots,r\} \backslash I \\ 0 & i = 0 \end{cases}$$

and is the oracle regression function for P_I and L^2-loss. The corresponding family of loss functions for our setup is $(\gamma_I)_{I \in H_r}$ with $\gamma_I(x,y) := (g_I(x) - y)^2$.

Now look at adjacent models, conditional on the training data. Due to symmetry, it suffices to consider $I \subset \{2,\ldots,r\}$ and $J = \{1\} \cup I$, so that P_I and P_J differ exactly on $\mathcal{X}_1$. Then for any probability P_K in the hypercube H, we have the risk difference

$$\begin{aligned} & E_{X,Y}\left[\gamma_I(X,Y) - \gamma_J(X,Y)\right] \\ = \; & p_1 \cdot E_{X,Y|X\in\mathcal{X}_1}\left[(g_I(X) - Y)^2 - (g_J(X) - Y)^2\right] \\ = \; & p_1 \cdot E_{X|X\in\mathcal{X}_1}\left[(g_I(X) - g_K(X))^2 - (g_J(X) - g_K(X))^2\right] \\ = \; & p_1 \cdot (p_3 y_1 - (p_2 1\{1 \in K\} + p_3 1\{1 \notin K\})y_1)^2 \\ & -p_1 \cdot (p_2 y_1 - (p_2 1\{1 \in K\} + p_3 1\{1 \notin K\})y_1)^2 \\ = \; & \pm p_1 (p_3 - p_2)^2 y_1^2 \;, \end{aligned}$$

with a "+" if $1 \in K$ and a "−" if not. This generalizes to

$$E_{X,Y}\left[\gamma_I(X,Y) - \gamma_K(X,Y)\right] = \Delta(I,K) \cdot p_1 (p_3 - p_2)^2 y_1^2 \;,$$

where $\Delta(I, J)$ is the Hamming distance $|I \Delta J|$ of the index sets I and J. Now consider Birgé's version of the Assouad Lemma ([11], Lemma 2 of Section 5), in which the Hellinger affinity $\rho(P_u, P_v) = \int \sqrt{dP_u dP_v}$ is involved:

Lemma 2.3.1 (Assouad,1983)**.** *Let $\{P_\delta, \delta \in \mathcal{D}\}$ be a family of distributions indexed by $\mathcal{D} = \{0,1\}^D$ and $X_1, \ldots, X_n$ an iid sample from a distribution in the family. Assume that $\rho(P_\delta, P_{\delta'}) \geq \bar{\rho}$ for each pair $(\delta, \delta') \in \mathcal{D}^2$ such that $\Delta(\delta, \delta') = 1$. Then for any estimator $\hat{\delta}(X_1, \ldots, X_n)$ with values in $\mathcal{D}$,*

$$\sup_{\delta \in \mathcal{D}} E_\delta[\Delta(\hat{\delta}(X_1, \ldots, X_n), \delta)] \geq \frac{D}{2}[1 - \sqrt{1 - \bar{\rho}^{2n}}] \geq \frac{D\bar{\rho}^{2n}}{4} \ ,$$

where E_δ denotes the expectation when the X_i have the distribution P_δ.

This lemma combines with our computation of risk differences to imply

$$\begin{aligned} & \sup_{I \in H_r} E_{P_I^{\otimes n}} \left[E_{P_I} \left[\gamma_{\hat{h}(X_1, Y_1, \ldots, X_n, Y_n)}(X, Y) - \gamma_I(X, Y) \right] \right] \\ \geq \quad & \frac{r \cdot \bar{\rho}^{2n}}{4} \cdot p_1 (p_3 - p_2)^2 y_1^2 \end{aligned}$$

for any model selection estimator $\hat{h} : \mathcal{Z}^n \to H_r$ and for some lower bound $\bar{\rho}$ of the Hellinger affinity of Hamming-adjacent probabilities $I, J : |I \Delta J| = 1$ in H. Now the Hellinger affinity of any two adjacent distributions in our hypercube is

$$\begin{aligned} & (1 - p_1) + p_1 \cdot (\sqrt{p_2 p_3} + \sqrt{(1 - p_2)(1 - p_3)}) \\ = \quad & 1 - p_1 (1 - \sqrt{p_2 p_3} - \sqrt{(1 - p_2)(1 - p_3)}) \ ; \end{aligned}$$

this we can take as our $\bar{\rho}$. What now remains is to set values for p_1, p_2, p_3 and y_1 such that the following conditions are satisfied:

- Margin condition: there exists a non-negative constant C_1 such that
$$C_1 \cdot P_I[\gamma_J - \gamma_I] \geq P_I[(\gamma_J - \gamma_I)^2]$$
for all $I, J \in H_r$. (In the fixed-design case, which we do not regard here, this would be implied by the finiteness of the (conditional)

variance of Y.) The LHS of the margin condition (for this random-design case) reduces to

$$C_1 \cdot p_1(p_3 - p_2)^2 \cdot y_1^2 \cdot |I \Delta J| \ ;$$

the RHS becomes

$$\begin{aligned} & y_1^4 \cdot p_1(p_3 - p_2)^2 \cdot \big((p_2 + p_3)^2 \cdot |I \Delta J| \\ + \quad & 4(1 - p_2 - p_3)(p_2 \cdot |I \setminus J| + p_3 \cdot |J \setminus I|)\big) \ . \end{aligned}$$

Simplifying this, and using that $|I \Delta J| \geq |I \setminus J|, |J \setminus I|$, we obtain the (sufficient) margin condition

$$y_1^2 \cdot (p_2 + p_3)\big(p_2 + p_3 + 4|1 - p_2 - p_3|\big) \leq C_1 \ .$$

- Moment condition: $\gamma_I(X, Y) \leq \Gamma(X, Y) \ \forall I \in H_r, \ E\left[\Gamma^s\right] \leq C_2$ for some $s > 1$. In other words, the loss functions have an envelope that possesses some finite moment. Here the minimal envelope for the family of losses quite clearly is

$$\Gamma(X, Y) := \begin{cases} 0 & X \in \mathcal{X}_0 \\ (p_2 \vee p_3)^2 \cdot y_1^2 & X \notin \mathcal{X}_0 \wedge Y = 0 \\ ((1 - p_2) \vee (1 - p_3))^2 \cdot y_1^2 & X \notin \mathcal{X}_0 \wedge Y = y_1^2 \end{cases}$$

  which possesses a finite s-th moment if

$$r p_1 \cdot [(p_2 \vee p_3)^{3s} + ((1 - p_2) \vee (1 - p_3))^{3s}] \cdot y_1^{2s} \leq C_2 \ .$$

- Rate lower bound condition: we need our lower bound on risk difference to possess a lower bound of the right order, ie.

$$\frac{r}{4} \cdot \bar{\rho}^{2n} \cdot p_1(p_3 - p_2)^2 y_1^2 \geq C_3 \cdot n^{-1+1/s} \ .$$

Let us first establish what order in n the open parameters may take for the conditions to be satisfied.

As $\bar{\rho} \in [0, 1]$, the term $\bar{\rho}^{2n}$ in the rate lower bound condition will converge to zero unless $\bar{\rho}$ converges to 1 quickly enough. If we take $\bar{\rho} = 1 - \mathcal{O}(1/n)$, then the sequence $(\bar{\rho}^{2n})_{n \in \mathbb{N}}$ has a strictly positive lower bound. For $\bar{\rho}$ to have this form itself requires that $p_1 = \mathcal{O}(1/n)$; so we choose $p_1 := 1/n$. In the moment condition, the term involving p_2

and p_3 has the non-zero lower bound $1/2^{3s+1}$, and thus y_1^{2s} can increase only as p_1 decreases - that is, the quickest rate of increase of y_1 is for $y_1 = \mathcal{O}(n^{1/2s})$.

With this rate, we find that the moment condition is fulfilled. Now for the rate lower bound condition we want $p_3 - p_2$ not to converge to zero; this can be attained by choosing $p_2 := 0$ and $p_3 := 1$, for example. The rates of p_1 and y_1 then ensure that the rate lower bound condition holds. The margin condition, however, is not so easily fulfilled: as $y_1 = n^{1/2s}$, it inevitably becomes $\mathcal{O}(n^{1/s}) \leq C_1$, which is only satisfied if we allow the margin parameter to depend on n, in the form of $C_1 = \mathcal{O}(n^{1/s})$. But as the margin parameter only appears in the oracle inequality (upper bound) as a coefficient of n^{-1}, rather than of $n^{-1+1/s}$, this not change the overall rate of the correction term. As it happens, the rate $n^{1/s}$ is the fastest-growing rate of the margin parameter that is still acceptable and does not worsen the oracle inequality rate.

If we now generalize to $y_1 := \alpha \cdot n^{1/2s}$, while keeping $p_1 = 1/n$, $p_2 = 0$ and $p_3 = 1$, and if we also (due to the previous remarks) take the margin parameter as $A \cdot n^{1/s}$, then the three conditions become (sufficiently)

$$\begin{aligned} \alpha^2 \cdot n^{1/s} &\leq A \cdot n^{1/s} \quad \text{(Margin condition)} \\ 2r \cdot \alpha^{2s} &\leq C_2 \quad \text{(Moment condition)} \\ \frac{r\alpha^2}{4}(1-1/n)^{2n} \cdot n^{-1+1/s} &\geq C_3 \cdot n^{-1+1/s} \quad \text{(Lower bound)} , \end{aligned}$$

where it is also obvious that $(1-1/n)^{2n} \geq 1/16$ for $n \geq 2$. Then we can take

$$\alpha := \sqrt{\min\{A, (C_2/2r)^{1/s}\}}$$

and then the three conditions above hold for

$$C_3 := \frac{r\alpha^2}{64} = \frac{r \cdot \min\{A, (C_2/2r)^{1/s}\}}{64} .$$

All in all, we have proven the following theorem:

Theorem 2.3.2. *Let integers n, r, s and positive constants A and B be given, and let $\mathcal{Z}$ be a space that can be written as $\mathcal{X} \times \mathbb{R}$ for some space $\mathcal{X}$ of cardinality at least $r+1$. Then there exists a family $(\gamma_h)_{h \in H}$ of loss functions of cardinality 2^r, such that for any data dependent model*

selection function $\hat{h} : \mathcal{Z}^n \to H$ *there exists a probability distribution* $P(X, Y)$ *on* $\mathcal{Z}$ *for which the margin condition*

$$P[\gamma_h^2] \leq \frac{A}{n^{1/s}} \cdot P[\gamma_h]$$

holds, a finite s-th moment

$$P[\Gamma^s] \leq B$$

exists for the envelope of the loss functions, and for which the oracle lower bound

$$P^{\otimes n} P[\hat{\gamma}] \geq P[\gamma_*] + C \cdot n^{-1+1/s}$$

holds for a constant

$$C := \frac{r \cdot \min\{A, (B/2r)^{1/s}\}}{64} \ .$$

2.4 Further steps

We have just seen that our first oracle inequality (Thm. 2.2.3) attains the optimal rate of model selection (it is also non-sharp, a fact which is generally only corrigible by using aggregation instead of selection methods). We have so far, however, only used a quadratic margin condition, which does suffice for some examples, but not for others (e.g. classification). Next we shall present a refined risk inequality that uses the truncation argument in a different way and which works for a more general margin condition. This risk inequality of Thm. 2.7.2 leads to an oracle inequality that generalizes the previous one, in that under a quadratic margin condition it attains the same (optimal) rate. We start with an inequality that will be necessary to describe the concentration of empirical risk – used for estimation – around its mean, the true risk, which is optimized by the oracle.

2.5 Bernstein's inequality

Bernstein's inequality for a single average is well known, and the extension of Bernstein's probability inequality to a uniform probability

inequality over p averages is completely straightforward. The result can be seen as the simplest version of a concentration inequality in the spirit e.g. of Bousquet [13] (emphasizing how tight these general concentration inequalities are). The moment inequality for the maximum of p averages is perhaps less known.

For all j, we let

$$\gamma_j^c(\cdot) := \gamma_j(\cdot) - P\gamma_j$$

denote the centered loss functions. To obtain our results, we make assumptions on the tails of the centered excess losses $\gamma_j^c - \gamma_*^c$ or of their envelope $\Gamma := \max_{1\le j\le p} \left|\gamma_j^c - \gamma_*^c\right|$ as follows:

Definition 2.5.1. We say that the excess losses $\gamma_j - \gamma_*$ satisfy the exponential moment condition for some $K > 0$ if

$$P\left|\gamma_j^c - \gamma_*^c\right|^m \le \frac{m!}{2}(2K)^{m-2} d^2(f_j, f_*) \tag{2.8}$$

for all $m = 2, 3, \ldots$ and for all $j = 1, \ldots, p$.

We say that the envelope function Γ has power tails of order $s > 1$ if there exists an $M \in (0, \infty)$ such that

$$\mathbf{P}\left(\Gamma > K\right) \le \left(\frac{M}{K}\right)^s \quad \forall K > 0\ . \tag{2.9}$$

Here $d(\cdot, \cdot)$ is the semi-metric introduced above on the underlying parameter space, that allows for different weighting of the procedures under consideration. As an important example of this define, for all γ, the variance

$$\sigma^2(\gamma) := P|\gamma^c|^2\ .$$

Then clearly (2.8) implies that

$$d^2(f_j, f_*) \ge \sigma^2(\gamma_j - \gamma_*)\ \forall\ j\ . \tag{2.10}$$

Moreover, if the bound $|\gamma_j - \gamma_*| \le 3K$ holds for all j, then (2.8) holds with

$$d^2(f_j, f_*) = \sigma^2(\gamma_j - \gamma_*)\ \forall\ j.$$

In the following sections, we will indeed often assume (2.8) with this value for $d(f_j, f_*)$, but we will also consider an extension. The choice

of the semi-metric d is intertwined with the margin behavior, which we consider in the next section. Furthermore, when applying the margin condition, we shall implicitly use Inequality (2.10). As we will make repeated use of Bernstein's inequality, and the term $2\log(2p)/n$ will appear frequently, we will henceforth denote this quantity by

$$\Delta := \frac{2\log(2p)}{n} \ .$$

Using this notation, the version of Bernstein's inequality that we will need in this chapter is:

Lemma 2.5.2. *(Bernstein's inequality for the maximum of p averages: weighted version) Assume that for some constant K, the exponential moment condition (2.8) holds. Then for all $t > 0$ and $\tau > 0$,*

$$\mathbf{P}\left(\max_{1\le j\le p}\frac{|P_n(\gamma_j^c-\gamma_*^c)|}{d(f_j,f_*)\vee\tau} \ge \sqrt{\Delta+2t/n}+\frac{K(\Delta+2t/n)}{\tau}\right) \le \exp[-t] \ .$$

Moreover, for all $1 \le m \le 1+\log p$,

$$\left(\mathbf{E}\max_{1\le j\le p}\left(\frac{|P_n(\gamma_j^c-\gamma_*^c)|}{d(f_j,f_*)\vee\tau}\right)^m\right)^{1/m} \le \sqrt{\Delta}+\frac{K\Delta}{\tau} \ .$$

Remark: The moment inequality is for moments of order $m \le 1+\log p$. It can be extended to hold for general m, provided a slight adjustment, depending on m, is made on the constants. Because we have the situation in mind where p is large, we have formulated the result for $m \le 1+\log p$ to facilitate the exposition.

2.6 Margin behavior

Definition 2.6.1. We say that the margin condition holds with the strictly convex and increasing margin function $G(\cdot)$ if

$$P(\gamma_j-\gamma_0) \ge G\left(d(f_j,f_0)\right), \ \forall\ j \ . \tag{2.11}$$

Furthermore, we say that the margin condition holds with constants $\kappa > 1/2$ and $C > 0$ if (2.11) holds with

$$G(u) = (u/C)^{2\kappa}, \ u > 0 \ .$$

The specific case of $G(u) = (u/C)^{2\kappa}$ is the one most typically used in the literature, with the semi-metric d taken to be the variance of the excess loss $\gamma_j - \gamma_0$. Such a margin condition can be found e.g. in Chesneau and Lecué [16] for regression and density estimation setups, with a comparable result as here for regression, and a result for a different example (squared loss) given for density estimation. Tsybakov [46] gives a similar margin condition for classification. In that paper, the use of 0-1-loss means that $d(f_j, f_0) = P_X(G_{f_j} \triangle G_{f_0})$, where G_f denotes the set $\{x : f(x) = 1\}$. The concept of a Bernstein class – as used by Bartlett and Mendelson [8] – is the same thing after a suitable reparametrization. As we shall see, $\kappa = 1$ in typical cases – but other, in particular larger, values can also occur.

Let us now consider some examples. In a regression or classification setup, we have i.i.d. random pairs $Z_i = (X_i, Y_i)$, with $Y_i \in \mathcal{Y} \subset \mathbf{R}$ a response variable, and $X_i \in \mathcal{X}$ a covariable, $i = 1, \ldots, n$. The quality of an estimator f of $\mathbf{E}[Y_i|X_i]$ can be measured by applying a loss function $\gamma : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$ to the true and the estimated response.

Example 2.6.2. *(Regression)* Suppose that $\{Z_i\}_{i=1}^n := \{(X_i, Y_i)\}_{i=1}^n$. Let $\mathbf{F}$ be a class of real-valued functions on $\mathcal{X}$, and for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let

$$\gamma_f(x, y) := \gamma(f(x), y), \ f \in \mathbf{F} \ .$$

Set

$$l(a, x) = \mathbf{E}(\gamma(a, Y_i)|X_i = x), \ a \in \mathbf{R} \ .$$

We moreover write $l_f(x) := l(f(x), x)$. As our target we take the overall minimizer

$$f_0(\cdot) := \arg\min_{a \in \mathbf{R}} l(a, \cdot) \ .$$

We now check whether the margin condition holds with $\kappa = 1$ and

$$d^2(f, f_0) := K_2^2 P|f - f_0|^2 \ ,$$

where K_2 is an appropriate constant.

Lemma 2.6.3. *Assume that for some $K_1 > 0$, and all $|f - f_0| \leq K_1$,*

$$l_f - l_{f_0} \geq (f - f_0)^2/C_0^2 \ . \tag{2.12}$$

Then

$$P(\gamma_f - \gamma_{f_0}) \geq d^2(f, f_0)/C^2 \ ,$$

where $C^2 := C_0^2 K_2^2$. *If we moreover assume (for* $i = 1, \ldots, n$*) that*

$$\text{var}(\gamma_f(Z_i) - \gamma_{f_0}(Z_i)) \leq K_2^2 \mathbf{E}(f(X_i) - f_0(X_i))^2 \ , \tag{2.13}$$

then for all $\|f - f_0\|_\infty \leq K_1$, *we have*

$$\sigma^2(\gamma_f - \gamma_{f_0}) \leq d^2(f, f_0) \ .$$

If $l(a, \cdot)$ has two derivatives near $a = f_0(\cdot)$, and the second derivatives are positive and bounded away from zero, then $l(a, \cdot)$ behaves quadratically near its minimum, i.e., then (2.12) holds for some $K_1 > 0$.

It also also clear that (2.13) holds as soon as $\gamma(\cdot, y)$ is Lipschitz for all y, with Lipschitz constant L. Then we may take $K_2 = L$. When $\gamma(\cdot, y)$ is not Lipschitz (e.g., quadratic loss), it may be useful to define

$$e_f(Z_i) := \gamma(f(X_i), Y_i) - l_f(X_i) \ .$$

Then obviously

$$\text{var}(\gamma_f(Z_i) - \gamma_{f_0}(Z_i)) = \text{var}(e_f(Z_i) - e_{f_0}(Z_i)) + \text{var}(l_f(X_i) - l_{f_0}(X_i)) \ . \tag{2.14}$$

Note that for fixed designs, the second term in (2.14) vanishes.

Quadratic loss:
In the case of least squares, the loss function is

$$\gamma(f, y) := (y - f)^2 \ ,$$

Then

$$l_f - l_{f_0} = |f - f_0|^2 \ ,$$

and

$$e_f(Z_i) - e_{f_0}(Z_i) = 2\epsilon_i(f(X_i) - f_0(X_i)) \ ,$$

with $\epsilon_i := Y_i - f_0(X_i)$. Assuming that the conditional variance is bounded by some constant σ_ϵ, i.e.,

$$\max_{1 \leq i \leq n} \text{var}(Y_i | X_i) \leq \sigma_\epsilon^2 \ , \tag{2.15}$$

we may conclude the following:

Least squares with fixed design:
The margin condition holds with $\kappa = 1$ and $C^2 = 4\sigma_\epsilon^2$.

Least squares with random design:
If $\|f_j - f_0\|_\infty \leq K_1$ for all j, the margin condition holds with $\kappa = 1$ and $C^2 = 4\sigma_\epsilon^2 + K_1^2$.

Example 2.6.4. *(Classification)* Suppose that $Z_i = (X_i, Y_i)$, with $Y_i \in \mathcal{Y} := \{0, 1\}$ a label, $i = 1, \ldots, n$. Let $\mathbf{F}$ be a class of functions $f : \mathcal{X} \to \{0, 1\}$. We consider 0/1-loss

$$\gamma_f(x, y) = \gamma(f(x), y) \quad := \quad (1 - y)f(x) + y(1 - f(x)), \\ f \in \mathbf{F},\ (x, y) \in \mathcal{X} \times \{0, 1\} .$$

For $a \in [0, 1]$, write

$$l(a, x) := \mathbf{E}(\gamma(a, Y_i)|X_i = x)$$

$$= (1 - \eta)a + \eta(1 - a) = a(1 - 2\eta) + \eta ,$$

where $\eta(x) = \mathbf{E}(Y_i|X_i = x)$. The target is again the overall minimizer

$$f_0(\cdot) := \arg \min_{a \in \{0,1\}} l(a, \cdot) .$$

It is clear that f_0 is the Bayes rule

$$f_0 = \mathrm{l}\{\eta - 1/2 > 0\} .$$

We moreover have

$$P(\gamma_f - \gamma_{f_0}) = P|(f - f_0)(1 - 2\eta)| .$$

Consider the function

$$H_1(v) \leq vP\mathrm{l}\{|1 - 2\eta| < v\},\ v \in [0, 1]$$

and its convex conjugate

$$G_1(u) = \max_v \{uv - H_1(v)\},\ u \in [0, 1]$$

(assuming the maximum exists).

Lemma 2.6.5. *The inequality*

$$P(\gamma_f - \gamma_{f_0}) \geq G\Big(\sigma(\gamma_f - \gamma_{f_0})\Big)$$

holds with $G(u) = G_1(u^2)$, $u \in [0, 1]$.

If $H_1(v) = 0$ for $v \leq C_1$, we take $G_1(u) = C_1 u$. In [46], the margin condition for classification is also formulated as

$$P\left[|\eta - 1/2| \leq t\right] \leq C_\eta t^\alpha$$

for some $C_\eta > 0$, $\alpha > 0$ and for all t within some interval $(0, t_*] \subseteq (0, 1/2]$. This implies that we may take

$$H_1(v) = v(C_\eta v)^\alpha \ ,$$

Then we have

$$G_1(u) = u^{1+1/\alpha}/C^{1+1/\alpha} \ ,$$

where

$$C = C_\eta^{\frac{\alpha}{1+\alpha}} \alpha^{\frac{1}{1+\alpha}} (1 + 1/\alpha) \ .$$

Thus, then the margin condition holds with this value of C and with $\kappa = 1 + 1/\alpha$ (for $d(f_j, f_0) = \sigma(\lambda_j - \lambda_0)$, $\forall\ j$).

Example 2.6.6. *(Maximum likelihood)* Suppose that $\{Z_i\}_{i=1}^n$ are iid. with density $f_0 := dP/d\mu$, where μ is a σ-finite dominating measure. Let $\mathbf{F}$ be a convex class of densities w.r.t. μ, containing f_0. Consider the transformed log-likelihood loss

$$\gamma_f(\cdot) := \gamma(f(\cdot)),$$

where $\gamma(a) = -\log(a)/2$. Define

$$\bar{f} = (f + f_*)/2, \ f \in \mathbf{F} \ .$$

The squared Hellinger distance of densities f and $\tilde{f}$ is

$$h^2(f, \tilde{f}) = \frac{1}{2} \int \left(\sqrt{f} - \sqrt{\tilde{f}} \right)^2 d\mu, \ f, \tilde{f} \in \mathbf{F} \ .$$

We now check the margin and power tail conditions for a distance measure $d(f, f_0)$ which is a multiple of $h(f, f_0)$, namely $d(f, f_0) = L \cdot h(f, f_0)$ for a scaling factor $L \geq 4$ that is involved in bounding the approximation error.

Lemma 2.6.7. *For all densities f, we have*

$$P(\gamma_f - \gamma_{f_0}) \geq h^2(f, f_0) \ .$$

Moreover, under the assumption

$$\sqrt{\frac{f_0}{f_*}} \leq \frac{L}{2\sqrt{2}} ,$$

we have

$$P|\gamma_{\bar{f}} - \gamma_{f_*}|^m \leq \frac{m!}{2} L^2 h^2(\bar{f}, f_*) .$$

This lemma contains the exponential moment condition (2.8) for $K = 1/2$, and also allows us to deduce the margin condition

$$C \cdot \left[P(\gamma_{\bar{f}} - \gamma_{f_0})\right]^{1/2} \geq d(\bar{f}, f_0)$$

for margin constants $\kappa = 1$ and $C = L$.

2.7 Main results

If we assume exponential tails on the loss functions, we are able to obtain a result for a wide range of margin conditions:

Theorem 2.7.1. *Let G be a strictly convex and increasing function with G(0)=0. Suppose that the margin condition holds for G. Let H be the convex conjugate of G, i.e.*

$$H(v) = \sup_{u \geq 0}[uv - G(u)] \quad \forall v \geq 0 .$$

Assume that for some $m \leq 1 + \log p$, the function $H(v^{\frac{1}{m}})$, $v > 0$, is concave. Assume moreover that the exponential moment condition (2.8) holds for some $K > 0$ and for $d(f_j, f_*) := G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_*)$. Then for all $0 < \delta < 1$, and $\varepsilon > 0$, we have

$$(1-\delta)\mathbf{E}\hat{\mathcal{E}} \leq 2H\left(\frac{\sqrt{\Delta}}{\delta} + \frac{K\Delta}{2\delta G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) .$$

The next theorem focuses on the common family of margin functions $G(u) = u^{2\kappa}/C^{2\kappa}, u > 0, \kappa \geq 1$, but also relaxes the exponential tail condition to a power tail condition. Note that for this family of

margin functions, the corresponding convex conjugates are $H(v)$ of order $\mathcal{O}(v^{\frac{2\kappa}{2\kappa-1}})$, and thus Theorem 2.7.1 gives an oracle inequality with correction term rate $\mathcal{O}(\Delta^{\frac{\kappa}{2\kappa-1}})$, which agrees with the rates found in the literature and in the next theorem:

Theorem 2.7.2. *(i) Suppose that the margin condition holds for the loss functions γ_j with constants $\kappa \geq 1$ and $C > 0$ and some d satisfying $d(f_j, f_0) \geq \sigma(\gamma_j - \gamma_0)$, $\forall\ j$. Also assume that the envelope Γ has power tails in the form of (2.9), of order $s > 1$ and for some $M > 0$. Then for all m in the interval $[2\kappa, \min(2s\kappa, 1 + \log(p))[$ and for all $\tau > 0$, we have the following inequality:*

$$\left\| \left(\hat{\mathcal{E}}\right)^{\frac{1}{2\kappa}} \right\|_m \leq (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2}$$

$$+\xi(\kappa, s, m) \cdot M^{\frac{s}{m} \cdot \frac{\alpha}{\alpha+\beta}} \cdot \Delta^{\frac{\alpha\beta}{\alpha+\beta}} \cdot (\mathcal{E}_* \vee \tau)^{-\frac{1}{2\kappa} \cdot \frac{\alpha\beta}{\alpha+\beta}} \ ,$$

where

$$\alpha := \frac{1}{2\kappa - 1} \ , \quad \beta := \frac{s}{m} - \frac{1}{2\kappa} \ ,$$

$$A(\kappa) := \frac{1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}}{\kappa^{\frac{1}{2\kappa-1}}}$$

and

$$\begin{aligned} \xi(\kappa, s, m) \quad &:= \quad A(\kappa)^{\frac{\beta}{\alpha+\beta}} \cdot 2^{\frac{1}{2\kappa} \cdot \frac{\alpha}{\alpha+\beta}} \cdot \left(\frac{2s\kappa}{2s\kappa - m}\right)^{\frac{\alpha}{\alpha+\beta} \cdot \frac{1}{m}} \\ &\quad \cdot \left(\left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} + \left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \right) \ . \end{aligned}$$

(ii) Furthermore, if the excess losses satisfy the exponential moment condition (2.8) for some constant $K > 0$ (from which $d(f_j, f_0) \geq \sigma(\gamma_j - \gamma_0)$ follows), then

$$\left\| \left(\hat{\mathcal{E}}\right)^{\frac{1}{2\kappa}} \right\|_m \leq (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa) \cdot \left(C \cdot \sqrt{\Delta} + \frac{K\Delta}{(\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}}} \right)^{\alpha}$$

for all m in the interval $[2\kappa, 1 + \log(p)[$. In this case we also have

tail bounds

$$\mathbf{P}\Bigg(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} \geq (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa)\Big(C\sqrt{\Delta + 2t/n} + \frac{K(\Delta + 2t/n)}{(\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}}}\Big)^{\alpha}\Bigg) \leq \mathrm{e}^{-t}$$

for all $t > 0$.

These statements lead to simpler ones if we use that $\tau \leq \mathcal{E}_* \vee \tau \leq \mathcal{E}_* + \tau$ and then optimize over τ, trading off the summand with positive exponent $1/(2\kappa)$ and the one with negative exponent $-1/(2\kappa) \cdot \alpha\beta/(\alpha + \beta)$. This yields the main result of this chapter:

Corollary 2.7.3. *In the setup of Theorem 2.7.2, we have the inequalities*

$$(i) \quad \Big\|\Big(\hat{\mathcal{E}}\Big)^{\frac{1}{2\kappa}}\Big\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2} + \tilde{\xi}(\kappa, s, m) \cdot M^{\frac{s}{m} \cdot \frac{\alpha}{\alpha+\beta+\alpha\beta}} \cdot \Delta^{\frac{\alpha\beta}{\alpha+\beta+\alpha\beta}}$$

when the loss envelope Γ *has power tails (2.9) (*$\tilde{\xi}(\kappa, s, m)$ *is a constant depending only on* κ*,* s *and* m*), and*

$$(ii) \quad \Big\|\Big(\hat{\mathcal{E}}\Big)^{\frac{1}{2\kappa}}\Big\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2} + \Big(A(\kappa)^{2\kappa-1} \cdot K\Delta\Big)^{\frac{1}{2\kappa}}$$

when the excess losses satisfy the exponential moment condition (2.8). In the latter case we also have the tail bound

$$\mathbf{P}\Big(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} \geq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot (\Delta + 2t/n)^{\alpha/2} + \Big(A(\kappa)^{2\kappa-1} \cdot K(\Delta + 2t/n)\Big)^{\frac{1}{2\kappa}}\Big) \leq \mathrm{e}^{-t}$$

for all $t > 0$.

Note: These risk inequalities yield oracle inequalities in quite a natural manner: In general, if we have an inequality

$$\Big\|\Big(\hat{\mathcal{E}}\Big)^{\frac{1}{2\kappa}}\Big\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + \xi$$

that holds for a range of values of m including $m = 2\kappa$, then this latter choice of m gives a further inequality

$$\left(\mathbf{E}\hat{\mathcal{E}}\right)^{\frac{1}{2\kappa}} \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + \xi \ . \tag{2.16}$$

For $a, b > 0$ and $\delta > 0$ we can formulate the general inequality $(a+b)^{2\kappa} \leq (1+\delta) \cdot a^{2\kappa} + c(\delta, \kappa) \cdot b^{2\kappa}$, and choose $c(\delta, \kappa)$ to be the minimal value for which this inequality is true over all $a, b > 0$:

$$\begin{aligned} c(\delta, \kappa) &:= \min_{a,b>0} \frac{(a+b)^{2\kappa} - (1+\delta)a^{2\kappa}}{b^{2\kappa}} \\ &= \min_{a>0} \left[(a+1)^{2\kappa} - (1+\delta)a^{2\kappa}\right] \ . \end{aligned}$$

Lemma 2.7.4. *We have the following properties of $c(\delta, \kappa)$:*

- *For all $\delta > 0$ and $\kappa \geq 1$,*

$$\begin{aligned} c(\delta, \kappa) &= \left(\frac{1}{(1+\delta)^{\frac{1}{2\kappa-1}} - 1} + 1\right)^{2\kappa} - \frac{1+\delta}{\left((1+\delta)^{\frac{1}{2\kappa-1}} - 1\right)^{2\kappa}} \\ &= \frac{1+\delta}{\left((1+\delta)^{\frac{1}{2\kappa-1}} - 1\right)^{2\kappa-1}} \end{aligned}$$

- *In particular, for $\kappa = 1$ (corresponding to the quadratic margin condition):*

$$c(\delta, 1) = 1 + \frac{1}{\delta}$$

- *When $\kappa \geq 1$ is fixed, the limit behaviour of $c(\delta, \kappa)$ for $\delta \to 0+$ is: $c(\delta, \kappa) = \mathcal{O}(\delta^{1-2\kappa})$.*
- *For each fixed value of $\kappa \geq 1$, we have $\lim_{\delta \to \infty} c(\delta, \kappa) = 1$.*

In combination with Inequality 2.16, we now have the oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta) \cdot \mathcal{E}_* + c(\delta, \kappa) \cdot \xi^{2\kappa} \ .$$

Corollary 2.7.3 naturally also leads to statements about risk ratios. Under the exponential moment condition, for example, we can see that when

$$\mathcal{E}_* \gg \max\{A(\kappa)^{2\kappa} \cdot C^{\frac{2\kappa}{2\kappa-1}} \cdot \Delta^{\frac{\kappa}{2\kappa-1}}, A(\kappa)^{2\kappa-1} \cdot K\Delta\} \ ,$$

we have have the ratio inequality

$$\mathbf{E}\left|\frac{\hat{\mathcal{E}}}{\mathcal{E}_*}\right|^{\frac{m}{2\kappa}} \to 1$$

for all $m \in [1, 1 + \log(p)]$.

The results of Corollary 2.7.3 constitute a generalization of other, similar, results to be found in the literature. For instance, the rate $\mathcal{O}(\Delta^{\frac{\kappa}{2\kappa-1}})$ we obtain for exponential tails and a margin condition of order $\kappa \geq 1$ is similar to that described by Lecué [30] for classification using Tsybakov's margin condition; the only difference is that there the rate also depends on that of the oracle, i.e. the rate at which $\mathcal{E}_*$ tends to zero as Δ does. For bounded losses, Chesneau and Lecué [16] give a general oracle inequality that they subsequently apply to examples of density estimation and bounded regression. Their most general oracle inequality also has the rate $\mathcal{O}(\Delta^{\frac{\kappa}{2\kappa-1}})$ when the oracle rate is not too large.

2.8 Application to examples

We can apply Corollary 2.7.3 to the (more restricted) cases described in the previous sections:

2.8.1 Quadratic margin, exponential tails

The quadratic margin condition corresponds to taking $\kappa = 1$. Taking the second part of Corollary 2.7.3 for this value of κ yields the oracle inequality

$$P\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \left(1 + \frac{1}{\delta}\right) \cdot 2\left(C + \sqrt{K}\right)^2 \cdot \Delta \qquad (2.17)$$

for all $\delta > 0$, when the losses satisfy the exponential moment condition.

Example 2.8.1. *(Maximum Likelihood)*

Take the setup of Example 2.6.6 and assume that

$$\sqrt{\frac{f_0}{f_*}} \leq \frac{L}{2\sqrt{2}} \tag{2.18}$$

for some $L \geq 2\sqrt{2}$. Define new parameters $\bar{f}_j = (\hat{f}_j + f_*)/2$ and Kullback-Leibler information numbers

$$\hat{\mathcal{K}} := P(\gamma_{(\hat{f}+f_*)/2} - \gamma_{f_0}) = \hat{\mathcal{E}}_{1/2}$$

and

$$\mathcal{K}_* := P(\gamma_{f_*} - \gamma_{f_0}) = \mathcal{E}_* \ .$$

In Lemma 2.6.7, we have already shown the margin and exponential moment conditions for the transformed parameters $\bar{f}_j$ and the scaled Hellinger distance $d(f, f') := Lh(f, f')$. The parameters there are $C = L$, $K = 1/2$ and $\kappa = 1$, and thus we obtain the oracle inequality

$$\mathbf{E}\hat{\mathcal{K}} \leq (1+\delta)\mathcal{K}_* + 2(L + 1/\sqrt{2})^2 \left(1 + \frac{1}{\delta}\right) \cdot \Delta \ . \tag{2.19}$$

This involves the density $(\hat{f} + f_*)/2$, which is *not* an estimator. We can however use this oracle inequality to deduce a risk inequality for the estimator $\hat{f}$ using the following lemma about the Hellinger distance:

Lemma 2.8.2. *Let f,f' and f_0 be densities with respect to the measure μ. Then we have the following inequality:*

$$h(f, f_0) \leq (2+\sqrt{2})h\left(\frac{f+f'}{2}, f_0\right) + (1+\sqrt{2})h(f', f_0) \ .$$

By the first part of Lemma 2.6.7, we have $\hat{\mathcal{K}} \geq h^2(\bar{f}, f_0)$ and $\mathcal{K}_* \geq h^2(f_*, f_0)$. Combining this with the oracle inequality (2.19) and with Lemma 2.8.2, we obtain the risk inequality

$$\begin{aligned}
\mathbf{E}h^2(\hat{f}, f_0) &\leq 2(2+\sqrt{2})^2 \mathbf{E}h^2(\bar{f}, f_0) + 2(1+\sqrt{2})^2 h^2(f_*, f_0) \\
&\leq 2(2+\sqrt{2})^2 \mathbf{E}\hat{\mathcal{K}} + 2(1+\sqrt{2})^2 \mathcal{K}_* \\
&\leq \left[2(2+\sqrt{2})^2(1+\delta) + 2(1+\sqrt{2})^2\right] \cdot \mathcal{K}_* \\
&\quad +4(L+1/\sqrt{2})^2(2+\sqrt{2})^2 \left(1 + \frac{1}{\delta}\right) \cdot \Delta \ .
\end{aligned}$$

We cannot expect to obtain an oracle inequality involving $\mathcal{E}_*$, however, as there is no general bound of the Kullback-Leibler distance of densities by their Hellinger distance.

2.8.2 Quadratic margin, power tails:

Here $\kappa = 1$ and thence $\alpha = 1$, $\beta = s/m - 1/2$ and $A(\kappa) = 2$. Corollary 2.7.3 thus implies

$$\left\| \sqrt{\hat{\mathcal{E}}} \right\|_m \leq \sqrt{\mathcal{E}_*} + 2C \cdot \sqrt{\Delta} + \tilde{\xi}(1, s, m) \cdot \sqrt{M} \cdot \Delta^{\frac{1}{2} - \frac{m}{4s}} \ ,$$

and for $m = 2$ and any $\delta > 0$, the oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \left(1 + \frac{1}{\delta}\right) \cdot 2 \left(2C^2 \cdot \Delta + \tilde{\xi}(1, s, 2)^2 \cdot M \cdot \Delta^{1 - \frac{1}{s}}\right) \quad (2.20)$$

holds.

Example 2.8.3. *(Regression)*

Upper bounds:

In Example 2.6.2, we saw that least-squares regression satisfies a quadratic margin condition, i.e. one with $\kappa = 1$. For instance, we have the margin parameter $C := 2\sigma_\epsilon$ in the fixed-design case. If furthermore we assume that the errors ϵ_i possess some finite moment of order $2s > 2$ – a less restrictive assumption than the Gaussianity often required – then the loss has power tails of order $s > 1$:

$$\gamma_f(x, y) = \gamma(f(x), y) = (y - f(x))^2 = (\epsilon + f_0(x) - f(x))^2$$

$$\begin{aligned} \Rightarrow \mathbf{E}[\Gamma^s] &\leq 2^s \mathbf{E} \left[\sup_{f \in F} \left| \gamma_f^c(X, Y) \right|^s \right] \\ &\leq 2^{4s-1} \cdot \mathbf{E} \left[|\epsilon|^{2s} + \sup_{f \in F} |f_0(X) - f(X)|^{2s} \right] \\ &= 2^{4s-1} \cdot \left(\mathbf{E}|\epsilon|^{2s} + \mathbf{E} \sup_{f \in F} |f_0(X) - f(X)|^{2s} \right) =: M \ , \end{aligned}$$

and so by Chebyshev,

$$P\left(\{\Gamma > K\}\right) \leq \frac{\mathbf{E}\left[\Gamma^s\right]}{K^s} \leq \left(\frac{M}{K}\right)^s \quad \forall K > 0 \ .$$

Thus the oracle inequality (2.20) holds here.

Lower bounds:

Consider the fixed-design case with double Pareto tails of order $s > 2$, i.e. the distribution of the ϵ_i is symmetric around 0, and

$$P(|\epsilon_i| \leq u) = 1 - \frac{1}{(1+u)^s}, \ u > 0 \ .$$

Fix some $p \in \mathbf{N}$, $n + 1 \geq p \geq 2$, and define $f_p := f_0 \equiv 0$,

$$f_j(x) = \mathrm{l}\{x = X_j\} n^{\frac{1}{2s}}, \ x \in \mathcal{X}, \ j = 1, \ldots, p-1 \ .$$

Thus $f_* \equiv 0$ and $\mathcal{E}_* = 0$, too.

Lemma 2.8.4. *The margin condition holds with $\kappa = 1$ and $C^2 = 8/((s-2)(s-1))$, and when $p \geq \sqrt{n} + 1$, the power tail condition (2.9) holds with $M = 2$. For $n \geq 2^{2s}$, moreover, we have*

$$\hat{\mathcal{E}} \geq n^{-\frac{s-1}{s}}$$

with probability at least $1 - \exp[-2^{-1} \cdot (p-1)/\sqrt{n}]$.

Remark: We can easily extend the lower bound result to $p > n$, because we can add, as candidates, any number of bounded functions f_j, say with $\|f_j\|_\infty \leq 1$, without necessitating an increase in the scale parameter M of the power tail condition. These additional functions may be selected by the least squares estimator, but if they all have norm $Pf_j^2 \geq n^{-\frac{s-1}{s}}$, selecting one of these still gives the same lower bound.

Combining this lower bound with the oracle inequality (2.20), we find that for $n \geq p \geq \sqrt{n} + 1$, we have

$$\frac{1}{3} n^{-\frac{s-1}{s}} \leq \mathbf{E}\hat{\mathcal{E}} \leq C'(s) \cdot \left(\frac{\log(n)}{n}\right)^{-\frac{s-1}{s}}$$

for some constant $C'(s)$ that depends only on s, which shows the rate-optimality – up to a logarithmic factor – of the upper bound. If p is small compared to $\sqrt{n}$, however, things look very different:

Lemma 2.8.5. *We have*

$$\left\| \sqrt{\hat{\mathcal{E}}} \right\|_s \leq \sqrt{\mathcal{E}_*} + C c_s p^{1/s} M / \sqrt{n} \ ,$$

where

$$c_s := 2\sqrt{\frac{2}{\pi}} \Gamma^{1/s} \left(\frac{s+1}{2} \right) \ .$$

(Γ denotes the gamma function here, not the loss envelope).

This leads to a (non-sharp) oracle inequality whose correction term has the order $p^{2/s}/n$. If $p \ll \sqrt{n}$, then $p^{2/s}/n \ll n^{-\frac{s-1}{s}}$, i.e. a lower bound of order $n^{-\frac{s-1}{s}}$ for $\mathbf{E}\hat{\mathcal{E}}$ will not hold.

2.8.3 General margin, exponential tails

The risk bound in this case was given in Part (ii) of Corollary 2.7.3, whose correction term is of order $\mathcal{O}(\Delta^{1/(4\kappa-2)})$. Taking $m = 2\kappa$, this leads to an oracle inequality of the form

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + c(\delta,\kappa)\mathcal{O}\left(\Delta^{\frac{\kappa}{2\kappa-1}}\right)$$

for all $\delta > 0$.

Example 2.8.6. *(Classification)*

In Example 2.6.4, we saw the margin condition for $\kappa = 1 + 1/\alpha$ and

$$C = C_\eta^{\frac{\alpha}{1+\alpha}} \alpha^{\frac{1}{1+\alpha}} (1 + 1/\alpha) \ ,$$

where $\alpha > 0$, as a consequence of Tsybakov's margin condition. Furthermore,

$$\begin{aligned} & P\left|\gamma_f^c - \gamma_{f_0}^c\right|^m \\ = \ & P\left|(f(X) - f_0(X)) \cdot (1-2Y) - P\left|(f-f_0)(1-2\eta)\right|\right|^m \\ \leq \ & 2^{m-2} \cdot P\left|\gamma_f^c - \gamma_{f_0}^c\right|^2 = 2^{m-2} \cdot \sigma^2\left(\gamma_f - \gamma_{f_0}\right) \end{aligned}$$

for all f in this example, which means that the excess losses have exponential moments (2.8) with $K = 1$. Thus we have an oracle inequality

$$\begin{aligned} \mathbf{E}\hat{\mathcal{E}} &\leq (1+\delta)\mathcal{E}_* + c(\delta,\kappa)\left(\tilde{A}_1(C_1,\alpha)\cdot\Delta^{\frac{\alpha+1}{\alpha+2}} + \tilde{A}_2\cdot\Delta\right) \\ &= (1+\delta)\mathcal{E}_* + c(\delta,\kappa)\mathcal{O}(\Delta^{\frac{\alpha+1}{\alpha+2}}) \end{aligned}$$

for all $\delta > 0$, and for constants $\tilde{A}_1(C,\alpha)$ and $\tilde{A}_2$.

2.9 Proofs

2.9.1 Proofs for Section 2.5

Proof of Lemma 2.5.2. Without loss of generality, suppose that for all i and j, we have $\mathbf{E}\gamma_j(Z_i) = 0$. Furthermore we can reduce to the case where $\gamma_* \equiv 0$ and all $d(f_j, f_*) = 1$ by looking at a new set of loss functions $(\gamma_j - \gamma_*)/(d(f_j, f_*) \vee \tau)$, where $\tau > 0$ is arbitrary. Thus it suffices to show that under the condition that

$$P|\gamma_j|^m \leq \frac{m!}{2}(2K)^{m-2},\ m = 2, 3, , \ldots$$

for centered loss functions γ_j, the inequality

$$\mathbf{P}\left(\max_{1\leq j\leq p}|P_n\gamma_j| \geq \sqrt{\frac{2(\log(2p)+t)}{n}} + \frac{2K(\log(2p)+t)}{n}\right) \leq \exp[-t] \tag{2.21}$$

holds for all $t > 0$, and that for all $1 \leq m \leq 1 + \log p$,

$$\left(\mathbf{E}\left(\max_{1\leq j\leq p}|P_n\gamma_j|\right)^m\right)^{1/m} \leq \sqrt{\frac{2\log(2p)}{n}} + \frac{2K\log(2p)}{n}\ . \tag{2.22}$$

Bernstein's probability inequality says that for all $t > 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n \gamma_j(Z_i) \geq 2Kt + \sqrt{2t}\right) \leq \exp\left[-nt\right], \forall\ j\ . \tag{2.23}$$

This inequality follows from the intermediate result

$$\mathbf{E}\exp[\sum_{i=1}^n \gamma_j(Z_i)/L] \leq \exp\left[\frac{n}{2(L^2 - 2LK)}\right],\ \forall\ j\ , \tag{2.24}$$

which holds for all $L > 2K$. Inequality (2.21) follows immediately from (2.23).

To prove (2.22), we apply Lemma 2.9.1 to the function $g : x \mapsto (L \cdot \log(x+1))^m$, which is increasing on $[0,\infty)$ and concave on $[\mathrm{e}^{m-1} - 1, \infty)$. We then obtain for all $L > 0$ and all m that

$$\begin{aligned} & \mathbf{E}\left(\max_j |\sum_{i=1}^n \gamma_j(Z_i)|^m\right) \\ \leq \quad & L^m \log^m \left[\mathbf{E}\exp[\max_j |\sum_{i=1}^n \gamma_j(Z_i)|/L - 1] + \mathrm{e}^{m-1} - 1\right]. \end{aligned}$$

From (2.24), and invoking $\mathrm{e}^{|x|} \leq \mathrm{e}^x + \mathrm{e}^{-x}$, we obtain for $L > 2K$,

$$\begin{aligned} & L^m \log^m \left[\mathbf{E}\exp[\max_j |\sum_{i=1}^n \gamma_j(Z_i)|/L - 1] + \mathrm{e}^{m-1} - 1\right] \\ & \leq L^m \log^m \left[p\{2\exp\left[\frac{n}{2(L^2 - 2LK)}\right] - 1\} + \mathrm{e}^{m-1}\right] \\ & \leq L^m \log^m \left[(2p + \mathrm{e}^{m-1} - p)\exp\left[\frac{n}{2(L^2 - 2LK)}\right]\right] \\ & = \left(L\log(2p + \mathrm{e}^{m-1} - p) + \left[\frac{n}{2(L - 2K)}\right]\right)^m. \end{aligned}$$

Now take

$$L = 2K + \sqrt{\frac{n}{2\log(2p + \mathrm{e}^{m-1} - p)}}$$

and use the extra restriction $m \leq 1 + \log p$ to get the desired result. ■

Lemma 2.9.1. *(Jensen's inequality for partly concave functions) Let X be a real-valued random variable, and let g be an increasing function on $[0,\infty)$, which is concave on $[c,\infty)$ for some $c \geq 0$. Then*

$$\mathbf{E}g(|X|) \leq g\left[\mathbf{E}|X| + c\mathbf{P}(|X| < c)\right].$$

Proof. We have

$$\mathbf{E}g(|X|) = \mathbf{E}g(|X|)1\{|X| \geq c\} + \mathbf{E}g(|X|)1\{|X| < c\}$$

$$\leq \mathbf{E}g(|X|)\mathrm{l}\{|X| \geq c\} + g(c)\mathbf{P}(|X| < c)$$
$$= \mathbf{E}\left[g(|X|)\middle||X| \geq c\right]\mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c) \ .$$

We now apply Jensen's inequality to the term on the left, and then use the concavity on $[c, \infty)$ to incorporate the term on the right:

$$\mathbf{E}g(|X|) \leq g\left[\mathbf{E}\left(|X|\middle||X| \geq c\right)\right]\mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c)$$
$$\leq g\Big[\mathbf{E}|X| + c\mathbf{P}(|X| < c)\Big] \ .$$

■

2.9.2 Proofs for Section 2.6

Proof of Lemma 2.6.3. This follows from

$$P(\gamma_f - \gamma_{f_0}) = P(l_f - l_{f_0}) \ .$$

■

Proof of Lemma 2.6.5. For all $v > 0$, we have

$$P|(f - f_0)(1 - 2\eta)| \geq vP|f - f_0|\mathrm{l}\{|1 - 2\eta| \geq v\}$$
$$\geq v\left(P|f - f_0| - P\mathrm{l}\{|1 - 2\eta| < v\}\right) = uv - H_1(v) \ ,$$

with $u = P|f - f_0|$. Since this is true for all v, we may maximize over v to obtain

$$P|(f - f_0)(1 - 2\eta)| \geq G_1\Big(P|f - f_0|\Big) \geq G_1\Big(P(f - f_0)^2\Big) \ ,$$

as

$$P|f - f_0| \geq P(f - f_0)^2 \ .$$

Moreover,

$$|\gamma_f(y) - \gamma_{f_0}(y)| = |(f - f_0)(1 - 2y)| \leq |f - f_0| \ ,$$

so that

$$\sigma^2(\gamma_f - \gamma_{f_0}) \leq P(\gamma_f - \gamma_{f_0})^2 \leq P(f - f_0)^2 \ .$$

■

Proof of Lemma 2.6.7. As the excess risk is a Kullback-Leibler distance to the true distribution, the first statement of the lemma is just the classical lower bound by the Hellinger distance:

$$\begin{aligned} P(\gamma_f - \gamma_{f_0}) &= -\int_{f_0>0} \log \sqrt{\frac{f}{f_0}} f_0 d\mu \\ &\geq -\int_{f_0>0} (\sqrt{\frac{f}{f_0}} - 1) f_0 d\mu \\ &= 1 - \int \sqrt{f f_0} d\mu = h^2(f, f_0) \ . \end{aligned}$$

For the second part, we can use Lemma 7.2 in van de Geer [48], which says that

$$\exp|\gamma_{\bar{f}} - \gamma_{f_*}| - |\gamma_{\bar{f}} - \gamma_{f_*}| - 1 \leq 4(\sqrt{\frac{\bar{f}}{f_*}} - 1)^2. \tag{2.25}$$

We moreover have

$$|\gamma_{\bar{f}} - \gamma_{f_*}|^m \leq \frac{m!}{2}\{\exp|\gamma_{\bar{f}} - \gamma_{f_*}| - |\gamma_{\bar{f}} - \gamma_{f_*}| - 1\} \ .$$

Thus

$$P|\gamma_{\bar{f}} - \gamma_{f_*}|^m \leq 2m! \int (\sqrt{\bar{f}} - \sqrt{f_*})^2 \frac{f_0}{f_*} d\mu \leq \frac{m!}{2} L^2 h^2(\bar{f}, f_*) \ .$$

■

2.9.3 Proofs for Section 2.7

Preparatory lemmas

We begin with two simple results (without proofs) for ease of reference.

Lemma 2.9.2. *If the loss envelope* Γ *has power tails (2.9), then for all* $m < 2s$ *and* $K > 0$*,*

$$P\Gamma^{m/2} 1\{\Gamma > K\} \leq \frac{2s}{2s-m} M^s K^{-(2s-m)/2} \ .$$

Lemma 2.9.3. *For positive constants a, b, α and β, the function*

$$g(x) := ax^{\alpha} + bx^{-\beta},\ x > 0$$

is minimized at

$$x_0 := \left(\frac{b\beta}{a\alpha}\right)^{\frac{1}{\alpha+\beta}} ,$$

and there attains a minimum of

$$g(x_0) = \tilde{\mathcal{C}}(\alpha,\beta) \cdot a^{\frac{\beta}{\alpha+\beta}} b^{\frac{\alpha}{\alpha+\beta}} ,$$

where

$$\tilde{\mathcal{C}}(\alpha,\beta) := \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} + \left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} .$$

Next we need a couple of auxiliary lemmas:

Lemma 2.9.4. *For all $0 \le z \le 1$, we have the inequality*

$$(1-z)^{2\kappa} \le 1 - 2\kappa z^{2\kappa-1} + (2\kappa-1)z^{2\kappa} ,$$

and for all $z \ge 0$,

$$(1+z)^{2\kappa} \ge 1 + 2\kappa z^{2\kappa-1} + z^{2\kappa} .$$

Proof. The second part is clear, as it involves the omission only of positive summands from the LHS to the RHS. For the first part, we write

$$f(z) := 1 - 2\kappa z^{2\kappa-1} + (2\kappa-1)\, z^{2\kappa} - (1-z)^{2\kappa}$$

and note that

$$\begin{aligned} f(z) &= 1 - z^{2\kappa} - (1-z)\cdot 2\kappa z^{2\kappa-1} - (1-z)^{2\kappa} \\ &= (1-z)\cdot\left(\sum_{i=0}^{2\kappa-1} z^j - 2\kappa z^{2\kappa-1} - (1-z)^{2\kappa-1}\right) \\ &= (1-z)^2\cdot\left(\sum_{j=0}^{2\kappa-2}(j+1)\, z^j - (1-z)^{2\kappa-2}\right) \\ &=: (1-z)^2\cdot\tilde{f}(z) . \end{aligned}$$

Now as $\tilde{f}(0) = 0$ and for $0 \le z \le 1$,

$$\left(\tilde{f}\right)'(z) = \sum_{j=1}^{2\kappa-2} j(j+1)z^{j-1} + (2\kappa-2)\cdot(1-z)^{2\kappa-3} \ge 0\ ,$$

we know that $\tilde{f}(z)$, and thus $f(z)$, is non-negative on $[0,1]$. ∎

Lemma 2.9.5. *Let a, b and c be positive, let $\kappa \ge 1$, and assume that*

$$a \le b + c\cdot\left(a^{\frac{1}{2\kappa}} + b^{\frac{1}{2\kappa}}\right)\ .$$

Then

$$a^{\frac{1}{2\kappa}} \le \left(1 + (2\kappa-1)^{\frac{1}{2\kappa-1}}\right)\cdot\left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}} + b^{\frac{1}{2\kappa}}\ .$$

Proof. First note that if $a^{1/2\kappa} \le (c/2\kappa)^{1/(2\kappa-1)}$, then the desired inequality automatically holds. Thus we can restrict ourselves to the case where $a^{1/2\kappa} > (c/2\kappa)^{1/(2\kappa-1)}$. Applying the first part of Lemma 2.9.4 for $z = (c/2\kappa)^{1/(2\kappa-1)}/a^{1/2\kappa}$ – which now is less than 1 – gives us the inequality

$$\left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} - \left(\frac{1}{2\kappa-1}\right)\left(\frac{c}{2\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} \le a - c\cdot a^{\frac{1}{2\kappa}} \le b + c\cdot b^{\frac{1}{2\kappa}}\ ,$$

and thus

$$\begin{aligned}\left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} &\le b + c\cdot b^{\frac{1}{2\kappa}} + (2\kappa-1)\left(\frac{c}{2\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} \\ &\le b + (2\kappa-1)\,c\cdot b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa}\cdot c\right)^{\frac{2\kappa}{2\kappa-1}}\ ,\end{aligned}$$

where in the second step we used that $\kappa \ge 1$. Now applying part 2 of Lemma 2.9.4 to $z = \left(\frac{2\kappa-1}{2\kappa}\cdot c\right)^{1/2\kappa-1}/b^{1/2\kappa}$ yields

$$\begin{aligned}&\left(b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa}\cdot c\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} \\ \ge\ & b + (2\kappa-1)\,c\cdot b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa}\cdot c\right)^{\frac{2\kappa}{2\kappa-1}} \\ \ge\ & \left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa}\ ,\end{aligned}$$

from which the stated inequality follows. ∎

Main proofs

Proof of Theorem 2.7.1. Define

$$\mathbf{Z} := \frac{|(P_n - P)(\hat{\gamma} - \gamma_*)|}{G^{-1}(\hat{\mathcal{E}}) + G^{-1}(\mathcal{E}_* \vee \varepsilon)} \ .$$

By the definition of the convex conjugate H, we have

$$H\left(\frac{\mathbf{Z}}{\delta}\right) \geq \mathbf{Z} \cdot G^{-1}(\hat{\mathcal{E}}) - \delta \cdot \hat{\mathcal{E}}$$

and

$$H\left(\frac{\mathbf{Z}}{\delta}\right) \geq \mathbf{Z} \cdot G^{-1}(\mathcal{E}_* \vee \varepsilon) - \delta \cdot (\mathcal{E}_* \vee \varepsilon) \ .$$

Then

$$\begin{aligned} \hat{\mathcal{E}} &\leq |(P_n - P)(\hat{\gamma} - \gamma_*)| + \mathcal{E}_* \\ &= \mathbf{Z}G^{-1}(\hat{\mathcal{E}}) + \mathbf{Z}G^{-1}(\mathcal{E}_* \vee \varepsilon) + \mathcal{E}_* \\ &\leq \delta\hat{\mathcal{E}} + 2\delta H\left(\frac{\mathbf{Z}}{\delta}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ . \end{aligned}$$

By the concavity of $v \mapsto H(v^{1/m})$, it follows that

$$\begin{aligned} (1-\delta)\mathbf{E}\hat{\mathcal{E}} &\leq 2\delta\mathbf{E}H\left(\frac{\mathbf{Z}}{\delta}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \\ &\leq 2H\left(\mathbf{E}\left(\frac{\mathbf{Z}}{\delta}\right)^m\right)^{1/m} + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ . \end{aligned}$$

Now as $d(f_j, f_*) = G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_*) \leq G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)$, we have the upper bound

$$P\left|\frac{(\gamma_j - \gamma_*)}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right|^{\tilde{m}} \leq \frac{\tilde{m}!}{2}\left(2\frac{K}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right)^{\tilde{m}-2}$$

for all j and for all $\tilde{m} \geq 2$. Thus we can apply Lemma 2.5.2 to obtain the moment bound

$$\|\mathbf{Z}\|_m \leq \left\|\sup_j \frac{|(P_n - P)(\gamma_j - \gamma_*)|}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right\|_m \leq \sqrt{\Delta} + \frac{K\Delta}{G^{-1}(\mathcal{E}_* \vee \varepsilon)} \ .$$

Altogether then

$$(1-\delta)\mathbf{E}\hat{\mathcal{E}} \leq 2H\left(\sqrt{\frac{\Delta}{\delta^2}} + \frac{K\Delta}{\delta G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ .$$

■

Proof of Theorem 2.7.2. (i) In the power tail case, we define

$$\mathcal{E}_*^\tau := \mathcal{E}_* \vee \tau \ ,$$

where τ is a strictly positive number, and

$$\mathbf{Z} := \frac{|P_n\left((\hat{\gamma}^c - \gamma_*^c)\,1\left\{\Gamma \le K\right\}\right)^c|}{C\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}\right)} \ .$$

Then we have

$$\begin{aligned}
\hat{\mathcal{E}} &\le |(P_n - P)(\hat{\gamma} - \gamma_*)| + \mathcal{E}_* \\
&= |P_n\left(\hat{\gamma}^c - \gamma_*^c\right)| + \mathcal{E}_* \\
&\le |P_n\left((\hat{\gamma}^c - \gamma_*^c)\,1\left\{\Gamma \le K\right\}\right)^c| + |P\left((\hat{\gamma}^c - \gamma_*^c)1\left\{\Gamma \le K\right\}\right)| \\
&\quad + |P_n\left((\hat{\gamma}^c - \gamma_*^c)\,1\left\{\Gamma > K\right\}\right)| + \mathcal{E}_* \\
&\le C\mathbf{Z}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}\right) + \mathcal{E}_* + (P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right) \\
&\le C\mathbf{Z}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + \left(\mathcal{E}_*^\tau + (P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}}\right) \\
&\quad + \mathcal{E}_*^\tau + (P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right) \ .
\end{aligned}$$

Using Lemma 2.9.5, we obtain the inequality

$$\begin{aligned}
\hat{\mathcal{E}}^{\frac{1}{2\kappa}} &\le \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa - 1}}\right)\left(\frac{C\mathbf{Z}}{2\kappa}\right)^{\frac{1}{2\kappa - 1}} \\
&\quad + \left(\mathcal{E}_*^\tau + (P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}} \\
&\le \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa - 1}}\right)\left(\frac{C\mathbf{Z}}{2\kappa}\right)^{\frac{1}{2\kappa - 1}} \\
&\quad + (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}} + \left((P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}} \ ,
\end{aligned}$$

where for the second step we used the elementary observation $a^{2\kappa} + b^{2\kappa} \le (a+b)^{2\kappa}$ for $a, b \ge 0$, $\kappa > 1/2$. Now we will first compute the moments of $\mathbf{Z}$ by an application of Bernstein's inequality. We know that

$$P\left|\left(\gamma_j^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right|^m \le K^{m-2}P\left[\left(\left(\gamma_j^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right)^2\right]$$

and

$$\begin{aligned} P\left[\left(\left(\gamma_j^c - \gamma_*^c\right) 1\left\{\Gamma \leq K\right\}\right)^2\right] &= P\left[\left(\gamma_j^c - \gamma_*^c\right)^2 1\left\{\Gamma \leq K\right\}\right] \\ &\leq P\left[\left(\gamma_j^c - \gamma_*^c\right)^2\right] \\ &= \sigma^2\left(\gamma_j - \gamma_*\right) \\ &= \sigma^2\left(\left(\gamma_j - \gamma_0\right) - \left(\gamma_* - \gamma_0\right)\right) \\ &\leq \left(\sigma\left(\gamma_j - \gamma_0\right) + \sigma\left(\gamma_* - \gamma_0\right)\right)^2 , \end{aligned}$$

which by the margin condition

$$\begin{aligned} &\leq \left(C \cdot \left(P\left(\gamma_j - \gamma_0\right)\right)^{1/2\kappa} + C \cdot \left(P\left(\gamma_* - \gamma_0\right)\right)^{1/2\kappa}\right)^2 \\ &= C^2 \cdot \left(\mathcal{E}_j^{1/2\kappa} + \mathcal{E}_*^{1/2\kappa}\right)^2 . \end{aligned}$$

Thus for all j,

$$\begin{aligned} P\left|\frac{\left(\gamma_j^c - \gamma_*^c\right) 1\left\{\Gamma \leq K\right\}}{C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right|^m &\leq \left(\frac{K}{C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right)^{m-2} \\ &\leq \left(\frac{K}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right)^{m-2} , \end{aligned}$$

from which

$$\Rightarrow P\left|\frac{\left(\left(\gamma_j^c - \gamma_*^c\right) 1\left\{\Gamma \leq K\right\}\right)^c}{C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right|^m \leq 2 \cdot \left(\frac{2K}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right)^{m-2}$$

follows, and we can apply Lemma 2.5.2 for loss functions

$$\left(\left(\gamma_j^c - \gamma_*^c\right) 1\left\{\Gamma \leq K\right\}\right)^c$$

and parameter distances

$$d(f_j, f_*) := C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right) \leq C \cdot \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}$$

to obtain

$$\begin{aligned} \left\|\mathbf{Z}\right\|_m &= \left\|\frac{P_n\left(\left(\hat{\gamma}^c - \gamma_*^c\right) 1\left\{\Gamma \leq K\right\}\right)^c}{C\left(\hat{\mathcal{E}}^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right\|_m \\ &\leq 2\left(\sqrt{\Delta} + \frac{K\Delta}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right) . \end{aligned}$$

Now to compute the moments of

$$(P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)^{\frac{1}{2\kappa}} \ ,$$

we proceed as follows for $m \geq 2\kappa$ (using that $\kappa \geq 1/2$):

$$\begin{aligned}
& \left\| \left((P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{1/2\kappa} \right\|_m \\
= \ & \left(\mathbf{E}\left[\left((P_n + P)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{m/2\kappa} \right] \right)^{1/m} \\
\leq \ & \left(2^{m/2\kappa - 1} \mathbf{E}\left[(P_n + P)\left(\Gamma^{m/2\kappa} 1\left\{\Gamma > K\right\} \right) \right] \right)^{1/m} \\
= \ & 2^{1/2\kappa} \left(P\left(\Gamma^{m/2\kappa} 1\left\{\Gamma > K\right\} \right) \right)^{1/m} \ .
\end{aligned}$$

By Lemma 2.9.2, for $m < 2s\kappa$, this has an upper bound in

$$2^{1/2\kappa} \left(\frac{2s\kappa}{2s\kappa - m} \right)^{1/m} M^{s/m} K^{1/2\kappa - s/m} \ .$$

Thus we find that for $m \in [2\kappa, \min\{1 + \log(p), 2s\kappa\})$,

$$\left\| \left(\hat{\mathcal{E}} \right)^{\frac{1}{2\kappa}} \right\|_m \leq (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\frac{1}{2\kappa - 1}} \cdot \left(\sqrt{\Delta} + \frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}} \right)^{\frac{1}{2\kappa - 1}}$$

$$+ B(\kappa, s, m) \cdot M^{s/m} K^{1/2\kappa - s/m} \ ,$$

where

$$A(\kappa) := \frac{1 + (2\kappa - 1)^{\frac{1}{2\kappa - 1}}}{\kappa^{\frac{1}{2\kappa - 1}}} \ ,$$

$$B(\kappa, s, m) := 2^{1/2\kappa} \left(\frac{2s\kappa}{2s\kappa - m} \right)^{\frac{1}{m}} \ .$$

If we now apply the straightforward bound

$$\left(\sqrt{\Delta} + \frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}} \right)^{\frac{1}{2\kappa - 1}} \leq \left(\sqrt{\Delta} \right)^{\frac{1}{2\kappa - 1}} + \left(\frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}} \right)^{\frac{1}{2\kappa - 1}}$$

and minimize the upper bound over $K \geq 0$ (using Lemma 2.9.3), we obtain the desired oracle inequality for the power tail case.

(ii) If we assume the exponential moment condition instead of power tails, we can omit truncation and take

$$\mathbf{Z} := \frac{|P_n\left(\hat{\gamma}^c - \gamma_*^c\right)|}{C\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + \left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}}\right)}$$

as our renormalized empirical process.

Similarly to Part (i), we obtain the inequality

$$\hat{\mathcal{E}} \quad \leq \quad C\mathbf{Z}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + \left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}}\right) + \mathcal{E}_*^\tau \ ,$$

which after applying Lemma 2.9.5 yields

$$\begin{aligned}\hat{\mathcal{E}}^{\frac{1}{2\kappa}} \quad &\leq \quad \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}\right)\left(\frac{C\mathbf{Z}}{2\kappa}\right)^{\frac{1}{2\kappa-1}} \\ &\quad + \left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}} \ .\end{aligned}$$

Now by the exponential moment condition 2.8,

$$\begin{aligned}P\left|\gamma_j^c - \gamma_*^c\right|^m \quad &\leq \quad \frac{m!}{2}(2K)^{m-2}d^2(f_j, f_*) \\ &\leq \quad \frac{m!}{2}(2K)^{m-2}\left(d(f_j, f_0) + d(f_*, f_0)\right)^2 \ ,\end{aligned}$$

which by the margin condition for κ is bounded above by

$$\frac{m!}{2}(2K)^{m-2}C^2\left(\mathcal{E}_j^{\frac{1}{2\kappa}} + \mathcal{E}_*^{\frac{1}{2\kappa}}\right)^2 \ .$$

Thus for all j,

$$P\left|\frac{\left(\gamma_j^c - \gamma_*^c\right)}{C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right|^m \quad \leq \quad \frac{m!}{2}\left(\frac{2K}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right)^{m-2} \ ,$$

and we can apply Lemma 2.5.2 for loss functions

$$\gamma_j^c - \gamma_*^c$$

and parameter distances

$$d(f_j, f_*) := C\left(\mathcal{E}_j^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right) \leq C \cdot \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}$$

to obtain

$$\|\mathbf{Z}\|_m = \left\| \frac{P_n\left(\hat{\gamma}^c - \gamma_*^c\right)}{C\left(\hat{\mathcal{E}}^{1/2\kappa} + \left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)} \right\|_m \quad \leq \quad \sqrt{\Delta} + \frac{K\Delta}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}} \ .$$

■

2.9.4 Proofs for Section 2.8

Proof of Lemma 2.8.2. Regard the term

$$\frac{\sqrt{a}+\sqrt{b}}{\sqrt{\frac{a+b}{2}}+\sqrt{b}}$$

for $a, b \geq 0$. Some simple calculus shows that for fixed $a > 0$, this ratio attains its maximum for $b = 0$; thus

$$\frac{\sqrt{a}+\sqrt{b}}{\sqrt{\frac{a+b}{2}}+\sqrt{b}} \leq \sqrt{2}$$

for all $(a, b) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \backslash (0, 0)$. Using this inequality, and the definition of the Hellinger distance, we can now compute

$$\begin{aligned} h^2(\bar{f}, f) &= \frac{1}{2}\int \left(\sqrt{\frac{f+f_*}{2}} - \sqrt{f}\right)^2 d\mu \\ &= \frac{1}{8}\int \left(\sqrt{f_*} - \sqrt{f}\right)^2 \cdot \frac{\left(\sqrt{f}+\sqrt{f_*}\right)^2}{\left(\sqrt{\frac{f+f_*}{2}}+\sqrt{f}\right)^2} d\mu \\ &\leq \frac{1}{4}\int \left(\sqrt{f_*} - \sqrt{f}\right)^2 d\mu \\ &= \frac{1}{2}h^2(f, f_*) \ . \end{aligned}$$

The triangle inequality now gives us

$$\begin{aligned} h(f,f_0) &\leq h(f,\bar{f}) + h(\bar{f},f_0) \\ &\leq \frac{1}{\sqrt{2}} h(f,f_*) + h(\bar{f},f_0) \\ &\leq \frac{1}{\sqrt{2}} h(f,f_0) + \frac{1}{\sqrt{2}} h(f_*,f_0) + h(\bar{f},f_0) \\ \Rightarrow \left(1 - \frac{1}{\sqrt{2}}\right) h(f,f_0) &\leq \frac{1}{\sqrt{2}} h(f_*,f_0) + h(\bar{f},f_0) \ , \end{aligned}$$

from which the statement of the lemma follows. ■

Proof of Lemma 2.8.4. Clearly, we have $f_* = f_p = f_0$, and $\mathcal{E}_* = 0$.

The margin condition holds with $\kappa = 1$, $C^2 = 4\sigma_\epsilon^2$, $\sigma_\epsilon^2 := E\epsilon^2 = 2/((s-2)(s-1))$, and $d(f,f_0) \geq \sigma^2(\gamma_f - \gamma_{f_0})$. When $p \geq \sqrt{n}$, moreover, the power tail condition holds with $M = 2$, since

$$\begin{aligned} \Gamma(Z_i) &= \max_{1\leq j\leq p} |\gamma_j^c(Z_i) - \gamma_*^c(Z_i)| = \max_{1\leq j\leq p} 2|\epsilon_i f_j(X_i)| \\ &= 2|\epsilon_i| n^{\frac{1}{2s}} \ \{1 \leq i \leq p-1\} \end{aligned}$$

and thus

$$\begin{aligned} P(\{\Gamma > K\}) &\leq \frac{1}{p-1} P(2|\epsilon| n^{\frac{1}{2s}} > K) = \frac{1}{p-1} \left(\frac{1}{1 + K/(2n^{\frac{1}{2s}})}\right)^s \\ &\leq 2^s K^{-s} \ . \end{aligned}$$

We also have for all $u > 0$, and $n \geq 2^{2s}$,

$$\mathbf{P}(\max_{1\leq j\leq p-1} 2\epsilon_j \leq (1+u)n^{\frac{1}{2s}}) = \left(1 - \frac{1}{2}\left(\frac{1}{1 + (1+u)n^{\frac{1}{2s}}/2}\right)^s\right)^{p-1}$$

$$= \left(1 - \frac{1}{2}\left(\frac{1}{n^{\frac{1}{2s}}(n^{-\frac{1}{2s}} + (1+u)/2)}\right)^s\right)^{p-1}$$

$$\leq \left(1 - \frac{1}{2}\left(\frac{1}{n^{\frac{1}{2s}}(1+u/2)}\right)^s\right)^{p-1} \leq \exp[-2^{-1}(1+u/2)^{-s} \cdot (p-1)/\sqrt{n}].$$

It follows that with probability at least $1 - \exp[-2^{-1}(1+u/2)^{-s} \cdot (p-1)/\sqrt{n}]$,

$$\min_{1\leq j\leq p} P_n(\gamma_j) < P_n(\gamma_0) - un^{\frac{1}{2s}}.$$

Thus with probability at least $1 - \exp[-2^{-1} \cdot (p-1)/\sqrt{n}]$, we have that $\hat{\gamma} \neq \gamma_0$.

But for $\gamma_j \neq \gamma_0$,

$$P(\gamma_j - \gamma_0) = P|f_j|^2 = \frac{1}{n}\sum_{i=1}^{n} f_j^2(X_i) = \frac{1}{n} f_j^2(X_j) = n^{-\frac{s-1}{s}}.$$

■

Proof of Lemma 2.8.5. Like in the proof of Theorem 2.7.1, define

$$\mathbf{Z} := \frac{|(P_n - P)(\hat{\gamma} - \gamma_*)|}{C(\sqrt{\hat{\mathcal{E}}})} ,$$

whenever $\hat{\mathcal{E}} > 0$. Then

$$\sqrt{\hat{\mathcal{E}}} \leq \sqrt{\mathcal{E}_*} + C\mathbf{Z} .$$

For any n constants $(b_1, \ldots, b_n)$, we know that

$$\left\| \frac{1}{n}\sum_{i=1}^{n} b_i \epsilon_i \right\|_s \leq \frac{c_s M}{n} \sqrt{\sum_{i=1}^{n} b_j^2}$$

(see Whittle [54] or Appendix A of van der Vaart and Wellner [52]). Hence

$$\|\mathbf{Z}\|_s \leq c_s p^{1/s} M/\sqrt{n},$$

and thus

$$\left\| \sqrt{\hat{\mathcal{E}}} \right\|_s \leq \sqrt{\mathcal{E}_*} + C c_s p^{1/s} M/\sqrt{n} .$$

■

Chapter 3

Model Selection using Cross-Validation

Cross-validation is a very important and widely applied family of model selection methods. Its key feature, the use of multiple splits to train and test the candidate models, renders it somewhat more difficult to handle in a theoretical way. Nevertheless, we shall show that a carefully crafted fundamental risk inequality opens the door to oracle inequalities for cross-validation too. In the course of this, we have to pay careful attention to the exact choice of the splits of our data, especially when retraining the selected model to obtain our final estimator.

We follow essentially the same strategy to obtain an oracle inequality, first deriving a fundamental inequality that connects the estimator and oracle risks via an empirical process, and then analyzing this empirical process in more depth, deriving bounds on its expected supremum.

This chapter is based on part of [32], namely the part where a discrete, rather than continuous, family of candidate estimators is being used for model selection via cross-validation.

3.1 Basics of cross-validation

Cross-validation is a widely-applied solution to the problem of data-driven model selection. It generalizes the splitting of data into one training set and one test set, as seen in the previous chapter, to the use of several splits simultaneously. A (finite) number of candidate estimators are pitted against each other, and for each split into training and test data, the training data is used to fit the estimators, and the test data to evaluate their average performance in terms of some loss function (e.g. squared loss for regression, 0-1 loss for classification). The overall performance of an estimation procedure can then be estimated by averaging its performance over all the splits. For the loss function used in evaluation, this overall performance is an estimate of the out-of-sample risk –confounded somewhat by the multiple use of the same data.

To use and analyze cross-validation procedures, we now introduce some notation. Let n be an integer, and V a divisor of n. Assume that we have the data set $D^n = Z_1, \ldots, Z_n$ with values in $\mathbb{R}$. We split this set into V subsets of equal size $n_C = n/V$, namely

$$B_k = (Z_{(k-1)n_C+1}, \ldots, Z_{kn_C}) \ , \tag{3.1}$$

which shall be test sets, and their complements

$$C_k = \cup_{j=1:j\neq k}^{V} B_j, \forall k = 1, \ldots, V \ , \tag{3.2}$$

the corresponding training sets. Note that C_k is a data set of size $n_V := n - n_C = n(1 - 1/V)$.

Let $\gamma(Z, f)$ be a loss function whose arguments are a data point Z and a parameter $f \in \mathcal{F}$. We consider the empirical risk on the set B_k for a fixed parameter value f,

$$R_{n,V}^{(k)}(f) = \frac{1}{n_C} \sum_{i=(k-1)n_C+1}^{kn_C} \gamma(Z_i, f) \ .$$

For a sequence $\hat{f} = (\hat{f}^{(n)} : \mathbb{R}^n \to \mathcal{F})_{n\in\mathbb{N}}$ of estimators depending on increasing sample sizes, we define the *V-fold CV empirical risk* to be

$$R_{n,V}(\hat{f}) = \frac{1}{V} \sum_{k=1}^{V} R_{n,V}^{(k)}(\hat{f}^{(n_V)}(C_k)) \ .$$

The *V-fold cross-validation procedure* is the procedure

$$\bar{f}_{VCV} = (\bar{f}^{(n)}_{VCV})_{n\in\mathbb{N}}$$

defined, for any n, by

$$\bar{f}^{(n)}_{VCV}(D^n) = \hat{f}^{(n)}_{\hat{\jmath}(D^n)}(D^n) \text{ s.t. } \hat{\jmath}(D^n) \in \arg\min_{j\in\{1,\dots,p\}} R_{n,V}(\hat{f}_j) \ . \quad (3.3)$$

Perhaps the oldest, and certainly the most frequently studied, cross-validation scheme is n-fold or *leave-one-out* cross-validation. It forms the intersection between the class of V-fold and the class of *leave-m-out* cross-validation schemes, defined by

$$\bar{f}^{(n)}_{lmo}(D^n) = \hat{f}^{(n)}_{\hat{\jmath}(D^n)}(D^n) \text{ s.t. } \hat{\jmath}(D^n) \in \arg\min_{j\in\{1,\dots,p\}} R_{n,-m}(\hat{f}_j), \quad (3.4)$$

where we take the *leave-m-out empirical risk* $R_{n,-m}$ to be

$$R_{n,-m}(\hat{f}) = \binom{n}{m}^{-1} \sum_{\substack{C\subset\{1,\dots,n\}:\\ |C|=m}} \frac{1}{m}\sum_{i\in C} \gamma(Z_i, \hat{f}^{(n-m)}((Z_k)_{k\in\{1,\dots,n\}\setminus C})) \ .$$

This method does however become very computationally inadequate as soon as m grows, as there are far too many subsets of $\{1,\dots,n\}$ to average over. One possible solution for this is *balanced incomplete cross-validation* , where cross-validation is treated as a block design and the available pieces of data are all used equally often for training, and equally often for testing. Alternatively, we could use *Monte Carlo cross-validation*, where the training and testing subsets are drawn randomly –without replacement– from the available data. See [42] for a discussion of all these cross-validation schemes.

We can place all deterministic cross-validation schemes into one general framework as follows. For any subset $C \subset \{1,\dots,n\}$ of indices, write $D_{(C)}$ for $(Z_i)_{i\in C}$. Assume that a fixed value n_C be given (the size of test sets), and define $n_V = n - n_C$. Let $C_1,\dots,C_{N_C}$ be a family of N_C subsets of $\{1,\dots,n\}$, each of size n_V. Now for any statistic $\hat{f}$ define the *CV risk*

$$R_{nC}(\hat{f}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i\notin C_k} \gamma(Z_i, \hat{f}^{(n_V)}(D_{(C_k)})) \ ,$$

and its minimizer by

$$\hat{f}^{(n)}_{CV}(D^n(n)) = \hat{f}^{(n)}_{\hat{\jmath}(D^n)}(D^n) \text{ s.t. } \hat{\jmath}(D^n) \in \arg\min_{j\in\{1,\ldots,p\}} R_{n_C}(\hat{f}_j) \ . \quad (3.5)$$

3.2 Cross-validation in the literature

The idea of cross-validation goes back a long way, and is often attributed to Larson [28], e.g. by Stone [43]. In the following years, it was given further consideration. However, it was in the 1960s and 1970s that this idea gained prominence. A good summary of these developments can be found in Stone [43]. In 1968, Tukey and Mosteller [47] briefly discuss cross-validation as a means for testing statistical procedures and avoiding the over-optimism associated with training error. They distinguish two types of cross-validation, namely simple and double – the difference being that double cross-validation assesses performance using data that was not used to determine the model and choose the variables. Thus their emphasis is still very much on model validation and not on selection. Stone ([43], [44]) already places a firm emphasis on using cross-validation to select from a class of predictors. This use of cross-validation as a kind of automated model selection technique (automated, as it is applicable in a very standardized way to a vast range of problems, although the outcome is sometimes called into question) is very much way it is understood today, as the growing number of statistical models in use, and the availability of raw computing power, have made the widespread use of such non-parametric model selection techniques both possible and necessary.

The types of split considered under cross-validation varied at the outset. Arlot and Celisse [1] report that Larson [28] used random splits, Tukey and Mosteller [47] consider 2-fold cross-validation to be the classical case (but also acknowledges the 10-fold procedure), while Stone [44] takes leave-one-out cross-validation as a standard, not considering "generalizations" such as leave-p-out, but acknowledging that they "may be of particular value" for special conditions or purposes. Special conditions where leave-p-out cross-validation is computationally feasible are given by Celisse and Robin [15], where a variant of it (which we shall refer to as averaged cross-validation in the next section) is computed exactly for histogram and kernel methods, leading to an exact quantification of the ensuing L_2-risk.

A further type of cross-validation used at an early stage is generalized cross-validation (GCV), described by Craven and Wahba [17]. It is introduced in a linear regression setup, where it approximates leave-one-out cross-validation using the trace of the hat matrix (the projection matrix H obtained by linear regression, with which the fitted values $\hat{y} = Hy$ can be computed using the observed responses y). An advantage of GCV is that it "in smoothing problems, [it] can alleviate the tendency of cross-validation to undersmooth" –cf. Hastie et al. [23], p. 217. It is furthermore similar to the AIC(which is equal to C_p in this model, linear regression by least squares)– also cf. Hastie et al. [23], p. 217.

In the "classical" case of leave-one-out cross-validation for least-squares regression, Li [34] already gives conditions under which it is asymptotically optimal (where asymptotic optimality means the convergence of the ratio estimator risk/oracle risk to 1). Shao [42], however, shows the inconsistency of leave-one-out cross-validation when the model class is fixed in size, and the need for the proportion n_C/n of the data assigned to testing to tend to 1 as n tends to infinity. Yang [56] also gives conditions under which leave-p-out cross-validation is (selection) consistent; these include the unboundedness of training and test sizes, and also a condition depending on the relationship of L_4 and L_2 risks.

A recent survey of cross-validation methods and the state of knowledge on them is provided by Arlot and Celisse [1]. There a choice of training ratio $n_V = \lambda n$ for $\lambda \in (0, 1)$ is identified from the literature as having good asymptotic properties. This would also suggest fixing V regardless of sample size. However, this choice is relativated by looking at the oracle inequalities that can be obtained. We will see more of this in Section 3.8.

3.3 Fundamental inequality

In the first step, the derivation of fundamental risk inequalities involving empirical processes, we shall not yet make any assumptions on the behaviour of the procedures under consideration when the training sample size varies. As a general cross-validation procedure performs model selection using $n_V < n$ data in each training step, we must use an es-

timator trained on n_V data for the time being. To this end, we define the *modified CV procedure*

$$\bar{f}^{(n)}_{mCV}(D^n) = \hat{f}^{(n_V)}_{\hat{\jmath}(D^n)}(D^{n_V}) \ . \tag{3.6}$$

We will require some simple (fixed sample size) properties on the estimators $\hat{f}_1, \ldots, \hat{f}_p$ to obtain an oracle inequality for the modified CV procedure.

Definition 3.3.1. We say that a statistic $\hat{f} = (\hat{f}^{(n)})_n$ is *exchangeable* if for any integer n, for any permutation $\phi : \{1, \ldots, n\} \longmapsto \{1, \ldots, n\}$ and for P_{D^n}-almost any vector $(z_1, \ldots, z_n) \in \mathcal{Z}^n$, we have

$$\hat{f}^{(n)}(z_1, \ldots, z_n) = \hat{f}^{(n)}(z_{\phi(1)}, \ldots, z_{\phi(n)}) \ .$$

In this thesis, we are interested in statistics in a setup where all data are available simultaneously. In such a setup, exchangeability is often a given. If, however, the data become available sequentially and the estimation procedures make use of this, statistics are unlikely to have this property.

Another modification of the general cross-validation estimator that trains on n_V data is given by the *averaged version of the modified CV procedure*,

$$\bar{f}^{(n)}_{amCV}(D^n) = \frac{1}{N_C} \sum_{k=1}^{N_C} \hat{f}^{(n_V)}_{\hat{\jmath}(D^n)}(D_{(C_k)}) \ , \tag{3.7}$$

or *averaged CV* for short.

In what follows we shall assume that the cross-validation splits C_k are deterministic, i.e. fixed from the beginning. In principle, we could also choose them randomly with certain requirements on independence from the data used in estimation, as is noted in [51], and obtain the same results.

Along the lines of Chapter 2, we denote the global risk minimizer by

$$f_0 := \arg\min_{f \in \mathcal{F}} P\gamma(\cdot, f) \ ,$$

and the excess risk at a given parameter f by

$$\mathcal{E}(f) := P\left(\gamma(\cdot, f) - \gamma(\cdot, f_0)\right) \ .$$

The following lemma shows that for both averaged and modified CV, supremum bounds on the shifted empirical process for the trained estimates $\hat{f}_j^{(n_V)}(D^{n_V})$ are sufficient in deriving oracle inequalities for the corresponding cross-validation procedures:

Lemma 3.3.1. *[Fundamental lemma] If the estimator $\bar{f}^{(n)}(D^n)$ we are considering is either*

- *the averaged CV procedure with splits that come from V-fold cross-validation and a risk that is convex, or*
- *the modified CV procedure without extra assumptions on the splits or the risk,*

and each candidate procedure $\hat{f}_j^{(n_V)}$ is an exchangeable statistic, then for any constant $a \geq 0$, we have the fundamental inequality

$$\mathbb{E}_{D^n}\left(\mathcal{E}(\bar{f}^{(n)}(D^n))\right) \leq (1+a) \min_{j=1,\ldots,p}\left[\mathbb{E}_{D^{n_V}}\mathcal{E}(\hat{f}_j^{(n_V)}(D^{n_V}))\right] + \mathbb{E}_{D^n} \max_{j=1,\ldots,p}\left[(P-(1+a)P_{n_C})\left(\gamma(\cdot,\hat{f}_j^{(n_V)}(D^{n_V})) - \gamma(\cdot,f_0)\right)\right] .$$

3.4 Oracle inequalities

As we have now established fundamental risk inequalities that serve as proto-oracle inequalities, the next step for us is to find bounds for the empirical process part of Lemma 3.3.1 and to thus obtain oracle inequalities for the modified and averaged modified cross-validation procedures. At this stage, the precise nature of the underlying model and the loss function come into play; both margin and tail properties and the loss (as applied to the parameter range $\mathcal{F}$) will be key here.

We will define two alternative sets of assumptions. In these, the behaviour at the margin (which describes the loss when it is small; cf. Mammen and Tsybakov [35], Tsybakov [46] and Bartlett and Mendelson [8]) and the behaviour in the tails (which describes the loss when it is large) complement each other in their contribution towards controlling the empirical process. The stronger the assumptions we are willing to make about the margin, the weaker the tail conditions are that we must require, and vice versa.

3.4.1 Assumptions

(A1) *There exist constants $\kappa \geq 1$ and $K_0, K_1 > 0$ such that the following holds: For any parameter $f \in \mathcal{F}$, we have the inequalities*

1. $\mathcal{E}(f) \leq K_0$,
2. $\left\|\gamma(\cdot, f) - \gamma(\cdot, f_0)\right\|_{L_{\psi_1}(\pi)} \leq K_1 \mathcal{E}(f)^{1/2\kappa}$.

(A2) *There exist constants $\kappa \geq 1$ and $K_0, K_1 > 0$, and a metric $d(\cdot, \cdot)$ on $\mathcal{F}$, such that:*

1. $\|\gamma(\cdot, f_1) - \gamma(\cdot, f_2)\|^m_{L_m(\pi)} \leq \frac{m!}{2} K_0^{m-2} d^2(f_1, f_2)$ *holds for all integers $m \geq 2$ and all $f_1, f_2 \in \mathcal{F}$, and*
2. $d(f, f_0) \leq K_1 \mathcal{E}(f)^{1/2\kappa}$ *holds for all $f \in \mathcal{F}$.*

Note that if

1. $\|\gamma(\cdot, f) - \gamma(\cdot, f_0)\|_{L_\infty(\pi)} \leq K_0$,
2. $\left\|\gamma(\cdot, f) - \gamma(\cdot, f_0)\right\|_{L_2(\pi)} \leq K_1' \mathcal{E}(f)^{1/2\kappa}$,

then $d(f, f_0) := \|\gamma(\cdot, f) - \gamma(\cdot, f_0)\|_{L_2(\pi)}$ and $d(f_1, f_2) := K_0$ (for $f_1, f_2 \neq f_0$) can be chosen and (A2) holds.

In both (A1) and (A2), the first condition controls the behavior of large losses, either by giving a shape for the loss tails (A2) or an average bound (A1). This is followed by margin conditions which relate small risks to some distance (the general $d(\cdot, \cdot)$ or the ψ_1-Orlicz distance) of the estimator from the oracle. Note that the exponential moment condition in (A2) directly implies that

$$\|\gamma(\cdot, f) - \gamma(\cdot, f_0)\|_{L_2(\pi)} \leq d(f, f_0) \ ,$$

and thus that an L_2-margin condition holds. The (stronger) margin condition using the ψ_1-Orlicz norm allows for a weaker condition on the tails of the losses. Assuming a margin condition means that even estimators close to (but distinct from) the oracle can most often be distinguished from the oracle by their empirical risk. Such an identifiability property is crucial for the use of empirical risk minimization.

3.4.2 Maximal inequalities for shifted empirical processes

In addition to the fundamental inequality already derived, we shall need the following maximum bounds for shifted empirical processes in order to obtain an oracle inequality:

Lemma 3.4.1. *Let $\mathcal{Q} := \{Q_1, \ldots, Q_p\}$ be a set of p measurable functions defined on $(\mathcal{Z}, \mathcal{T})$. Let $Z, Z_1, \ldots, Z_m$ be i.i.d. random variables with values in $(\mathcal{Z}, \mathcal{T})$ such that $\forall Q \in \mathcal{Q}, \mathbb{E}Q(Z) \geq 0$. Assume the existence of constants $C_0, C_1 > 0$ and $\kappa \geq 1$ such that*

$$\forall Q \in \mathcal{Q}, \|Q(Z)\|_{L_{\psi_1}} \leq C_0 (\mathbb{E}Q(Z))^{1/2\kappa} \text{ and } \mathbb{E}Q(Z) \leq C_1 \ . \tag{3.8}$$

Now let a shift parameter $a > 0$ be given. If assumption (3.8) holds, we then have the inequality

$$\mathbb{E} \max_{Q \in \mathcal{Q}} \left(\mathbb{E}Q(Z) - (1+a) \frac{1}{m} \sum_{i=1}^{m} Q(Z_i) \right) \leq \tilde{C}_0 \left(\frac{\log p}{m} \right)^{\frac{\kappa}{2\kappa - 1}} + \tilde{C}_1 \left(\frac{\log p}{m} \right)$$

for constants

$$\tilde{C}_0 := 2\tilde{C}_2^{\frac{\kappa}{2\kappa-1}} \ , \quad \tilde{C}_1 := 2\tilde{C}_2 \ , \quad \tilde{C}_2 := \frac{3C_0(1+a) \vee 9C_0^2(1+a)^2}{(a \wedge 1/C_1)^{1/(2\kappa)}} \ .$$

Lemma 3.4.2. *Let $\mathcal{Q} := \{Q_1, \ldots, Q_p\}$ be a set of p measurable functions defined on $(\mathcal{Z}, \mathcal{T})$. Let $Z, Z_1, \ldots, Z_m$ be i.i.d. random variables with values in $(\mathcal{Z}, \mathcal{T})$ such that $\forall Q \in \mathcal{Q}, \mathbb{E}Q(Z) \geq 0$. Assume constants v_Q indexed in $Q \in \mathcal{Q}$, and that the margin condition*

$$v_Q \leq C_0^2 (\mathbb{E}Q(Z))^{1/\kappa}$$

and the exponential moment bound

$$\mathbb{E}|Q(Z)|^m \leq \frac{m!}{2} (2C_1)^{m-2} v_Q \quad \forall m \geq 2$$

hold for constants C_0, C_1 and for all $Q \in \mathcal{Q}$.

Now let a shift parameter $a > 0$ be given. Then we have the inequality

$$\mathbb{E} \max_{Q \in \mathcal{Q}} \left(\mathbb{E}Q(Z) - (1+a) \frac{1}{m} \sum_{i=1}^{m} Q(Z_i) \right) \leq c \left(\frac{\log p}{m} \right)^{\frac{\kappa}{2\kappa - 1}} \ ,$$

where c is a constant which is independent of m and p.

3.4.3 Oracle inequalities for subsample-retrained estimators

Up to now, we have not introduced any conditions on how a candidate statistic $\hat{f}$ behaves when its training sample size changes, i.e. about the relationship of $\hat{f}_m$ and $\hat{f}_n$ for $m \neq n$. As the usual application of cross-validation involves retraining the selected model using *all* the available data to obtain a final estimator, such assumptions are crucial for avoiding such pathological "counter-examples" such as the one we shall see in Example 3.5.2. The first result we formulate is a simple case where even after selection involving estimation with training size n_V, we still only use training samples of size n_V to build the final estimator. In the next section, we then focus on transferring this result to the case where we retrain on all available data after performing cross-validation-based model selection.

We can combine the fundamental inequality and empirical process bounds obtained so far for the following oracle inequality for the averaged and modified CV procedures:

Theorem 3.4.1. Let $\hat{f}_1, \ldots, \hat{f}_p$ be p exchangeable statistics, and assume either Assumption (A1) or (A2). Assume that the risk function $f \longmapsto \mathcal{E}(f)$ is convex. Then both the averaged and the modified cross-validation procedures satisfy the following oracle inequality:

$$\mathbb{E}_{D^n}\Big(\mathcal{E}(\bar{f}^{(n)}(D^n))\Big) \leq \quad (1+a) \min_{j=1,\ldots,p} \Big[\mathbb{E}_{D^{n_V}} \mathcal{E}(\hat{f}_j^{(n_V)}(D^{n_V}))\Big] + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} .$$

3.5 Retraining on the full sample

In Part 1 of Theorem 3.4.1, we make the assumption that the risk $\mathcal{E}(\cdot)$ is convex –for which e.g. the conditional convexity of the contrast function $\gamma(z, f)$, for all z, would suffice– and thereafter in Part 2 we assume that our candidate statistics are exchangeable. To derive a result for a cross-validation estimator retrained on the full data D^n, we shall combine and strengthen these two assumptions.

Regard the modified CV procedure, whose final estimator

$$\bar{f}^{(n)}_{mCV}(D^n) = \hat{f}^{(n_V)}_{\hat{\jmath}(D^n)}(D^{n_V})$$

is retrained on the first n_V pieces of data. For symmetry reasons, Part 2 of Theorem 3.4.1 remains true if we replace $\bar{f}^{(n)}_{mCV}(D^n)$ by

$$\bar{f}^{(n)}_{mCV,k}(D^n) = \hat{f}^{(n_V)}_{\hat{\jmath}(D^n)}(D_{(C_k)})$$

using the training set C_k from the $k-$th split. Now assume that the statistics $\hat{f}_1, \ldots, \hat{f}_p$ can all be written as functionals on the cumulative distribution function of the data, i.e. that there exist functionals $G_1, \ldots, G_p$ such that

$$\hat{f}^{(m)}_j(D^{(m)}) = G_j(F_{D^{(m)}}) \ , \ j = 1, \ldots, p, m \in \mathbb{N} \ , \tag{3.9}$$

where $F_{D^{(m)}}(x) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{Z_i \leq x\}$. (This assumption automatically implies the exchangeability of the statistics.) Obviously

$$F_{D^n}(z) = \frac{1}{V}\sum_{k=1}^{V} F_{D_{(C_k)}}(z) \ .$$

Thus if we make the right convexity assumptions, we can combine the upper bounds for the estimators $\bar{f}^{(n)}_{mCV,k}(D^n)$ to obtain a bound for V-fold cross-validation (of which modified CV is a variant). The full functional convexity assumption that would suffice is:

(B1) *The excess risk $\mathcal{E}(\cdot)$ is convex, and there exist functionals $G_1, \ldots, G_p$ such that*

1. $$\hat{f}^{(m)}_j(D^{(m)}) = G_j(F_{D^{(m)}}) \ , \ j = 1, \ldots, p, m \in \mathbb{N}$$

2. *For all $m \in \mathbb{N}$, the expected estimator risks*

$$E_{D^m}\left[\mathcal{E}(G_j(F_{D^{(m)}}))\right] = E_{D^m}\left[\mathcal{E}(\hat{f}^{(m)}_j(D^{(m)}))\right]$$

are convex as functionals

$$F \mapsto E_{D^m}\left[\mathcal{E}(G_j(F))\right] \ .$$

Under Assumption (B1), we can compute

$$\begin{aligned} E_{D^n}\left[\mathcal{E}\left(\bar{f}^{(n)}_{VCV}(D^n)\right)\right] &= E_{D^n}\left[\mathcal{E}\left(G_{\hat{\jmath}(D^n)}\left(F_{D^n}\right)\right)\right] \\ &= E_{D^n}\left[\mathcal{E}\left(G_{\hat{\jmath}(D^n)}\left(\frac{1}{V}\sum_{k=1}^{V} F_{D_{(C_k)}}\right)\right)\right] \\ &\leq E_{D^n}\left[\frac{1}{V}\sum_{k=1}^{V}\mathcal{E}\left(G_{\hat{\jmath}(D^n)}\left(F_{D_{(C_k)}}\right)\right)\right] \\ &= E_{D^n}\left[\frac{1}{V}\sum_{k=1}^{V}\mathcal{E}\left(\bar{f}^{(n)}_{mCV,k}(D^n)\right)\right], \end{aligned}$$

and thus

$$\begin{aligned} & \mathbb{E}_{D^n}\Big(\mathcal{E}(\bar{f}^{(n)}_{VCV}(D^n))\Big) \\ \leq\ & \frac{1}{V}\sum_{k=1}^{V}\mathbb{E}_{D^n}\Big(\mathcal{E}(\bar{f}^{(n)}_{mCV,k}(D^n))\Big) \\ \leq\ & (1+a)\min_{j=1,\ldots,p}\Big[\mathbb{E}_{D^{n_V}}\mathcal{E}(\hat{f}^{(n_V)}_j(D^{n_V}))\Big] + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \end{aligned}$$

by Part 2 of Theorem 3.4.1.

In many situations, such a convexity condition is too much to ask. As it works by directly replacing a summand in the oracle inequality, Condition (B1) need only hold *up to a small correction term or a constant factor*. Thus either of the following conditions (involving V) suffice:

(B2) There exists a constant $c \geq 0$ such that for all $n \in V\mathbb{N}$, we have the inequality

$$\begin{aligned} & \mathbb{E}_{D^n}\left[\sup_j\left(\mathcal{E}\left(\hat{f}^{(n)}_j(D^n)\right) - \frac{1}{V}\sum_{k=1}^{V}\mathcal{E}\left(\hat{f}^{(n)}_j(D_{(C_k)})\right)\right)\right] \\ \leq\ & c\left(\frac{\log(p)}{n_C}\right)^{\frac{\kappa}{2\kappa-1}}. \end{aligned}$$

(B3) There exists a constant $c \geq 0$ such that for all $n \in V\mathbb{N}$, we

have the inequality

$$\begin{aligned} & \mathbb{E}_{D^n} \sup_j \left[\mathcal{E}\left(\hat{f}_j^{(n)}(D^n)\right) \Big/ \left(\frac{1}{V} \sum_{k=1}^{V} \mathcal{E}\left(\hat{f}_j^{(n)}(D_{(C_k)})\right)\right)\right] \\ \leq \quad & c \, . \end{aligned}$$

We now have the following theorem:

Theorem 3.5.1. Let $\hat{f}_1, \ldots, \hat{f}_p$ be p exchangeable statistics. Assume that at least one of the Assumptions (A1) and (A2) hold, as well as at least one of Assumptions (B1), (B2) and (B3), and let $a > 0$. Then for the V-fold cross-validation procedure, we have the oracle inequality

$$\begin{aligned} \mathbb{E}_{D^n}\left(\mathcal{E}(\bar{f}_{VCV}^{(n)}(D^n))\right) \quad \leq \quad & c_1(1+a) \min_{j=1,\ldots,p} \left[\mathbb{E}_{D^{n_V}} \mathcal{E}(\hat{f}_j^{(n_V)}(D^{n_V}))\right] \\ & + c_2 \left(\frac{\log p}{n_C}\right)^{\frac{\kappa}{2\kappa-1}} \end{aligned}$$

for some $c_1 \geq 1$, $c_2 \geq 0$. If Assumption (B1) or (B2) holds, c_1 can be taken as 1.

The reason why we need extra assumptions such as (B1), (B2) or (B3) is that the computation of the index $\hat{\jmath}(D^n)$ only involves the performances of the estimators for n_V observations ($R_{n_C}(\hat{f})$ depends only on $\hat{f}^{(n_V)}$). Without extra assumptions, it is thus easy to find counter-examples for which $\hat{f}^{(n_V)}$ performs well and $\hat{f}^{(n)}$ performs badly:

Example 3.5.2. Fix an integer V and a sample size $n_0 > 1$ that is a multiple of V. We will construct a set $\mathcal{F} = \{\hat{f}_1, \hat{f}_2\}$ of two estimators (which are functionals of the training data) for which V-fold cross-validation does *not* satisfy the oracle inequality from Theorem 3.5.1.

We consider the classification problem with $0-1$ loss

$$\gamma(Z, f) = \gamma((X, Y), f) = \mathbb{1}_{f(X) \neq Y} \, .$$

Assume that $Y \equiv 1$ a.s. and X is uniformly distributed on $[0, 1]$. The Bayes rule is thus given by $f_0(x) = \mathbb{P}(Y = 1 | X = x) = 1, \forall x \in [0, 1]$. We define statistics $\hat{f}_1 = (\hat{f}_1^{(n)})_n$ and $\hat{f}_2 = (\hat{f}_2^{(n)})_n$ by

$$\hat{f}_1^{(n)} \equiv \begin{cases} 0 & \text{if} \quad 1 \leq n \leq n_0 - 1 \\ 1 & \text{if} \quad n \geq n_0 \end{cases}$$

and

$$\hat{f}_2^{(n)} \equiv \begin{cases} 1 & \text{if} \quad 1 \leq n \leq n_0 - 1 \\ 0 & \text{if} \quad n \geq n_0 \end{cases} .$$

It is easy to see that $\hat{\jmath}(D^n) = \arg\min_{j \in \{1,2\}} R_{n,V}(\hat{f}_j)$ is always equal to 2. Thus the V-fold CV procedure is

$$\bar{f}_{VCV}^{(n_0)}(D^{n_0}) = \hat{f}_{\hat{\jmath}(D^{n_0})}^{(n_0)}(D^{n_0}) = \hat{f}_2^{(n_0)}(D^{n_0}) .$$

Set $\mathcal{F} = \{\hat{f}_1, \hat{f}_2\}$. For any $1 \leq n \leq n_0$, it is easy to check that

$$\min_{\hat{f} \in \mathcal{F}} \mathbb{E}_{D^{n_0}}[\mathcal{E}(\hat{f}^{(p)}(D^{n_0}))] = 0$$

and

$$\mathbb{E}_{D^{n_0}}[\mathcal{E}(f_{VCV}^{(n_0)}(D^{n_0}))] = 1 .$$

As we can do this for arbitrarily high sample sizes n_0, V-fold cross-validation is not even risk-consistent at this level of generality –and certainly does not satisfy any meaningful oracle inequalities. Convexity and near-convexity conditions such as (B1)–(B3) are necessary.

3.6 Verifying the classical conditions (A1) and (A2) in applications

In order to obtain either of Theorems 3.4.1 and 3.5.1, we first need to verify the classical conditions (A1) and (A2). These conditions fulfil the same roles as in Chapter 2, namely controlling the tail and margin behaviour for loss functions of trained estimators. Unlike in Chapter 2, we shall only be able to handle exponential-tailed loss functions here, rather than loss functions with power tails. Condition (A2) corresponds to the exponential tail case in Chapter 2, and we have already seen it in some examples; Condition (A1), on the other hand, is formulated in terms of ψ_1-Orlicz norms, which we will look at more depth here.

3.6.1 Regression

Regard the following general Gaussian regression model: let (X, Y) be a random pair, with X taking on values in $\mathbb{R}^d$, and with $Y = \beta_*^T X + \sigma\varepsilon$

for a variance $\sigma \geq 0$ and a standard normal error ε independent of X. We would like to select between linear regression estimators $\beta^T X$, and the loss function we use for this is the squared loss $\gamma(x, y; \beta) := (y - \beta^T x)^2$. Our candidate estimators can depend on n independent data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ which are all distributed identically to (X, Y). To apply our results up to and including Theorem 3.5.1 to this example, we need to show

- either Assumption (A1) or Assumption (A2) on the distribution of $\gamma(X, Y)$, i.e. a margin and a risk bound condition –possibly only with high probability
- the representability of each candidate estimator by a functional satisfying condition (B1), or at least one of the conditions (B2) and (B3) .

Margin condition for the ψ_1-Orlicz norm: We start with the more difficult margin condition, namely

$$\left\| \gamma(\cdot, \hat{f}^{(m)}(D^n(m))) - \gamma(\cdot, f_0) \right\|_{L_{\psi_1}(\pi)} \leq K_1 \big(\mathcal{E}(\hat{f}^{(m)}(D^n(m))) \big)^{1/2\kappa} ,$$

a bound on an exponential Orlicz norm. This we can obtain with the help of the following lemma:

Lemma 3.6.1. *Assume that X and $\beta \in \mathbb{R}^p$ are such that*

$$\frac{\langle X, \beta \rangle^2}{\mathbb{E}\langle X, \beta \rangle^2} \in L_{\psi_1} , \qquad \left\| \frac{\langle X, \beta \rangle^2}{\mathbb{E}\langle X, \beta \rangle^2} \right\|_{\psi_1} \leq K_2 \tag{3.10}$$

$$\text{and} \quad \mathbb{E}\langle X, \beta - \beta_* \rangle^2 \leq K_0 . \tag{3.11}$$

Then we have

$$\begin{aligned} & \|(Y - X^t\beta)^2 - (Y - X^t\beta_*)^2\|_{\psi_1} \\ \leq \quad & K_1 (E\left[(Y - X^t\beta)^2 - (Y - X^t\beta_*)^2\right])^{1/2}, \end{aligned}$$

where $K_1 := 10\sqrt{K_2} \cdot \max(\sqrt{K_0 K_2}, \sigma)$ is an absolute constant depending only on K_0, σ and K_2.

Thus the problem of showing the ψ_1 margin condition is reduced to that of verifying conditions (3.10) and (3.11) on the regression design (obviously only random designs are of material interest here).

Example 3.6.2. *Assume that the random vector X has range $\mathbb{R}^d$ and follows a multivariate normal distribution with mean 0 and the positive definite covariance matrix Σ. Write $\tilde{\beta}$ for $\beta - \beta_0$. Then $\langle X, \tilde{\beta} \rangle$ follows a standard normal distribution with mean 0, and thus*

$$Z := \frac{\langle X, \tilde{\beta} \rangle}{\sqrt{E\left[\langle X, \tilde{\beta} \rangle^2\right]}}$$

is standard normal. Consequently,

$$\left\| \frac{\langle X, \tilde{\beta} \rangle^2}{E\left[\langle X, \tilde{\beta} \rangle^2\right]} \right\|_{\psi_1} = \left\| Z^2 \right\|_{\psi_1} = \|Z\|_{\psi_2}^2 = \frac{8}{3} \ .$$

This holds for all values of $\tilde{\beta}$.

Risk tails and oracle inequalities: Assume that X has a finite second moment K_3 (which is already necessary for the margin condition in (A1) or (A2) to hold), and that the distribution $\tilde{P} = P^{\otimes n}$ of the training data is such that

$$\tilde{P}\left[\max_{\lambda, \mu \in [0,1]} |\beta(\lambda) - \beta(\mu)|_2 > t \right] \leq K_4 \cdot \exp(-K_5 t) \ .$$

Then we have

$$\begin{aligned} \mathbb{E}Q((X,Y),\lambda) = \mathbb{E}|Y - X\beta(\lambda)|^2 &= \sigma^2 + \mathbb{E}\langle X, \beta(\lambda^*) - \beta(\lambda) \rangle^2 \\ &\leq \sigma^2 + K_3^2 t^2 \end{aligned}$$

with probability at least $1 - K_4 \cdot \exp(-K_5 t)$. Thus for any given $C_1 > \sigma^2$, we have

$$\begin{aligned} \tilde{P}\left[\mathbb{E}Q(Z,\lambda) \leq C_1 \ \forall \lambda \in \mathcal{G}\right] &\geq 1 - pK_4 \exp\left(-K_5 \sqrt{\frac{C_1 - \sigma^2}{K_3^2}} \right) \\ &= 1 - pK_4 \exp\left(-\frac{K_5}{K_3} \sqrt{C_1 - \sigma^2} \right) \ . \end{aligned}$$

This is also a lower bound on the probability that

$$\begin{aligned} & \mathbb{E} \max_{Q \in \mathcal{Q}} \left(\mathbb{E}Q(Z) - (1+a) \frac{1}{m} \sum_{i=1}^{m} Q(Z_i) \right) \\ \leq \quad & \tilde{C}_0 \left(\frac{\log |\mathcal{G}|}{m} \right)^{\frac{\kappa}{2\kappa - 1}} + \tilde{C}_1 \left(\frac{\log |\mathcal{G}|}{m} \right) \end{aligned}$$

for constants

$$\tilde{C}_0 := 2\tilde{C}_2^{\frac{\kappa}{2\kappa-1}} \ , \quad \tilde{C}_1 := 2\tilde{C}_2 \ , \quad \tilde{C}_2 := \frac{3C_0(1+a) \vee 9C_0^2(1+a)^2}{(a \wedge 1/C_1)^{1/(2\kappa)}} \ .$$

Quadratic margin condition: We have already looked at the quadratic margin condition for Gaussian linear regression in Example 2.6.2. Essentially there, a quadratic margin condition easily holds when we have a Lipschitz loss function, are performing fixed-design regression, or else a supremum-bounded class of candidate regression functions. As part 1 of Condition (A2) is a boundedness condition, it is all that needs to hold for Condition (A2) to apply. Then we obtain the oracle inequality of Theorem 3.5.1.

3.6.2 Density estimation

In Chapter 2, Example 2.6.6 (and subsequently, Section 2.8.1) looked at moment and margin conditions for maximum-likelihood density estimation, where the loss is given by $\gamma_f(x) := -\log(f)/2$. The conditions shown there were almost the same as those required by Assumption (A2), with the small change that Assumption (A2) now only requires conditions w.r.t. the true density f_0, rather than the oracle f_*. Thus the approximability condition $\sqrt{f_0/f^*} \leq L/2\sqrt{2}$ is not necessary. Even if it were necessary, e.g. if the moment condition is only true over the oracle, then kernel density estimation (for example) with a kernel supported by the full underlying space would suffice to obtain high-probability results.

Thus as in Example 2.6.6, the risk inequality

$$\begin{aligned}\mathbf{E}h^2(\hat{f}_{VCV}, f_0) &\leq 2(2+\sqrt{2})^2\mathbf{E}h^2(\bar{f}, f_0) + 2(1+\sqrt{2})^2 h^2(f_*, f_0) \\ &\leq 2(2+\sqrt{2})^2\mathbf{E}\hat{\mathcal{K}} + 2(1+\sqrt{2})^2\mathcal{K}_* \\ &\leq \left[2(2+\sqrt{2})^2(1+\delta) + 2(1+\sqrt{2})^2\right] \cdot \mathcal{K}_* \\ &\quad +4(L+1/\sqrt{2})^2(2+\sqrt{2})^2\left(1+\frac{1}{\delta}\right)\cdot\left(\frac{\log(p)}{n_C}\right)\end{aligned}$$

holds for any true density f_0 and any density estimators $\hat{f}_j$ when $\hat{f}$ is the modified cross-validation procedure or averaged V-fold cross-validation.

3.6.3 Classification

Example 2.6.4 and Section 2.8.3 look at classification with 0-1 loss and arbitrary measurable classifiers, and derive exponential moment conditions in such a case, as well as margin conditions for some margin parameter $\kappa = 1 + 1/\alpha > 1$ when the Tsybakov margin condition

$$P\left[|\eta - 1/2| \leq t\right] \leq C_\eta t^\alpha$$

holds for the true class probability

$$\eta = \mathbf{E}(Y_i|X_i = \cdot) = P(Y_i = 1|X_i = \cdot)\ .$$

Thus whenever such a Tsybakov margin condition holds (such as when η is differentiable in a neighbourhood of the margin set $\{\eta = 1/2\}$), Assumption (A2) directly follows.

3.7 Verifying conditions (B1) – (B3)

After verifying classical conditions (A1) or (A2), like those in Chapter 2, we obtain oracle inequalities for the modified and averaged cross-validation procedures, which only train on subsets of size n_V, however. Now we should investigate to what extent Conditions (B1), (B2), or (B3) hold. These conditions tell of the extent to which raising the sample size used for retraining the final estimator inflates the loss of this estimator. If at least one of these conditions holds in a given example, we obtain

an oracle inequality for the fully re-trained cross-validation estimator, the type that is most widely used in practice. In some cases theoretical considerations suffice to obtain one of these conditions, and in other cases, simulations may shed some light onto their veracity.

3.7.1 Simple case: Location model

Regard the location model $X = \mu + \sigma\varepsilon$, where μ is a real parameter, σ is a positive parameter, and $\varepsilon \sim \mathcal{N}(0,1)$ is standard Gaussian noise. Let n observations $X_1, \ldots, X_n$ of this variable be given. The M-estimator of the location μ for square loss is $\hat{\mu} = \bar{X}$, the arithmetic mean of the observations. We can write $\hat{\mu}$ as a functional $\hat{\mu}(F_{D^n})$ of the CDF of the training data. Its excess risk is

$$\mathcal{E}\left(\hat{\mu}(F_{D^n})\right) = (\mu - \bar{X})^2 \ ,$$

and for V-fold cross-validation, the discrepancy between the risk when training on the full data set and the risks when training on the partial training sets C_k is

$$\begin{aligned}
&\mathcal{E}\left(\mu(F_{D^n})\right) - \frac{1}{V}\sum_{k=1}^{V} \mathcal{E}\left(\mu(F_{C_k})\right) \\
&\quad = (\mu - \bar{X})^2 - \frac{1}{V}\sum_{k=1}^{V}(\mu - \frac{1}{n_C}\sum_{i \in C_k} X_i)^2 \\
&= -\frac{1}{V-1}\sum_{k=1}^{V}\sum_{i \neq j \in B_k} 2\varepsilon_i\varepsilon_j + \frac{1}{(V-1)^2}\sum_{k \neq \ell = 1}^{V}\sum_{i \in B_k, j \in B_\ell} 2\varepsilon_i\varepsilon_j \\
&\quad - \frac{1}{V-1}\sum_{i=1}^{n}\varepsilon_i^2 \\
&\leq -\frac{1}{(V-1)^2}\left(\sum_{i=1}^{n}\varepsilon_i\right)^2 \\
&\leq 0 \ .
\end{aligned}$$

Thus

$$\mathcal{E}\left(\mu(\sum_{k=1}^{V}\frac{1}{V}F_{C_k})\right) = \mathcal{E}\left(\mu(F_{D^n})\right) \leq \frac{1}{V}\sum_{k=1}^{V}\mathcal{E}\left(\mu(F_{C_k})\right) \ ,$$

which means that the (excess) risk is convex as a functional $\mathcal{E}(\mu(\cdot))$ of the training data, when the latter is written in the form of its CDF. This convexity property is exactly Condition (B1).

3.7.2 Condition (B1) for kernel density estimation

Assume that we are estimating a density f_0 on $\mathbb{R}$ from real-valued data $X_1, \ldots, X_n$ using a kernel density estimator

$$\hat{f}_h(D^n; x) := \frac{1}{nh} \sum_{i=1}^{n} \phi(\frac{x - X_i}{h}) \ ,$$

where ϕ is the probability density used as a kernel and h is the bandwidth. If the collection of splits C_k (each of size n_C) we are using for cross-validation is a balanced one, with each index appearing in equally many C_k, then the average estimate over all the splits is exactly the estimate for the full data set, due to the linear nature of the estimator:

$$\frac{1}{V} \sum_{k=1}^{V} \hat{f}_h(D_{(C_k)}; x) = \hat{f}_h(D^n; x) \quad \forall x \ .$$

Now we can use any loss $\gamma_f(x) := \gamma(f(x))$ derived from a convex function $\gamma(\cdot)$, and we have the convexity property

$$\begin{aligned}
& \mathcal{E}(\hat{f}_h(D^n)(\\
= \ & \mathbb{E}[\gamma \left(\frac{1}{V} \sum_{k=1}^{V} \hat{f}_h(D_{(C_k)}) \right) - \gamma(f_0)] \\
\leq \ & \mathbb{E}[\frac{1}{V} \sum_{k=1}^{V} \gamma \left(\hat{f}_h(D_{(C_k)}) \right) - \gamma(f_0)] \\
= \ & \frac{1}{V} \sum_{k=1}^{V} \mathcal{E} \left(\hat{f}_h(D_{(C_k)}) \right) \ .
\end{aligned}$$

Thus if our family of candidate estimators contains only kernel density estimators with an arbitrary combination of kernels and bandwidths, and we perform "balanced" cross-validation (such as V-fold and leave-p-out) using a convex loss, Condition (B1) holds.

3.7.3 Conditions (B1) and (B3) for classification

Let $\ell(\cdot)$ be a convex function, and let $\hat{f} = \hat{f}(D^n; X)$ be a trained function that gives rise to a classifier $\hat{Y} := \text{sign}(\hat{f}(D^n; X))$. Evaluate this classifier using the loss $\gamma_{\hat{f}}(x, y) := \ell(y \cdot \hat{f}(x))$. If the discriminant function $\hat{f}$ is additive, i.e. $\hat{f}(D^n) \equiv \frac{1}{V}\sum_{k=1}^{V} \hat{f}(D_{(C_k)})$, then we have

$$\begin{aligned}\mathcal{E}(\hat{f}(D^n)) &= \mathbb{E}\ell(Y \cdot \hat{f}(X)) = \mathbb{E}\ell(Y \cdot \frac{1}{V}\sum_{k=1}^{V} \hat{f}(D_{(C_k)}; X)) \\ &\leq \frac{1}{V}\ell(Y \cdot \sum_{k=1}^{V} \hat{f}(D_{(C_k)}; X)) = \frac{1}{V}\mathcal{E}(\hat{f}(D_{(C_k)}))\ ,\end{aligned}$$

and thus Condition (B1) holds directly. This occurs e.g. when we are performing linear discriminant analysis for known covariance structure and overall class ratios (and thus only the group means are estimated), and we use hinge loss $\ell(x) = (1 - x)_+$.

If we are using 0-1 loss (as in Chapter 2), we have no convexity properties of it. The risk ratio in Condition (B3) is then

$$\frac{\mathcal{E}(\hat{f})}{\frac{1}{V}\sum_{k=1}^{V} \mathcal{E}(\hat{f})} = \frac{P_{X,Y}\left[\hat{f}(D^n; X) \neq Y\right]}{\frac{1}{V}\sum_{k=1}^{V} P_{X,Y}\left[\hat{f}(D_{(C_k)}; X) \neq Y\right]}\ , \tag{3.12}$$

and Condition (B3) becomes

$$E_{D^n}\left[\sup_j \left(\frac{P_{X,Y}\left[\hat{f}(D^n; X) \neq Y\right]}{\frac{1}{V}\sum_{k=1}^{V} P_{X,Y}\left[\hat{f}(D_{(C_k)}; X) \neq Y\right]}\right)\right] \leq c \tag{3.13}$$

for some constant $c > 0$.

A sufficient condition for Inequality 3.13 to hold is given by the following covering assumption. Denote the estimate, using training data D, of the class with label 1 by $\hat{K}_1(D)$. Assume that all candidate classifiers are such that for any two index sets J_1 and J_2 with $J_1 \cup J_2 = \{1, \ldots, n\}$, $\hat{K}_1(D^n) \subseteq \hat{K}_1(D_{J_1}) \cup \hat{K}_1(D_{J_2})$. Then as any two distinct sub-datasets $D_{(C_{k_1})}$ and $D_{(C_{k_2})}$ cover D^n, the sum of probabilities in the denominator of Condition (B3) is less than the probability in its enumerator, and thus Condition (B3) holds with ratio bound 2. This covering condition is satisfied e.g. by 1-nearest neighbours classification.

3.7.4 Condition (B3) for least-squares regression

Assume the following model: Let $Y = \beta_0^T X + \varepsilon$, where X is a p-dimensional vector whose components X_j are i.i.d. with mean μ and variance τ^2, and the noise ε is Gaussian with mean zero and variance σ^2. Then the excess risk $\mathcal{E}(\hat{\beta})$ of a linear fit $\hat{\beta}$ is

$$\begin{aligned}\mathcal{E}(\hat{\beta}) &= E\left[(Y - \hat{\beta}^T X)^2\right] - E\left[(Y - \beta_0^T X)^2\right] = E\left[((\beta_0 - \hat{\beta})^T X)^2\right] \\ &= \sum_{j,k=1}^{p} (\beta_{0,j} - \hat{\beta}_j)(\beta_{0,k} - \hat{\beta}_k) E\left[X_j X_k\right] \\ &= \mu^2 \cdot \sum_{j,k=1}^{p} (\beta_{0,j} - \hat{\beta}_j)(\beta_{0,k} - \hat{\beta}_k) + \tau^2 \cdot \sum_{j=1}^{p} (\beta_{0,j} - \hat{\beta}_j)^2 \ ,\end{aligned}$$

which can be written as

$$(\beta_0 - \hat{\beta})^T A (\beta_0 - \hat{\beta})$$

with $A_{jk} = \mu^2$ for $j \neq k$ and $A_{jj} = \mu^2 + \tau^2$ for all j.

As A is a fixed matrix, rapidly-converging estimation procedures may stabilize the risk ratio

$$\left[\mathcal{E}\left(\hat{\beta}_j^{(n)}(D^n)\right)\right] \Big/ \left[\frac{1}{V}\sum_{k=1}^{V} \mathcal{E}\left(\hat{\beta}_j^{(n)}(D_{(C_k)})\right)\right]$$

in a uniform way, but such properties are not obvious to see in a theoretical manner.

3.7.5 Simulating Conditions (B1), (B2) and (B3)

In some situations, such as the regression example above, it is practical to utilise simulations to provide evidence that condition (B1), (B2) or (B3) holds, and thus to show that retraining on the full sample does not weaken the rate $\frac{\kappa}{2\kappa-1}$ found in the oracle inequalities for subsample-trained estimators. For (B2), as an example, the simplest approach is simply to simulate the cross-validation empirical process

$$\left|\mathcal{E}\left(G_j\left(F_{D^n}\right)\right) - \frac{1}{V}\sum_{k=1}^{V} \mathcal{E}\left(G_j\left(F_{D_{(C_k)}}\right)\right)\right|_{j \in \{1,\ldots,p\}}$$

that it features, and from that estimate its rate. Formally, this means that for some suitably large approximation sample size N_A, and after taking a sample of D^n, we take independent samples $Z_{i,1}, \ldots, Z_{i,N_A}$, $Z'_{i,1}, \ldots, Z'_{i,N_A}$ of Z. Then we first approximate the cross-validation empirical process by

$$\left| \frac{1}{N_A} \sum_{i=1}^{N_A} \gamma(Z_{i,j}, \hat{f}_j(D^n)) - \frac{1}{V} \sum_{k=1}^{V} \frac{1}{N_A} \sum_{i=1}^{N_A} \gamma\left(Z'_{i,j}, \hat{f}_j(D_{(C_k)})\right) \right|_{j \in \{1,\ldots,p\}}.$$

We now take the absolute maximum of this empirical process, and repeat for "outer samples" D^n_ℓ of D^n with ℓ indexed in $\{1, \ldots, N_B\}$ for some reasonably large sample size N_B, obtaining the simulated values

$$C_\ell := \sup_{j \in \{1,\ldots,p\}} \left| \frac{1}{N_A} \sum_{i=1}^{N_A} \gamma(Z_{i,j}, \hat{f}_j(D^n)) - \frac{1}{V} \sum_{k=1}^{V} \frac{1}{N_A} \sum_{i=1}^{N_A} \gamma\left(Z'_{i,j}, \hat{f}_j(D_{(C_k)})\right) \right|.$$

The excess risks simulated for this computation can also be used to approximate both sides of Conditions (B1) and (B3), and thus to check those two conditions.

Elastic Net parameter selection using squared loss. We already looked at the classical conditions (A1) and (A2) for regression models that we select by minimize the empirical mean squared error. When we now simulate the cross-validation empirical process for this general example, we profit from the simplification of knowing the excess risk $\mathcal{E}$ exactly: if we denote the true linear coefficient vector by β, the excess risk of an estimate $\hat{\beta}(D^n)$ is then $\mathcal{E}(\hat{\beta}(D^n)) = \|\hat{\beta}(D^n) - \beta\|_2^2$. Thus only samples of D^n are needed.

Simulating suprema C_ℓ of the empirical process repeatedly, and then doing the same again for a number of different sample sizes n, we typically get results as shown in Figure 3.2 (here on a log-log scale), where the empirical process supremum eventually follows a power law. The class of estimators used for this simulation consists of Elastic Net estimators for a fixed grid of ℓ_1- and ℓ_2-penalty coefficients. See Appendix A for details.

Regressing to estimate rates. Write the rough model

$$\log(C_\ell) = \alpha_0 - \alpha \log(n_\ell) + \varepsilon \ ,$$

where n_ℓ is the total sample size used in the simulation of C_ℓ, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some unknown variance σ^2. This emanates from the model

$$C_\ell \approx \eta_0 \cdot n_\ell^{-\alpha} \ ,$$

and the multiplicative errors it contains are justified if we consider that our empirical process suprema are all non-negative. If we assume this model, we can use simple linear regression to estimate the rate $n^{-\alpha}$. If the variance has not yet stabilized for low sample sizes, we remove these and regard only higher sample sizes. Then we apply linear regression to the remaining data to obtain a rate estimate, such as that shown in Figure 3.2, where the estimated rate is $n^{-\alpha} = n^{-0.73\pm0.01}$ (see Appendix A for details). In that situation our heuristics suggest that the risk inequality in Theorem 3.5.1 does not necessarily retain the full rate from before retraining. However, Appendix A gives several variations on the same example and these sometimes give better rates, especially for low signal-to-noise ratios.

3.8 Choice of cross-validation procedure

In the Elastic Net example used above, the behaviour of the supremum of the cross-validation empirical process from Assumptions (B1)–(B3) for various choices of split number ($V = 2$, 5, 10 and 20) is illustrated in Figure 3.3. This combines simulations as in Figure 3.1 for those different values of V. Thus Conditions (B1)–(B3) appear not to influence the choice of the number of splits in this example at least.

We can, however, use the oracle inequalities we have gained for some heuristical reasoning. If we have a family of estimators whose excess risk when trained on n_C data is $\mathcal{O}(n_C^{-\beta})$, and the correction term in the oracle inequality (Theorem 3.4.1) implies that selecting the smallest loss using n_V pieces of training data leads to an extra risk $\mathcal{O}(n_V^{-\frac{\kappa}{2\kappa-1}})$, we find an overall rate $\mathcal{O}(n_C^{-\beta}) + \mathcal{O}(n_V^{\frac{\kappa}{2\kappa-1}})$. (If we re-train on the full data set, condition (B2) is supposed to ensure this rate is not affected.) This combined rate is optimal when its two constituent rates are equal. For

V-fold cross-validation, $n_C := n(1-1/V)$ and $n_V := n/V$, and thus the condition is

$$\mathcal{O}(n^{-\beta}(1-1/V)^{-\beta}) = \mathcal{O}(n^{-\frac{\kappa}{2\kappa-1}} V^{\frac{\kappa}{2\kappa-1}}) \ .$$

As $V \geq 2$ at all times, the factor $(1-1/V)^{-\beta}$ drops away again. For the rates to be equal, we need to have $\beta \leq \frac{\kappa}{2\kappa-1}$; then choosing $V = \mathcal{O}(n^{\frac{\kappa}{2\kappa-1}-\beta})$ suffices, giving us the same overall rate $n^{-\beta}$ for the estimator risk as for the individual risks.

For example, if we have parametric regression procedures (differing e.g. by the complexity of the model involved), then a common rate at which the individual excess risks will decay is $\mathcal{O}(n^{-1})$, giving us $\beta = 1$. Using a quadratic loss function for cross-validation then leads to a quadratic margin condition ($\kappa = 1$); then $\frac{\kappa}{2\kappa-1} - \beta = 0$, meaning that a constant value of V is best. If, however, our individual estimators (e.g. non-parametric) have risk decay rate $\mathcal{O}(n^{-2/3})$, then $\frac{\kappa}{2\kappa-1} - \beta = 1/3$, and we need to increase V proportionally to $n^{1/3}$ to optimize the behaviour of the oracle inequality.

3.9 Proofs

Proof of Lemma 3.3.1. First define the modified risk

$$\tilde{R}_{N_C} := R_{n_C}(\hat{f}_{\hat{\jmath}(D^n)}) = \frac{1}{N_C} \sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i \notin C_k} \gamma(Z_i, \hat{f}^{(n_V)}_{\hat{\jmath}(D^n)}(D_{(C_k)})) \ , \quad (3.14)$$

regardless of the estimator involved. Obviously we have $\tilde{R}_{n_C} \leq R_{n_C}(\hat{f}_j)$ for all indices $j \in \{1, \ldots, p\}$.

Using this inequality, we have the following basic inequality for all datasets D^n and indices j:

$$\begin{aligned}
& \mathcal{E}(\bar{f}^{(n)}(D^n)) \\
= \ & (1+a)(\tilde{R}_{n_C} - R_{n_C}(f_0)) + (\mathcal{E}(\bar{f}^{(n)}(D^n))) \\
& -(1+a)(\tilde{R}_{n_C} - R_{n_C}(f_0)) \\
\leq \ & (1+a)(R_{n_C}(\hat{f}_j) - R_{n_C}(f_0)) \\
& + \Big(\mathcal{E}(\bar{f}^{(n)}(D^n)) - (1+a)(\tilde{R}_{n_C} - R_{n_C}(f_0))\Big) \ .
\end{aligned}$$

Now for every j, the i.i.d. property of $Z_1, \ldots, Z_n$ tells us that

$$\begin{aligned}
& \mathbb{E}_{D^n} R_{n_C}(\hat{f}_j) - \mathbb{E}_{D^n} R_{n_C}(f_0) \\
= \; & \frac{1}{N_C} \sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i \notin C_k} \left(\mathbf{E}_{D^n} \gamma(Z_i, \hat{f}_j^{(n_V)}(D_{(C_k)})) - \mathbf{E}_{D^n} \gamma(Z_i, f_0) \right) \\
= \; & \frac{1}{N_C} \sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i \notin C_k} \left(\mathbf{E}_{D_{(C_k)}} \mathcal{E}(\hat{f}_j^{(n_V)}(D_{(C_k)})) \right) \\
= \; & \mathbb{E}_{D^{n_V}} \mathcal{E}(\hat{f}_j^{(n_V)}(D^{n_V})) \; .
\end{aligned}$$

Combining these two parts, we obtain a proto-oracle inequality

$$\begin{aligned}
\mathbb{E}_{D^n} \mathcal{E}(\bar{f}^{(n)}(D^n)) \leq (1+a) \min_{j=1,\ldots,p} & \left[\mathbb{E}_{D^{n_V}} \mathcal{E}(\hat{f}_j^{(n_V)}(D^{n_V})) \right] \\
& + \mathbb{E}_{D^n} \left[\mathcal{E}(\bar{f}^{(n)}(D^n)) - (1+a)(\tilde{R}_{n_C} - R_{n_C}(f_0)) \right] .
\end{aligned}$$

Now we reach a key stage, at which we must use some properties of the risk or of the estimator to allow the conversion of the second summand into an empirical process term.

If $\bar{f} = \bar{f}_{amCV}$ and the risk $\mathcal{E}$ is convex, we have

$$\begin{aligned}
\mathbb{E}_{D^n} \mathcal{E}(\bar{f}_{amCV}^{(n)}(D^n)) &= \mathbb{E}_{D^n} \mathcal{E}(\frac{1}{N_C} \sum_{k=1}^{N_C} \hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D_{(C_k)})) \\
&\leq \frac{1}{N_C} \sum_{k=1}^{N_C} \mathbb{E}_{D^n} \mathcal{E}(\hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D_{(C_k)})) \\
&= \mathbb{E}_{D^n} \mathcal{E}(\hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D^{n_V})) \\
&= \mathbb{E}_{D^n} P\gamma(\cdot, \hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D^{n_V})) \; ,
\end{aligned}$$

where we use the particular form of the family $(C_k)_{k=1,\ldots,N_C}$ of index sets that is given by V-fold cross-validation. We specifically use that for each index k there exists some permutation of $\{1, \ldots, n\}$ which maps $D_{(C_k)}$ to D^{n_V} and maps the family $(C_k)_{k=1,\ldots,N_C}$ to itself. Such a permutation easily exists for V-fold cross-validation (a cyclical shift by n_C, for example. It does not change $\hat{\jmath}(D^n)$, and thus $\mathbb{E}_{D^n} \mathcal{E}(\hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D_{(C_k)}))$ is equal to $\mathbb{E}_{D^n} \mathcal{E}(\hat{f}_{\hat{\jmath}(D^n)}^{(n_V)}(D^{n_V}))$.

Furthermore, if $\bar{f} = \bar{f}_{mCV}$, we have

$$\begin{aligned}\mathbb{E}_{D^n}\mathcal{E}(\bar{f}^{(n)}_{mCV}(D^n)) &= \mathbb{E}_{D^n}\mathcal{E}(\hat{f}^{(n_V)}_{\hat{j}(D^n)}(D^{n_V})) \\ &= \mathbb{E}_{D^n}P\gamma(\cdot, \hat{f}^{(n_V)}_{\hat{j}(D^n)}(D^{n_V}))\end{aligned}$$

more directly.

Using this estimator-dependent step, the i.i.d. property of our data, and the exchangeability of all the $\hat{f}_j$, we can reduce the second summand of our proto-oracle inequality as follows:

$$\begin{aligned}&\mathbb{E}_{D^n}\left[\mathcal{E}(\bar{f}^{(n)}(D^n)) - (1+a)(\tilde{R}_{n_C} - R_{n_C}(f_0))\right] \\ &\leq \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}_{D^n}\Bigg[P\gamma(\cdot, \hat{f}^{(n_V)}_{\hat{j}(D^n)}(D^{n_V})) - P\gamma(\cdot, f_0) \\ &\quad - \frac{1+a}{n_C}\sum_{i\notin C_k}\Big(\gamma(Z_i, \hat{f}^{(n_V)}_{\hat{j}(D^n)}(D_{(C_k)})) - \gamma(Z_i, f_0)\Big)\Bigg] \\ &\leq \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}_{D^n}\max_{j=1,\ldots,p}\Bigg[P\gamma(\cdot, \hat{f}^{(n_V)}_{j}(D^{n_V})) - P\gamma(\cdot, f_0) \\ &\quad - \frac{1+a}{n_C}\sum_{i\notin C_k}\Big(\gamma(Z_i, \hat{f}^{(n_V)}_{j}(D_{(C_k)})) - \gamma(Z_i, f_0)\Big)\Bigg] \\ &= \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}_{D^n}\max_{j=1,\ldots,p}\Bigg[P\gamma(\cdot, \hat{f}^{(n_V)}_{j}(D^{n_V})) - P\gamma(\cdot, f_0) \\ &\quad - \frac{1+a}{n_C}\sum_{i=n_V+1}^{n}\Big(\gamma(Z_i, \hat{f}^{(n_V)}_{j}(D^{n_V})) - \gamma(Z_i, f_0)\Big)\Bigg] \\ &= \mathbb{E}_{D^n}\max_{j=1,\ldots,p}\Bigg[P\gamma(\cdot, \hat{f}^{(n_V)}_{j}(D^{n_V})) - P\gamma(\cdot, f_0) \\ &\quad \frac{1+a}{n_C}\sum_{i=n_V+1}^{n}\Big(\gamma(Z_i, \hat{f}^{(n_V)}_{j}(D^{n_V})) - (Z_i, f_0)\Big)\Bigg] \\ &= \mathbb{E}_{D^n}\max_{j=1,\ldots,p}\left[(P - (1+a)P_{n_C})\left(\gamma(\cdot, \hat{f}^{(n_V)}_{j}(D^{n_V})) - \gamma(\cdot, f_0)\right)\right] ,\end{aligned}$$

which now gives the desired result.

■

Proof of Lemma 3.4.1. For any $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
\leq\ &\sum_{Q\in\mathcal{Q}} \mathbb{P}\Big[\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \geq \delta\Big] \\
\leq\ &\sum_{Q\in\mathcal{Q}} \mathbb{P}\Big[\mathbb{E}Q(Z) - \frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \geq \frac{\delta + a\mathbb{E}Q(Z)}{1+a}\Big] \ .
\end{aligned}$$

Now we apply Lemma 1.7.4 to the random variables $Q(Z_1), \ldots, Q(Z_m)$ and $Q(Z)$ (which all lie in the ψ_1-Orlicz space) to obtain

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
\leq\ &\exp\left(-\frac{m}{2}\left(\left(\frac{\delta + a\mathbb{E}Q(Z)}{4(1+a)\|Q(Z)\|_{L_{\psi_1}}}\right)^2, \frac{\delta + a\mathbb{E}Q(Z)}{4(1+a)\|Q(Z)\|_{L_{\psi_1}}}\right)\right) \ .
\end{aligned}$$

We then combine this with the margin condition

$$\|Q(Z)\|_{L_{\psi_1}} \leq C_0(\mathbb{E}Q(Z))^{1/2\kappa}$$

to get the inequality

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
\leq\ &\sum_{Q\in\mathcal{Q}} \exp\Big[-C_2 m\left(\Big(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}}\Big)^2 \wedge \Big(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}}\Big)\right)\Big] \ ,
\end{aligned}$$

where we use the constant

$$C_2 := \frac{1}{8C_0(1+a)(1 \vee 4C_0(1+a))} \ .$$

By using the inequality $u + v \geq u^{(2\kappa-1)/(2\kappa)} v^{1/(2\kappa)}, \forall u, v > 0$ (which follows from looking at the cases $u \leq v$ and $u \geq v$ separately), it is easy to see that

$$\Big(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}}\Big)^2 \geq \delta^{2-\frac{1}{\kappa}} \cdot a^{\frac{1}{2\kappa}} \ .$$

As furthermore

$$\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}} \geq \delta \cdot C_1^{-\frac{1}{2\kappa}} ,$$

we thus have the inequality

$$C_2 \left(\left(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}} \right)^2 \wedge \left(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}} \right) \right) \geq C_3(\delta^{2-1/\kappa} \wedge \delta)$$

when

$$C_3 := C_2 \cdot \left(a \wedge \frac{1}{C_1} \right)^{1/(2\kappa)} .$$

So for any $\delta > 0$,

$$\mathbb{P}\Big[\max_{Q \in \mathcal{Q}} \Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \Big) \geq \delta \Big] \leq p \exp(-C_3 m(\delta^{2-1/\kappa} \wedge \delta)) .$$

Now we can use the fact that

$$\int_a^\infty \exp(-bt^\alpha)dt \leq (\alpha b a^{\alpha-1})^{-1} \exp(-ba^\alpha)$$

for any $\alpha \geq 1$ and $a, b > 0$ to get, for any $u > 0$ and $v > 0$,

$$\begin{aligned}
&\mathbb{E}\Big[\max_{Q \in \mathcal{Q}} \Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \Big) \Big] \\
&\leq \int_0^\infty \mathbb{P}\Big[\max_{Q \in \mathcal{Q}} \Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \Big) \geq \delta \Big] d\delta \\
&\leq u + p\int_u^1 \exp(-C_3 m\delta^{2-1/\kappa})d\delta + v + p\int_{1+v}^\infty \exp(-C_3 m\delta)d\delta \\
&\leq u + p\frac{\exp(-C_3 m u^{2-1/\kappa})}{C_3 m u^{1-1/\kappa}} + v + p\frac{\exp(-C_3 m v)}{C_3 m} . \qquad (3.15)
\end{aligned}$$

We denote by $\mu(p)$ the unique solution of $\mu = p\exp(-\mu)$. For this quantity, we have the inequality $(\log p)/2 \leq \mu(p) \leq \log p$. Take u such that $C_3 m u^{2-1/\kappa} = \mu(p)$; then

$$u + p\frac{\exp(-C_3 m u^{2-1/\kappa})}{C_3 m u^{1-1/\kappa}} \leq 2\left(\frac{\mu(p)}{C_3 m} \right)^{\frac{\kappa}{2\kappa-1}} \leq 2\left(\frac{\log p}{C_3 m} \right)^{\frac{\kappa}{2\kappa-1}} .$$

Now take v such that $C_3 m v = \mu(p)$ to obtain

$$v + p\frac{\exp(-C_3 m v)}{C_3 m} \leq \frac{2\mu(p)}{C_3 m} \leq \frac{2\log p}{C_3 m} \ .$$

Then by plugging these values of u and v in Equation (3.15), we obtain the result.

■

Proof of Lemma 3.4.2. As in the previous proof, we have

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
&\leq \sum_{Q\in\mathcal{Q}} \mathbb{P}\Big[\mathbb{E}Q(Z) - \frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \geq \frac{\delta + a\mathbb{E}Q(Z)}{1+a}\Big]
\end{aligned}$$

for any $\delta > 0$. Now we apply Lemma 1.7.3 to the random variables $Q(Z), Q(Z_1), \ldots, Q(Z_m)$ (using the exponential moment bound) to obtain

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
&\leq \sum_{Q\in\mathcal{Q}} \exp\left(-\frac{m\left(\frac{\delta + a\mathbb{E}Q(Z)}{1+a}\right)^2}{2\left(v_Q + 2C_1\frac{\delta + a\mathbb{E}Q(Z)}{1+a}\right)}\right) \ .
\end{aligned}$$

After applying the margin condition, we obtain upper bounds

$$\begin{aligned}
&\sum_{Q\in\mathcal{Q}} \exp\left(-\frac{m(\delta + a\mathbb{E}Q(Z))^2}{2C_0^2(1+a)^2(\mathbb{E}Q(Z))^{1/\kappa} + 4C_1(1+a)(\delta + a\mathbb{E}Q(Z))}\right) \\
\leq &\sum_{Q\in\mathcal{Q}} \exp\left(-C_2 m \cdot \left(\frac{(\delta + a\mathbb{E}Q(Z))^2}{(\mathbb{E}Q(Z))^{1/\kappa}} \wedge (\delta + a\mathbb{E}Q(Z))\right)\right),
\end{aligned}$$

where we use the constant

$$C_2 := \frac{1}{2(1+a)(C_0(1+a) \vee 2C_1)} \ .$$

As in the previous proof, we have the inequality

$$C_2 \cdot \left(\frac{(\delta + a\mathbb{E}Q(Z))^2}{(\mathbb{E}Q(Z))^{1/\kappa}} \wedge (\delta + a\mathbb{E}Q(Z)) \right) \geq C_3(\delta^{2-1/\kappa} \wedge \delta)$$

with the constant

$$C_3 := \frac{C_2}{(C_1)^{1/\kappa}} \ .$$

So as above,

$$\mathbb{P}\Big[\max_{Q \in \mathcal{Q}} \Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \Big) \geq \delta \Big] \leq p \exp(-C_3 m(\delta^{2-1/\kappa} \wedge \delta))$$

for any $\delta > 0$ (albeit with slightly different constants), and thus the statement of the lemma follows. ∎

Proof of Theorem 3.4.1. Using the fundamental inequality in Lemma 3.3.1, all we have to show is the empirical process upper bound

$$\mathbb{E}_{D^n} \quad \max_{j=1,\ldots,p} \Bigg[(P - (1+a)P_{n_C}) \Big(\quad \gamma(\cdot, \hat{f}_j^{(n_V)}(D^{n_V})) \\ -\gamma(\cdot, f_0) \Big) \Bigg]$$

$$\leq c \Big(\frac{\log p}{n_C} \Big)^{\frac{\kappa}{2\kappa - 1}} \ . \tag{3.16}$$

Splitting up the expectation $\mathbb{E}_{D^n}$ into $\mathbb{E}_{D^{n_V}} \mathbb{E}_{D^{n_C}}$, we then condition on D^{n_V}. We define

$$Q_j(z) := \gamma(z, \hat{f}_j^{(n_V)}(D^{n_V})) - \gamma(z, \hat{f}_0)$$

for all $j = 1, \ldots, p$ and each data point z. Note that the excess risk $\mathcal{E}(\hat{f}_j(D^{n_V})$ is just the expectation of $Q_j(Z)$.

With this set $\mathcal{Q} = \{Q_1, Q_2, \ldots, \}$, Assumption (A1) contains exactly the conditions for Lemma 3.4.1, which under this assumption then gives us the final result.

Now let Assumption (A2) hold, and define $v_{Q_j} := d^2(\hat{f}_j(D^{n_V}), f_0)$ for all $j = 1, \ldots, p$. Then the exponential moment condition (Part 1 of (A2)) can be applied to $f_1 = \hat{f}_j(D^{n_V})$ and $f_2 = f_0$ to yield the corresponding condition as required by Lemma 3.4.2. This lemma then completes the proof. ∎

Proof of Lemma 3.6.1. Let $\beta \in \mathbb{R}^p$ be given, and assume that it satisfies the conditions of this lemma. We want to derive a bound on the ψ_1-Orlicz norm of the excess loss $\tilde{\gamma}(X,Y) := (Y - \beta^T X)^2 - (Y - \beta_*^T X)^2$. For this we need the bounds on expectation given by the following sub-lemma:

Lemma 3.9.1. *For any constant $x \in \mathbb{R}$ and $c > 0$, we have*

$$E\left[\exp\left(\frac{x(x-2\sigma\varepsilon)}{c}\right)\right] = \exp\left(x^2 \cdot \left(\frac{1}{c} + \frac{2\sigma^2}{c^2}\right)\right) \ .$$

Proof.

$$\begin{aligned} E\left[\exp\left(\frac{x(x-2\sigma\varepsilon)}{c}\right)\right] &= \exp\left(\frac{x^2}{c}\right) \cdot E\left[\exp\left(-\frac{2\sigma x\varepsilon}{c}\right)\right] \\ &= \exp\left(\frac{x^2}{c}\right) \cdot \exp\left(-\left(\frac{2\sigma x}{c}\right)^2\right) \\ &= \exp\left(x^2 \cdot \left(\frac{1}{c} + \frac{2\sigma^2}{c^2}\right)\right) \ . \end{aligned}$$

■

Rewriting $\tilde{\gamma}(X,Y)$ as $\langle X, \beta - \beta_* \rangle \cdot (\langle X, \beta - \beta_* \rangle - 2\sigma\varepsilon)$, we can condition on X and apply the lemma for $x := \langle X, \beta - \beta_* \rangle$ to obtain the upper bound

$$E\left[\exp\left(\frac{\tilde{\gamma}(X,Y)}{C}\right)\right] \leq E\left[\exp\left(\langle X, \beta - \beta_* \rangle^2 \cdot \left(\frac{1}{C} + \frac{2\sigma^2}{C^2}\right)\right)\right] \ .$$

Now we can set $C := K_1\sqrt{E\left[\tilde{\gamma}\right]} = K_1\sqrt{E\left[\langle X, \beta - \beta_* \rangle^2\right]}$ for an as yet undetermined constant $K_1 > 0$ and apply the loss bound $\mathbb{E}\langle X, \beta -$

$\beta^*\rangle^2 \leq K_0$, which gives us

$$\begin{aligned} & E\left[\exp\left(\frac{\tilde{\gamma}(X,Y)}{K_1\sqrt{E\left[\langle X,\beta-\beta_*\rangle^2\right]}}\right)\right] \\ \leq\ & \frac{1}{2}E\left[\exp\left(\frac{2\sqrt{K_0}\langle X,\beta-\beta_*\rangle^2}{K_1 E\left[\langle X,\beta-\beta_*\rangle^2\right]}\right)\right] \\ & +\frac{1}{2}E\left[\exp\left(\frac{4\sigma^2\langle X,\beta-\beta_*\rangle^2}{K_1^2 E\left[\langle X,\beta-\beta_*\rangle^2\right]}\right)\right] . \end{aligned}$$

The first summand here is bounded by 1 if $K_1/2\sqrt{K_0} \geq K_2$, and the second summand is bounded by 1 if $K_1^2/4\sigma^2 \geq K_2$. Thus their sum is at most 2 when we choose $K_1 := \max(2K_2\sqrt{K_0}, 2\sigma\sqrt{K_2})$. As

$$\begin{aligned} & E\left[\exp\left(\frac{|\tilde{\gamma}(X,Y)|}{K_1\sqrt{E\left[\langle X,\beta-\beta_*\rangle^2\right]}}\right)\right] \\ \leq\ & 2E\left[\exp\left(\frac{\tilde{\gamma}(X,Y)}{K_1\sqrt{E\left[\langle X,\beta-\beta_*\rangle^2\right]}}\right)\right] , \end{aligned}$$

we deduce that

$$E\left[\exp\left(\frac{|\tilde{\gamma}(X,Y)|}{K_1\sqrt{E\left[\langle X,\beta-\beta_*\rangle^2\right]}}\right)\right] \leq 4$$

for the chosen value of K_1. Finally, Lemma 1.6.2 tells us that

$$\|\tilde{\gamma}(X,Y)\|_{\psi_1} \leq 5K_1\sqrt{E\left[\tilde{\gamma}\right]} ,$$

which concludes the proof. ■

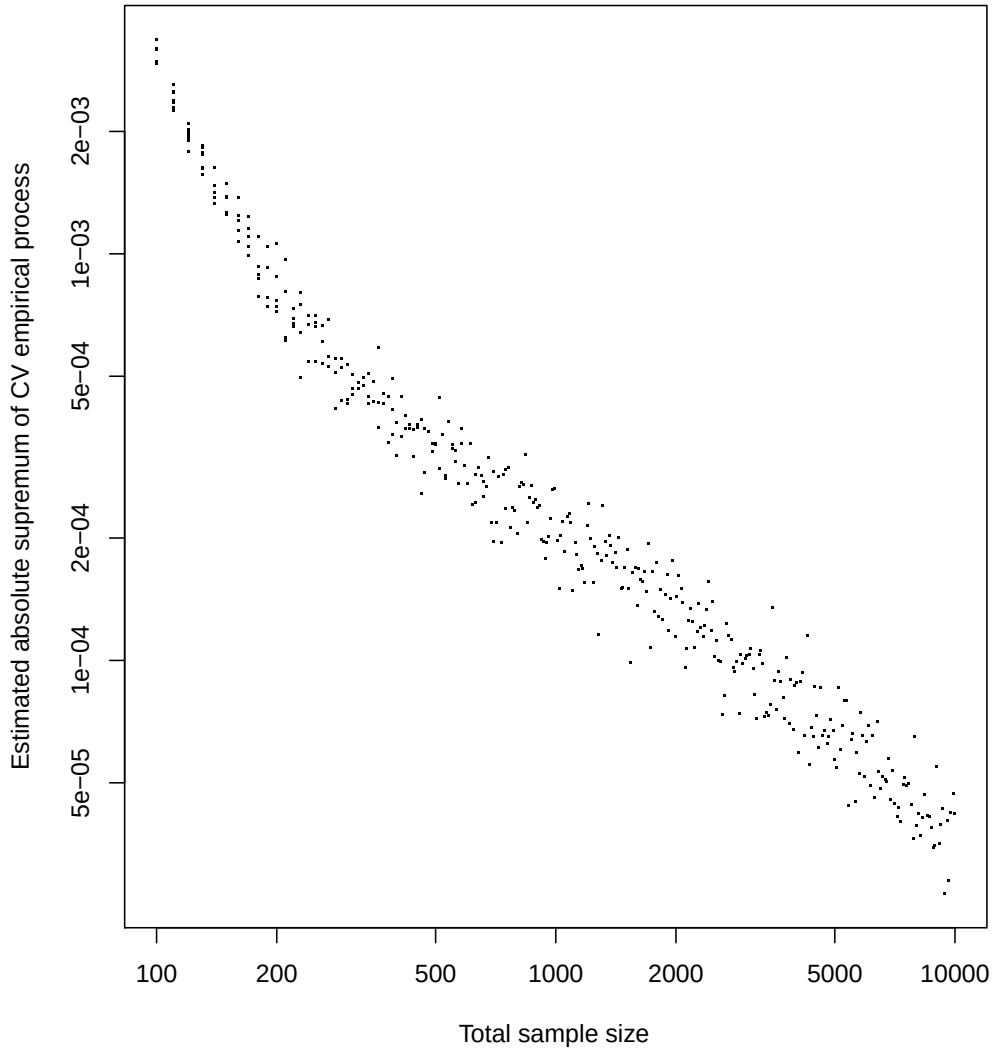

Figure 3.1: *CVA empirical process for an Elastic Net example.*

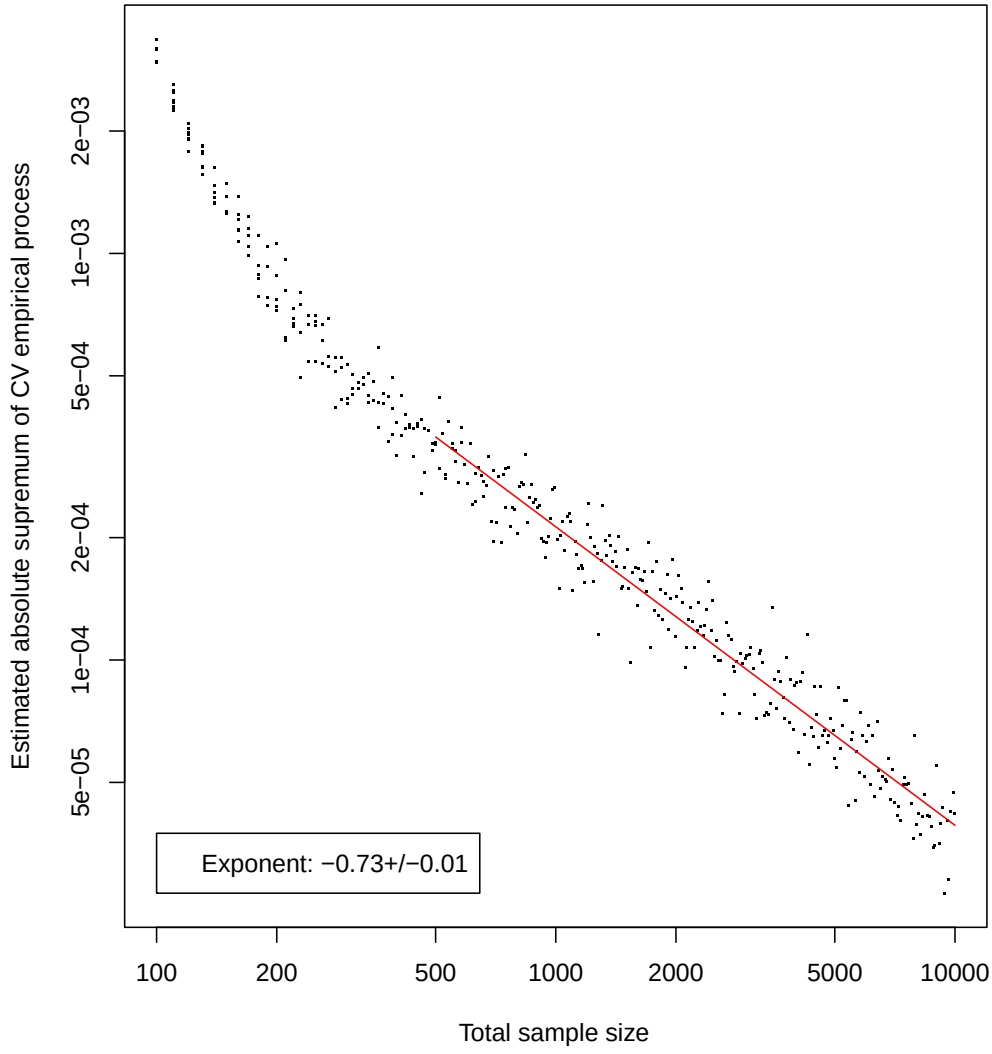

Figure 3.2: *CVA empirical process for an Elastic Net example. Large-sample trend line included.*

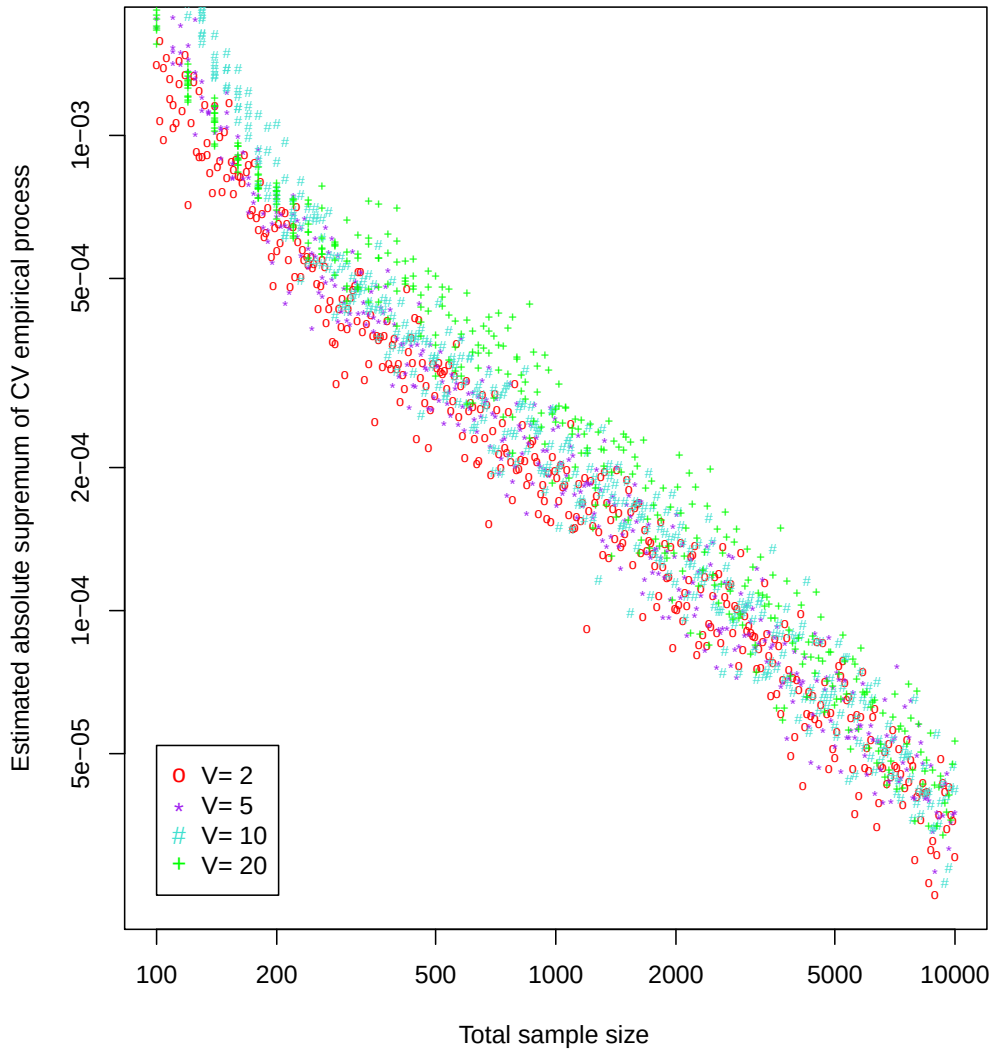

Figure 3.3: *Samples of the supremum of the cross-validation empirical process for different numbers of splits*

Appendix A

Simulations

In this section, we briefly describe the simulations which were implemented to check empirical process conditions (B1)-(B3) of Chapter 3 in situations where theoretical results are not readily available. We performed these simulations in the specific example of linear regression, and display the results for Condition (B2). As they do not always uphold the full rate n^{-1} we would like to see in our oracle inequalities for regression, we present them in an appendix rather than the main text.

In the simulations verifying Condition (B2), we simulated the cross-validation empirical process term

$$\left| \mathcal{E}\left(G_j\left(F_{D^n}\right)\right) - \frac{1}{V}\sum_{k=1}^{V} \mathcal{E}\left(G_j\left(F_{D_{(C_k)}}\right)\right) \right|_{j\in\{1,\ldots,p\}}$$

described in Section 3.7.

A.1 Model

For the simulation runs given here, we start out from a base model and look at the development of the CVA empirical process for this model and other ones that differ from it in exactly one way. The distribution

sampled from is the Gaussian linear model

$$\begin{gathered} Y = X\beta_0 + \varepsilon \ , \ X = (X^{(1)}, \dots X^{(d)}) \ , \\ X^{(\ell)} \text{ i.i.d.} \sim \text{Unif}[0,1] \ , \ell \in \{1, \dots, d\} \ , \\ \beta_0 = (\underbrace{1/\sqrt{s}, \dots, 1/\sqrt{s}}_{s \text{ dimensions}}, \underbrace{0, \dots, 0}_{d-s \text{ dimensions}}) \in \mathbb{R}^d \ , \\ \varepsilon \sim \mathcal{N}(0, \sigma) \ . \end{gathered}$$

Our estimators of the model parameter β_0, amongst which we perform model selection, are Elastic Net estimators as described by Zou and Hastie [57], Sections 2.2 and 3.2. For parameters $\lambda_1, \lambda_2 \geq 0$ and data pairs $(X_1, Y_1), \dots (X_n, Y_n)$ that are sampled i.i.d. to (X, Y), the Elastic Net estimator is defined as

$$\hat{\beta}_{EN} := (1 + \lambda_2)\hat{\beta}_{\text{naïve}} \ ,$$

where

$$\hat{\beta}_{\text{naive}} := \arg\min_{\beta \in \mathbb{R}^d} \left[\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda_2 |\beta|_2 + \lambda_1 |\beta|_1 \right]$$

is the naïve Elastic Net estimator. As in [57], we compute $\hat{\beta}_{EN}$ as

$$\hat{\beta}_{EN} := (1 + \lambda_2)\hat{\beta}^* \ ,$$

where

$$\hat{\beta}^* := \arg\min_{\beta \in \mathbb{R}^d} \left[(Y^* - X^*\beta)^2 + \frac{\lambda_1}{\sqrt{1 + \lambda_2}} |\beta|_1 \right]$$

is the naïve Elastic Net estimator computed using augmented data

$$X^* := \begin{pmatrix} X_1^{(1)} & \cdots & X_1^{(d)} \\ \vdots & & \vdots \\ X_n^{(1)} & \cdots & X_n^{(d)} \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \ , \quad Y^* := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n+d} \ .$$

Model selection was performed from amongst a family of Elastic Net estimators with a grid of possible values of the normalised penalty

parameters λ_1 (for the ℓ_1-penalty) and λ_2 (for the ℓ_2-penalty). In all of the models, the grid of penalty parameters consisted of $\lambda_1 \in \Lambda_1 = \{0, 0.1, \ldots, 1\}$ and $\lambda_2 \in \Lambda_2 = \{0, 0.5, 1\}$. The grid $\Lambda_1 \times \Lambda_2$ is an approximation to the true (metric) parameter space $[0, 1]^2$, and as such it is to expected that the logarithm of the grid size appears in the upper bound of the CV empirical process increments, at a rate matching that of the inverse sample size $1/n$. We thus refrained from varying the grid size, but instead only investigated the convergence rates of the increment upper bounds as dependent on n.

Further parameters used in the base model were:

Dimension of covariates	d	20
No. of covariates in true model	s	2
Error variance	σ^2	1/12
Type of aggregation used		Model selection by cross-validation
No. of splits used	V	10

The error variance was chosen so as to ensure a signal-to-noise ratio of 1 at all times. The sample sizes regarded started out at 10 to 10000 in the base model, with 400 sample sizes chosen with even logarithmic spacing, but in several models, the range of sample sizes was changed for reasons of computability. In this model, the exact risk of a trained estimator could be computed, leaving only one step of random sampling (namely that of the training data).

A.2 Results

The simulation results were as follows:

Base model: Figure 3.2 shows the simulated suprema of the CV empirical process under the base model. As the behaviour for small sample sizes is non-linear, the slope was estimated from different cutoff points onwards. We can see that linear behaviour is found, although not at the target rate of -1.0. Thus this simulation only indicates the existence of slightly weaker oracle inequalities for the re-trained cross-validation procedure in the regression model.

Increasing the number of covariates: The base model was computed using a constant dimension d for the covariates. If this dimension is varied as $\sqrt{n}$ (while the sparsity, i.e. the number s of covariates actually involved in the true model, is increased as $n^{1/4}$), then the increments of the CV empirical process decrease steeply, as seen in Figure A.1. For a logarithmic increase in sparsity ($s = \log(n)$), we have a similar rate of decrease, cf. Figure A.2.

Changing the number of splits: Our model selection has one key parameter, namely V, the number of splits used in cross-validation. Our base model uses $V = 10$, which is a very typical value. Changing V has little impact on the outcome in terms of the decay rate of suprema of the CV empirical process. The results of such changes are shown in Figure A.3 for $V = 2$, Figure A.4 for $V = 5$ and Figure A.5 for $V = 20$.

Changing the signal-to-noise ratio: By increasing the error variance σ^2, we make the underlying linear estimation problem more difficult. We would expect the candidate models to perform very similarly when the signal-to-noise ratio is extremely low or extremely high, as linear regression is then very easy or practically impossible, respectively. However, in simulation runs with σ reducing by a factor of 5, the CV empirical process performs worse than for the base model (see Figure A.6), and vice versa when σ is multiplied by 5.

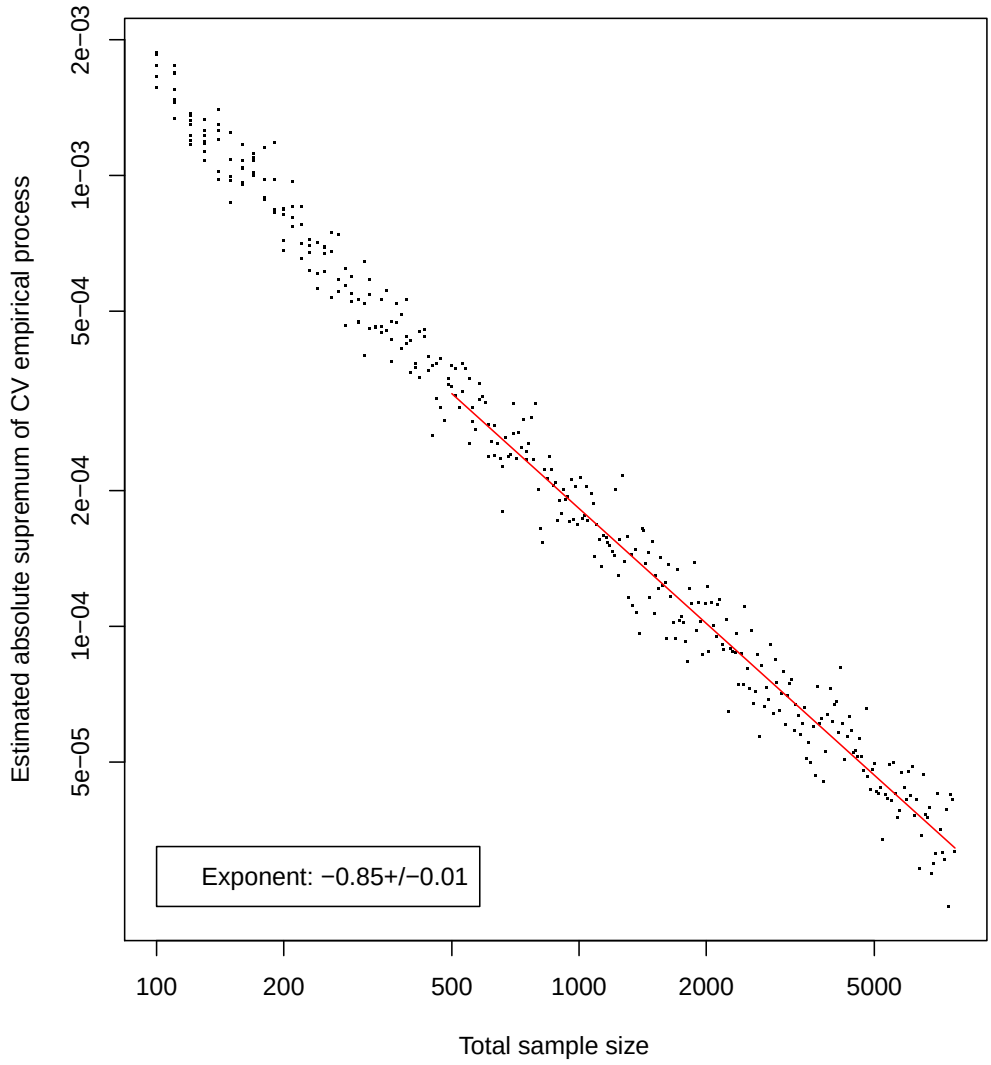

Figure A.1: *CVA empirical process when increasing dimension of covariates linearly with sample size. Covariate dimension increasing as* $n^{1/2}$*, sparsity increasing as* $n^{1/4}$*.*

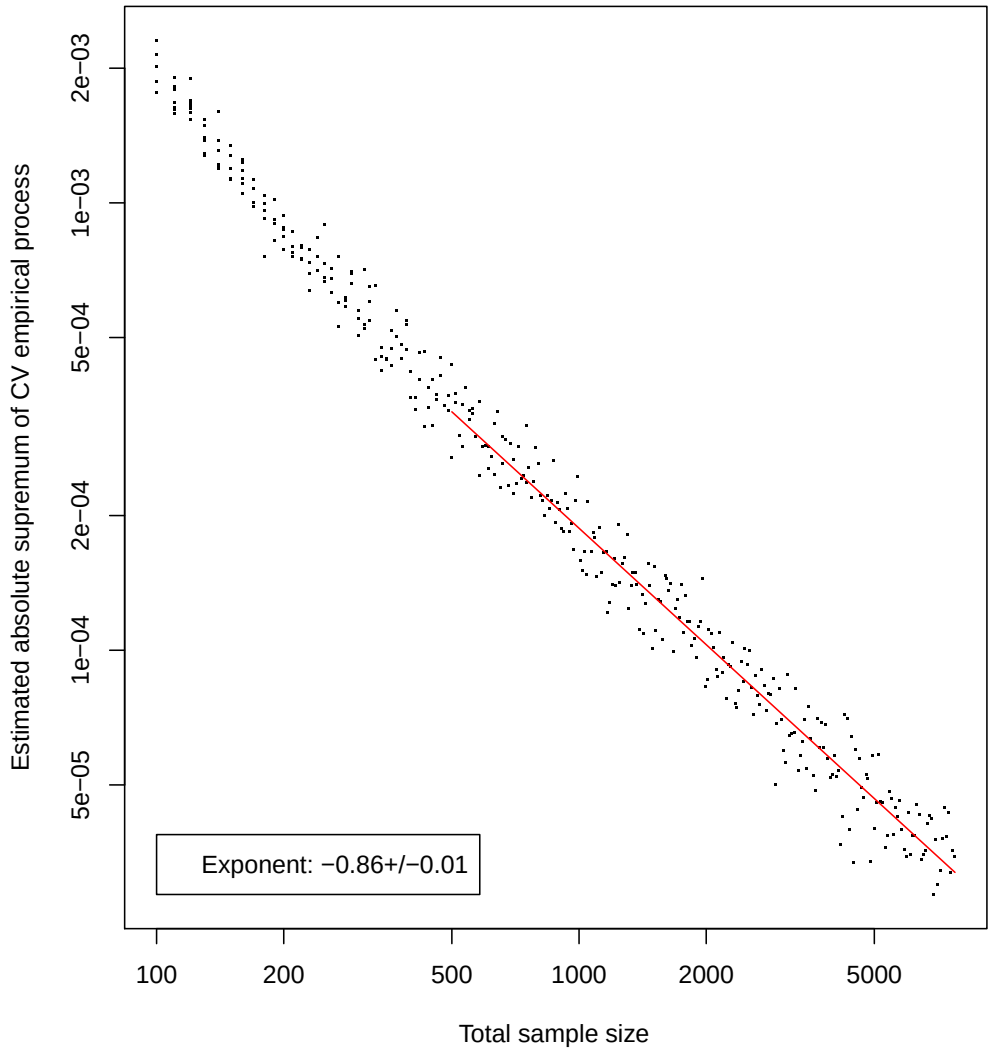

Figure A.2: *CVA empirical process when increasing dimension of covariates linearly with sample size. Covariate dimension increasing as* $n^{1/2}$*, sparsity increasing as* $\log(n)$*.*

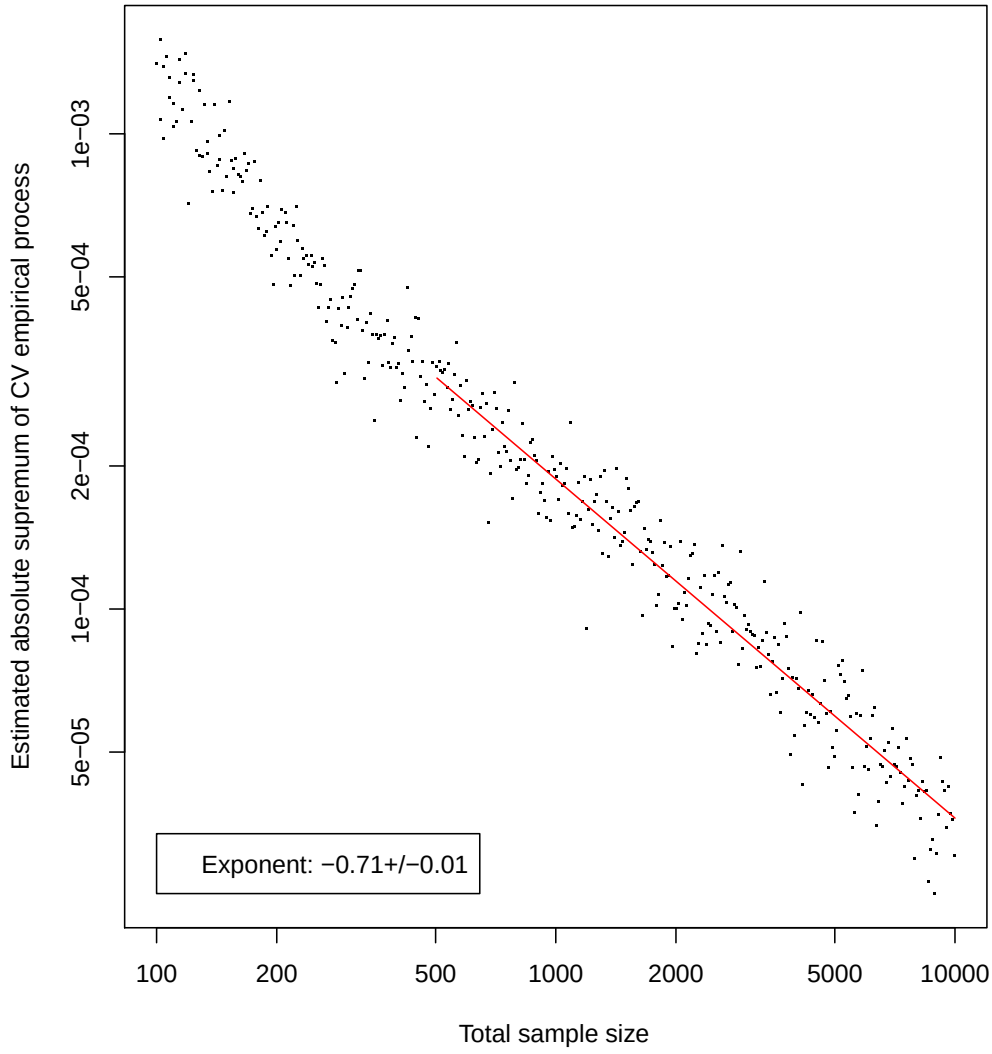

Figure A.3: *CVA empirical process for 2-fold cross-validation.*

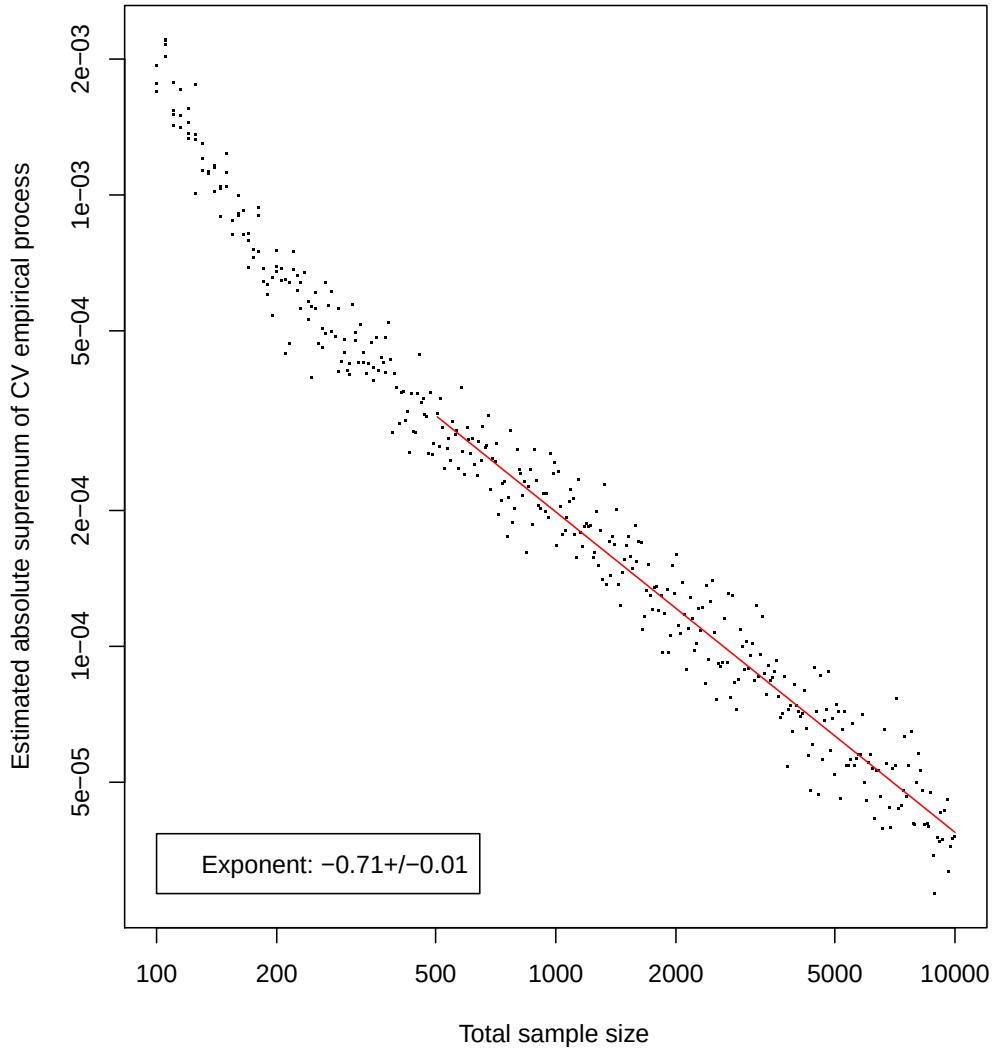

Figure A.4: *CVA empirical process for 5-fold cross-validation.*

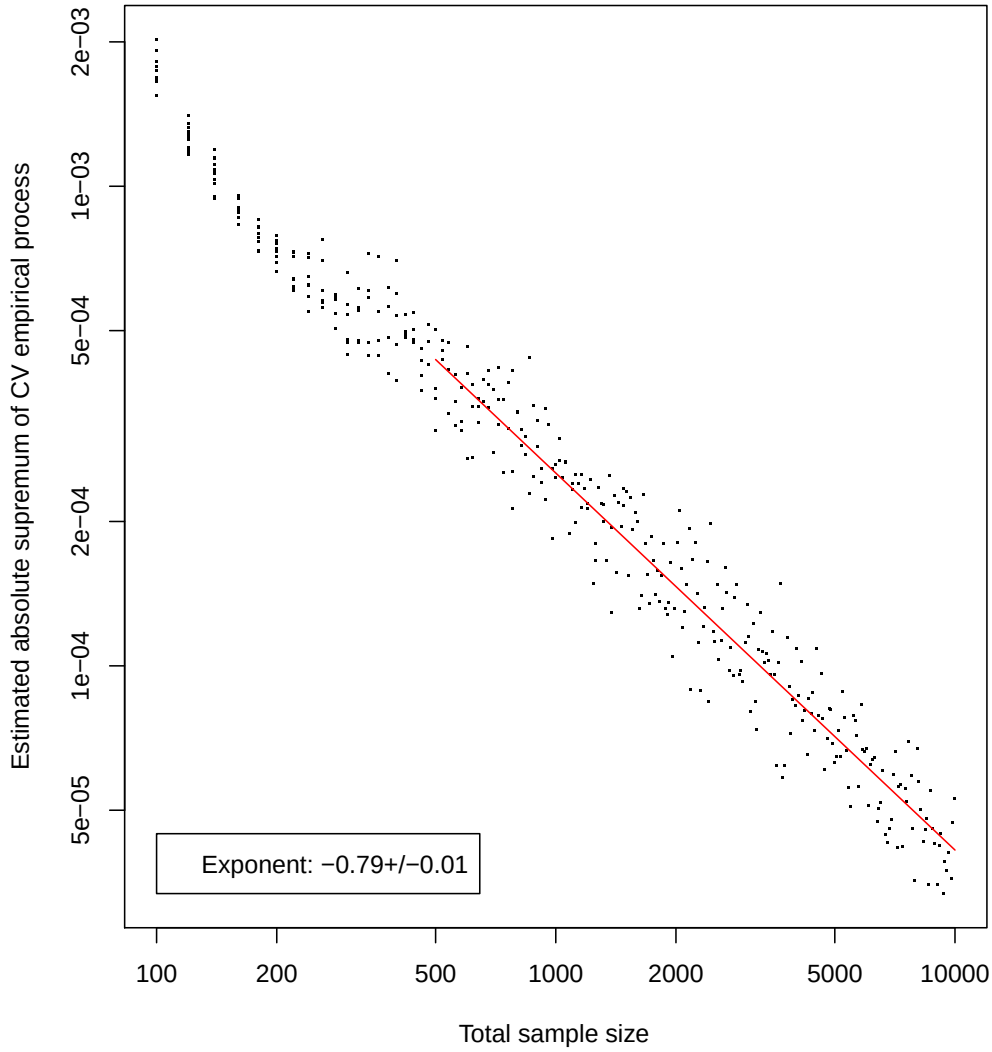

Figure A.5: *CVA empirical process for 20-fold cross-validation.*

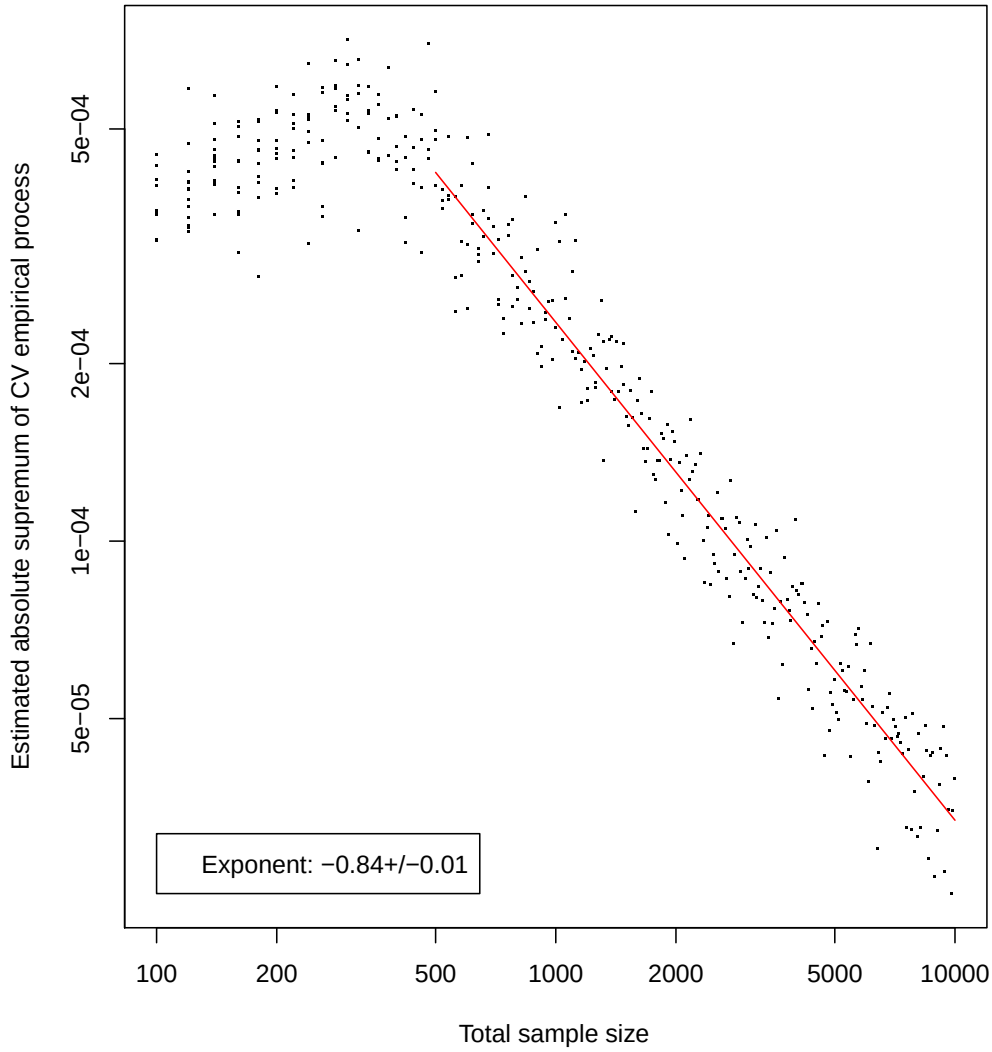

Figure A.6: *CVA empirical process for higher signal-to-noise ratio.*

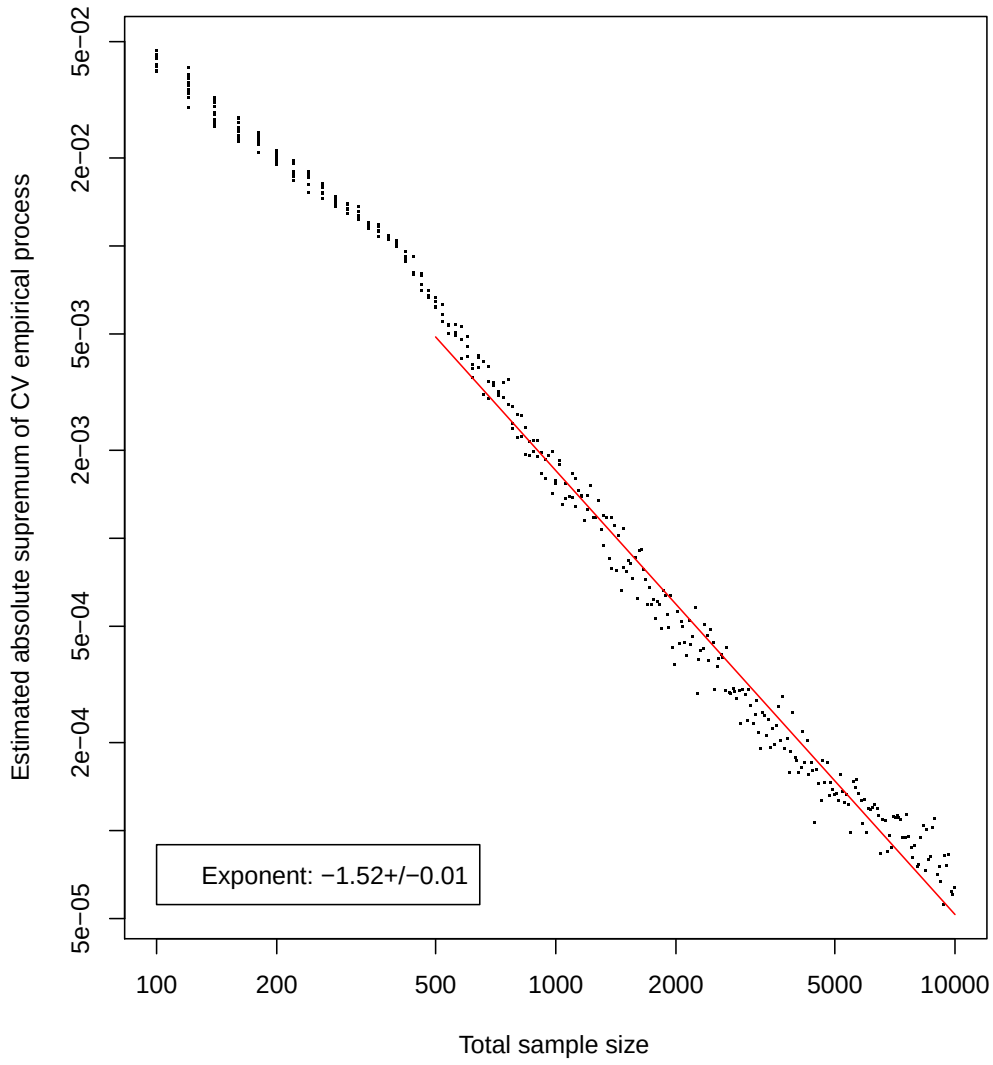

Figure A.7: *CVA empirical process for lower signal-to-noise ratio.*

Bibliography

[1] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. ArXiv preprint 0907.4728v1, July 2009.

[2] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, 296(23):1021–1024, 1983.

[3] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. Preprint no. 805, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, March 2003.

[4] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. Technical Report 06-20, CERTIS, ENPC, 2006.

[5] J.-Y. Audibert. A randomized online learning algorithm for better variance control. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 392–407, 2006.

[6] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 2007.

[7] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

[8] P. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[9] S. Bernstein. Über eine Modifikation einer Ungleichung von Tschebycheff und uber die Abweichung der Laplaceschen Formel. *Charkov Ann. Sci.*, 1:38–49, 1924.

[10] P. J. Bickel and K. A. Doksum. *Mathematical Statistics; Basic Ideas and Selected Topics*. Holden-Day Inc., Oakland, 1977.

[11] L. Birgé. Statistical estimation with model selection. Brouwer Lecture 2005. Available at `http://arxiv.org/pdf/math/0605187`.

[12] L. Birgé and P. Massart. From model selection to adaptive estimation. Festschrift for Lucien Le Cam, 1997.

[13] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Acad. Sci. Paris*, 334(6):495–500, 2002.

[14] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, August 2007.

[15] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[16] C. Chesneau and G. Lecué. Adapting to Unknown Smoothness by Aggregation of Thresholded Wavelet Estimators. ArXiv preprint math.ST/0612546, 2006.

[17] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[18] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[19] R. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929, Dec. 1978.

[20] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–989, Nov. 1984.

[21] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, 2002.

[22] L. Gyorfi and M. Wegkamp. Quantization for Nonparametric Regression. *Information Theory, IEEE Transactions on*, 54(2):867–874, 2008.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[24] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

[25] A. Juditsky, P. Rigollet, and A. Tsybakov. Learning by mirror averaging, 2008.

[26] A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96, 2005.

[27] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. 2004 IMS Medallion Lecture, July 2005.

[28] S. Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45–55, 1931.

[29] G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. ArXiv preprint math.ST/0603526, March 2006.

[30] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

[31] G. Lecué. Suboptimality of penalized empirical risk minimization in classification. In *Proceedings of the 20th Annual Conference On Learning Theory*, volume 4539 of *Lecture Notes in Artificial Intelligence*, pages 142–156. Springer, Heidelberg, 2007.

[32] G. Lecué and C. Mitchell. Adaptivity and aggregation by cross-validation. In preparation.

[33] W. Lee, P. Bartlett, and R. Williamson. The importance of convexity in learning with squared loss. *Information Theory, IEEE Transactions on*, 44(5):1974–1980, 1998.

[34] K. Li. Asymptotic Optimality for C_p,C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–-975, 1987.

[35] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, December 1999.

[36] S. Mendelson. Lower bounds for the empirical minimization algorithm, 2008.

[37] S. Mendelson. On weakly bounded empirical processes. *Mathematische Annalen*, 340(2):293–314, 2008.

[38] C. Mitchell and S. van de Geer. General oracle inequalities for model selection. *Electronic Journal of Statistics*, 3:176–204, 2009.

[39] D. Pollard. Asymptotics via empirical processes. *Statistical Science*, 4(4):341–354, Nov. 1989.

[40] D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. IMS, 1990.

[41] P. Rigollet. *Inégalités d'oracle, agrégation et adaptation*. PhD thesis, Université Paris-VI, 2006.

[42] J. Shao. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[43] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–147, 1974.

[44] M. Stone. Cross-validation: a review. *Math. Operationsforsch. Statist. Ser. Statist.*, 9(1):127–139, 1978.

[45] A. Tsybakov. Optimal rates of aggregation. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: Proceedings*, 2003.

[46] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

[47] J. Tukey and F. Mosteller. Data analysis, including statistics. *The handbook of social psychology*, 1:133–160, 1968.

[48] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[49] S. van de Geer. Oracle inequalities and regularization. In: Lecture Notes EMS Summer School: Empirical Processes: Theory and Statistical Applications, 2004.

[50] A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[51] A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24:351–371, 2006.

[52] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[53] V. Vapnik and A. Chervonenkis. Theory of Pattern Recognition, 1974.

[54] P. Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications*, 5(3):302–305, 1960.

[55] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10 (1):25–47, 2004.

[56] Y. Yang. How powerful can any regression learning procedure be? In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[57] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.