Probab. Theory Relat. Fields (2018) 171:459–502
https://doi.org/10.1007/s00440-017-0784-y

Learning without concentration for general loss functions

Shahar Mendelson[1,2]

Received: 6 April 2016 / Revised: 8 May 2017 / Published online: 12 June 2017

Abstract We study the performance of empirical risk minimization in prediction and estimation problems that are carried out in a convex class and relative to a sufficiently smooth convex loss function. The framework is based on the small-ball method and thus is suited for heavy-tailed problems. Moreover, among its outcomes is that a well-chosen loss, calibrated to fit the noise level of the problem, negates some of the ill-effects of outliers and boosts the confidence level—leading to a gaussian like behaviour even when the target random variable is heavy-tailed.

Mathematics Subject Classification 62G99 · 60G25 · 68T05

1 Introduction

Prediction and estimation problems play a major role in modern mathematical statistics. The aim is to approximate, in one way or another, an unknown random variable Y by a function from a given class F, defined on a probability space (Ω, μ). The given data is a random sample $(X_i, Y_i)_{i=1}^N$, distributed according to the N-product of the joint distribution of μ and Y, endowed on the product space $(\Omega \times \mathbb{R})^N$.

The notion of approximation may change from problem to problem. It is reflected by different choices of *loss functions*, which put a price tag on predicting $f(X)$ instead of Y. Although it is not the most general form possible, we also assume throughout

Partially supported by ISF Grant 707/14.

✉ Shahar Mendelson
shahar@tx.technion.ac.il

[1] Department of Mathematics, Technion — Israel Institute of Technology, Haifa, Israel

[2] Mathematical Sciences Institute, The Australian National University, Canberra, Australia

this article that if ℓ is the loss function, the cost of predicting $f(X)$ instead of Y is $\ell(f(X)-Y)$:

Definition 1.1 A loss is a real-valued function that is even, increasing in $\mathbb{R}_+$ and convex, and vanishes at 0. We will assume that it is sufficiently smooth—for example, that it has a second derivative, except, perhaps at $\pm x_0$ for some fixed x_0—although, as will be clear from what follows, this assumption can be relaxed further.

Once the loss is selected, one can define the best element in the class, namely, a function in F that minimizes the average loss, or *risk*, $\mathbb{E}\ell(f(X)-Y)$ (with the obvious underlying assumption that the minimizer exists). We denote that minimizer by f^*.

Next, one may choose a procedure that uses the data $(X_i, Y_i)_{i=1}^N$ to produce a (random) function $\hat{f} \in F$. The effectiveness of $\hat{f}$ may be measured in several ways, and the two we focus on here lead to the notions of *prediction* and *estimation*.

Problem 1.2 *Given a procedure $\hat{f}$, find the 'smallest' functions $\mathcal{E}_p$ and $\mathcal{E}_e$ possible for which the following holds. If $F \subset L_2(\mu)$ is a class of functions and Y is the unknown target, then with probability at least $1-\delta$ over samples $(X_i, Y_i)_{i=1}^N$,*

$$\mathbb{E}\left(\ell(\hat{f}(X)-Y)\middle|(X_i,Y_i)_{i=1}^N\right) \le \inf_{f\in F} \mathbb{E}\ell(f(X)-Y) + \mathcal{E}_p.$$

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

$$\left\|\hat{f}-f^*\right\|_{L_2}^2 = \mathbb{E}\left((\hat{f}-f^*)^2(X)\middle|(X_i,Y_i)_{i=1}^N\right) \le \mathcal{E}_e.$$

The functions $\mathcal{E}_p$ and $\mathcal{E}_e$ may depend on the structure of F, the sample size N, the confidence level δ, some 'global' properties of Y (e.g., its L_q norm), etc.

The prediction error $\mathcal{E}_p$ measures the 'predictive capabilities' of $\hat{f}$, specifically whether $\hat{f}$ is likely to be almost as effective as the best possible in the class—the latter being f^*. The estimation error $\mathcal{E}_e$ measures the distance between $\hat{f}$ and f^*, with respect to the underlying $L_2(\mu)$ metric.

Literature devoted to the study of prediction and estimation is extensive and goes well beyond what can be reasonably surveyed here. We refer the reader to the manuscripts [2,4,6,9,15,24,25] as possible starting points for information on the history of Problem 1.2, as well as for more recent progress.

The procedure we focus on here is empirical risk minimization (ERM), in which $\hat{f}$ is selected to be a function in F that minimizes the empirical risk

$$P_N\ell_f \equiv \frac{1}{N}\sum_{i=1}^N \ell(f(X_i)-Y_i);$$

here, and throughout the article, P_N denotes the empirical mean associated with the random sample $(X_i, Y_i)_{i=1}^N$.

Since it is impossible to obtain nontrivial information on the performance of *any* procedure, including ERM, without imposing some assumptions on the class F, the target Y and the loss ℓ, one has to select a framework that, on the one hand, is general enough to include natural problems that one would like to study, but on the other, still allows one to derive significant results on prediction and estimation. Unfortunately, some of the assumptions that are commonly used in literature are highly restrictive, though seemingly benign. And among the more harmful assumptions are that the loss is a Lipschitz function and that functions in F and Y are uniformly bounded.

The origin of these assumptions is technical: they are an outcome of the 'classical' method of analysis used to tackle Problem 1.2. The method itself is based on tools from empirical processes theory, most notably, on contraction and concentration arguments that are simply false without imposing the right assumptions on the class, the target and the loss. However, these assumptions leave a large number of natural problems out of reach.

To explain why concentration and contraction arguments are so appealing, let us outline the standard method of analyzing data driven procedures like ERM.

The basic underlying assumption behind such procedures, and in particular, behind ERM, is that sampling mimics reality. Since one's goal is to identify the function f^* which minimizes in F the functional $f \to \mathbb{E}\ell(f(X) - Y)$, a natural course of action is to compare empirical means of the loss functional to the actual means in the hope that an empirical minimizer will be close to the true minimizer.

To that end, one may consider the excess loss functional associated with $f \in F$

$$\mathcal{L}_f(X, Y) = \ell(f(X) - Y) - \ell(f^*(X) - Y),$$

observe that for every $f \in F$, $\mathbb{E}\mathcal{L}_f \geq 0$ and that if f^* is unique, equality is achieved only by f^*. Also, since $\mathcal{L}_{f^*} = 0$, it is evident that the empirical minimizer $\hat{f}$ satisfies that $P_N\mathcal{L}_{\hat{f}} \leq 0$; thus, for every sample, the empirical minimizer belongs to the random set

$$\left\{f \in F : P_N\mathcal{L}_f \leq 0\right\}. \tag{1.1}$$

The key point in the analysis of ERM is that the random set of potential minimizers consists of functions for which sampling behaves in an a-typical manner: $P_N\mathcal{L}_f \leq 0$ while $\mathbb{E}\mathcal{L}_f > 0$. The hope is that one may identify the set by exploiting the discrepancy between the 'empirical' and 'actual' behaviour of means. For example, a solution to the prediction problem follows if the set (1.1) consists only of functions with 'predictive capabilities' that are close to the optimal in F, while the estimation problem may be resolved if (1.1) consists only of functions that are close to f^* with respect to the L_2 distance.

What makes the nature of the set $\{f : P_N\mathcal{L}_f \leq 0\}$ rather elusive is not only the fact that it is random, but also that one has no real knowledge of the functions $\ell_f = \ell(f(X) - Y)$ and $\mathcal{L}_f = \ell_f - \ell_{f^*}$: the two have unknown components—the target Y and the true minimizer f^*.

Concentration and contraction help in identifying the set (1.1). By applying concentration results to a well-chosen subset of excess loss functions $\{\mathcal{L}_f : f \in F'\}$, one may show that there is a large subset $F' \subset F$, on which $P_N\mathcal{L}_f$ cannot be too far from $\mathbb{E}\mathcal{L}_f$ (or, for more sophisticated results, that the ratios $P_N\mathcal{L}_f/\mathbb{E}\mathcal{L}_f$ cannot be too far from 1). Since $\mathbb{E}\mathcal{L}_f > 0$ if $f \neq f^*$, this forces $f \to P_N\mathcal{L}_f$ to be positive on F' and thus $\hat{f} \in F \backslash F'$.

Naturally, concentration results come at a cost, and for estimates such as

$$\sup_{f \in F'} \left| P_N\mathcal{L}_f - \mathbb{E}\mathcal{L}_f \right| < \varepsilon \text{ or } \sup_{f \in F'} \left| \frac{P_N\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} - 1 \right| < \varepsilon \tag{1.2}$$

to hold with high probability requires strong assumptions on the random variables involved—for example, that functions in F and Y are uniformly bounded (see the books [3,11] for more details on concentration of measure phenomena).

Finally, and what is possibly the most costly step in the classical method of analysis, is contraction. The contraction argument is based on the fact that class members and the target are uniformly bounded functions and that the loss is Lipschitz on the ranges of the functions $f(X) - Y$. It implies that the supremum of the empirical process indexed by the (unknown!) excess loss class may be controlled in terms of the supremum of an empirical process indexed by functions of the form $f - f^*$ (see, for example, [1,9,18] for more details).

One result that is based on the classical method and that utilizes the full strength of the two assumptions—that class members and the target are uniformly bounded and that the loss is Lipschitz—is Theorem 1.3 below, proved originally in [1]. It serves as a preliminary benchmark for our discussion.

Assume that F is a class of functions that are bounded by 1 and let Y be the target random variable that is also bounded by 1. Let ℓ be a Lipschitz function with constant $\|\ell\|_{\text{lip}}$ on $[-2, 2]$, which is an interval containing all the ranges of $f(X) - Y$. Assume further that f^* exists and is unique and that for every $f \in F$, $\|f - f^*\|_{L_2}^2 \leq B\mathbb{E}\mathcal{L}_f$, which is the significant part of the so-called Bernstein condition (see, e.g., [13,16,17]).

A standard example in which all these conditions hold is when F is a closed, convex class consisting of functions into $[-1, 1]$, Y also maps into $[-1, 1]$, and $\ell(t) = t^2$. In that case it is straightforward to show that $B = 1$ and $\|\ell\|_{\text{lip}} = 4$.

Let D_{f^*} be the $L_2(\mu)$ ball of radius 1, centred at f^*. Thus, $\{f \in F : \|f - f^*\|_{L_2} \leq r\} = F \cap rD_{f^*}$. For every $r > 0$, let

$$k_N(r) = \sup_{f \in F \cap rD_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right|, \tag{1.3}$$

and

$$\bar{k}_N(r) = \mathbb{E} \sup_{f \in F \cap rD_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right|. \tag{1.4}$$

Here $(\varepsilon_i)_{i=1}^N$ are independent, symmetric, $\{-1, 1\}$-valued random variables that are independent of $(X_i)_{i=1}^N$ and the expectation is taken with respect to both $(X_i)_{i=1}^N$ and $(\varepsilon_i)_{i=1}^N$. Finally, set

$$k_N^*(\gamma, \delta) = \inf \left\{ r > 0 : Pr\left(k_N(r/\|\ell\|_{\text{lip}}) \leq \gamma r^2 \sqrt{N} \right) \geq 1 - \delta \right\}$$

and

$$\bar{k}_N^*(\gamma) = \inf\left\{r > 0 : \bar{k}_N(r/\|\ell\|_{\text{lip}}) \le \gamma r^2\sqrt{N}\right\}.$$

Theorem 1.3 *There exist absolute constants c_1 and c_2 for which the following holds. If F, Y and ℓ are as above, then for every $0 < \delta < 1$, with probability at least $1-\delta$*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \le c_1 \max\left\{\left(k_N^*\left(c_2\left(B\,\|\ell\|_{\text{lip}}\right)^{-1}, \delta\right)\right)^2, \frac{\|\ell\|_{\text{lip}}^2 B}{N}\right\}, \tag{1.5}$$

and

$$\mathbb{E}\mathcal{L}_{\hat{f}} \le c_1 \max\left\{\left(\bar{k}_N^*\left(c_2(B\,\|\ell\|_{\text{lip}})^{-1}\right)\right)^2, (\|\ell\|_{\text{lip}}^2 B)\frac{\log(1/\delta)}{N}\right\}. \tag{1.6}$$

Recalling that $\|f-f^*\|_{L_2}^2 \le B\mathbb{E}\mathcal{L}_f$ for every $f \in F$, analogous results hold for the estimation problem.

The proof of Theorem 1.3 and results of a similar nature may be found in [1,9,18].

Although Theorem 1.3 is one of the main benchmarks of the performance of ERM, it truly requires rather restrictive assumptions, as do all the results that are based on the concentration-contraction mechanism. For example, Theorem 1.3 cannot be used to tackle one of the most fundamental problems in Statistics—linear regression in $\mathbb{R}^n$ relative to the squared loss and with independent additive gaussian noise:

Example 1.4 Let $\ell(x) = x^2$. Given $T \subset \mathbb{R}^n$, set $F_T = \left\{\left\langle t, \cdot\right\rangle : t \in T\right\}$ to be the class of linear functionals on $\mathbb{R}^n$ associated with T. Let μ be a probability measure on $\mathbb{R}^n$ and set X to be a random vector distributed according to μ. Let W be a standard gaussian variable that is independent of X and the target is $Y = \left\langle t_0, \cdot\right\rangle + W$ for some fixed but unknown $t_0 \in T$.

Observe that

- Y is not bounded (because of the gaussian noise).
- Unless μ is supported in a bounded set in $\mathbb{R}^n$, functions in F_T are not bounded.
- The loss $\ell(x) = x^2$ satisfies a Lipschitz condition in $[-a, a]$ with a constant $2a$. Unless μ has a bounded support and Y is bounded, ℓ does not satisfy a Lipschitz condition on an interval containing the ranges of the functions $f(X) - Y$.

Each one of these observations is enough to place linear regression with independent additive gaussian noise outside the scope of Theorem 1.3, and what is equally alarming is that the same holds even if μ is the standard gaussian measure on $\mathbb{R}^n$, regardless of T, the choice of noise or even its existence.

An additional downside of Theorem 1.3 is that even in situations that do fall within its scope, resulting bounds are often less than satisfactory (see, for example, the discussion in [19]).

The suboptimal behaviour of Theorem 1.3 and, in fact, of the entire concentration-contraction mechanism, happens to be endemic: it is caused by the nature of the complexity parameter used to govern the rates $\mathcal{E}_p$ and $\mathcal{E}_e$. Indeed, when considering likely sources of error in prediction and estimation problems, two generic reasons come to mind:

- $(X_1, .., X_N)$ is merely a sample and two functions in F can coincide on that sample but still be far from one another in $L_2(\mu)$. This leads to the notion of the *version space*: a random subset of F, defined by
$$\left\{ f \in F : f(X_i) = f^*(X_i) \text{ for every } 1 \leq i \leq N \right\}$$
and which measures the way a random sample can be used to distinguish between class members. Clearly, the $L_2(\mu)$ diameter of the version space is an intrinsic property of the class F and has nothing to do with the noise[1] $\xi = f^*(X) - Y$. Standard arguments show (see, e.g. [10]) that even in noise-free problems, when $Y = f_0(X)$ for some $f_0 \in F$, it is impossible to construct a learning procedure whose error rate consistently outperforms the L_2 diameter of the version space.
- Measurements are noisy: one does not observe $f^*(X_i)$ but rather Y_i. Known results as well as common sense indicate that the 'closer' Y is to F, the better the behaviour of $\mathcal{E}_p$ and $\mathcal{E}_e$ should be. Thus, it stands to reason that $\mathcal{E}_p$ and $\mathcal{E}_e$ should depend on the 'noise level' of the problem, i.e., on a natural distance between the target and the class.

With that in mind, it is reasonable to conjecture that $\mathcal{E}_p$ and $\mathcal{E}_e$ exhibit two regimes, captured by two different complexity parameters. Firstly, a 'low noise' regime, in which the 'noise' $\xi = f^*(X) - Y$ is sufficiently close to zero in the right sense, and the behaviour of ERM is similar to its behaviour in the noise-free problem—essentially the L_2 diameter of the version space. Secondly, a 'high noise' regime, in which mistakes occur because of the way the loss affects the interaction between class members and the noise.

Theorem 1.3 yields only one regime, and that regime is governed by a single complexity parameter. This parameter does not depend on the noise $\xi = f^*(X) - Y$, except via a trivial L_∞ bound; rather, it depends on the correlation of the set $\{(f(X_i))_{i=1}^N : f \in F\}$ (the so-called random coordinate projection of F) with a generic random noise model, represented by a random point in $\{-1, 1\}^N$. Obviously, the generic noise may have nothing to do with the actual noise one faces.

The main goal of this article is to address Problem 1.2 by showing that $\mathcal{E}_p$ and $\mathcal{E}_e$ indeed have two regimes. Each one of those regimes is captured by a different parameter: firstly, an 'intrinsic parameter' that governs low-noise problems (when Y is sufficiently close to F) and depends only on the class and not on the target or on the loss; secondly, an external parameter that captures the interaction of the class with the noise and with the loss, and dominates in high-noise situations, when Y is far from F.

Moreover, a solution to Problem 1.2 has to hold without the restrictive assumptions of the concentration-contraction mechanism, namely:

- The class F need not be bounded in L_∞, but rather satisfies significantly weaker tail conditions.
- The target Y need not be bounded (in fact, $Y \in L_2$ suffices in most cases).

[1] We refer to $f^*(X) - Y$ as the noise of the problem. This name makes perfect sense when $Y = f_0(X) - W$ for a mean-zero random variable W that is independent of X, and we use the term 'noise' even when the target does not have that particular form.

- The loss function ℓ need not be Lipschitz on an interval containing the ranges of $f(X) - Y$.

The problem of ERM's estimation error relative to the squared loss was studied in [19], leading to a satisfactory solution in that case. The aim here is to explore more general loss functions, not merely for the sake of generality, but because of a real side-effect of the squared loss: the combination of its rapid growth with heavy-tailed sampling inevitably leads to outliers. Roughly put, outliers are sample points that are misleading because they capture some a-typical behaviour of the sampled object. Outliers occur more frequently when dealing with heavy-tailed functions, and have a significant impact on ERM when the loss grows quickly. It is highly desirable to find a way of removing the ill-effects of outliers, and thanks to the general theory we develop here we are able to do just that: we show that if one chooses a loss that is calibrated to fit the noise level and the intrinsic structure of the underlying class, the ill-effects of outliers caused by the 'noise' are removed.

1.1 Basic definitions and some notation

Throughout the article, absolute constants are denoted by $c_1, c_2, \ldots$; their values may change from line to line. We write $A \lesssim B$ if there is an absolute constant c_1 for which $A \leq c_1 B$, and $A \sim B$ if $c_1 A \leq B \leq c_2 A$ for absolute constants c_1 and c_2. $A \lesssim_r B$ or $A \sim_r B$ means that the constants depend on some parameter r. κ_0, κ_1,..., denote constants whose values remain unchanged.

Given a probability measure μ, set $D = B(L_2(\mu))$ to be the unit ball of $L_2(\mu)$, let rD be the $L_2(\mu)$ ball of radius r and put rD_f to be the $L_2(\mu)$ ball centred at f and of radius r; $S(L_2)$ denotes the unit sphere in $L_2(\mu)$. From this point onward we do not specify the L_2 space to which the functions in question belong, as that will be clear from the context.

Given $1 \leq p < \infty$, let $B_p^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n |x_i|^p \leq 1\}$ be the unit ball in the space $\ell_p^n = (\mathbb{R}^n, \| \ \|_p)$, with the obvious modification when $p = \infty$; set S^{n-1} to be the Euclidean unit sphere in $\mathbb{R}^n$.

Let $\{G_f : f \in F\}$ be the canonical gaussian process indexed by F with a covariance structure endowed by $L_2(\mu)$ (see, e.g., [7] for a detailed survey on gaussian processes) and set

$$\mathbb{E}\|G\|_F = \sup \left\{ \mathbb{E} \sup_{h \in H} G_h : H \subset F, \ H \text{ is finite} \right\}.$$

Put $d_F(L_2) = \sup_{f \in F} \|f\|_{L_2}$ and let

$$k_F = \left(\frac{\mathbb{E}\|G\|_F}{d_F(L_2)} \right)^2,$$

which is an extension of the celebrated *Dvoretzky–Milman dimension* of a convex body in $\mathbb{R}^n$. We refer the reader to [22,23] for more details on the Dvoretzky–Milman dimension and its role in asymptotic geometric analysis.

For $\alpha \geq 1$, L_{ψ_α} is the Orlicz space of all measurable functions for which the ψ_α norm, defined by

$$\|f\|_{\psi_\alpha} = \inf\left\{c > 0 : \mathbb{E}\exp\left(|f/c|^\alpha\right) \leq 2\right\},$$

is finite. Some basic facts on Orlicz spaces may be found, for example, in [25].

The most important Orlicz norm in our context is the ψ_2 norm, which calibrates the subgaussian tail behaviour of a function. A class is L-subgaussian if the ψ_2 and L_2 norms are L-equivalent on F, that is, if for every $f, h \in F \cup \{0\}$, $\|f-h\|_{\psi_2} \leq L\|f-h\|_{L_2}$. In particular, such norm equivalence implies that there are absolute constants c_1 and c_2 such that for every $p \geq 2$, $\|f-h\|_{L_p} \leq c_1 L\sqrt{p}\|f-h\|_{L_2}$ and for $u \geq 1$, $Pr(|f-h| > c_2 u L\|f-h\|_{L_2}) \leq 2\exp(-u^2/2)$.

A class of functions H is star-shaped around 0 if for every $h \in H$ and every $\lambda \in [0,1]$, $\lambda h \in H$. In other words, if $h \in H$ then H contains the entire interval connecting h to 0.

It is straightforward to verify that if F is convex and $f \in F$ then $H_f = F - f = \{h - f : h \in F\}$ is star-shaped around 0.

A class that is star-shaped around zero has some regularity. The star-shape property implies that if $r < \rho$, then $H \cap rS(L_2)$ contains a 'scaled-down' version of $H \cap \rho S(L_2)$. Indeed, if $h \in H \cap \rho S(L_2)$ and since $r/\rho \in [0,1]$, it follows that $(r/\rho)h \in H \cap rS(L_2)$. In particular, normalized 'layers' of a star-shaped class become richer the closer the layer is to zero.

Finally, if A is a finite set, we denote by $|A|$ its cardinality.

2 Beyond the squared loss

As noted previously, our goal is to extend the results established in [19] from the squared loss $\ell(t) = t^2$ to a general smooth convex loss. For reasons we clarify later, the most interesting choices of loss functions satisfy some strong convexity property—either globally or in a neighbourhood of 0.

Definition 2.1 A function ℓ is strongly convex with a constant $c_0 > 0$ in the interval I if

$$\ell(y) \geq \ell(x) + \ell'(x)(y-x) + \frac{c_0}{2}(y-x)^2$$

for every $x, y \in I$.

Note that if $\inf_{x \in \mathbb{R}} \ell''(x) \geq c > 0$ then ℓ is strongly convex in $\mathbb{R}$ with a constant c, and one such example is the squared loss $\ell(t) = t^2$.

If one wishes the loss to be convex, as we do, its growth from any point must be at least linear. Therefore, it seems natural to consider loss functions that are strongly convex in an interval around zero, thus mimicking the local behaviour of the squared loss, while away from zero exhibit a linear, or almost linear growth, hopefully limiting the negative effect of outliers.

Typical examples of such losses are the Huber loss with parameter γ, defined by

$$\ell_\gamma(t) = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \le \gamma \\ \gamma|t| - \frac{\gamma^2}{2} & \text{if } |t| > \gamma, \end{cases} \tag{2.1}$$

and a version of the log-loss[2]

$$\ell(t) = -\log\left(\frac{4\exp(t)}{(1+\exp(t))^2}\right), \tag{2.2}$$

which is strongly convex in any bounded interval, but with a constant that decays exponentially to zero with the interval's length (because $\ell''(t) = 2\exp(t)/(\exp(t)+1)^2$).

As noted previously, the analysis of ERM is usually based on exclusion: showing that a large (random) part of the class cannot contain the empirical minimizer because the empirical risk functional is positive on functions that belong to that part. Strong convexity properties of the loss come in handy when exploring the decomposition of $\mathcal{L}_f(X,Y)$ via a Taylor expansion—a decomposition that leads naturally to an exclusion argument:

If ℓ has a second derivative then for every (X,Y) there is a mid-point Z for which

$$\begin{aligned}\mathcal{L}_f(X,Y) &= \ell(f(X)-Y) - \ell\left(f^*(X)-Y\right) \\ &= \ell'(\xi)(f-f^*)(X) + \frac{1}{2}\ell''(Z)(f-f^*)^2(X) = (1)+(2).\end{aligned}$$

One may exclude $H \subset F$ by showing that the empirical mean of the quadratic term (2) is positive on H, while the empirical mean of the multiplier component (1) cannot be very negative there. For such functions, $P_N\mathcal{L}_f > 0$, implying that $\hat{f} \in F \backslash H$.

If ℓ does not have a second derivative everywhere, one may modify this decomposition by noting that for every x_1 and x_2,

$$\ell(x_2) - \ell(x_1) = \int_{x_1}^{x_2} \ell'(w)dw = \ell'(x_1)(x_2-x_1) + \int_{x_1}^{x_2} \left(\ell'(w) - \ell'(x_1)\right)dw.$$

Therefore, when applied to (X,Y) and a fixed $f \in F$, and setting $\xi = f^*(X) - Y$, the quadratic component in the decomposition is

$$\int_{\xi}^{\xi+(f-f^*)(X)} \left(\ell'(w) - \ell'(\xi)\right)dw.$$

[2] The log-loss is more commonly used in the context of binary classification problems rather than in the type of real-valued problems we study here. However, because of its convexity properties it is an interesting example of the phenomenon we explore.

In particular, it is straightforward to show that if ℓ is twice differentiable in $\mathbb{R}$, except, perhaps, at $\pm x_0$, then for every X, Y one has

$$\mathcal{L}_f(X,Y) \geq \ell'(\xi)(f-f^*)(X) + \frac{1}{16}\ell''(Z)(f-f^*)^2(X)$$

for a well-chosen midpoint Z.

Definition 2.2 For every $f, f^* \in F$ and (X, Y), set

$$\mathcal{M}_{f-f^*}(X,Y) = \ell'(\xi)(f-f^*)(X),$$

and put

$$\mathcal{Q}_{f-f^*}(X,Y) = \int_{\xi}^{\xi+(f-f^*)(X)} \big(\ell'(w)-\ell'(x)\big)\,dw,$$

representing the multiplier and quadratic components of the excess loss $\mathcal{L}_f(X,Y)$.

Observe that a nontrivial, uniform, lower bound on ℓ'' significantly simplifies the question of lower bounding the empirical means of the quadratic component. In fact, in such a case the problem reverts to the study of the quadratic component of the squared loss, explored in [19]. Our focus is on situations in which no such lower bound exists.

2.1 The exclusion argument in prediction and estimation

A structural assumption that is needed throughout this exposition is the following:

Assumption 2.1 Assume that for every $f \in F$,

$$\mathbb{E}\ell'(\xi)(f-f^*)(X) \geq 0.$$

Assumption 2.1 is not really restrictive:

- If $\xi(X,Y) = f^*(X) - Y$ is independent of X (e.g., when $Y = f_0(X) - W$ for some unknown $f_0 \in F$ and an independent, mean-zero random variable W), then $\mathbb{E}\ell'(\xi)(f-f^*)(X) = 0$, because ℓ' is odd.
- If F is a convex class of functions and ℓ satisfies minimal integrability conditions, then $\mathbb{E}\ell'(\xi)(f-f^*)(X) \geq 0$ for every $f \in F$. Indeed, if there is some $f_1 \in F$ for which $\mathbb{E}\ell'(\xi)(f_1-f^*)(X) < 0$, then by considering $f_\lambda = \lambda f_1 + (1-\lambda)f^* \in F$ for λ close to 0,

$$\mathbb{E}\ell\left(f_\lambda(X)-Y\right) < \mathbb{E}\ell\left(f^*(X)-Y\right),$$

  which is impossible.

Under Assumption 2.1, if ℓ is twice differentiable then given a sample $(X_i, Y_i)_{i=1}^N$ and $f \in F$, there are mid-points Z_i that belong to the interval whose end points are $f^*(X_i) - Y_i$ and $f(X_i) - Y_i = (f - f^*)(X_i) + \xi_i$, such that

$$\begin{aligned} P_N \mathcal{L}_f =& \frac{1}{N} \sum_{i=1}^N \ell(f(X_i) - Y_i) - \ell\left(f^*(X_i) - Y_i\right) \\ \geq & \frac{1}{N} \sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i) + \frac{1}{2N} \sum_{i=1}^N \ell''(Z_i)(f - f^*)^2(X_i) \\ \geq & - \left| \frac{1}{N} \sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i) - \mathbb{E}\ell'(\xi)(f - f^*) \right| + \mathbb{E}\ell'(\xi)(f - f^*) \\ & + \frac{1}{2N} \sum_{i=1}^N \ell''(Z_i)(f - f^*)^2(X_i). \end{aligned} \tag{2.3}$$

Assume that on a high-probability event $\mathcal{A}$, for every $f \in F$,

$$\left| \frac{1}{N} \sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i) - \mathbb{E}\ell'(\xi)(f - f^*)(X) \right| \leq \frac{\theta}{4} \max\left\{ \|f - f^*\|_{L_2}^2, r_M^2 \right\},$$

for well chosen values r_M and θ. Assume further that on a high probability event $\mathcal{B}$, for every $f \in F$ with $\|f - f^*\|_{L_2} \geq r_Q$,

$$\frac{1}{N} \sum_{i=1}^N \ell''(Z_i)(f - f^*)^2(X_i) \geq \theta \left\| f - f^* \right\|_{L_2}^2.$$

Theorem 2.3 *If F satisfies Assumption 2.1, then on the event $\mathcal{A} \cap \mathcal{B}$, $\|\hat{f} - f^*\|_{L_2} \leq \max\{r_M, r_Q\}$.*

Proof By Assumption 2.1 and (2.3),

$$\begin{aligned} P_N \mathcal{L}_f \geq & \frac{1}{2N} \sum_{i=1}^N \ell''(Z_i)(f - f^*)^2(X_i) \\ & - \left| \frac{1}{N} \sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i) - \mathbb{E}\ell'(\xi)(f - f^*)(X) \right|. \end{aligned}$$

Hence, on the event $\mathcal{A} \cap \mathcal{B}$, if $\|f - f^*\|_{L_2} \geq \max\{r_M, r_Q\}$ then $P_N \mathcal{L}_f \geq (\theta/4)\|f - f^*\|_{L_2}^2 > 0$, and f cannot be an empirical minimizer. □

Theorem 2.3 implies that to bound the performance of ERM in the estimation problem it suffices to identify r_M and r_Q for which the event $\mathcal{A} \cap \mathcal{B}$ is sufficiently large.

Turning to the prediction problem, there is an additional assumption that is needed, namely, that $\mathbb{E}\mathcal{Q}_{f-f^*}$ does not increase too quickly when f is close to f^*.

Assumption 2.2 Assume that there is a constant β for which, for every $f \in F$ with $\|f - f^*\|_{L_2} \le \max\{r_M, r_Q\}$, one has

$$\mathbb{E}\mathcal{Q}_{f-f^*} \le \beta \left\|f - f^*\right\|_{L_2}^2.$$

Clearly, if ℓ' is a Lipschitz function, one may take $\beta = \|\ell'\|_{\text{lip}}$; hence, if ℓ'' exists everywhere and is a bounded function, $\beta \le \|\ell''\|_{L_\infty}$. Moreover, even when ℓ'' is not bounded, such a β exists if the functions $f - f^*$ have well behaved tails relative to the growth of ℓ''. Since the analysis required in these cases is rather obvious, we will not explore this issue further.

Theorem 2.4 *Assume that the loss ℓ is twice differentiable and satisfies Assumption 2.1 and Assumption 2.2. Using the notation introduced above, on the event $\mathcal{A} \cap \mathcal{B}$ one has*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \le 2(\theta + \beta)\max\left\{r_M^2, r_Q^2\right\}.$$

Proof Fix a sample in $\mathcal{A} \cap \mathcal{B}$. By Theorem 2.3, $\|\hat{f} - f^*\|_{L_2} \le \max\{r_M, r_Q\}$. Thus, it suffices to show that if $\|f - f^*\|_{L_2} \le \max\{r_M, r_Q\}$ and $\mathbb{E}\mathcal{L}_f \ge 2(\theta+\beta)\max\{r_M^2, r_Q^2\}$, then $P_N\mathcal{L}_f > 0$; in particular, such a function cannot be an empirical minimizer.

Note that for every $f \in F$, $\mathcal{L}_f = \mathcal{M}_{f-f^*} + \mathcal{Q}_{f-f^*}$ and thus either $\mathbb{E}\mathcal{L}_f \le 2\mathbb{E}\mathcal{M}_{f-f^*} = 2\mathbb{E}\ell'(\xi)(f - f^*)(X)$ or $\mathbb{E}\mathcal{L}_f \le 2\mathbb{E}\mathcal{Q}_{f-f^*}$.

However, if f satisfies the above, only the first option is possible; indeed, if $\mathbb{E}\mathcal{Q}_{f-f^*}$ is dominant, then by Assumption 2.2,

$$\mathbb{E}\mathcal{L}_f \le 2\mathbb{E}\mathcal{Q}_{f-f^*} \le 2\beta\left\|f - f^*\right\|_{L_2}^2 \le 2\beta\max\left\{r_M^2, r_Q^2\right\},$$

which is impossible by the choice of f. Therefore, it suffices to treat the case in which $\mathbb{E}\mathcal{L}_f \le 2\mathbb{E}\ell'(\xi)(f - f^*)(X)$.

Fix such an $f \in F$. Since ℓ is convex, $P_N\mathcal{L}_f \ge \frac{1}{N}\sum_{i=1}^N \xi_i(f - f^*)(X_i)$, and on $\mathcal{A} \cap \mathcal{B}$,

$$\begin{aligned} P_N\mathcal{L}_f \ge& \mathbb{E}\ell'(\xi)(f - f^*)(X) - \left|\frac{1}{N}\sum_{i=1}^N \ell'(\xi)(f - f^*)(X_i) - \mathbb{E}\ell'(\xi)(f - f^*)(X)\right| \\ \ge& \frac{1}{2}\mathbb{E}\mathcal{L}_f - \frac{\theta}{4}\max\left\{r_M^2, \left\|f - f^*\right\|_{L_2}^2\right\} \\ \ge& (\theta + \beta)\max\left\{r_M^2, r_Q^2\right\} - \frac{\theta}{4}\max\left\{r_M^2, r_Q^2\right\} > 0. \end{aligned}$$

□

Remark 2.5 When ℓ is twice differentiable except perhaps at $\pm x_0$, then Theorem 2.3 and Theorem 2.4 are still true though with modified constants.

In the following sections we develop the necessary machinery leading to a uniform lower estimate on the quadratic term $f \to P_N \mathcal{Q}_{f-f^*}$ and to an upper estimate on the multiplier term $f \to P_N \mathcal{M}_{f-f^*}$. Combining the two, we identify the values r_Q and r_M, as well as the right choice of θ.

3 Preliminary estimates

Let $(Z_i)_{i=1}^N$ be independent copies of a random variable Z and set $(Z_i^*)_{i=1}^N$ to be the monotone non-increasing rearrangement of $(|Z_i|)_{i=1}^N$. Below we obtain upper and lower estimates on various functions of $(Z_i^*)_{i=1}^N$. All the estimates we present are straightforward applications of either a concentration inequality for $\{0,1\}$-valued random variables (*selectors*) with mean δ, or, alternatively, a rather crude binomial estimate. Indeed, given a property $\mathcal{P}$ set $\delta_i = \mathbb{1}_{\{Z_i \in \mathcal{P}\}}$—the characteristic function of the event that Z_i satisfies property $\mathcal{P}$. Let $\delta = Pr(Z \in \mathcal{P})$ and note that $|\{i : Z_i \in \mathcal{P}\}| = \sum_{i=1}^N \delta_i$. By Bernstein's inequality (see, for example, [3,25]),

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^N \delta_i - \delta\right| \le t\right) \ge 1 - 2\exp\left(-cN\min\left\{t^2/\delta, t\right\}\right)$$

for a suitable absolute constant c. Hence, taking $t = u\delta$,

$$N\delta(1-u) \le |\{i : Z_i \in \mathcal{P}\}| \le N\delta(1+u) \tag{3.1}$$

with probability at least $1 - 2\exp(-cN\delta\min\{u^2, u\})$.

The binomial estimate we employ is equally simple: it is based on the fact that

$$Pr\left(|\{i : Z_i \in \mathcal{P}\}| \ge k\right) \le \binom{N}{k} Pr^k(Z \in \mathcal{P}) \le \left(\frac{eN}{k} \cdot Pr(Z \in \mathcal{P})\right)^k.$$

3.1 Tail-based upper estimates

Assume that one has information on $\|Z\|_{L_q}$ for some $q > 2$ and set $L = \|Z\|_{L_q}/\|Z\|_{L_2}$. Applying Chebyshev's inequality,

$$Pr\left(|Z| \ge w\,\|Z\|_{L_2}\right) \le \frac{\mathbb{E}\,|Z|^q}{\|Z\|_{L_2}^q\, w^q} = \frac{L^q}{w^q}.$$

Hence, if $\mathcal{P} = \{|Z| < w\|Z\|_{L_2}\}$ it follows that $Pr(Z \in \mathcal{P}) \ge 1 - (L/w)^q$, which can be made arbitrarily close to 1 by selecting w that is large enough. This observation implies that with high probability, an arbitrary large proportion of $\{|Z_1|, \ldots, |Z_N|\}$ are not very large.

Lemma 3.1 *There exists absolute constants c_1 and c_2 for which the following holds. Let $Z \in L_2$. For every $0 < \varepsilon < 1$, with probability at least $1 - 2\exp(-c_1 \varepsilon N)$ there exists a subset $I \subset \{1, \ldots, N\}$, $|I| \geq (1-\varepsilon)N$, and for every $i \in I$,*

$$|Z_i| \leq c_2 \varepsilon^{-1/2} \|Z\|_{L_2}.$$

Proof Fix ε as above and note that $Pr(|Z| \geq 2\|Z\|_{L_2}/\sqrt{\varepsilon}) \leq \varepsilon/4$. Hence, by a binomial estimate,

$$\begin{aligned} & Pr\left(\left|\left\{i : |Z_i| \geq 2\|Z\|_{L_2}/\sqrt{\varepsilon}\right\}\right| \geq N\varepsilon\right) \leq \binom{N}{\varepsilon N} Pr^{\varepsilon N}\left(|Z| \geq 2\|Z\|_{L_2}/\sqrt{\varepsilon}\right) \\ & \leq \left(\frac{e}{\varepsilon}\right)^{N\varepsilon} \cdot \left(\frac{\varepsilon}{4}\right)^{N\varepsilon} \leq \exp(-cN\varepsilon), \end{aligned}$$

for a suitable absolute constant c. □

Given a vector $a = (a_i)_{i=1}^N$, the L_q norm of a, when considered as a function on $\Omega = \{1, \ldots, N\}$ endowed with the uniform probability measure, is

$$\|a\|_{L_q^N} = \left(\frac{1}{N}\sum_{i=1}^N |a_i|^q\right)^{1/q}.$$

The weak-L_q norm of the vector a is

$$\|a\|_{L_{q,\infty}^N} = \inf\left\{c > 0 : d_a(t) \leq (c/t)^q \text{ for every } t > 0\right\},$$

where $d_a(t) = N^{-1}|\{i : |a_i| > t\}|$.

The next observation is that sampling preserves the L_q structure of Z, in the sense that if $Z \in L_q$, then with high probability, $\|(Z_i)_{i=1}^N\|_{L_{q,\infty}^N} \lesssim \|Z\|_{L_q}$.

Lemma 3.2 *Let $1 \leq q \leq r$. If $Z \in L_r$, $u \geq 2$ and $1 \leq k \leq N/2$, then*

$$Z_k^* \leq u(N/k)^{1/q} \|Z\|_{L_r}$$

with probability at least $1 - u^{-kr}\left(\frac{eN}{k}\right)^{-k((r/q)-1)}$.

In particular, with probability at least $1 - 2u^{-r}N^{-((r/q)-1)}$,

$$\|(Z_i)\|_{L_{q,\infty}^N} \leq u\|Z\|_{L_r}.$$

Proof Let $\eta = (r/q) - 1$, fix $1 \leq k \leq N/2$ and set $v > 0$ to be named later. The binomial estimate implies that

$$Pr\left(Z_k^* \geq v(eN/k)^{(1+\eta)/r}\, \|Z\|_{L_r}\right) \leq \binom{N}{k} Pr^k\left(|Z| \geq v(eN/k)^{(1+\eta)/r}\, \|Z\|_{L_r}\right)$$
$$\leq \left(\frac{eN}{k}\right)^k \left(\frac{k}{eN}\right)^{(1+\eta)k} \cdot v^{-kr} = \left(\frac{eN}{k}\right)^{-\eta k} v^{-kr}.$$

In particular, for $v = u(eN/k)^{1/q-(1+\eta)/r}$,

$$Z_k^* \leq u(N/k)^{1/q}\, \|Z\|_{L_r}$$

with probability at least

$$1 - u^{-kr}\left(\frac{eN}{k}\right)^{k((r/q)-1)}.$$

The second part of the claim follows by summing up the probabilities for $k \leq N/2$, using that $Z_k^* \leq Z_{N/2}^*$ for $k \geq N/2$ and that $(u^{-kr})_{k=1}^{N/2}$ is a geometric progression. □

Remark 3.3 Note that similar statements to Lemma 3.2 are true if one simply assumes that $Pr(|Z| \geq t) < \varepsilon$, even without moment assumptions. Of course, under such an assumption one has no information whatsoever on the largest εN coordinates of $(|Z_1|, \ldots, |Z_N|)$, but rather, only on a certain proportion that is slightly smaller than $(1-\varepsilon)N$ of the smallest coordinates.

Also, observe that $\|(Z_i^*)_{i \geq j}\|_{L_{q,\infty}^N} \lesssim \|Z\|_{L_q}$ with a probability estimate that improves exponentially in j.

3.2 Lower estimates using a small-ball property

A similar line of reasoning to the one used above is true for lower estimates, and is based on a small-ball condition:

Definition 3.4 A random variable Z satisfies a small-ball condition with constants $\kappa > 0$ and $0 < \varepsilon < 1$ if

$$Pr\left(|Z| \geq \kappa\, \|Z\|_{L_2}\right) \geq \varepsilon.$$

A class of functions F satisfies a small-ball property with constants κ and $0 < \varepsilon < 1$ if for every $f \in F$,

$$Pr\left(|f| \geq \kappa\, \|f\|_{L_2}\right) \geq \varepsilon.$$

Remark 3.5 Because the applications considered below require that many of the $|Z_i|$'s are at least of the order of $\|Z\|_{L_2}$, the L_2 norm is used as a point of reference in the definition of the small-ball condition—though the notion of 'small-ball' can be modified to fit other norms, as well as in situations in which Z need not even be integrable.

This small-ball condition was introduced in the context of estimation problems in [19], and is the most important feature of our presentation. It is a rather weak assumption that is almost universally satisfied and eliminates the need to guarantee two-sided concentration, which is a far-more restrictive phenomenon. Unlike concentration, a small-ball condition captures that behaviour of a random variable close to zero: by considering $h = Z/\|Z\|_{L_2}$, it quantifies the weight that h assigns to a neighbourhood of zero. In particular, if $Z \in L_2$ and $Z \not\equiv 0$, then it satisfies a small-ball condition for some constants ε and κ. It is also important to emphasize that a small-ball condition has nothing to do with the tail behaviour of Z (other than the obvious assumption that Z is square-integrable). For example, if Z is mean-zero, variance 1, random variable that has a density bounded by M, it satisfies a small-ball condition with constants that depend only on M—though it is possible that Z does not have any moment beyond the second one.

The proofs we present below are based on the assumption that the class satisfies a small-ball property, which is simply a small-ball condition with fixed constants that holds for every function in the class[3].

Naturally, for such an approach to be of any use, one must show that there are enough natural situations in which a small-ball property holds, and as indications let us describe several cases involving classes of linear functionals on $\mathbb{R}^n$.

Let μ be a probability measure on $\mathbb{R}^n$ and let X be distributed according to μ. If there are constants κ and ε such that for every $t \in \mathbb{R}^n$,

$$Pr\left(\left|\left\langle t, X\right\rangle\right| \geq \kappa \left\|\left\langle t, X\right\rangle\right\|_{L_2}\right) \geq \varepsilon \tag{3.2}$$

then it immediately follows that any class of functions $\{\langle t, \cdot\rangle : t \in T\}$ satisfies the small-ball property with constants κ and ε.

The simplest situation in which (3.2) holds for every $t \in \mathbb{R}^n$ is when there is some $q > 2$ for which the L_q and L_2 norms of linear functionals are equivalent; that is,

$$\sup_{t \in S^{n-1}} \frac{\left\|\left\langle t, X\right\rangle\right\|_{L_q}}{\left\|\left\langle t, X\right\rangle\right\|_{L_2}} \leq L. \tag{3.3}$$

A straightforward application of the Paley-Zygmund inequality (see, e.g. [5]) shows that (3.2) holds for constants κ and ε that depend only on q and L, and in particular, are independent of the dimension n of the underlying space. As an extreme example, if $0 < \alpha \leq 2$ and X is a Ψ_α random vector in $\mathbb{R}^n$ (e.g., if X is subgaussian or log-concave) then (3.2) holds—though this is an obvious overkill: a small-ball condition like in (3.2) is true under a minimal norm equivalence like (3.3) and does not require the high moment equivalence $\|\langle X, t\rangle\|_{L_p} \leq cp^{1/\alpha}\|\langle X, t\rangle\|_{L_2}$ that is ensured by a $\Psi_\alpha - L_2$ norm equivalence.

To indicate yet again the clear difference between a small-ball condition and tail estimates, let us construct random vectors that need not have any moment beyond the

[3] Let us mention that it is possible to modify the arguments and tackle situations in which the constants κ and ε are not uniform, but to keep this article at a reasonable length we defer this to future work.

second one and still satisfy (3.2). One way of generating such random vectors is based on the following observation:

Lemma 3.6 *There are absolute constants c_0 and c_1 for which the following holds.*

(1) Let z be a random variable that satisfies a small-ball condition with constants κ and ε. Let $z_1, \ldots, z_n$ to be independent copies of z and put $Z = (z_i)_{i=1}^n$. Then for every $t \in \mathbb{R}^n$, $\langle t, Z\rangle$ satisfies a small-ball condition with constants $c_0\sqrt{\varepsilon}\kappa$ and $c_1\varepsilon$.

(2) Let $Z = (z_1, \ldots, z_n)$ be an unconditional random vector (that is, Z has the same distribution as $(\varepsilon_i z_i)_{i=1}^n$ for any choice of signs $(\varepsilon_i)_{i=1}^n$). Assume that for every $1 \le i \le n$, $\mathbb{E}z_i^2 = 1$ and $|z_i| \le M$ almost surely. Then, for every $t \in \mathbb{R}^n$, $\langle t, Z\rangle$ satisfies a small-ball condition with constants c_2 and c_3 depending only on M.

The proof of Lemma 3.6 is rather standard and we only sketch it for the sake of completeness. The idea is that for every $t \in S^{n-1}$, the random variable $\langle t, Z\rangle = \sum_{i=1}^n t_i z_i$ has the same distribution as $\sum_{i=1}^n \varepsilon_i |t_i z_i|$, where $(\varepsilon_i)_{i=1}^n$ are independent random signs that are also independent of $(z_i)_{i=1}^n$. Note that conditioned on $(z_i)_{i=1}^n$, $\sum_{i=1}^n \varepsilon_i |t_i z_i|$ satisfies a small-ball condition with absolute constants; indeed, this is an immediate outcome of the Paley-Zygmund inequality and the fact that there is an absolute constant L such that for every fixed $(a_i)_{i=1}^n \in \mathbb{R}^n$, $\| \sum_{i=1}^n \varepsilon_i a_i \|_{L_2} \le L \| \sum_{i=1}^n \varepsilon_i a_i \|_{L_1}$. Therefore, there are constants $c_1(L)$ and $c_2(L)$ such that

$$Pr_\varepsilon \left(\Big| \sum_{i=1}^n \varepsilon_i \, |t_i z_i| \Big| > c_1 \left(\sum_{i=1}^n t_i^2 z_i^2 \right)^{1/2} \right) \ge c_2$$

and all that remains is to show that with nontrivial probability, $\sum_{i=1}^n t_i^2 z_i^2 \ge c \sum_{i=1}^n t_i^2 \mathbb{E} z_i^2$.

In (2), let $w = \sum_{i=1}^n t_i^2 z_i^2$. By the L_∞ bound on the z_i's and since $\mathbb{E}z_i^2 = 1$,

$$\mathbb{E}w^2 = \mathbb{E}\left(\sum_{i=1}^n t_i^2 z_i^2 \right)^2 \le M^4 \left(\sum_{i=1}^n t_i^2 \right)^2 = M^4 \left(\mathbb{E}\left(\sum_{i=1}^n t_i^2 z_i^2 \right) \right)^2 = M^4 \left(\mathbb{E}\,|w| \right)^2.$$

Hence $\|w\|_{L_2} \le M^2 \|w\|_{L_1}$ and the required estimate follows from the Paley-Zygmund inequality.

In (1), by the small-ball property of the z_i's,

$$\sum_{i=1}^n t_i^2 z_i^2 \ge \kappa^2 \sum_{i=1}^n t_i^2 (\mathbb{E}z_i^2) \cdot \mathbb{1}_{\{z_i^2 \ge \kappa \mathbb{E} z_i^2\}} = \kappa^2 \sum_{i=1}^n t_i^2 (\mathbb{E}z_i^2) \cdot \delta_i,$$

where $(\delta_i)_{i=1}^n$ independent selectors with mean at least ε. Again, one concludes by invoking the Paley-Zygmund inequality for $\sum_{i=1}^n t_i^2 (\mathbb{E}z_i^2) \cdot \delta_i$. □

Lemma 3.7 *There exists an absolute constant c for which the following holds. Assume that Z satisfies a small-ball condition with constants κ and ε and let $(Z_i)_{i=1}^N$ be* *independent copies of Z. Then, with probability at least $1 - 2\exp(-cN\varepsilon)$, there is a subset I of $\{1, \dots, N\}$ of cardinality at least $(3/4)\varepsilon N$ and for every $i \in I$, $|Z_i| \geq \kappa \|Z\|_{L_2}$.*

The proof, which we omit, is an immediate application of Bernstein's inequality for the i.i.d. selectors $\delta_i = \mathbb{1}_{\{|Z_i| \geq \kappa \|Z\|_{L_2}\}}$ for the choice of $u = 1/4$ in (3.1). Naturally, at a price of a weaker probability estimate the constant $3/4$ in Lemma 3.7 can be made arbitrarily close to 1.

Combining the upper estimate from Lemma 3.1 and lower one from Lemma 3.7 yields the following corollary:

Corollary 3.8 *There exist absolute constants c_1 and c_2 for which the following holds. Assume that $Z \in L_2$ and that it satisfies a small-ball condition with constants κ and ε. Then, with probability at least $1 - 2\exp(-c_1 \varepsilon N)$, there is $J \subset \{1, \dots, N\}$, $|J| \geq \varepsilon N/2$ and for every $j \in J$,*

$$\kappa \|Z\|_{L_2} \leq \left|Z_j\right| \leq c_2 \|Z\|_{L_2} / \sqrt{\varepsilon}.$$

Corollary 3.8 allows one to control the behaviour of $(Z_i)_{i=1}^N$ on a subset of $\{1, \dots, N\}$ of cardinality $\sim \varepsilon N$, and with exponentially high probability. Moreover, by modifying c_1 and c_2, the cardinality of J can be made arbitrarily close to εN.

Remark 3.9 Note that by the union bound, a version of Corollary 3.8 holds uniformly for a collection of $\exp(c_1 N\varepsilon/2)$ random variables with probability at least $1 - 2\exp(-c_1 N \varepsilon/2)$.

3.3 Uniform estimates on a class

Next, we use the estimates obtained above and study the structure of a typical coordinate projection of a class H, $P_\sigma H = \{(h(X_i))_{i=1}^N : h \in H\}$. We show that with high probability, for every function in H of sufficiently large L_2 norm, most of the coordinates of $P_\sigma h$ are of the order of $\|h\|_{L_2}$. Such a result is an extension of Corollary 3.8—from a single function to a class of functions—and the class we focus on in what follows is $H_{f^*} = \{f - f^* : f \in F\}$.

To quantify what is meant by "sufficiently large L_2 norm", recall that for $H \subset L_2(\mu)$, $\{G_h : h \in H\}$ is the canonical gaussian process indexed by H with a covariance structure endowed by $L_2(\mu)$.

Definition 3.10 Given a class of functions $H \subset L_2(\mu)$, a sample size N and positive constants ζ_1 and ζ_2 set

$$r_{1,Q}(H, N, \zeta_1) = \inf \left\{ r > 0 : \mathbb{E} \|G\|_{H \cap rD} \leq \zeta_1 r \sqrt{N} \right\}, \tag{3.4}$$

and put

$$r_{2,Q}(H, N, \zeta_2) = \inf \left\{ r > 0 : \mathbb{E} \sup_{H \cap rD} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| \leq \zeta_2 r \sqrt{N} \right\}, \tag{3.5}$$

where D is, as always, the unit ball of $L_2(\mu)$.

When the class H and sample size N are obvious from the context, we denote the fixed points by $r_{1,Q}(\zeta_1)$ and $r_{2,Q}(\zeta_2)$ respectively.

Finally, set

$$r_Q(\zeta_1,\zeta_2)=r_Q(H,N,\zeta_1,\zeta_2)=\max\left\{r_{1,Q}(\zeta_1),r_{2,Q}(\zeta_2)\right\}.$$

If H is star-shaped around 0, it is straightforward to show that when $r>r_{1,Q}(\zeta_1)$, one has $\mathbb{E}\|G\|_{H\cap rD}\leq\zeta_1 r\sqrt{N}$, while if $r<r_{1,Q}(\zeta_1)$, $\mathbb{E}\|G\|_{H\cap rD}\geq\zeta_1 r\sqrt{N}$. A similar observation is true for $r_{Q,2}(\zeta_2)$.

Theorem 3.11 *There exist absolute constants c_0,c_1,c_2,c_3,c_4 and c_5 for which the following holds. Let H be a class of functions that is star-shaped around 0 and that satisfies a small-ball property with constants κ_0 and ε. If $\zeta_1=c_1\kappa_0\varepsilon^{3/2}$, $\zeta_2=c_2\kappa_0\varepsilon$ and $r>r_Q(\zeta_1,\zeta_2)$, there is $V_r\subset H\cap rS(L_2)$ and an event Ω' of probability at least $1-2\exp(-c_0\varepsilon^2 N)$, such that:*

(1) $|V_r|\leq\exp(c_3\varepsilon N)$ *and* $c_3\leq 1/1000$.

(2) *On the event Ω', for every $v\in V_r$ there is a subset $I_v\subset\{1,\ldots,N\}$, $|I_v|\geq\varepsilon N/2$ and for every $i\in I_v$,*

$$\kappa_0 r\leq|v(X_i)|\leq c_4 r/\sqrt{\varepsilon}.$$

(3) *On the event Ω', for every $h\in H\cap rS(L_2)$ there is some $v\in V_r$ and a subset $J_h\subset I_v$, consisting of at least* 3/4 *of the coordinates of I_v (and in particular, $|J_h|\geq\varepsilon N/4$), and for every $j\in J_h$,*

$$(\kappa_0/2)\,\|h\|_{L_2}\leq\left|h(X_j)\right|\leq c_5\left(\kappa_0+1/\sqrt{\varepsilon}\right)\|h\|_{L_2}$$

and

$$\mathrm{sgn}(h(X_j))=\mathrm{sgn}(v(X_j)).$$

The idea of the proof is to find an appropriate net in $H\cap rS(L_2)$ (the set V_r), and show that each point in the net has many 'well-behaved' coordinates in the sense of (2). Also, if πh denotes the best approximation of $h\in H\cap rS(L_2)$ in V_r with respect to the L_2 norm, then

$$\sup_{h\in H\cap rS(L_2)}\frac{1}{N}\sum_{i=1}^{N}|h-\pi h|(X_i)$$

is not very big, implying that $|(h-\pi h)(X_i)|$ cannot have too many large coordinates. Since $h(X_i)=(\pi h)(X_i)+(h-\pi h)(X_i)$, the first term is dominant on a proportional number of coordinates, leading to (3).

Proof Recall that by Corollary 3.8, if $Z \in L_2$ satisfies the small-ball condition with constants κ_0 and ε then with probability at least $1 - 2\exp(-c_1 \varepsilon N)$, there is $I \subset \{1, \ldots, N\}$, $|I| \geq \varepsilon N/2$ and for every $i \in I$,

$$\kappa_0 \left\|Z\right\|_{L_2} \leq |Z_i| \leq c_2 \left\|Z\right\|_{L_2} / \sqrt{\varepsilon}.$$

Fix ζ_1 and ζ_2 to be named later, let $r > r_Q(\zeta_1, \zeta_2)$ and set $c_1' = \min\{c_1, 1/500\}$. Let $V_r \subset H \cap rS(L_2)$ be a maximal separated set whose cardinality is at most $\exp(c_1' \varepsilon N/2)$ and denote its mesh by η. Therefore, by Corollary 3.8 and the union bound, it follows that with probability at least $1 - 2\exp(-c_1 \varepsilon N/2)$, for every $v \in V_r$ there is a subset I_v as above, i.e., $|I_v| \geq \varepsilon N/2$ and for every $i \in I_v$,

$$\kappa_0 r = \kappa_0 \left\|v\right\|_{L_2} \leq |v(X_i)| \leq c_2 \left\|v\right\|_{L_2} / \sqrt{\varepsilon} = c_2 r / \sqrt{\varepsilon}.$$

Moreover, by Sudakov's inequality (see, e.g. [7,12,23]) and because $r \geq r_{Q,1}(\zeta_1)$, it follows that

$$\eta \leq c_3 \frac{\mathbb{E}\left\|G\right\|_{H \cap rS(L_2)}}{\sqrt{c_1' N\varepsilon/2}} \leq \left(c_4 \zeta_1 / \sqrt{\varepsilon}\right) r$$

for $c_4 = \sqrt{2} c_3 / \sqrt{c_1'}$.

For every $h \in H \cap rS(L_2)$, let $\pi h \in V_r$ such that $\|h - \pi h\|_{L_2} \leq \eta$, set $u_h = \mathbb{1}_{\{|h - \pi h| > \kappa_0 r/2\}}$ and put

$$U_r = \{u_h : h \in H \cap rS(L_2)\}.$$

Let $\phi(t) = t/(\kappa_0 r/2)$ and note that pointwise, for every $u_h \in U_r$, $u_h(X) \leq \phi(|h - \pi h|(X))$.

Applying the Giné–Zinn symmetrization theorem (see, e.g., [25]) and recalling that $r > r_{Q,2}(\zeta_2)$, one has

$$\begin{aligned}
\mathbb{E} \sup_{u_h \in U_r} \frac{1}{N} \sum_{i=1}^{N} u_h(X_i) &\leq \mathbb{E} \sup_{h \in H \cap rS(L_2)} \frac{1}{N} \sum_{i=1}^{N} \phi\left(\left|h - \pi h\right|(X_i)\right) \\
&\leq \mathbb{E} \sup_{h \in H \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^{N} \phi\left(\left|h - \pi h\right|(X_i)\right) - \mathbb{E}\phi\left(\left|h - \pi h\right|(X_i)\right) \right| \\
&\quad + \sup_{h \in H \cap rS(L_2)} \mathbb{E}\phi\left(\left|h - \pi h\right|\right) \\
&\leq \frac{4}{\kappa_0 r} \cdot \left(\mathbb{E} \sup_{h \in H \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (h - \pi h)(X_i) \right| + \sup_{h \in H \cap rS(L_2)} \left\|h - \pi h\right\|_{L_2} \right) \\
&\leq \frac{4}{\kappa_0 r} \cdot (2\zeta_2 r + \eta) \leq \frac{\varepsilon}{32},
\end{aligned}$$

provided that $\zeta_1 \sim \kappa_0 \varepsilon^{3/2}$ and $\zeta_2 \sim \kappa_0 \varepsilon$.

Let $\psi(X_1,\dots,X_N) = \sup_{u\in U_r}\frac{1}{N}\sum_{i=1}^N u(X_i)$. By the bounded differences inequality (see, for example, [3]), with probability at least $1-\exp(-c_5t^2)$,

$$\psi(X_1,\dots,X_N) \le \mathbb{E}\psi + \frac{t}{\sqrt{N}}.$$

Thus, for $t=\varepsilon\sqrt{N}/32$, with probability at least $1-\exp(-c_6\varepsilon^2N)$, $\psi(X_1,\dots,X_N)\le\varepsilon/16$, implying that for every $h\in H\cap rS(L_2)$,

$$|\{i : |h-\pi h|\,(X_i)\le(\kappa_0/2)r\}| \ge \left(1-\frac{\varepsilon}{16}\right)N.$$

Recall that $\pi h\in V_r$ and that $|I_{\pi h}|\ge\varepsilon N/2$. Let

$$J_h=\left\{j : |h-\pi h|\,(X_j)\le(\kappa_0/2)r\right\}\cap I_{\pi h}$$

and thus J_h consists of at least 3/4 of the coordinates in $I_{\pi h}$ and $|J_h|\ge\varepsilon N/4$. Moreover, for every $j\in J_h$,

$$\left|h(X_j)\right| \ge \left|\pi h(X_j)\right| - \left|(h-\pi h)(X_j)\right| \ge \kappa_0 r-(\kappa_0/2)r=(\kappa_0/2)r, \qquad (3.6)$$

which also shows that $\mathrm{sgn}(h(X_j))=\mathrm{sgn}(\pi h(X_j))$.

The upper estimate follows from a similar argument, using that $|h(X_j)|\le|\pi h(X_j)|+|(h-\pi h)(X_j)|$. □

Remark 3.12 Observe that by the star-shape property of H, if $\rho_1>\rho_2$, then

$$\frac{1}{\rho_1}\left(H\cap\rho_1S(L_2)\right)\subset\frac{1}{\rho_2}\left(H\cap\rho_2S(L_2)\right).$$

Therefore, certain features of $H\cap\rho_2S(L_2)$ are automatically transferred to $H\cap\rho_1S(L_2)$, and in particular, a version of Theorem 3.11 holds uniformly for every level that is 'larger' than $2r_Q(\zeta_1,\zeta_2)$. Indeed, assume that one has chosen $\rho_2=2r_Q$ in Theorem 3.11 and fix $h\in H\cap\rho_1S(L_2)$. By applying Theorem 3.11 to $h'=(\rho_2/\rho_1)h\in H\cap\rho_2S(L_2)$ it follows that on the event Ω' there is a subset J of $\{1,\dots,N\}$ of cardinality at least $\varepsilon N/4$ on which

$$\left|h(X_j)\right|\ge(\kappa_0/2)\rho_1 \ \text{ and } \ \mathrm{sgn}(h(X_j))=\mathrm{sgn}\left(\pi h'\right)(X_j).$$

Next, let $F\subset L_2$ be a convex set, fix $f^*\in F$ and put $H_{f^*}=\{f-f^* : f\in F\}$. Since H_{f^*} is clearly star-shaped around 0 and $H_{f^*}\subset F-F$ one has:

Corollary 3.13 *If F is a convex class of functions, $F-F$ satisfies a small-ball property with constants κ_0 and ε, and $r=2r_Q(F-F,N,\zeta_1,\zeta_2)$, then with probability at least $1-2\exp(-c_0\varepsilon^2N)$, the following holds. For every $f_1,f_2\in F$ that satisfy*

$\|f_1 - f_2\|_{L_2} \geq r$, there is a subset $J_{f_1,f_2} \subset \{1,\ldots,N\}$ of cardinality at least $\varepsilon N/4$ and for every $j \in J_{f_1,f_2}$,

$$\left|(f_1 - f_2)(X_j)\right| \geq (\kappa_0/2)\,\|f_1 - f_2\|_{L_2}\,.$$

In particular, on the same event,

$$\inf_{\left\{f\in F:\|f-f^*\|_{L_2}\geq 2r_Q\right\}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f-f^*}{\|f-f^*\|_{L_2}}\right)^2(X_i) \geq \frac{\varepsilon\kappa_0^2}{16}.$$

4 The quadratic component of the loss

Following the path of the exclusion argument, the aim is to show that the quadratic component of the loss is sufficiently positive. And, although the results are formulated below in full generality, there are three examples that one should keep in mind: First, when ℓ is strongly convex; second, when ℓ is a general loss function and $Y = f_0(X)+W$ for some $f_0 \in F$ and a symmetric random variable W that is independent of X; finally, a situation that is, in some sense, a mixture of the two: a loss function that is guaranteed to be strongly convex only in a neighbourhood of 0 (for example, the loss functions (2.1) and (2.2)), and without assuming that the noise $Y - f^*(X)$ is independent of X.

Throughout this section we assume that F is a convex class of functions and that $F-F = \{f-h : f,h \in F\}$ satisfies a small-ball property with constants κ_0 and ε. Set $r_Q = r_Q(F-F, N, \zeta_1, \zeta_2)$ with the choice of ζ_1 and ζ_2 as in Theorem 3.11—namely, $\zeta_1 \sim \kappa_0\varepsilon^{3/2}$ and $\zeta_2 \sim \kappa_0\varepsilon$. Also assume that ℓ'' exists everywhere except perhaps at $\pm x_0$ and set for every $0 < t_1 < t_2$

$$\rho(t_1,t_2) = \inf\left\{\ell''(x) : x \in [t_1,t_2],\ x \neq \pm x_0\right\}. \tag{4.1}$$

The following lower bound on the quadratic component in the strongly convex case is an immediate application of Corollary 3.13 and the fact that $\mathcal{Q}_{f-f^*}(X,Y) \gtrsim \ell''(Z)(f-f^*)^2(X)$ for an appropriate mid-point Z. Its proof is omitted.

Theorem 4.1 *There exists an absolute constant c_1 for which the following holds. If $\inf_{x\in\mathbb{R}\setminus\{\pm x_0\}}\ell''(x) \geq 2c_0$, then with probability at least $1-2\exp(-c_1 N\varepsilon^2)$, for every $f \in F$ with $\|f-f^*\|_{L_2} \geq 2r_Q$,*

$$P_N\mathcal{Q}_{f-f^*} \geq \frac{c_0\varepsilon\kappa_0^2}{16}\,\|f-f^*\|_{L_2}^2\,.$$

Theorem 4.1 generalizes a similar result from [19] for the squared loss.

Turning to the more difficult (and interesting) problem of a loss that need not be strongly convex, we begin with the case of independent noise.

Assumption 4.1 Assume that $Y = f_0(X) + W$ for $f_0 \in F$ and a symmetric random variable $W \in L_2$ that is independent of X and for which

$$Pr\left(|W| \leq \kappa_1 \|W\|_{L_2}\right) \leq \varepsilon/1000. \tag{4.2}$$

Clearly, $f^* = f_0$; moreover, (4.2) is a rather minimal assumption, as a small-ball condition for a single function at one level holds when the function is absolutely continuous, by selecting the right value κ_1.

Given a sample $(X_i, Y_i)_{i=1}^N$ let $W_i = Y_i - f^*(X_i)$ and set Z_i to be the mid-points in the lower bound on the quadratic component of ℓ—again using the fact that for the losses in question, $\mathcal{Q}_{f-f^*}(X_i, Y_i) \gtrsim \ell''(Z_i)(f - f^*)^2(X_i)$.

Theorem 4.2 *There exist absolute constants* c_1, c_2, c_3 *and* c_4 *for which the following holds. Let* F *and* W *be as above. With probability at least* $1 - 2\exp(-c_1\varepsilon^2 N)$*, for every* $f \in F$ *that satisfies* $\|f - f^*\|_{L_2} \geq 2r_Q$ *one has*

$$P_N \mathcal{Q}_{f-f^*} \geq c_2 \varepsilon \kappa_0^2 \rho(t_1, t_2) \left\|f - f^*\right\|_{L_2}^2.$$

where

$$t_1 = \kappa_1 \|W\|_{L_2} \text{ and } t_2 = c_3 \varepsilon^{-1/2} \|W\|_{L_2} + c_4 \left(\kappa_0 + \varepsilon^{-1/2}\right) \left\|f - f^*\right\|_{L_2}.$$

The proof of Theorem 4.2 is based on several observations leading to accurate information on the 'location' of the midpoints Z_i in the lower bound on $\mathcal{Q}_{f-f^*}(X_i, Y_i)$. For every (X, Y), the corresponding mid-point belongs to interval whose end-points are $(f - f^*)(X) - W$ and $-W$. If I_f is the set of coordinates on which $|(f - f^*)(X_i)|$ is of the order of $\|f - f^*\|_{L_2}$, and since X and W are independent and W is symmetric, then on roughly half of these coordinates the signs of $(f - f^*)(X_i)$ coincide with the signs of $-W_i$. For those coordinates,

$$|Z_i| \in \left[|W_i|, |W_i| + \left|\left(f - f^*\right)(X_i)\right|\right].$$

Moreover, by excluding a further, sufficiently small proportion of the coordinates in I_f it follows that $|W_i| \sim \|W\|_{L_2}$, as long as W is not highly concentrated around zero.

The difficulty is in making this argument uniform, in the sense that it should hold for every $f \in F$ rather than for a specific choice of f. The first step towards such a uniform result is the following lemma.

Lemma 4.3 *Let* $1 \leq k \leq m/40$ *and set* $\mathcal{S} \subset \{-1, 0, 1\}^m$ *of cardinality at most* $\exp(k)$*. For every* $s = (s(i))_{i=1}^m \in \mathcal{S}$ *set* $I_s = \{i : s(i) \neq 0\}$ *and assume that* $|I_s| \geq 40k$*. If* $(\varepsilon_i)_{i=1}^m$ *are independent, symmetric* $\{-1, 1\}$*-valued random variables then with probability at least* $1 - 2\exp(-k)$*, for every* $s \in \mathcal{S}$*,*

$$|\{i \in I_s : \operatorname{sgn}(s(i)) = \varepsilon_i\}| \geq |I_s|/3.$$

Proof For every fixed $s \in \mathcal{S}$, the event $|\{i : (s(i))_{i \in I_s} = \varepsilon_i\}| \geq \ell$ has the same distribution as $|\{i \in I_s : \varepsilon_i = 1\}| \geq \ell$. If $(\delta_i)_{i \in I_s}$ are selectors of mean $1/2$ then

$$Pr\left(|\{i \in I_s : s(i) = \varepsilon_i\}| \geq \ell\right) = Pr\left(\sum_{i=1}^{|I_s|} \delta_i \geq \ell\right) = (*).$$

Applying Bernstein's inequality for $\ell = |I_s|/3$,

$$(*) \geq 1 - Pr\left(\left|\frac{1}{|I_s|}\sum_{i=1}^{|I_s|} \delta_i - \frac{1}{2}\right| \geq \frac{1}{6}\right) \geq 1 - 2\exp\left(-\left|I_s\right|/20\right),$$

and by the union bound,

$$Pr\left(\text{for every } s \in \mathcal{S},\ |\{i \in I_s : s(i) = \varepsilon_i\}| \geq \left|I_s\right|/3\right) \geq 1 - 2\left|\mathcal{S}\right| \exp\left(-|I_s|/20\right) \geq 1 - 2\exp(-k).$$

□

Let r be as in Theorem 3.11 for the class $H_{f^*} = F - f^*$ and set Ω' to be the event on which the assertion of Theorem 3.11 holds. Using the notation of that theorem, consider $r = 2r_Q$ and the set V_r. For every $v \in V_r$ and a sample $(X_1, \ldots, X_N) \in \Omega'$, let $I_v = \{i : \kappa_0 r \leq |v(X_i)| \leq c_1 r/\sqrt{\varepsilon}\}$,

$$s_v = \left(\operatorname{sgn}(v(X_i)) \cdot \mathbb{1}_{I_v}(X_i)\right)_{i=1}^N \text{ and } \mathcal{S} = \{s_v : v \in V_r\} \subset \{-1, 0, 1\}^N.$$

By Theorem 3.11, $Pr(\Omega') \geq 1 - 2\exp(-c_2\varepsilon^2 N)$ and on Ω',

$$|\mathcal{S}| \leq \exp\left(\varepsilon N/1000\right) \text{ and } \min_{v \in V} |I_v| \geq \varepsilon N/2.$$

Lemma 4.4 *Conditioned on Ω', with probability at least $1 - 2\exp(-c_0\varepsilon N)$ with respect to the uniform measure on $\{-1, 1\}^N$, the following holds. For every $h \in H_{f^*}$ with $\|h\|_{L_2} \geq r$, there is subset $\mathcal{I}_h \subset \{1, \ldots, N\}$ of cardinality at least $\varepsilon N/24$, and for every $i \in \mathcal{I}_h$,*

$$(\kappa_0/2)\left\|h\right\|_{L_2} \leq |h(X_i)| \leq c_1\left(\kappa_0 + 1/\sqrt{\varepsilon}\right)\left\|h\right\|_{L_2} \text{ and } \operatorname{sgn}(h(X_i)) = \varepsilon_i.$$

Proof Fix $h \in H$ with $\|h\|_{L_2} = r$ and let $\pi h = v \in V_r$ be as in Theorem 3.11. Recall that there is a subset $J_h \subset I_v$ consisting of at least $3/4$ of the coordinates of I_v, such that for every $j \in J_h$,

$$(\kappa_0/2)r \leq \left|h(X_j)\right| \leq c_1\left(\kappa_0 + 1/\sqrt{\varepsilon}\right)r \text{ and } \operatorname{sgn}(h(X_j)) = \operatorname{sgn}(v(X_j)).$$

Applying Lemma 4.3 to the set $\mathcal{S} = \{s_v : v \in V_r\}$ for $k = \varepsilon N/1000$, and noting that for every $s_v \in \mathcal{S}$, $|\{i : s_v(i) \neq 0\}| \geq \varepsilon N/2 \geq 40k$, it follows that with probability at least $1 - 2\exp(-c_2\varepsilon N)$ (relative to the uniform measure on $\{-1, 1\}^N$), for every $v \in V_r$, $s_v(i) = \varepsilon_i$ on at least $1/3$ of the coordinates that belong to I_v.

Since the set J_h contains at least $3/4$ of the coordinates of I_v and $v(X_i) = \varepsilon_i$ on at least a $1/3$ of the coordinates of I_v it follows that on the coordinates that belong to the intersection of these two sets (at least $1/12$ of the coordinates in I_v), both conditions hold, as claimed.

Finally, the claim is positive homogeneous and because H_{f^*} is star-shaped around 0, it holds on the same event when $\|h\|_{L_2} \geq r$. □

Corollary 4.5 *There exist absolute constants c_0 and c_1 for which the following holds. Let F and W be as above. With probability at least $1 - 2\exp(-c_0\varepsilon^2 N)$ with respect to the product measure $(X \otimes W)^N$, for every $f \in F$ with $\|f - f^*\|_{L_2} \geq 2r_Q$ there is a (random) subset $\mathcal{J}_f \subset \{1, \ldots, N\}$ of cardinality at least $\varepsilon N/100$, and for every $j \in \mathcal{J}_f$,*

1. $(\kappa_0/2)\|f - f^*\|_{L_2} \leq |(f - f^*)(X_j)| \leq c_1(\kappa_0 + 1/\sqrt{\varepsilon})\|f - f^*\|_{L_2}$,
2. $\mathrm{sgn}((f - f^*)(X_i)) = \mathrm{sgn}(-W)$, *and*
3. $\kappa_1\|W\|_{L_2} \leq |W_j| \leq c_2\|W\|_{L_2}/\sqrt{\varepsilon}$.

Proof Recall that W is symmetric and therefore it has the same distribution as $\eta|W|$ for a symmetric $\{-1, 1\}$-valued random variable η that is independent of $|W|$ and of X.

Considering $(\eta_i|W_i|)_{i=1}^N$, a direct application of Lemma 4.4 shows that with probability at least $1 - 2\exp(-c_0\varepsilon^2 N)$, if $\|f - f^*\|_{L_2} \geq 2r_Q$, there is a subset $\mathcal{I}_f \subset \{1, \ldots, N\}$ of cardinality at least $\varepsilon N/24$, and for every $i \in \mathcal{I}_f$,

$$(\kappa_0/2)\left\|f - f^*\right\|_{L_2} \leq \left|(f - f^*)(X_j)\right| \leq c_1\left(\kappa_0 + 1/\sqrt{\varepsilon}\right)\left\|f - f^*\right\|_{L_2}$$

and

$$\mathrm{sgn}\left((f - f^*)(X_i)\right) = \mathrm{sgn}(-\eta_i).$$

The final component is that for many of the coordinates in $\mathcal{I}_f$, $|W_i| \sim \|W\|_{L_2}$. Indeed, by excluding the largest and smallest $\varepsilon N/200$ coordinates of $(|W_i|)_{i \in \mathcal{I}_f}$, one obtains a subset $\mathcal{J}_f \subset \mathcal{I}_f$ of cardinality at least $\varepsilon N/100$, and for every $j \in \mathcal{J}_f$,

$$W^*_{N(1-\varepsilon/200)} \leq \left|W_j\right| \leq W^*_{\varepsilon N/200},$$

where $(W_i^*)_{i=1}^N$ is the non-increasing rearrangement of $(|W_i|)_{i=1}^N$.

Observe that by Lemma 3.1 applied to $\varepsilon' = \varepsilon/200$, with probability at least $1 - 2\exp(-c_2 N\varepsilon)$,

$$W^*_{\varepsilon N/200} \leq c_3\varepsilon^{-1/2}\left\|W\right\|_{L_2}.$$

Moreover, recalling that $Pr(|W| \leq \kappa_1\|W\|_{L_2}) \leq \varepsilon/1000$, a straightforward application of a binomial estimate shows that with probability at least $1 - 2\exp(-c_4 N\varepsilon)$, there are at most $\varepsilon N/200$ W_i's that satisfy $|W_i| < \kappa_1\|W\|_{L_2}$. Therefore, on that event,

$$W^*_{(1-\varepsilon/200)N} \geq \kappa_1 \|W\|_{L_2},$$

completing the proof. □

Proof of Theorem 4.2 The convexity of ℓ implies that $\mathcal{Q}_{f-f^*}$ is nonnegative. Consider the event from Corollary 4.5, and given $f \in F$ for which $\|f - f^*\|_{L_2} \geq 2r_Q$ let $\mathcal{J}_f \subset \{1, \ldots, N\}$ be the set of coordinates as above: that is, for every $j \in \mathcal{J}_f$,

$$(\kappa_0/2)\left\|f - f^*\right\|_{L_2} \leq \left|(f - f^*)\right|(X_j) \leq c_1\left(\kappa_0 + 1\sqrt{\varepsilon}\right)\left\|f - f^*\right\|_{L_2}$$

and if $j \in \mathcal{J}_f$, $-W_j$ and $(f - f^*)(X_j)$ share the same sign—say positive. Thus, the mid-point Z_j belongs to the interval whose end-points are $t_1 = \kappa_1\|W\|_{L_2}$ and $t_2 = c_1(\kappa_0 + 1\sqrt{\varepsilon})\|f - f^*\|_{L_2} + c_2\|W\|_{L_2}/\sqrt{\varepsilon}$, implying that

$$P_N\mathcal{Q}_{f-f^*} \geq \frac{1}{N}\sum_{j\in\mathcal{J}_f} \ell''(Z_i)(f - f^*)^2(X_i) \geq c_3\varepsilon\rho(t_1, t_2)\kappa_0^2\left\|f - f^*\right\|_{L_2}^2.$$

□

Next, consider the general noise scenario, in which $\xi = f^*(X) - Y$ need not be independent of X, nor does it necessarily satisfy a small-ball condition.

It is straightforward to verify that the only place in the proof above where the assumption that ξ and X are independent has been used, was to find a large subset of $\{1, \ldots, N\}$ on which $(f - f^*)(X_i)$ and ξ_i share the same sign. Also, the assumption that the noise satisfies a small-ball condition is only used to show that many of the $|\xi_i|$'s are sufficiently large—at least of the order of $\|\xi\|_{L_2}$. Both components are not needed if one wishes to show that $|Z_i| \leq c(\kappa_0, \varepsilon)(\|\xi\|_{L_2} + \|f - f^*\|_{L_2})$ for a proportional number of coordinates.

Indeed, with high probability, if $\|f - f^*\|_{L_2} \geq 2r_Q$, there is a subset of $\{1, \ldots, N\}$ of cardinality at least $\varepsilon N/100$ on which

$$|Z_i| \lesssim \left(\kappa_0 + 1/\sqrt{\varepsilon}\right)\cdot\left(\left\|f - f^*\right\|_{L_2} + \|\xi\|_{L_2}\right).$$

Formally, one has:

Theorem 4.6 *There exist absolute constants c_0, c_1 and c_2 for which the following holds. Let F be as above, set $Y \in L_2$ and put $\xi = f^*(X) - Y$. Then, with probability at least $1 - 2\exp(-c_0\varepsilon^2 N)$, for every $f \in F$ with $\|f - f^*\|_{L_2} \geq 2r_Q$,*

$$P_N\mathcal{Q}_{f-f^*} \geq c_1\varepsilon\kappa_0^2\rho(0, t)\left\|f - f^*\right\|_{L_2}^2,$$

for $t = c_2(\kappa_0 + \varepsilon^{-1/2})\cdot(\|f - f^*\|_{L_2} + \|\xi\|_{L_2})$.

Remark 4.7 (1) The assumption in Theorem 4.2 that the noise is independent of X allows one to obtain a positive lower bound on t_1 of the order of the variance $\|W\|_{L_2}$. This is significant when the loss function is not strongly convex in a large enough neighbourhood of zero (for example, when $\ell(t) = t^p$ for $p > 2$).

(2) When F is bounded in L_2, one may use the trivial bound $\|f - f^*\|_{L_2} \leq \operatorname{diam}(F, L_2)$ and replace t_2 by $c(\varepsilon, \kappa_0)(\|\xi\|_{L_2} + \operatorname{diam}(F, L_2))$. This is of little importance when ℓ'' decreases slowly, but matters a great deal when, for example, it has a compact support. Consider, for example, the Huber loss with parameter γ. If $\gamma \sim \|\xi\|_{L_2} + \operatorname{diam}(F, L_2)$ then $\rho(0, t_2) = 1$, but as stated, for a smaller value of γ, $\rho(0, t) = 0$ leading to a useless estimate on the quadratic component.
It turns out that one may improve Theorem 4.6 dramatically by ruling-out functions in F for which $\|f - f^*\|_{L_2}$ is significantly larger than $\|\xi\|_{L_2}$ as potential empirical minimizers, implying that one may set $t = c(\kappa_0, \varepsilon)\|\xi\|_{L_2}$ rather than $t = c(\kappa_0, \varepsilon)(\|\xi\|_{L_2} + \operatorname{diam}(F, L_2))$. We present this additional exclusion argument in Section 5.1.

5 Error estimates and oracle inequalities

Let us explore the multiplier component of the process, defined by $f \to \frac{1}{N}\sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i)$.

The following complexity term may be used to control the multiplier process, and is similar to the one used in [19].

Definition 5.1 Given a loss function ℓ, let $\phi_N^\ell(r)$ be the random function

$$\phi_N^\ell(r) = \frac{1}{\sqrt{N}} \sup_{\left\{f \in F : \|f - f^*\|_{L_2} \leq r\right\}} \left| \sum_{i=1}^N \varepsilon_i \ell'(\xi_i)(f - f^*)(X_i) \right|.$$

Note that the randomness of $\phi_N^\ell(r)$ is in $(\xi_i)_{i=1}^N$, $(X_i)_{i=1}^N$ and the independent random signs $(\varepsilon_i)_{i=1}^N$.

Set

$$r'_M(\kappa, \delta) = \inf\left\{r > 0 : Pr\left(\phi_N^\ell(r) \leq r^2\kappa\sqrt{N}\right) \geq 1 - \delta\right\},$$

recall that $H_{f^*} = F - f^*$, put

$$r_0(\kappa) = \inf\left\{r : \sup_{h \in H_{f^*} \cap rD} \left\|\ell'(\xi)h(X)\right\|_{L_2} \leq \sqrt{N}\kappa r^2/4\right\}$$

and let

$$r_M(\kappa, \delta) = r'_M(\kappa, \delta) + r_0(\kappa).$$

The function $\phi_N^\ell(r)$ and the definition of r_M arise naturally from a symmetrization argument. Indeed, it is well known that

$$
\begin{aligned}
&Pr\left(\sup_{h\in H_{f^*}\cap rD}\left|\frac{1}{N}\sum_{i=1}^{N}\ell'(\xi_i)h(X_i)-\mathbb{E}\ell'(\xi)h(X)\right|>x\right)\\
&\quad\le 2Pr\left(\sup_{h\in H_{f^*}\cap rD}\left|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\ell'(\xi_i)h(X_i)\right|>\frac{x}{4}\right), \qquad (5.1)
\end{aligned}
$$

provided that $x \geq 4N^{-1/2}\sup_{h\in H_{f^*}\cap rD}\|\ell'(\xi)h(X)\|_{L_2}$ (see, e.g. [8])

Lemma 5.2 *If F is a convex class of functions and $r = 2r_M(\kappa/4,\delta/2)$ then with probability at least $1-\delta$, for every $f\in F$ such that $\|f-f^*\|_{L_2}\geq r$, one has*

$$
\left|\frac{1}{N}\sum_{i=1}^{N}\ell'(\xi_i)(f-f^*)(X_i)-\mathbb{E}\ell'(\xi)(f-f^*)(X)\right|\leq\kappa\max\left\{\|f-f^*\|_{L_2}^2,r^2\right\}.
$$

Proof The class F is convex and therefore $H_{f^*}=F-f^*$ is star-shaped around 0. Also, $r\geq r_0$, implying that $4N^{-1/2}\sup_{h\in H_{f^*}\cap rD}\|\ell'(\xi)h(X)\|_{L_2}\leq\kappa r^2$. By (5.1) for $x=\kappa r^2$, one has

$$
\begin{aligned}
&Pr\left(\sup_{h\in H_{f^*}\cap rD}\left|\frac{1}{N}\sum_{i=1}^{N}\ell'(\xi_i)h(X_i)-\mathbb{E}\ell'(\xi)h(X)\right|>\kappa r^2\right)\\
&\quad\le 2Pr\left(\phi_N^{\ell}(r)>\frac{\kappa r^2}{4}\right)\le\delta,
\end{aligned}
$$

because $r>r'_M(\kappa/4,\delta/2)$.

Using once again that H_{f^*} is star-shaped around 0, it follows that if $\|f-f^*\|_{L_2}\geq r$ then $r(f-f^*)/\|f-f^*\|_{L_2}\in H_{f^*}\cap rS(L_2)$; thus,

$$
\left|\frac{1}{N}\sum_{i=1}^{N}\ell'(\xi_i)(f-f^*)(X_i)-\mathbb{E}\ell'(\xi)(f-f^*)(X)\right|\leq\kappa\left\|f-f^*\right\|_{L_2}^2.
$$

□

The function $\phi_N^{\ell}(r)$ is a natural geometric parameter: it is the ‘weighted width’ of the coordinate projection of $H_{f^*}\cap rD$ in a random direction selected according to a symmetrized noise vector, which is a point-wise product of a vector that belongs to the combinatorial cube $\{-1,1\}^N$ and the ‘noise multipliers’ $(\ell'(\xi_i))_{i=1}^N$ for $\xi_i=f^*(X_i)-Y_i$.

The more standard counterparts of $\phi_N^{\ell}(r)$, appearing in Theorem 1.3 and in similar results of that flavour, are the random function

$$
\sup_{f\in F\cap rD_{f^*}}\left|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(f-f^*)(X_i)\right|;
$$

the conditional expectation—the so-called *Rademacher average*

$$\mathbb{E}_\varepsilon \left(\sup_{f \in F \cap r D_{f^*}} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right| \, \left| (X_i)_{i=1}^N \right. \right);$$

and its expectation with respect to both $(\varepsilon_i)_{i=1}^N \otimes (X_i)_{i=1}^N$. Those represent the width or average width of

$$P_\sigma(H_{f^*} \cap r D) = \left\{ \left((f - f^*)(X_i) \right)_{i=1}^N : f \in F, \ \left\| f - f^* \right\|_{L_2} \le r \right\}$$

relative to a generic noise model, given by the random vector $(\varepsilon_i)_{i=1}^N$ weighted by the constant $\|\ell\|_{\text{lip}}$. A contraction argument shows that the typical width relative to a 'generic noise' vector dominates the typical width relative to the natural noise vector $(\varepsilon_i \ell'(\xi_i))_{i=1}^N$, implying that ϕ_N^ℓ is inherently superior to the generic complexity term.

Remark 5.3 It is straightforward to verify that when F consists of heavy-tailed random variables or if Y is a heavy-tailed random variable then the random sets

$$\left\{ \left(\ell'(\xi_i)(f - f^*)(X_i) \right)_{i=1}^N \ : \ f \in F, \ \left\| f - f^* \right\|_{L_2} \le r \right\}$$

are weakly bounded. The unfortunate byproduct is that $\phi_N^\ell(r)$ may exhibit rather poor concentration around its conditional mean or its true mean. This is why r_M is defined using ϕ_N^ℓ rather than by the conditional mean or by its true mean: unlike bounded problems or subgaussian ones, there might be a substantial gap between the fixed point defined using ϕ_N^ℓ and the one defined using those means.

Combining the bounds on the quadratic and multiplier terms with Theorems 2.3 and 2.4, one has the following:

Theorem 5.4 *For every κ_0 and $0 < \varepsilon < 1$ there exist constants c_0, c_1, c_2 and c_3 that depend only on κ_0 and ε, and an absolute constant c_4 for which the following holds. Let F be a convex class of functions and assume that $F - F$ satisfies a small-ball property with constants κ_0 and ε. Set $t_1 = 0$, $t_2 = c_0(\kappa_0, \varepsilon)(\|\xi\|_{L_2} + \text{diam}(F, L_2))$, $\zeta_1 = c_1(\kappa_0, \varepsilon)$, $\zeta_2 = c_2(\kappa_0, \varepsilon)$ and put $\theta = c_3(\kappa_0, \varepsilon)\rho(t_1, t_2)$. Then,*

- *With probability at least $1 - \delta - 2\exp(-c_4 N \varepsilon^2)$,*

$$\left\| \hat{f} - f^* \right\|_{L_2} \le 2 \max \left\{ r_Q(\zeta_1, \zeta_2), r_M(\theta/16, \delta/2) \right\}.$$

- *If ℓ satisfies Assumption 2.2 with a constant β, then with the same probability estimate,*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \le 2(\theta + \beta) \max \left\{ r_Q(\zeta_1, \zeta_2), r_M(\theta/16, \delta/2) \right\}.$$

Proof By Theorem 4.6, there is an absolute constant c_0 and an event of probability at least $1-2\exp(-c_0\varepsilon^2 N)$, on which, if $\|f-f^*\|_{L_2} \geq 2r_Q$ then

$$P_N \mathcal{Q}_{f-f^*} \geq \theta \left\| f-f^* \right\|_{L_2}^2 .$$

And, by Lemma 5.2, on an event of probability at least $1-\delta$, if $\|f-f^*\|_{L_2} \geq 2r_M(\theta/16, \delta/2)$, then

$$\left| P_N \mathcal{M}_{f-f^*} \right| \leq (\theta/4) \left\| f-f^* \right\|_{L_2}^2 .$$

Using the notation of Theorems 2.3 and 2.4, the first event is $\mathcal{B}$, the second in $\mathcal{A}$ and the claim follow. □

Theorem 5.4 is close to the result we would like to establish, with one significant step still missing: t_2 is not of the order of $\|\xi\|_{L_2}$. This is of little significance in the strongly convex case, but requires an additional argument when dealing with a general convex loss. That final step is presented next.

5.1 The main results

Let us begin by showing how to improve the choice of t_2 from $c(\kappa_0, \varepsilon)(\|\xi\|_{L_2} + \mathrm{diam}(F, L_2))$ to the potentially much smaller $2c(\kappa_0, \varepsilon)\|\xi\|_{L_2}$. To that end, we show that with high probability, the empirical minimizer does not belong to the set

$$\left\{ f \in F : \left\| f-f^* \right\|_{L_2} \geq \max\left\{ \|\xi\|_{L_2}, 2r_Q \right\} \right\} .$$

Once that is established, the study of ERM in F may be reduced to the set $F \cap \max\{\|\xi\|_{L_2}, 2r_Q\} D_{f^*}$, and in which case, Theorem 5.4 may be used directly, as the diameter of the class in question is $\sim \max\{\|\xi\|_{L_2}, r_Q\}$.

Recall that

$$\mathcal{Q}_{f-f^*} = \int_{\xi}^{\xi+(f-f^*)(X)} \left(\ell'(w) - \ell'(\xi) \right) dw. \tag{5.2}$$

Using Theorem 4.6, there are absolute constants c_0, c_1 and c_2 for which, with probability at least $1-2\exp(-c_0\varepsilon^2 N)$, if $\|f-f^*\|_{L_2} \geq 2r_Q$, then

$$P_N \mathcal{Q}_{f-f^*} \geq c_1 \varepsilon \kappa_0^2 \rho(0,t) \left\| f-f^* \right\|_{L_2}^2 , \tag{5.3}$$

where $t = c_2(\kappa_0 + \varepsilon^{-1/2}) \cdot (\|f-f^*\|_{L_2} + \|\xi\|_{L_2})$.

Let $\theta = c_1\varepsilon\kappa_0^2\rho(0,t)$ for $t = 2c_2(\kappa_0 + \varepsilon^{-1/2}) \max\{\|\xi\|_{L_2}, r_Q\}$ and assume further that

$$r_M(\theta/16, \delta/2) \leq \max\left\{ \|\xi\|_{L_2}, 2r_Q \right\} .$$

Theorem 5.5 *On an event of probability at least* $1 - \delta - 2\exp(-c_0 N\varepsilon^2)$,

$$\left\| \hat{f} - f^* \right\|_{L_2} \leq \max\left\{ \|\xi\|_{L_2}, 2r_Q \right\}.$$

The proof of Theorem 5.5 is based on several observations.

Note that if $h \in F$ and $\|h - f^*\|_{L_2} > R$, there is some $\lambda > 1$ and $f \in F$ for which $\|f - f^*\|_{L_2} = R$ and $\lambda(f - f^*) = (h - f^*)$. Indeed, set $\lambda = \|h - f^*\|_{L_2}/R > 1$ and put $f = h/\lambda + (1 - 1/\lambda)f^*$; by convexity, $f \in F$. Hence, for every $R > 0$,

$$\begin{aligned}&\left\{ h - f^* : h \in F, \ \left\| h - f^* \right\|_{L_2} \geq R \right\} \\ &\quad \subset \left\{ \lambda(f - f^*) : \lambda \geq 1, \ f \in F, \ \left\| f - f^* \right\|_{L_2} = R \right\}. \end{aligned} \tag{5.4}$$

Lemma 5.6 *On the event where* (5.3) *holds, if* $\|f - f^*\|_{L_2} = \max\{\|\xi\|_{L_2}, 2r_Q\}$ *and* $\lambda \geq 1$ *then*

$$P_N \mathcal{Q}_{\lambda(f-f^*)} \geq \lfloor \lambda \rfloor \theta \max\left\{ \|\xi\|_{L_2}^2, 4r_Q^2 \right\}.$$

Proof Fix $a, x \in \mathbb{R}$ and observe that for every $\lambda \geq 1$,

$$\int_a^{a+\lambda x} \left(\ell'(w) - \ell'(a) \right) dw \geq \lfloor \lambda \rfloor \int_a^{a+x} \left(\ell'(w) - \ell'(a) \right) dw. \tag{5.5}$$

To see this, let $x > 0$ and write

$$\begin{aligned}\int_a^{a+\lambda x} \left(\ell'(w) - \ell'(a) \right) dw &= \sum_{j=0}^{\lfloor \lambda \rfloor - 1} \int_{a+jx}^{a+(j+1)x} \left(\ell'(w) - \ell'(a) \right) dw \\ &+ \int_{a+\lfloor \lambda \rfloor}^{a+\lambda x} \left(\ell'(w) - \ell'(a) \right) dw.\end{aligned}$$

Since $\ell'(w) - \ell'(a)$ is an increasing function in w, the first term in the sum is the smallest and

$$\int_a^{a+\lambda x} \left(\ell'(w) - \ell'(a) \right) dw \geq \lfloor \lambda \rfloor \int_a^{a+x} \left(\ell'(w) - \ell'(a) \right) dw.$$

The case where $x < 0$ is equally simple.

When (5.5) is applied to (5.2), it follows that pointwise,

$$\mathcal{Q}_{\lambda(f-f^*)} = \int_{\xi}^{\xi + \lambda(f-f^*)(X)} \left(\ell'(w) - \ell'(\xi) \right) dw \geq \lfloor \lambda \rfloor \mathcal{Q}_{f-f^*},$$

and by the lower bound on $P_N \mathcal{Q}_{f-f^*}$ the claim follows. □

Proof of Theorem 5.5 Recall that we assume that $r_M(\theta/16, \delta/2) \leq \max\{\|\xi\|_{L_2}, 2r_Q\}$; hence, with probability at least $1-\delta$, if $\|f - f^*\|_{L_2} \leq \max\{\|\xi\|_{L_2}, 2r_Q\}$ then

$$\left|P_N \mathcal{M}_{f-f^*}\right| = \left|\frac{1}{N}\sum_{i=1}^{N} \ell'(\xi_i)(f-f^*)(X_i)\right| \leq (\theta/4)\max\left\{\|\xi\|_{L_2}^2, 4r_Q^2\right\}.$$

Also, $P_N\mathcal{M}_{f-f^*}$ is linear in $f - f^*$ and it follows that for every $\lambda \geq 1$,

$$\begin{aligned}\left|P_N \mathcal{M}_{\lambda(f-f^*)}\right| &= \left|\frac{1}{N}\sum_{i=1}^{N} \ell'(\xi_i)\lambda(f-f^*)(X_i)\right| = \lambda\left|P_N\mathcal{M}_{f-f^*}\right| \\ &\leq \lambda(\theta/4)\max\left\{\|\xi\|_{L_2}^2, 4r_Q^2\right\}.\end{aligned}$$

Combining this with the lower bound on $P_N\mathcal{Q}_{f-f^*}$ shows that with probability at least $1-\delta-2\exp(-c_0\varepsilon^2 N)$, if $\|f-f^*\|_{L_2} = \max\{\|\xi\|_{L_2}, 2r_Q\}$ and $\lambda \geq 1$ then

$$P_N\mathcal{Q}_{\lambda(f-f^*)} - \left|P_N\mathcal{M}_{\lambda(f-f^*)}\right| \geq \lambda(\theta/2)\max\left\{\|\xi\|_{L_2}^2, 4r_Q^2\right\} > 0.$$

Thus, by (5.4), on that event the empirical minimizer belongs to the set

$$F \cap \max\left\{\|\xi\|_{L_2}, 2r_Q\right\} D_{f^*}.$$

□

Now we are finally ready to formulate and prove the main results of the article.

Theorem 5.7 *For every κ_0 and $0 < \varepsilon < 1$ there exist constants c_0, c_1, c_2 and c_3 that depend only on κ_0 and ε, and an absolute constant c_4 for which the following holds.*

Let F be a convex class of functions and assume that $F-F$ satisfies the small-ball property with constants κ_0 and ε. Set $t_1 = 0$ and $t_2 = c_0(\varepsilon, \kappa_0)\|\xi\|_{L_2}$, $\zeta_1 = c_1(\varepsilon, \kappa_0)$ and $\zeta_2 = c_2(\varepsilon, \kappa_0)$. Put $\theta = c_3(\varepsilon, \kappa_0)\rho(t_1, t_2)$.

If $r_M(\theta/16, \delta/2) \leq \max\{\|\xi\|_{L_2}, 2r_Q(\zeta_1, \zeta_2)\}$, then with probability at least $1-\delta-2\exp(-c_4N\varepsilon^2)$,

- $\|\hat{f} - f^*\|_{L_2} \leq 2\max\{r_Q(\zeta_1, \zeta_2), r_M(\theta/16, \delta/2)\}$.
- *If ℓ satisfies Assumption 2.2 with a constant β then with the same probability estimate,*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \leq 2(\theta+\beta)\max\left\{r_Q(\zeta_1, \zeta_2), r_M(\theta/16, \delta/2)\right\}.$$

- *If ξ is independent of X and satisfies a small-ball condition with constants κ_1 and ε, one may take $t_1 = c_5\kappa_1\|\xi\|_{L_2}$ for a constant $c_5 = c_5(\varepsilon)$, and the two assertions formulated above hold as well.*

Proof By the preliminary exclusion argument of Theorem 5.5, with probability at least $1-\delta-2\exp(-c_0N\varepsilon^2)$, $\|\hat{f}-f^*\|_{L_2} \leq \max\{\|\xi\|_{L_2}, 2r_Q\}$. If $\|\xi\|_{L_2} \leq 2r_Q$ then Theorem 5.5 suffices to prove the assertion of Theorem 5.7. Otherwise, the assertion follows from Theorem 5.4, applied to the class $F \cap \|\xi\|_{L_2}D_{f^*}$. □

The second main result deals with the case in which ℓ is strongly convex in a neighbourhood of zero.

Theorem 5.8 *For every κ_0 and $0<\varepsilon<1$ there exist constants c_0, c_1, c_2 and c_3 that depend only on κ_0 and ε, and an absolute constant c_4 for which the following holds.*

Assume that ℓ is strongly convex in the interval $[-\gamma,\gamma]$ with a constant κ_2 and that $\|\xi\|_{L_2} \leq c_0\gamma$.

Assume further that F is a convex class of functions and that $F-F$ satisfies the small-ball property with constants κ_0 and ε. Set $\zeta_1 = c_1(\kappa_0,\varepsilon)$, $\zeta_2 = c_2(\kappa_0,\varepsilon)$ and $\theta = c_3(\kappa_0,\varepsilon)\kappa_2$.

If $r_M(\theta/16,\delta/2) \leq \gamma$, then with probability at least $1-\delta-2\exp(-c_4N\varepsilon^2)$,

- $\|\hat{f}-f^*\|_{L_2} \leq 2\max\{r_Q(\zeta_1,\zeta_2), r_M(\theta/16,\delta/2)\}$.
- *If ℓ satisfies Assumption 2.2 with a constant β then with the same probability estimate,*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \leq 2(\theta+\beta)\max\left\{r_Q(\zeta_1,\zeta_2), r_M(\theta/16,\delta/2)\right\}.$$

The proof of Theorem 5.8 is almost identical to that of Theorem 5.7, with one difference: when $\gamma > 2r_Q$ then instead of considering the preliminary exclusion argument of Theorem 5.5 at the level $\sim \max\{\|\xi\|_{L_2}, 2r_Q\}$, one performs preliminary exclusion at the level γ, and with an identical proof. The rest of the argument remains unchanged and we omit its details.

At this point, let us return to the rather detailed ‘wish list’ that was outlined in the introduction regarding the parameters governing prediction and estimation problems and see where we stand.

Theorems 5.7 and 5.8 lead to bounds on $\mathcal{E}_p$ and $\mathcal{E}_e$ without assuming that the class consists of uniformly bounded or subgaussian functions, nor that the target is even in L_p for some $p>2$. And, under a minor smoothness assumption, ℓ need not be a Lipschitz function. Thus, all the restrictions of the classical method and of subgaussian learning have been bypassed successfully.

As for the complexity parameters involved, r_Q is indeed an intrinsic parameter of the class F and has nothing to do with the choice of the loss or with the target. It does measure (with the very high probability of $1-2\exp(-c\varepsilon^2N)$), the L_2 diameter of the version space of F associated with f^*, and thus corresponds to the solution of the noise-free problem.

The noise and loss influence the problem in two places. In the quadratic component, the loss has to be calibrated to fit the noise level: the loss must be strongly convex in the interval $[0, c_1(\kappa_0,\varepsilon)\|\xi\|_{L_2}]$, or, when the noise is independent, it suffices that the loss is strongly convex in the smaller interval $[c_2(\kappa_1,\varepsilon)\|\xi\|_{L_2}, c_1(\kappa_0,\varepsilon)\|\xi\|_{L_2}]$. The strong convexity constant in those intervals also determines the level θ appearing in the definition of the multiplier component.

Although the noise and loss affect the quadratic component, their main impact is seen in the multiplier component, and thus in the external complexity parameter r_M. Indeed, while the quadratic component and the level $\sim \theta$ of a given class will be exactly the same for any loss that has the same strong convexity constant in the interval $[0, c_1(\kappa_0, \varepsilon)\|\xi\|_{L_2}]$, the difference between losses is 'coded' in r_M. And, as expected, the interaction between the class, the noise and the loss is captured by a single parameter: the correlation (or width) of a random coordinate projection of the localized class with the random vector $(\varepsilon_i \ell'(\xi_i))_{i=1}^N$.

Therefore, the 'wish list' is satisfied in full: without any boundedness assumptions and for a rather general loss function, prediction and estimation problems exhibit the expected two-regime behaviour: a 'low-noise' regime captured by an intrinsic parameter and a 'high-noise' regime by an external one. The exact nature of the loss and noise determines the external parameter *only* through the multiplier vector $(\varepsilon_i \ell'(\xi_i))_{i=1}^N$, and this random vector also determines where the phase transition between the high-noise regime, in which the external parameter r_M is dominant, and the low-noise one, in which the intrinsic parameter r_Q is dominant, takes place.

The one remaining issue still left open is to show that selecting the loss with some care may be used to negate the effects of noise related outliers.

6 Loss functions and the removal of outliers

Damaging outliers appear when sample points are far from where one would expect them to be, *and* the loss assigns a large value to those points. This combination means that outliers actually have a true impact on the empirical mean $P_N \mathcal{L}_f$ and therefore on the identity of the empirical minimizer.

The reason why outliers are not an issue in problems that feature a strong concentration phenomenon is clear: no matter what the loss is (as long as it does not grow incredibly quickly) only an insignificant fraction of the sample points fall outside the 'right area', and thus their impact is negligible.

The situation is different when either the class consists of heavy-tailed functions or when the noise $Y - f^*(X)$ is heavy-tailed. In such cases, a more substantial fraction of a typical sample falls in a potentially misleading location, and if the effect is amplified by a fast-growing loss, outliers become a problem that has to be contended with. This problem may be partly resolved by ensuring that the loss is not very big outside the 'expected area' of $[-c\|\xi\|_{L_2}, c\|\xi\|_{L_2}]$, which already hints towards the 'right choice' of a loss.

As noted previously, as long as ℓ is strongly convex in $[0, c_1(\kappa_0, \varepsilon)\|\xi\|_{L_2}]$ (or in the smaller interval when the noise is independent of X) the effects of the loss and the noise are coded in the vector $(\varepsilon_i \ell'(\xi_i))_{i=1}^N$. If ℓ is the squared loss, this vector is likely to have relatively many large coordinates when ξ is heavy-tailed. However, losses that grow almost linearly in $[a, \infty)$ lead to bounded multipliers because $|\ell'(\xi)| \lesssim |\ell'(a)|$. This results in a better-behaved multiplier component and a smaller fixed point r_M.

As examples, we focus on the three losses mentioned earlier: the squared loss, the log loss (2.2) and the Huber loss (2.1).

- The squared loss is the canonical example of a strongly convex loss with a bounded second derivative. However, it is susceptible to the problem of noise related outliers because it grows rapidly: $\ell'(\xi) \sim \xi$.
- The log-loss (2.2) exhibits a strongly convex behaviour in any bounded interval, but with a constant that decreases exponentially quickly to zero with the length of the interval. At the same time, its growth becomes close to linear for large values (i.e., ℓ' tends to a constant). Therefore, on the one hand the lower estimate on the quadratic component becomes trivial as $\|\xi\|_{L_2}$ increases, but the multiplier component displays a better behaviour than the squared loss.
- The Huber loss with parameter γ is strongly convex in $(-\gamma, \gamma)$ and grows linearly outside that interval. Hence, for $\gamma \sim \|\xi\|_{L_2}$ one may hope to get the best of both worlds—a quadratic component that behaves like its counterpart for the squared loss, and a multiplier component that behaves as if ℓ were linear.

We show that the three losses respond to outliers in very different ways. Our focus is on situations when F is subgaussian but ξ is potentially heavy-tailed, and the goal is to see if despite the heavy-tailed ξ, ERM performs in the same way as if ξ were subgaussian. To simplify the presentation and to give each loss a level playing field, we consider targets of the form $Y = f_0(X) + W$ for $f_0 \in F$ and W that is symmetric an independent of X. The reason for this choice of targets is that for any ‘legal’ loss function, $f^* = f_0$. Indeed, for any $f \in F$,

$$\mathbb{E}\ell(f(X) - Y) - \mathbb{E}\ell(f_0(X) - Y) \geq \mathbb{E}\ell'(W)(f - f_0)(X) = 0,$$

because ℓ' is odd and W is symmetric and independent of X. Hence, if the minimizer is also unique (as we assume), that minimizer must be f_0. In particular, for all the three losses $\mathrm{argmin}_{f \in F}\mathbb{E}\ell(f(X) - Y) = f_0$.

We first present general estimates on the performance of ERM in terms the parameters r_Q and r_M for a general target Y, and then control the parameters for an arbitrary convex, L-subgaussian class and a heavy-tailed target. As an example we consider case of $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^n\}$, which is the class of all linear functionals on $\mathbb{R}^n$.

In what follows $F \subset L_2$ is a closed, convex class of functions and $F - F$ satisfies a small-ball property with constants κ_0 and ε. The target one wishes to estimate is $Y \in L_q$ for some $q > 2$ and for the sake of simplicity we assume at times that $q = 4$, though that is not really an essential assumption.

6.1 Some facts on multiplier processes

The following is an upper estimate on multiplier and empirical processes indexed by a class that is L-subgaussian.

Theorem 6.1 [20] *There exists an absolute constant c_0 and for every $L > 1$ there are constants c_1 and c_2 that depend only on L and for which the following holds.*

Assume that $\Lambda \in L_q$ for some $q > 2$ and that F is L-subgaussian. Recall that $k_F = (\mathbb{E}\|G\|_F / d_F(L_2))^2$ and let $N \geq k_F$. If $(X_i, \Lambda_i)_{i=1}^N$ are N independent copies of (X, Λ) then

- *for $u > c_0$, with probability at least* $1 - 2\exp(-c_1 u^2 k_F)$,

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} f(X_i) - \mathbb{E} f \right| \le c_2 u \frac{\mathbb{E}\,\|G\|_F}{\sqrt{N}}.$$

- *for $u, \beta > c_0$, with probability at least* $1 - 2\beta^{-q} N^{-((q/2)-1)} - 2\exp(-c_1 u^2 k_F)$,

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \Lambda_i f(X_i) - \mathbb{E} \Lambda f \right| \le c_2 \beta u \, \|\Lambda\|_{L_q} \frac{\mathbb{E}\,\|G\|_F}{\sqrt{N}}.$$

6.2 The squared loss

If $\ell(t) = t^2$ then for every t_1, t_2, $\rho(t_1, t_2) = 2$. Therefore, by Theorem 5.8 for an arbitrarily large γ, it follows that with probability at least $1 - \delta - 2\exp(-c_1 \varepsilon^2 N)$,

$$\left\| \hat{f} - f^* \right\|_{L_2} \le \max \left\{ r_Q(\zeta_1, \zeta_2), r_M(c_2/4, \delta/2) \right\},$$

for constants ζ_1, ζ_2 and c_2 that depend only on κ_0 and ε.

Clearly, $\|\ell''\|_{L_\infty} \le 2$, implying that with the same probability estimate,

$$\mathbb{E} \mathcal{L}_{\hat{f}} \le 2(c_2 + 1) \max \left\{ r_Q^2(\zeta_1, \zeta_2), r_M^2(c_2/4, \delta/2) \right\}.$$

When F is, in addition, an L-subgaussian class, one may identify the parameters r_M and r_Q. Recall that $\|f\|_{\psi_2} \sim \sup_{p \ge 2} \|f\|_{L_p}/\sqrt{p}$ (which, as noted previously, suffices to ensure that the small-ball property holds for $F - F$ with constants κ_0 and ε that depend only on L).

Recall that

$$r_{2,Q}(\zeta_2) = \inf \left\{ r > 0 : \mathbb{E} \sup_{f \in F \cap r D_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right| \le \zeta_2 r \sqrt{N} \right\};$$

setting $F_r = \{f - f^* : f \in F \cap r D_{f^*}\}$, it follows from Theorem 6.1 that

$$\mathbb{E} \sup_{f \in F \cap r D_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right| \le c_3 \mathbb{E}\,\|G\|_{F_r}.$$

Therefore,

$$r_Q(\zeta_1, \zeta_2) \le \inf \left\{ r > 0 : \mathbb{E}\,\|G\|_{F_r} \le c_4 \min\{\zeta_1, \zeta_2\} r \sqrt{N} \right\}.$$

Turning to r_M, one has to identify

$$r_0 = \inf \left\{ r > 0 : \sup_{f \in F \cap r D_{f^*}} \left\| \xi (f - f^*)(X) \right\|_{L_2} \leq \sqrt{N} r^2 (c_2/16) \right\}$$

and the 'lowest' level r for which

$$Pr \left(\frac{1}{\sqrt{N}} \sup_{f \in F \cap r D_{f^*}} \left| \sum_{i=1}^{N} \varepsilon_i \xi_i (f - f^*)(X_i) \right| \leq r^2 (c_2/4) \sqrt{N} \right) \geq 1 - \delta.$$

Applying the $L_4 - L_2$ norm equivalence, $\| f - f^* \|_{L_4} \leq 2L \| f - f^* \|_{L_2}$ and

$$\left\| \xi (f - f^*)(X) \right\|_{L_2} \leq 2L \, \| \xi \|_{L_4} \, r \lesssim_L \sqrt{N} r^2,$$

provided that $r \gtrsim \| \xi \|_{L_4} / \sqrt{N}$. Moreover, by Theorem 6.1 for $q = 4$, it follows that with probability at least $1 - 2/(\beta^4 N) - 2 \exp(-c_3(L) u^2 k_{F_r})$,

$$\sup_{f \in F_r} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \xi_i (f - f^*)(X_i) \right| \leq c_4(L) \beta u \, \| \xi \|_{L_4} \, \mathbb{E} \, \| G \|_{F_r} \, .$$

Fix $0 < \delta < 1$. If $k_{F_r} \geq \log(2/\delta)$ one may take $u = c_5(L)$ and if the reverse inequality is satisfied, one may set $u \sim_L (k_{F_r}^{-1} \log(2/\delta))^{1/2}$, leading to a probability estimate of $1 - \delta$. Therefore, if

$$u(r, \delta) = c_6(L) \left(1 + k_{F_r}^{-1} \log(2/\delta) \right)^{1/2}$$

and

$$\beta \sim \max \left\{ \frac{1}{(\delta N)^{1/4}}, 1 \right\},$$

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

$$\sup_{f \in F_r} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \xi_i (f - f^*)(X_i) \right| \leq c_7(L) \beta u \, \| \xi \|_{L_4} \, \mathbb{E} \, \| G \|_{F_r} \, ,$$

and

$$r_M(c_2/4, \delta/2) \leq \frac{\| \xi \|_{L_4}}{\sqrt{N}} + \inf \left\{ r > 0 : \mathbb{E} \, \| G \|_{F_r} \leq c_8(L) \sqrt{N} \, \| \xi \|_{L_4}^{-1} \, (\beta u)^{-1} r^2 \right\}. \tag{6.1}$$

Thus, $r_M(c_2/4, \delta/2)$ dominates $r_Q(\zeta_1, \zeta_2)$ as long as $\| \xi \|_{L_4}$ is not very small.

As a point of reference, consider the case in which the Dvoretzky–Milman dimension k_{F_r} is at least of the order of $\log(2/\delta)$, and thus the infimum in (6.1) is attained for a value r for which $u(r,\delta)=c(L)$. Therefore,

$$r_M(c_2/4,\delta/2)\leq\frac{\|\xi\|_{L_4}}{\sqrt{N}}+\inf\left\{r>0:\mathbb{E}\,\|G\|_{F_r}\leq c_9(L)\sqrt{N}\,\|\xi\|_{L_4}^{-1}\beta^{-1}r^2\right\}. \quad (6.2)$$

The difference between (6.2) and the analogous estimate in the purely subgaussian case (see [10]) is the factor β^{-1}. Since $\beta\sim\max\{1/(\delta N)^{1/4},1\}$, it causes a slower rate when the desired confidence level is high. Indeed, if $\delta\ll 1/N$, then $\beta\gg 1$ leading to a larger value of r_M than in a purely subgaussian problem, and to a weaker accuracy/confidence tradeoff that one would expect if ξ were subgaussian. This different accuracy/confidence tradeoff is caused by the noise-related outliers one encounters and is the price one pays for using the squared loss in a problem involving a potentially heavy-tailed noise ($\xi\in L_4$ rather than $\xi\in L_{\psi_2}$). As a result, the accuracy/confidence tradeoff has a polynomial dependence on $1/\delta$ rather than the logarithmic dependence ERM exhibits when ξ is subgaussian.

6.3 The log-loss

It is straightforward to verify that for $t\geq 0$, $\ell'(t)=1-2/(\exp(t)+1)$, and in particular $\ell'(t)\leq\min\{2t,1\}$. Also, $\ell''(t)=2\exp(t)/(\exp(t)+1)^2$, which is a decreasing function on $\mathbb{R}_+$ and is upper-bounded by 1. Therefore, ℓ satisfies Assumption 2.2, and one may verify that $\theta\sim\rho(t_1,t_2)\geq\exp(-c_1\|\xi\|_{L_2})$.

The difference between the log-loss and the squared loss can be seen in the behaviour of $r_M(\theta/16,\delta/2)$. While the multipliers for the squared loss are independent copies of $\ell'(\xi)=\xi$, for the log-loss one has $|\ell'(\xi)|\leq\min\{2|\xi|,1\}$. This almost linear growth of the loss outside [0, 1] helps one overcome the issue of noise-related outliers and leads to an improved accuracy/confidence tradeoff—a logarithmic dependence in $1/\delta$ rather than the polynomial one exhibited by the squared loss—but only when $\|\xi\|_{L_2}\sim 1$. The tradeoff deteriorates when $\|\xi\|_{L_2}$ is small (because of the multiplier component) and when $\|\xi\|_{L_2}$ is large (because of the quadratic component).

The improved tradeoff is based on a contraction argument: since $|\ell'(\xi)|\leq 1$,

$$\begin{aligned}&Pr\left(\sup_{f\in F\cap rD_{f^*}}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i\ell'(\xi_i)(f-f^*)(X_i)\right|>t\right)\\&\leq 2Pr\left(\sup_{f\in F\cap rD_{f^*}}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i(f-f^*)(X_i)\right|>\frac{t}{4}\right).\end{aligned}$$

Clearly, removing any dependence on the multipliers is a costly step when $\|\xi\|_{L_2}$ is very small, but when $\|\xi\|_{L_2}\sim 1$ the resulting estimate is a significant improvement over the estimate for the squared loss: if $u(r,\delta)$ is as defined above,

$$r_M(\theta/16, \delta/2) \le \frac{1}{\sqrt{N}} + \inf\left\{r > 0 : \mathbb{E}\,\|G\|_{F_r} \lesssim \theta u(r,\delta)\sqrt{N} r^2\right\};$$

hence, if $r_M \le \|\xi\|_{L_2}$, then with probability at least $1 - \delta - 2\exp(-c_0\varepsilon^2 N)$,

$$\left\|\hat{f} - f^*\right\|_{L_2} \lesssim 2\max\left\{r_Q, r_M\right\} \ \text{ and } \ \mathbb{E}\mathcal{L}_{\hat{f}} \lesssim \max\left\{r_Q^2, r_M^2\right\}.$$

Again, let us consider the case in which $\inf\{r > 0 : \mathbb{E}\|G\|_{F_r} \lesssim \theta u(r,\delta)\sqrt{N} r^2\}$ is attained by a value r for which $u(r,\delta) = c(L)$. Then,

$$r_M(\theta/16, \delta/2) \le \frac{1}{\sqrt{N}} + \inf\left\{r > 0 : \mathbb{E}\,\|G\|_{F_r} \lesssim \theta\sqrt{N} r^2\right\},$$

which is the same bound on r_M one has for the squared loss when ξ is a standard gaussian random variable that is independent of X, though we only know that $\|\xi\|_{L_2} \sim 1$.

The improved accuracy/confidence tradeoff occurs when $\|\xi\|_{L_2} \sim 1$ simply because the log-loss happens to be calibrated for that noise level. This is a mere coincidence rather than premeditation, and the accuracy/confidence tradeoff of ERM relative to the log-loss deteriorates when $\|\xi\|_{L_2}$ is either very large or very small.

6.4 The Huber loss

Let r_Q be as above for suitable constants ζ_1 and ζ_2. For $\zeta_3 = \min\{\zeta_1, \zeta_2\}$ one has

$$r_Q = \inf\left\{r > 0 : \mathbb{E}\,\|G\|_{F_r} \le c\zeta_3 r\sqrt{N}\right\}$$

where $c = c(L)$. Without loss of generality, one may assume that $c\zeta_3$ is smaller than a fixed constant which will be the constant $c_3 = c_3(L)$ defined below.

Let $\ell(t)$ be the Huber loss with parameter $\gamma = c_0(L)\max\{\|\xi\|_{L_2}, r_Q\}$ for a constant c_0 to be specified later.

Observe that $|\ell'(t)| = \min\{|t/2|, \gamma\}$, that ℓ' is a Lipschitz function with constant 1 and that Assumption 2.2 is verified. Moreover, this setup falls within the scope of Theorem 5.8 for $\theta = c_1(L)$.

Regarding the multiplier component, note that if $\|f - f^*\|_{L_2} \le r$ then

$$\left\|\ell'(\xi)(f - f^*)\right\|_{L_2} \le \gamma\left\|f - f^*\right\|_{L_2} \le \gamma r \le (c_1(L)/16) r^2\sqrt{N},$$

provided that $r \gtrsim_L \gamma/\sqrt{N}$. Moreover, a contraction argument shows that

$$\begin{aligned} & Pr\left(\sup_{f\in F\cap rD_{f^*}}\left|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\ell'(\xi_i)(f-f^*)(X_i)\right|>t\right) \\ & \leq 2Pr\left(\sup_{f\in F\cap rD_{f^*}}\left|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(f-f^*)(X_i)\right|>\frac{t}{4\gamma}\right) \end{aligned}$$

and if $u(r,\delta)$ is as defined above, one has

$$r_M \leq c_2(L)\left(\frac{\max\left\{\|\xi\|_{L_2}, r_Q\right\}}{\sqrt{N}} + \inf\left\{r>0 : \gamma\mathbb{E}\,\|G\|_{F_r} \leq c_1(L)u(r,\delta)\sqrt{N}r^2\right\}\right).$$

Again, let us consider the case in which $\inf\{r>0 : \gamma\mathbb{E}\|G\|_{F_r} \leq c_2u(r,\delta)\sqrt{N}r^2\}$ is attained for a value r for which $u(r,\delta)=c'(L)$. Hence, setting $c_3=c_1(L)u$,

$$r_M \leq c_2\left(c_0\frac{\max\left\{\|\xi\|_{L_2}, r_Q\right\}}{\sqrt{N}} + \inf\left\{r>0 : \gamma\mathbb{E}\,\|G\|_{F_r} \leq c_3\sqrt{N}r^2\right\}\right), \quad (6.3)$$

which matches the estimate on r_M for the squared loss when $\|\xi\|_{\psi_2}\sim\|\xi\|_{L_2}$.

If $r_Q \leq \|\xi\|_{L_2}$ then $\gamma = c_0\|\xi\|_{L_2}$, and since $c_3 \geq c\zeta_3$, $r=\|\xi\|_{L_2}$ belongs to the set $\{r>0 : c_0\|\xi\|_{L_2}\mathbb{E}\|G\|_{F_r} \leq c_3\sqrt{N}r^2\}$ in (6.3). Therefore,

$$r_M \leq c_2\left(1+\frac{c_0}{\sqrt{N}}\right)\|\xi\|_{L_2} \leq c_0\,\|\xi\|_{L_2} = \gamma$$

as long as $N \geq c_4(L)$ and for a well chosen c_0. Hence, the assumption of Theorem 5.8 is verified.

Otherwise, $\|\xi\|_{L_2} \leq r_Q$, and in which case, r_Q belongs to the set in (6.3) implying that

$$r_M \leq c_2\left(1+\frac{c_0}{\sqrt{N}}\right)r_Q \leq \gamma.$$

Therefore, by Theorem 5.8, with probability at least $1-\delta-2\exp(-c_0\varepsilon^2N)$,

$$\left\|\hat{f}-f^*\right\|_{L_2} \lesssim \max\left\{r_M, r_Q\right\} \text{ and } \mathbb{E}\mathcal{L}_{\hat{f}} \lesssim \max\left\{r_M^2, r_Q^2\right\}.$$

The right choice of γ in the Huber loss leads to the optimal interval of strong convexity $[0, c\max\{\|\xi\|_{L_2}, r_Q\}]$, and to a far better accuracy/confidence tradeoff than exhibited by the squared loss for such a heavy-tailed problem. In fact, it matches the behaviour of the squared loss when $\|\xi\|_{\psi_2}\sim\|\xi\|_{L_2}$.

Let us present a concrete example in which all these rates can be computed explicitly, and which shows how the rates are affected by the choice of the loss.

For $T\subset\mathbb{R}^n$, set $F_T=\{\langle t,\cdot\rangle : t\in T\}$ and assume that μ is an isotropic, L-subgaussian measure on $\mathbb{R}^n$; that is, its covariance structure coincides with the standard

ℓ_2^n distance on $\mathbb{R}^n$ (isotropicity), and for every $t \in S^{n-1}$ and every $p \geq 2$, $\|\langle X, t\rangle\|_{L_p} \leq L\sqrt{p}\|\langle X, t\rangle\|_{L_2}$ (L-subgaussian). In particular, $F - F$ satisfies a small-ball property with constants that depend only on L.

We focus on the case $T = \mathbb{R}^n$. Clearly, for every $r > 0$ and any possible $f^* \in F$, $F_r = rB_2^n$, implying that $\mathbb{E}\|G\|_{F_r} \sim r\sqrt{n}$. Therefore, $k_{F_r} \sim \sqrt{n}$ and $\mathbb{E}\|G\|_{F_r} \leq \alpha\sqrt{N}r^2$ when $r \geq \alpha^{-1}\sqrt{n/N}$.

- **The squared loss.** It is straightforward to verify that if $N \geq c_1(L)n$, then with probability at least $1 - 2\exp(-c_2(L)N)$, $r_Q = 0$. Also,

$$u(r, \delta) = c_3(L)\left(1 + \sqrt{\frac{\log(1/\delta)}{N}}\right).$$

Using the definition of r_M, it is evident that with probability at least $1 - \delta - 2\exp(-c_4(L)N)$,

$$\|\hat{f} - f^*\|_{L_2} \lesssim_L \max\left\{\frac{1}{(N\delta)^{1/4}}, 1\right\} \cdot \left(1 + \sqrt{\frac{\log(1/\delta)}{n}}\right) \cdot \|\xi\|_{L_4}\sqrt{\frac{n}{N}}, \quad (6.4)$$

exhibiting once again that the accuracy/confidence tradeoff has a polynomial dependence in $1/\delta$.

- **The log-loss.** Let $t_2 \sim_L \|\xi\|_{L_2}$ and therefore, $\theta \sim_L \exp(-c_1\|\xi\|_{L_2})$. One has to take $N \geq c_2(\theta)n$ to ensure a nontrivial bound on r_Q, and in which case, $r_Q = 0$. Therefore, and in a similar way to the squared loss, with probability at least $1 - \delta - 2\exp(-c_3(L)N)$

$$\left\|\hat{f} - f^*\right\|_{L_2} \lesssim_L \exp\left(c_1(L)\,\|\xi\|_{L_2}\right)\left(1 + \sqrt{\frac{\log(1/\delta)}{n}}\right) \cdot \sqrt{\frac{n}{N}},$$

which is better than (6.4) in terms of the dependence on δ when $\|\xi\|_{L_2}$ is of the order of a constant and $\delta \ll 1/N$, but does not scale correctly with $\|\xi\|_{L_2}$ when the norm is either very small or very large. This was to be expected from the 'calibration' of the log-loss, which only fits a constant noise level, when $\|\xi\|_{L_2} \sim 1$.

- **The Huber loss.** As noted previously, for a nontrivial bound on r_Q one must take $N \geq c_1(L)n$, and in which case, $r_Q = 0$.
Fix $\gamma = c_2(L)\max\{\|\xi\|_{L_2}, r_Q\} = c_2(L)\|\xi\|_{L_2}$ and $\theta = c_2(L)$. Using the definition of r_M (because $|\ell'(\xi)| \leq \gamma$), it is evident that with probability at least $1 - \delta - 2\exp(-c_3(L)N)$,

$$\left\|\hat{f} - f^*\right\|_{L_2} \leq c_4(L)\left(\sqrt{\frac{\log(1/\delta)}{n}} + 1\right) \cdot \|\xi\|_{L_2}\sqrt{\frac{n}{N}}.$$

This is the optimal accuracy/confidence tradeoff for any choice of $\|\xi\|_{L_2}$ and coincides with the optimal tradeoff for the squared loss when ξ is gaussian and independent of X (see, e.g. [10]).

Unlike the squared loss and the log-loss, the optimal rate is obtained thanks to this choice of the Huber loss and despite the noise being heavy-tailed. This phenomenon occurs specifically because the Huber loss can be calibrated to fit the noise level of the problem and the intrinsic complexity of the class. It indicates that the right choice of loss can be used to effectively remove outliers caused by a heavy-tailed noise.

7 Final remarks

Because we study general loss functions, one should take care when comparing the results on the accuracy/confidence tradeoff to the tradeoff exhibited by ERM for the squared loss. Specifically, one should keep in mind that different loss functions may lead to different minimizers, and $f^* = \operatorname{argmin}_{f \in F} \mathbb{E}\ell(f(X)-Y)$ need not be the same as $\operatorname{argmin}_{f \in F} \mathbb{E}(f(X)-Y)^2$. As noted previously, one generic situation in which the 'best in the class' is the same for every convex loss is when $Y = f_0(X) + W$ for $f_0 \in F$ and W that is symmetric and independent of X.

Our results indicate that the choice of the Huber loss with parameter $\gamma \sim \|\xi\|_{L_2}$ leads to the removal of outliers—with a dramatic effect when the class is 'well-behaved' and the cause of the outliers is the heavy-tailed noise. The effect is diminished when F itself happens to consist of heavy-tailed functions: although it is still true that

$$Pr\left(\sup_{f \in F \cap rD_{f^*}} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \ell'(\xi_i)(f-f^*)(X_i)\right| > t\right) \leq 2Pr\left(\sup_{f \in F \cap rD_{f^*}} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i (f-f^*)(X_i)\right| > t/4\gamma\right), \qquad (7.1)$$

the random variable $\sup_{f \in F \cap rD_{f^*}} \left|\sum_{i=1}^{N} \varepsilon_i (f-f^*)(X_i)\right|$ is likely to be large with a non-negligible probability. As a result, the accuracy/confidence tradeoff exhibited by ERM is far from the conjectured performance of an optimal procedure, even when using a well-calibrated Huber loss.

Unfortunately, the suboptimal accuracy/confidence tradeoff of ERM is a fact of life and not an artifact of the argument we use. The fundamental instability of ERM is caused by the oscillations in (7.1) and cannot be overcome by a wise choice of a loss. Thus, the optimal tradeoff has to be based on a different procedure and not on ERM. Recently, a procedure that attains the optimal accuracy/confidence tradeoff for the squared loss was introduced in [14], and the optimal tradeoff performance was guaranteed under minimal assumptions on the class and on the target.

Finally, it is natural to ask whether it is possible to choose the right calibration for the Huber loss in a data-dependent manner—for example, to use the given data to identify a suitable upper estimate on $\|\xi\|_{L_2}$. For the sake of brevity we will sketch the method one may use and only consider the case in which the true minimizer is the same as for the squared loss[4]. Data dependent estimates on $\|\xi\|_{L_2}$ are possible using arguments similar to those introduced in [21], where a high-probability, uniform, crude isomorphic estimator of L_2 distances was constructed. More precisely, one may construct a data-dependent functional ϕ, which, given a function in a class H whose L_2 norm is above a critical level r^*, returns $\phi(h)$ such that $\alpha\|h\|_{L_2} \leq \phi(h) \leq \beta\|h\|_{L_2}$ for suitable constants α and β, and when $\|h\|_{L_2} \leq r^*$ then $\phi(h) \leq \beta r^*$. Such a functional may be constructed for the class $H = \{f(X) - Y : f \in F\}$, and by selecting $\bar{h} = \operatorname{argmin}_{h \in H} \phi(h)$ and using a standard median-of-means estimator to bound $\mathbb{E}h^2$, one may obtain a high probability estimator for $\|\xi\|_{L_2}$.

References

1. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. **33**(4), 1497–1537 (2005)
2. Birgé, L., Massart, P.: Rates of convergence for minimum contrast estimators. Probab. Theory Relat. Fields **97**(1–2), 113–150 (1993)
3. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013). ISBN 978-0-19-953525-5
4. Bühlmann, P., van de Geer, S.: Statistics for high-dimensional data. In: Springer Series in Statistics. Methods, Theory and Applications, pp. xviii, 556. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20192-9
5. de la Peña, V.H., Giné, E.: Decoupling. From Dependence to Independence, Randomly Stopped Processes. U-Statistics and Processes. Martingales and Beyond. Probability and Its Applications. Springer, New York (1999)
6. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition, Volume 31 of Applications of Mathematics. Springer, New York (1996)
7. Dudley, R.M.: Uniform Central Limit Theorems, Volume 63 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1999)
8. Giné, E., Zinn, J.: Some limit theorems for empirical processes. Ann. Probab. **12**(4), 929–998 (1984). With discussion
9. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, Volume 2033 of Lecture Notes in Mathematics. Lectures from the 38th Probability Summer School Held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Springer, Heidelberg (2011)
10. Lecué, G., Mendelson, S.: Learning subgaussian classes: upper and minimax bounds. Technical Report, CNRS, Ecole polytechnique and Technion (2013)
11. Ledoux, M.: The Concentration of Measure Phenomenon. American Mathematical Society, Providence, RI (2001)
12. Ledoux, M., Talagrand, M.: Probability in Banach Spaces, volume 23 of Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Isoperimetry and Processes. Springer, Berlin (1991)
13. Lee, W.S., Bartlett, P.L., Williamson, R.C.: The importance of convexity in learning with squared loss. In: Proceedings of the Ninth Annual Conference on Computational Learning Theory, pp. 140–146. ACM Press (1996)
14. Lugosi, G., Mendelson, S.: Risk minimization by median-of-means tournaments. (2016). https://arxiv.org/abs/1608.00757
15. Massart, P.: Concentration Inequalities and Model Selection, Volume 1896 of Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory Held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Springer, Berlin (2007)

[4] One may show that under rather reasonable conditions, if f^*_γ is the true minimizer of the Huber loss with parameter γ and f^* is the true minimizer of the squared loss then $\|f^*_\gamma(X) - Y\|_{L_2} \lesssim \|f^*(X) - Y\|_{L_2}$.

16. Mendelson, S.: Improving the sample complexity using global data. IEEE Trans. Inf. Theory **48**(7), 1977–1991 (2002)
17. Mendelson, S.: Obtaining fast error rates in nonconvex situations. J. Complex. **24**(3), 380–397 (2008)
18. Mendelson, S.: Learning without concentration for general loss functions. Technical Report, Technion (2014). http://arxiv.org/abs/1410.3192
19. Mendelson, S.: Learning without concentration. J. ACM **62**(3), Art. 21, 25 (2015)
20. Mendelson, S.: Upper bounds on product and multiplier empirical processes. Stoch. Process. Appl. **126**(12), 3652–3680 (2016)
21. Mendelson, S.: On aggregation for heavy-tailed classes. Probab. Theory Relat. Fields (2016). doi:10.1007/s00440-016-0720-6
22. Milman, V.D., Schechtman, G.: Asymptotic Theory of Finite-Dimensional Normed Spaces, Volume 1200 of Lecture Notes in Mathematics. With an appendix by M. Gromov. Springer, Berlin (1986)
23. Pisier, G.: The Volume of Convex Bodies and Banach Space Geometry, Volume 94 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge (1989)
24. Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, New York (2009) (Revised and extended from the 2004 French original, Translated by Vladimir Zaiats)
25. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer Series in Statistics. Springer, New York (1996). (With applications to statistics)