JMLR: Workshop and Conference Proceedings vol 35:1–15, 2014

Learning without Concentration

Shahar Mendelson SHAHAR@TX.TECHNION.AC.IL
Department of Mathematics, Technion, I.I.T., Israel

Abstract

We obtain sharp bounds on the convergence rate of Empirical Risk Minimization performed in a convex class and with respect to the squared loss, without any boundedness assumptions on class members or on the target.

Rather than resorting to a concentration-based argument, the method relies on a 'small-ball' assumption and thus holds for heavy-tailed sampling and heavy-tailed targets. Moreover, the resulting estimates scale correctly with the 'noise level' of the problem.

When applied to the classical, bounded scenario, the method always improves the known estimates.

1. Introduction

The aim of this note is to study the error rate of Empirical Risk Minimization (ERM), performed in a convex class, and relative to the squared loss.

To be more precise, given a class of real-valued functions $\mathcal{F}$ on a probability space (Ω, μ) and an unknown target function Y, one would like to find some function in $\mathcal{F}$ that is 'closest' to Y in some sense.

A rather standard way of measuring how close Y is to $\mathcal{F}$ is by using the squared loss $\ell(t) = t^2$ to capture the 'point-wise distance' $(f(x) - y)^2$, and being 'close' is measured by averaging that point-wise distance. Hence, the goal of the learner is to identify the function $f^* \in \mathcal{F}$ that minimizes $\mathbb{E}\ell(f(X) - Y) = \|f(X) - Y\|_{L_2}^2$ in $\mathcal{F}$, assuming, of course, that such a minimizer exists.

Unlike questions in Approximation Theory, the point in prediction problems is to identify f^* using random data – an independent sample $(X_i, Y_i)_{i=1}^N$ selected according to the joint distribution defined by the underlying measure μ and the target Y.

A more 'statistical' way of describing this problem is finding the function that minimizes the average cost of a mistake that is incurred by predicting $f(X)$ instead of Y. If the cost of a mistake is $(f(X) - Y)^2$, the functional one would like to minimize is the average cost $\mathbb{E}(f(X) - Y)^2 = \|f(X) - Y\|_{L_2}^2$.

Of course, the choice of the squared loss for measuring the point-wise cost of an error is only one possibility out of many, and although the new results presented here focus on the squared loss, the claims extend far beyond that case (see the discussion in Section 7).

One way of using the given data $(X_i, Y_i)_{i=1}^N$ is by selecting a random element in $\mathcal{F}$, denoted by $\hat{f}$, that minimizes the empirical loss

$$P_N \ell_f = \frac{1}{N} \sum_{i=1}^N \ell(f(X_i) - Y_i)^2.$$

With this choice, $\hat{f}$ is called the *empirical minimizer*, and the procedure that selects $\hat{f}$ is Empirical Risk Minimization (ERM).

There are various ways in which one may measure the success of ERM and the effectiveness of the choice of $\hat{f}$, and here, we will focus on the following one: that with high probability over the random samples $(X_i, Y_i)_{i=1}^N$, ERM produces a function that is close in L_2 to the best approximation of the target Y in $\mathcal{F}$; that is, an upper estimate on $\|\hat{f} - f^*\|_{L_2}^2 = \mathbb{E}\left((\hat{f} - f^*)^2(X)|(X_i, Y_i)_{i=1}^N\right)$ for the empirical minimizer $\hat{f}$ selected according to the data $(X_i, Y_i)_{i=1}^N$, that holds with sufficiently high probability.

The other typical notion is measured by an *oracle inequality*. Although the method presented here may be used to establish an oracle inequality, we will not present it here. Rather, we refer the reader to Mendelson (b), in which this question, and others – related to different loss functionals are explored.

The starting point of this note is a well known result that deals with this very question: controlling the distance between the random function produced by ERM and f^*. Theorem 1.1 formulated below is from Bartlett et al. (2005) (see Corollary 5.3 there and also Theorem 5.1 in the survey Koltchinskii (2011)).

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

$$k_N(r) = \mathbb{E} \sup_{f \in \mathcal{F} \cap 2r\mathcal{D}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right|, \tag{1.1}$$

where $(\varepsilon_i)_{i=1}^N$ are independent, symmetric, $\{-1, 1\}$-valued random variables that are independent of $(X_i)_{i=1}^N$.

Set

$$k_N^*(\gamma) = \inf \left\{ r > 0 : k_N(r) \leq \gamma r^2 \sqrt{N} \right\}.$$

Theorem 1.1 *There exist absolute constants c_0, c_1 and c_2 for which the following holds. If $\mathcal{F}$ is a convex class of functions that are bounded by 1 and the target Y is also bounded by 1, then for every $t > 0$, with probability at least $1 - c_0 \exp(-t)$,*

$$\|\hat{f} - f^*\|_{L_2}^2 \leq c_1 \max \left\{ \left(k_N^*(c_2)\right)^2, \frac{t}{N} \right\}. \tag{1.2}$$

The proof of Theorem 1.1 relies heavily on the fact that $\mathcal{F}$ consists of functions that are bounded by 1 and that the target is also bounded by 1. Both are restrictive assumptions that exclude many natural prediction problems that one would like to consider.

1. **Gaussian noise and heavy-tailed noise:** arguably the most basic statistical problem is when $Y = f_0(X) + W$ for some $f_0 \in \mathcal{F}$ and W that is a centred gaussian variable with variance σ that is independent of X. Thus, the given data consists of 'noisy' measurements of f_0 relative to a gaussian noise.

   Since a gaussian random variable is unbounded, Theorem 1.1 cannot be used to address a prediction problem that involves gaussian noise, regardless of the choice of $\mathcal{F}$. For the same reason, any kind of a prediction problem that involves a heavy-tailed target Y cannot be treated using Theorem 1.1.

2. **Gaussian regression:** Let $\mathcal{F}$ be a class of linear functionals indexed by $T \subset \mathbb{R}^n$ – that is, $\mathcal{F} = \{\langle t, \cdot \rangle : t \in T\}$ for some $T \subset \mathbb{R}^n$. If the underlying measure μ is the standard gaussian measure on $\mathbb{R}^n$, then for every $t \in T$, $f_t(X) = \langle t, X \rangle$ is unbounded. Thus, regardless of the target, it is impossible to apply Theorem 1.1 to a prediction problem that involves the class $\mathcal{F}$.

3. **General regression:** A class of linear functionals on $\mathbb{R}^n$ is not bounded unless the underlying measure μ has a compact support. What is rather striking is that even in problems in which μ has a compact support and which seemingly belong to the bounded framework, Theorem 1.1 is far from optimal.

   For example, let $B_1^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n |x_i| \leq 1\}$, set $T_R = RB_1^n$ for some $R > 0$ and put $\mathcal{F} = \{\langle t, \cdot \rangle : \ t \in T_R\}$. Assume that μ is supported in $\kappa B_\infty^n = \{x \in \mathbb{R}^n : \max_i |x_i| \leq \kappa\}$, and that it is isotropic; that is, for every $t \in \mathbb{R}^n$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{\ell_2^n}^2$, where $\|t\|_{\ell_2^n}^2 = \sum_{i=1}^n t_i^2$. An example of such a random vector is $X = (x_i)_{i=1}^n$ whose coordinate are independent, mean-zero, variance one random variables that are bounded by κ.

   Assume further that $Y = \langle t_0, \cdot \rangle + W$ for some $t_0 \in T_R$ and that W is independent of X and is also bounded by κ.

   Despite the fact that the prediction problem associated with the target Y and the class $\mathcal{F} = \{\langle t, \cdot \rangle : t \in T_R\}$ belongs to the bounded framework, we will indicate later that the estimate resulting from Theorem 1.1 on the Euclidean distance $\|\hat{t} - t_0\|_{\ell_2^n}$ is far from optimal, and scales incorrectly both with R and with the variance of W.

   Moreover, one may show that this poor outcome is endemic, and Theorem 1.1 leads to suboptimal estimates in many generic learning problems, for example, in *compressed sensing* and *LASSO*.

The reason why Theorem 1.1 is restrictive is that its proof is based on concentration and contraction arguments. This forces one to deal with only classes of uniformly bounded functions and bounded targets, and the resulting error rate does not scale well with the 'noise level' – the distance between Y and f^*.

Here, we will address all these issues, and in particular show that there is no need for the coarse concentration or contraction methods that were used in the proof of Theorem 1.1 to obtain a bound on $\|\hat{f} - f^*\|_{L_2}$. The resulting bound holds in full generality – for almost every choice of convex class $\mathcal{F}$, measure μ and a target Y, and it scales correctly with the 'noise level'. Moreover, in the bounded framework, Theorem 2.2, formulated below, always improves the estimate from Theorem 1.1.

Finally, a word about the title of this note. The meaning of "Learning without concentration" is not that concentration methods are not used in the proof. Rather, it is there to highlight that prediction is possible even in situations where concentration is simply *false* – for example, when trying to obtain prediction bounds for classes of functions with 'heavy tails', e.g. that have a well behaved fourth moment, but are not in L_p for $p > 4$. An empirical mean of such a function concentrates poorly around its true mean, rendering standard methods of analysis useless. Despite that, learning is still possible: the key point is that concentration fails because of the 'upper tail' in the deviation estimate, while the lower one is true almost 'for free'. The new argument shows that this lower tail suffices when trying to prove prediction bounds.

2. The main result

To formulate the main result, one has to introduce the following complexity parameters.

Definition 2.1 *Let $\xi = f^*(X) - Y$, and given $(X_i, Y_i)_{i=1}^N$ set $\xi_i = f^*(X_i) - Y_i$. Let*

$$\phi_N(s) = \sup_{f \in \mathcal{F} \cap s\mathcal{D}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i (f - f^*)(X_i) \right|, \tag{2.1}$$

where $(\varepsilon_i)_{i=1}^N$ are, as always, independent, symmetric $\{-1,1\}$-valued random variables, that are also independent of $(X_i, Y_i)_{i=1}^N$. Set

$$\alpha_N^*(\kappa, \delta) = \inf \left\{ s > 0 : Pr\left(\phi_N(s) \leq \kappa s^2 \sqrt{N} \right) \geq 1 - \delta \right\}$$

and let

$$\beta_N^*(\kappa) = \inf \left\{ r > 0 : \mathbb{E} \sup_{f \in \mathcal{F} \cap r\mathcal{D}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \leq \kappa \sqrt{N} r \right\}.$$

The role of the two parameters will be explained in Section 3.

The key assumption on the class is the following:

Assumption 2.1 *Let $\mathcal{H}$ be a class of functions and set $Q_{\mathcal{H}}(u) = \inf_{h \in \mathcal{H}} Pr(|h| \geq u \|h\|_{L_2})$.*

Given a class of functions $\mathcal{F}$, let $\mathcal{F} - \mathcal{F} = \{f - h : f, h \in \mathcal{F}\}$. We will assume that there is some $u > 0$ for which $Q_{\mathcal{F}-\mathcal{F}}(u) > 0$.

The reason for the name 'weak small-ball' is that in the context of random vectors, a small-ball estimate refers to a lower bound on the measure of a ball (with respect to some norm) of radius ε. In the same context, all that is assumed here is a lower bound on the measure of each individual marginal.

In Section 6 we will present several generic examples showing that the weak 'small-ball' condition of Assumption 2.1 is indeed minimal, and in natural situations one may choose u and Q to be appropriate absolute constants. Also, Assumption 2.1 does not restrict the choice of possible targets, nor does it restrict the joint distribution of the target Y and X. It only implies that if $f - h \neq 0$, then $(f-h)/\|f-h\|_{L_2}$ cannot have 'too-much' weight arbitrarily close to zero.

Example. To give some sense of the meaning of Assumption 2.1, consider a measure μ on $\mathbb{R}^n$, and as normalization, assume that μ is isotropic (that is, if X is distributed according to μ then for every $t \in \mathbb{R}^n$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{\ell_2^n}^2$). As Assumption 2.1 is positive homogeneous, it suffices to consider $t \in S^{n-1}$, the unit Euclidean sphere in $\mathbb{R}^n$. Set $H_{t,u} = \left\{ x \in \mathbb{R}^n : |\langle t, x\rangle| \leq u \right\}$, and note that $H_{t,u} = t^\perp + \rho t$ for $t^\perp = \{x \in \mathbb{R}^n : \langle t, x\rangle = 0\}$ and $-u \leq \rho \leq u$; thus, $H_{t,u}$ is the 'slab' that is orthogonal to t and of 'width' u.

One may verify that if μ is absolutely continuous, then $V(t,u) = \mu(H_{t,u})$ is continuous in t and u. A compactness argument shows that there is some $u > 0$ for which $\sup_{t \in S^{n-1}} \mu(H_{t,u}) \leq 1/2$. Hence,

$$\inf_{t \in \mathbb{R}^n} \mu \left\{ |\langle t, \cdot \rangle| > u \|t\|_{\ell_2^n} \right\} \geq 1/2, \tag{2.2}$$

and the existence of a level u for which $Q_{\mathcal{F}-\mathcal{F}}(u) \geq 1/2$ for any class consisting of linear functionals on $\mathbb{R}^n$ follows from the absolute continuity of μ.

Naturally, there is no hope of obtaining a quantitative estimate on u from (2.2) with such a 'soft' assumption as absolute continuity. In what follows we will provide sufficient conditions ensuring that $Q_{\mathcal{F}-\mathcal{F}}(u)$ and u are well behaved.

2.1. The main result

Finally, let us formulate the result, with the notation used above.

Theorem 2.2 *Let $\mathcal{F} \subset L_2$ be a closed, convex class of functions and set $Y \in L_2$ to be the unknown target.*

Fix $\tau > 0$ for which $Q_{\mathcal{F}-\mathcal{F}}(2\tau) > 0$ and set $\kappa < \tau^2 Q_{\mathcal{F}-\mathcal{F}}(2\tau)/16$. For every $\delta > 0$, with probability at least $1 - \delta - \exp(-NQ^2_{\mathcal{F}-\mathcal{F}}(2\tau)/2)$,

$$\|\hat{f} - f^*\|_{L_2} \leq 2 \max\left\{ \alpha^*_N\left(\kappa, \delta/4\right), \beta^*_N\left(\frac{\tau Q_{\mathcal{F}-\mathcal{F}}(2\tau)}{16}\right)\right\}.$$

Highlights of the proof of Theorem 2.2 will be presented in Section 5.

While Theorem 2.2 is similar to Theorem 1.1 in the sense that it leads to a bound on $\|\hat{f} - f^*\|_{L_2}$ using a complexity parameter of the class $\mathcal{F}$, unlike Theorem 1.1, it holds with essentially no restrictions on the class or on the target. In particular, Theorem 2.2 is valid for an arbitrary target $Y \in L_2$ and for almost any class $\mathcal{F}$ that one may consider, and in particular, it may be applied in all the examples described in the introduction that fall outside the scope of Theorem 1.1, once Assumption 2.1 is satisfied.

To illustrate the clear advantages Theorem 2.2 has over Theorem 1.1, we will formulate one concrete example in which the complexity parameters involved can be easily computed: the Persistence Problem (see e.g., Bartlett et al. (2012) and references therein).

In the persistence framework, one studies a family of linear regression problems in the set $T = RB_1^n$ (the ℓ_1^n ball of radius R, centred at 0). The point is to identify the correct way in which the error scales with the radius R and the dimension n.

If the dimension n and the radius R are allowed to grow with the sample size N, one has to find conditions on $n(N)$ and $R(N)$ that still ensure that $\|\hat{f} - f^*\|_{L_2}$ tends to zero with high probability.

Formally, for every $R \geq 1$ let $\mathcal{F}_R = \{\langle t, \cdot \rangle : \|t\|_{\ell_1^n} \leq R\}$. For the sake of simplicity, assume further that $X = (\zeta_i)_{i=1}^n$ has iid, mean-zero variance 1 coordinates, that the unknown target is $Y = \langle t_0, \cdot \rangle + W$ for some $t_0 \in RB_1^n$ and a mean-zero, variance σ random variable W that is independent of X and which represents the noise. We will identify $\mathcal{F}_R$ with $RB_1^n = \{t \in \mathbb{R}^n : \|t\|_{\ell_1^n} \leq R\}$ in the natural way.

Question 2.3 *If $\hat{t} \in RB_1^n$ is selected by ERM using an N-sample $(X_i, Y_i)_{i=1}^N$, find a function $\rho(N, n, R, \sigma, \delta)$ for which $\|\hat{t} - t_0\|_{\ell_2^n} \leq \rho$ with probability at least $1 - \delta$.*

The following two statements summarize the outcomes of Theorem 1.1 and Theorem 2.2 in this case.

First, let us present the estimate resulting from Theorem 1.1.

Theorem 2.4 *For every $\kappa > 1$ there exist constants c_1, c_2 and c_3 that depend only on κ for which the following holds. Assume that $\|\zeta\|_{L_\infty}, \|W\|_{L_\infty} \leq \kappa$. Set*

$$\rho_N = \begin{cases} \frac{R^2}{\sqrt{N}} \sqrt{\log\left(\frac{2c_1 n}{\sqrt{N}}\right)} & \text{if } N \leq c_1 n^2 \\ \\ \frac{R^2 n}{N} & \text{if } N > c_1 n^2. \end{cases}$$

With probability at least $1 - 2\exp(-c_2 N \rho_N / R^2)$, ERM produces $\hat{t}$ that satisfies $\|\hat{t} - t_0\|_{\ell_2^n}^2 \leq c_3 \rho_N$.

The 'killer' term is $R^2/\sqrt{N}$ when $n \gtrsim \sqrt{N}$, which is a direct consequence of the contraction argument used in the proof of Theorem 2.2. Not only does it display the wrong behaviour when R is large, but also when the noise is low.

In contrast, the next result follows from Theorem 2.2. To formulate it, set

$$\|W\|_{L_{2,1}} = \int_0^\infty \sqrt{Pr(|W| > t)} dt.$$

This norm falls between the L_2 norm and any L_q norm for $q > 2$. The proof of Theorem 2.5 shows that $\|W\|_{L_{2,1}}$, which is slightly larger than the L_2 norm, captures the noise level of the problem. Since in virtually all examples the L_2 and $L_{2,1}$ norms are equivalent, we will abuse notation and denote $\sigma = \|W\|_{L_{2,1}}$ rather than $\sigma = \|W\|_{L_2}$.

Theorem 2.5 *For every $\kappa > 1$ there exist constants c_1, c_2, c_3 and c_4 that depend only on κ for which the following holds. Assume that $\|\zeta\|_{L_\infty}, \|W\|_{L_\infty} \leq \kappa$ and that $\|W\|_{2,1} = \sigma < \infty$. Put*

$$v_1 = \begin{cases} \frac{R^2}{N} \log\left(\frac{2c_1 n}{N}\right) & \text{if } N \leq c_1 n, \\ \\ 0 & \text{if } N > c_2 n. \end{cases}$$

and

$$v_2 \geq \begin{cases} \frac{R\sigma}{\sqrt{N}} \sqrt{\log\left(\frac{2c_2 n \sigma}{\sqrt{N} R}\right)} & \text{if } N \leq c_2 n^2 \sigma^2 / R^2 \\ \\ \frac{\sigma^2 n}{N} & \text{if } N > c_2 n^2 \sigma^2 / R^2. \end{cases}$$

Then with probability at least $1 - 2\exp\left(-c_3 N v_2 \min\left\{\frac{1}{\sigma^2}, \frac{1}{R}\right\}\right)$, $\|\hat{t} - t_0\|_{\ell_2^n}^2 \leq c_4 \max\{v_1, v_2\}$.

Observe that Theorem 2.5 yields a much better dependence of the error $\|\hat{t} - t_0\|_{\ell_2^n}^2$ on the parameters involved than Theorem 2.4. Moreover, the results of Lecué and Mendelson (a) show that Theorem 2.5 is optimal in the minimax sense, when W is a gaussian variable.

If one is willing to settle for a weaker probability estimate, something that cannot be helped when dealing with heavy-tailed ensembles, Theorem 2.2 can be used to tackle a more general scenario. For example, the assumption that $\|W\|_{L_\infty} \leq \kappa$ is only used to obtain a high probability estimate and can be removed. In fact, Theorem 2.5 can be improved even further: it holds for a general target Y rather than just for $Y = \langle t_0, \cdot \rangle + W$, the assumption that X has iid coordinates can be relaxed, and even 'heavy-tailed' measures μ may be used.

Unfortunately, the proof of a more general version of Theorem 2.5 comes at a high technical cost, and since it is not our main focus, we will not explore this direction further.

Remark 2.6 *It is well known that accurate bound for $T = RB_1^n$ is also the essence of the estimates on compressed sensing and LASSO problems Chafaï et al. (2012); Lecué and Mendelson (b,c), an appropriate version of Theorem 2.5, which studies ERM in a heavy-tailed persistence framework, may be used to obtain optimal estimates for heavy-tailed compressed sensing and LASSO.*

3. Two parameters for two regimes

When one considers the performance of a learning procedure, it is reasonable to expect two different 'performance regimes', according to the difficulties the learner faces. As will be explained below, there are clear differences between 'low noise' problems and 'high noise' ones, and each one of the two regimes is naturally associated with one of the two complexity parameters defined above.

3.1. Controlling the version space - quadratic estimates using β_N^*

Observe that β_N^* measures when the Rademacher averages of the 'localized' set $\{f - f^* : f \in \mathcal{F} \cap r\mathcal{D}_{f^*}\}$ scale like r rather than like the normalization r^2 that is used in the definition of k_N^* and in Theorem 1.1. Thus, when β_N^* is nontrivial, in the sense that $\beta_N^* < 1$, it is much smaller than k_N^*.

Off-hand, the meaning and significance of β_N^* is not obvious. It is, perhaps, surprising that it captures properties of the 'version space' of the problem.

Recall that the version space is a random subset of $\mathcal{F}$ that consists of all the functions in the class that agree with f^* on the sample $(X_i)_{i=1}^N$. When the problem is noise-free ($Y = f^*(X)$), a learning procedure is likely to make significant mistakes only when there are functions in $\mathcal{F}$ that, despite being 'far-away' from f^*, still satisfy that $f(X_i) = f^*(X_i)$ for every $1 \leq i \leq N$. For example, in empirical minimization bad mistakes occur in a noise-free problem only when the version space is large.

It seems plausible that when the noise level is low rather than zero, the situation does not change significantly and mistakes are essentially due to a large version space.

Thus, when trying to bound the error of ERM, the first order of business is to identify a parameter that captures the 'size' of the version space, and β_N^* gives that and much more.

One may show that with high probability, if $\|f - f^*\|_{L_2} \geq \beta_N^*$, then

$$\frac{1}{N}\sum_{i=1}^N (f - f^*)^2(X_i) \geq c\|f - f^*\|_{L_2}^2$$

for an appropriate constant c, and with (almost) no assumption on $\mathcal{F}$ (see Theorem 5.2 for the exact formulation).

One immediate outcome of this fact is that the version space cannot include functions for which $\|f^* - f\|_{L_2} \geq \beta_N^*$, and that sampling is 'stable' for functions that are not too close to f^*.

3.2. Controlling the interaction with the noise via α_N^*

The second regime is encountered once the noise level increases, and mistakes happen for a totally different reason: the 'interaction' of the target Y with class members, a phenomenon which is governed by α_N^*.

Although α_N^* and k_N^* seem similar and share the same scaling, of $\sim \sqrt{N}s^2$, the two differ on one key issue: while the correlation in the definition on α_N^* is relative to the actual noise faced by

the learner, represented by the random vector $(\varepsilon_i\xi_i)_{i=1}^N$, the random process used in the definition of k_N^*,

$$\sup_{f\in\mathcal{F}\cap s\mathcal{D}_{f^*}}|\frac{1}{\sqrt{N}}\sum_{i=1}^N\varepsilon_i(f-f^*)(X_i)|,$$

measures the correlation of the same random set with a random vector $(\varepsilon_i)_{i=1}^N$ that represents 'generic' noise and has nothing to do with the specific noise that one has to deal with. The random noise $(\varepsilon_i\xi_i)_{i=1}^N$ captures the 'true impact' of the noise ξ on the problem, a property that is totally missed when measuring the correlation of the random set with $(\varepsilon_i)_{i=1}^N$.

Note that if $\mathcal{F}$ and Y are bounded by 1, then ξ is bounded by 2. A standard contraction argument (see, e.g. Ledoux and Talagrand (1991)) shows that with probability at least $1/2$,

$$\sup_{f\in\mathcal{F}\cap s\mathcal{D}_{f^*}}\left|\sum_{i=1}^N\varepsilon_i\xi_i(f-f^*)(X_i)\right|\leq 4\mathbb{E}\sup_{f\in\mathcal{F}\cap s\mathcal{D}_{f^*}}\left|\sum_{i=1}^N\varepsilon_i(f-f^*)(X_i)\right| \tag{3.1}$$

and α_N^* is trivially smaller than k_N^* in the bounded case.

Therefore, while k_N^* is insensitive to the noise-level, because the contraction argument used in (3.1) destroys any dependence on ξ, α_N^* is highly affected by it and may become very small when ξ is close to 0, as should be expected from a parameter that captures the interaction of class members with the noise.

Remark. Although it seems that the complexity parameters α_N^* and β_N^* (and k_N^* as well) depend on the unknown function f^*, in virtually all applications one may replace the indexing set with $(\mathcal{F}-\mathcal{F})\cap r\mathcal{D}$, where $\mathcal{D}$ is the unit ball in $L_2(\mu)$. Moreover, if $\mathcal{F}$ is centrally symmetric (i.e. if $f\in\mathcal{F}$ then $-f\in\mathcal{F}$) in addition to being convex, then $\mathcal{F}-\mathcal{F}\subset 2\mathcal{F}$ and the indexing set becomes $2\mathcal{F}\cap r\mathcal{D}$.

4. The method of analysis

Next, let us explain why the heuristic description of the roles of the parameters α_N^* and β_N^* from the previous section is reasonable, and why splitting the prediction problem to two components, each captured by one of the two parameters, is the first step in bypassing the concentration-contraction mechanism used in proof of Theorem 1.1.

The excess loss functional $\mathcal{L}_f$ is defined by $\mathcal{L}_f(X,Y)=\ell_f(X,Y)-\ell_{f^*}(X,Y)=(f(X)-Y)^2-(f^*(X)-Y)^2$, and satisfies

$$\begin{aligned}\mathcal{L}_f(X,Y)=&(f-f^*)^2(X)+2(f-f^*)(X)(f^*(X)-Y)\\=&(f-f^*)^2(X)+2\xi(f-f^*)(X),\end{aligned} \tag{4.1}$$

where f^* is the unique minimizer of $\mathbb{E}(f(X)-Y)^2$ in $\mathcal{F}$ and $\xi=f^*(X)-Y$.

Let $\mathcal{L}_{\mathcal{F}}=\{\mathcal{L}_f:f\in\mathcal{F}\}$ be the excess loss class and note that it has two important properties. First, because $\mathcal{L}_{\mathcal{F}}$ is a shift of the loss class $\{(f(X)-Y)^2:f\in\mathcal{F}\}$ by the fixed function ℓ_{f^*}, an empirical minimizer of the loss is an empirical minimizer of the excess loss. Moreover, since $0\in\mathcal{L}_{\mathcal{F}}$, the empirical minimizer $\hat{f}$ satisfies that $P_N\mathcal{L}_{\hat{f}}\leq 0$. Therefore, if $(X_i,Y_i)_{i=1}^N$ is a sample for which

$$\{f\in\mathcal{F}:\|f-f^*\|_{L_2}\geq\rho\}\subset\{f\in\mathcal{F}:P_N\mathcal{L}_f>0\},$$

then $\|\hat{f} - f^*\|_{L_2} < \rho$ – simply because an empirical minimizer cannot belong to the set $\{f \in \mathcal{F} : \|f - f^*\|_{L_2} \geq \rho\}$.

To identify the random set $\{f \in \mathcal{F} : P_N\mathcal{L}_f > 0\}$, observe that by (4.1),

$$P_N\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) + \frac{2}{N}\sum_{i=1}^N \xi_i(f-f^*)(X_i),$$

and one may bound $P_N\mathcal{L}_f$ from below by showing that with high probability, and in rather general situations,

1. *the 'version space condition' holds*: that is, for every $f \in \mathcal{F}$ that satisfies $\|f-f^*\|_{L_2} \geq \beta_N^*(\kappa_1)$, one has $\frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) \geq c_1\kappa_1\|f-f^*\|_{L_2}^2$.
2. *the 'noise interaction' condition holds*: that is, for every $f \in \mathcal{F}$ that satisfies $\|f-f^*\|_{L_2} \geq \alpha_N^*(\kappa_2,\delta)$, one has $\left|\frac{1}{N}\sum_{i=1}^N \xi_i(f-f^*)(X_i)\right| \leq c_2\kappa_2\|f-f^*\|_{L_2}^2$.

Therefore, if both (1) and (2) hold, and if κ_2 is chosen to be smaller than $c_1\kappa_1/2c_2$, then with probability at least $1-c_3\delta$, $P_N\mathcal{L}_f > 0$ when $\|f-f^*\|_{L_2} \geq \max\{\alpha_N^*(\kappa_2,\delta), \beta_N^*(\kappa_1)\}$; on that event,

$$\|\hat{f} - f^*\|_{L_2} \leq \max\{\alpha_N^*(\kappa_2,\delta), \beta_N^*(\kappa_1)\}.$$

Let us emphasize that one may obtain a very good lower bound on the quadratic component of $P_N\mathcal{L}_f$ (and thus on the version space condition), solely because the required estimate is *one sided*. A two-sided bound, which is an upper estimate on

$$\sup_{f\in\mathcal{F}\cap s\mathcal{D}_{f^*}} \left|\frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) - \mathbb{E}(f-f^*)^2\right|,$$

requires $\mathcal{F}$ to have considerably more structure than what is needed for the lower estimate (see Mendelson et al. (2007); Mendelson (2010); Mendelson and Paouris (2012, 2014) for two-sided estimates on the quadratic process). The fact that $\frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) \gtrsim \mathbb{E}(f-f^*)^2$ when $\|f-f^*\|_{L_2}$ is large enough is (almost) universally true, and the quadratic term may be bounded from below (almost) for free, while the 'upper tail' results are simply not true in general.

5. Highlights of the Proof of Theorem 2.2

We begin this section with a few definitions needed for the proof of Theorem 2.2.

Definition 5.1 *A class $\mathcal{H}$ is star-shaped around 0 if for every $h \in \mathcal{H}$ and any $0 < \lambda \leq 1$, $\lambda h \in \mathcal{H}$. For every $\kappa > 0$, set $\beta_N(\mathcal{H},\kappa) = \inf\left\{r > 0 : \mathbb{E}\sup_{h\in\mathcal{H}\cap r\mathcal{D}} \left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \leq \kappa r\right\}$.*

We will sometimes write $\beta_N(\kappa)$ instead of $\beta_N(\mathcal{H},\kappa)$.

Observe that $\beta_N^*(\kappa) = \beta_N(\mathcal{F}-f^*,\kappa)$. Also, it is straightforward to verify that if $\mathcal{H}$ is star-shaped around 0 and $r > \beta_N(\kappa)$ then $\mathbb{E}\sup_{h\in\mathcal{H}\cap r\mathcal{D}} \left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \leq \kappa r$ (see, for example, the discussion in Lecué and Mendelson (a)).

The main component in the proof of Theorem 2.2 is the following:

Theorem 5.2 *Let $\mathcal{F} \subset L_2$ be a closed, convex class and assume that there is some $\tau > 0$ for which $Q_{\mathcal{F}-\mathcal{F}}(2\tau) > 0$. For every $f^* \in \mathcal{F}$, set $\mathcal{H} = \mathcal{F} - f^*$. Then, for every $r > \beta_N(\mathcal{H}, \tau Q(2\tau)/16)$, with probability at least $1 - 2\exp(-NQ^2_{\mathcal{H}}(2\tau)/2)$, for every $f \in \mathcal{F}$ that satisfies that $\|f - f^*\|_{L_2} \geq r$, one has*

$$|\{i : |(f - f^*)(X_i)| \geq \tau \|f - f^*\|_{L_2}\}| \geq N \frac{Q_{\mathcal{F}-\mathcal{F}}(2\tau)}{4}. \tag{5.1}$$

In other words, Theorem 5.2 implies that on a proportional subset of coordinates, $|(f - f^*)(X_i)| \geq \tau \|f - f^*\|_{L_2}$, provided that $\|f - f^*\|_{L_2} > \beta_N$.

The complete proof of Theorem 5.2 is presented in the Mendelson (a). It is based on the following uniform empirical small-ball estimate – which is of a similar nature to the results from Koltchinskii and Mendelson and Mendelson (c).

Theorem 5.3 *Let $S(L_2)$ be the L_2 unit sphere and set $\mathcal{H} \subset S(L_2)$. Assume that there is some $\tau > 0$ for which $Q_{\mathcal{H}}(2\tau) > 0$. If*

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i h(X_i) \right| \leq \frac{\tau Q_{\mathcal{H}}(2\tau)}{16},$$

then with probability at least $1 - 2\exp(-NQ^2_{\mathcal{H}}(2\tau)/2)$,

$$\inf_{h \in \mathcal{H}} |\{i : |h(X_i)| \geq \tau\}| \geq N \frac{Q_{\mathcal{H}}(2\tau)}{4}.$$

Proof. Recall that $P_N f = \frac{1}{N} \sum_{i=1}^N f(X_i)$ and $Pf = \mathbb{E} f(X)$. Note that for every $h \in \mathcal{H}$ and $u > 0$, $|\{i : |h(X_i)| \geq u\}| = N P_N \mathbb{1}_{\{|h| \geq u\}}$. Also,

$$P_N \mathbb{1}_{\{|h| \geq u\}} = P \mathbb{1}_{\{|h| \geq 2u\}} + \left(P_N \mathbb{1}_{\{|h| \geq u\}} - P \mathbb{1}_{\{|h| \geq 2u\}} \right) = (*).$$

Let $\phi_u : \mathbb{R}_+ \to [0, 1]$ be the function

$$\phi_u(t) = \begin{cases} 1 & t \geq 2u, \\ (t/u) - 1 & u \leq t \leq 2u, \\ 0 & t < u, \end{cases}$$

and observe that for every $t \in \mathbb{R}$, $\mathbb{1}_{[u,\infty)}(t) \geq \phi_u(t)$ and $\phi_u(t) \geq \mathbb{1}_{[2u,\infty)}(t)$. Hence,

$$\begin{aligned} (*) &\geq P \mathbb{1}_{\{|h| \geq 2u\}} + P_N \phi_u(|h|) - P \phi_u(|h|) \\ &\geq \inf_{h \in \mathcal{H}} Pr(|h| \geq 2u) - \sup_{h \in \mathcal{H}} |P_N \phi_u(|h|) - P \phi_u(|h|)|. \end{aligned}$$

Let $Z(X_1, ..., X_N) = \sup_{h \in \mathcal{H}} |P_N \phi_u(|h|) - P \phi_u(|h|)|$. By the bounded differences inequality applied to Z (see, for example, Boucheron et al. (2013)), it follows that for every $t > 0$, with probability at least $1 - 2\exp(-2t^2)$,

$$\sup_{h \in \mathcal{H}} |P_N \phi_u(|h|) - P \phi_u(|h|)| \leq \mathbb{E} \sup_{h \in \mathcal{H}} |P_N \phi_u(|h|) - P \phi_u(|h|)| + \frac{t}{\sqrt{N}}.$$

Note that ϕ_u is a Lipschitz function that vanishes in 0 and with a Lipschitz constant $1/u$. Therefore, by the Giné-Zinn symmetrization theorem Giné and Zinn (1984) and the contraction inequality for Bernoulli processes (see, e.g. Ledoux and Talagrand (1991)),

$$\mathbb{E}\sup_{h\in\mathcal{H}}|P_N\phi_u(|h|)-P\phi_u(|h|)|\leq\frac{4}{u}\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i h(X_i)\right|.$$

Hence, with probability at least $1-2\exp(-2t^2)$, for every $h\in\mathcal{H}$,

$$P_N\mathbb{1}_{\{|h|\geq u\}}\geq\inf_{h\in\mathcal{H}}Pr(|h|\geq 2u)-\frac{4}{u}\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i h(X_i)\right|-\frac{t}{\sqrt{N}}.$$

If $Q_{\mathcal{H}}(2\tau)>0$, set $u=\tau$ and $t=\sqrt{N}Q_{\mathcal{H}}(2\tau)/2$. Recall that

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i h(X_i)\right|\leq\frac{\tau Q_{\mathcal{H}}(2\tau)}{16},$$

and thus, with probability at least $1-2\exp(-NQ^2_{\mathcal{H}}(2\tau)/2)$, $|\{i:|h(X_i)|\geq\tau\}|\geq N\frac{Q_{\mathcal{H}}(2\tau)}{4}$. ■

6. Some Examples

It turns out that the weak small-ball assumption needed in Theorem 2.2 holds even in the extreme end of the 'boundedness spectrum': classes in which one has almost no moment control. In particular, this leads to prediction bounds in cases that are completely out of reach for the concentration-contraction based 'bounded theory'.

Lemma 6.1 *Let $\mathcal{F}$ be a class of functions on a probability space (Ω,μ).*

1. *If $\|f_1-f_2\|_{L_2}\leq\kappa_1\|f_1-f_2\|_{L_1}$ for every $f_1,f_2\in\mathcal{F}$, then there are constants c_1 and c_2 that depend only on κ_1 for which $Q(c_1)\geq c_2$.*

2. *If there are $p>2$ and κ_2 for which $\|f_1-f_2\|_{L_p}\leq\kappa_2\|f_1-f_2\|_{L_2}$ for every $f_1,f_2\in\mathcal{F}$, then there are constants c_1 and c_2 that depend only on κ_2 and p for which $Q(c_1)\geq c_2$.*

Lemma 6.1 is an immediate outcome of the Paley-Zygmund inequality (see, e.g. de la Peña and Giné (1999)) and its proof is omitted.

Let ζ be a mean-zero, variance 1 random variable and set $X=(\zeta_1,...,\zeta_n)$ to be a vector with independent coordinates, distributed according to ζ. Clearly, such a random vector is isotropic, since $\mathbb{E}\langle X,t\rangle^2=\|t\|^2_{\ell_2^n}$ for every $t\in\mathbb{R}^n$.

Lemma 6.2 *Let ζ and X be as above.*

1. *Assume that there is some $\kappa_1>0$ for which $\|\zeta\|_{L_2}\leq\kappa_1\|\zeta\|_{L_1}$. Then $\|\langle t,X\rangle\|_{L_2}\leq c_1\|\langle t,X\rangle\|_{L_1}$ for every $t\in\mathbb{R}^n$, and for a constant c_1 that depends only on κ_1.*

2. *If $\|\zeta\|_{L_p}\leq\kappa_2$ for some $p>2$ then $\|\langle t,X\rangle\|_{L_p}\leq c_2\|\langle t,X\rangle\|_{L_2}$ for every $t\in\mathbb{R}^n$ and for a constant c_2 that depends only on κ_2 and p.*

The proof of Lemma 6.2 as well as examples of a similar flavour may be found in Mendelson (a).

Using Lemma 6.2, one may obtain a prediction bound in the following generic regression problem. Let $X = (\zeta_i)_{i=1}^n$ be a random vector as above, set $T \subset \mathbb{R}^n$ to be a closed, convex set and put $\mathcal{F} = \{\langle t, \cdot \rangle : t \in T\}$. Consider a square-integrable target Y and let $f^* = \langle t^*, \cdot \rangle$ be the unique minimizer in $\mathcal{F}$ of $f \to \mathbb{E}(f(X) - Y)^2$. Note that if X is isotropic then for every $f_t = \langle t, \cdot \rangle$, $\|f_t - f^*\|_{L_2} = \|t - t^*\|_{\ell_2^n}$.

Corollary 6.3 *If either one of the moment conditions of Lemma 6.2 holds, then with probability at least* $1 - \delta - \exp(-c_1 N)$*, ERM produced* $\hat{t} \in T$ *for which*

$$\|\hat{t} - t^*\|_{\ell_2^n} \leq 2 \max\{\alpha_N^*(c_2, \delta/4), \beta_N^*(c_3)\}$$

for appropriate constants c_1*,* c_2 *and* c_3 *that depend only on* κ_1 *or on* κ_2 *and* p*.*

Needless to say that the situation in Corollary 6.3 falls outside the scope of Theorem 1.1.

7. Concluding Remarks

The fact that one parameter (β_N^*) depends on the average over samples and the other (α_N^*) does not is just an artifact of our presentation. It is possible to replace β_N^* with a parameter that is not averaged using a slightly different argument (see Mendelson (b)). Obviously, one may replace α_N^* with an averaged version, but when doing that, the resulting high probability estimate will depend on concentration – something one would rather avoid when tackling a heavy-tailed learning scenario.

Another observation is that although the results presented here are formulated for the squared loss, with some effort they can be extended well beyond that case.

To explain why the choice of the squared loss is not essential for the method presented above, consider a smooth, increasing and even function ℓ that satisfies $\ell(0) = 0$. The point-wise cost of predicting $f(X)$ instead of Y is $\ell(f(X) - Y) \equiv \ell_f(X, Y)$. As above, set f^* to be a minimizer of the functional $\mathbb{E}\ell(f(X) - Y)$ in $\mathcal{F}$.

Assumption 7.1 *Assume that* f^* *is unique, and setting* $\xi = f^*(X) - Y$*, assume that* $\mathbb{E}\ell'(\xi)(f - f^*)(X) \geq 0$*.*

Assumption 7.1 is not really restrictive. It is straightforward to verify that it holds, for example, when ℓ is convex and $\mathcal{F}$ is closed and convex, or when $Y = f_0(X) + W$ for an arbitrary class $\mathcal{F}$, $f_0 \in \mathcal{F}$ and an independent noise W.

Recall that the excess loss functional is $\mathcal{L}_f = \ell_f - \ell_{f^*}$, and that it shares similar properties to the ones mentioned for the squared excess loss: an empirical minimizer of the loss is an empirical minimizer of the excess loss, and $P_N \mathcal{L}_{\hat{f}} \leq 0$.

Given the data $(X_i, Y_i)_{i=1}^N$, and since $\mathbb{E}\ell'(\xi)(f - f^*) \geq 0$, a straightforward application of Taylor's expansion around $\xi_i = f^*(X_i) - Y_i$ shows that for every $f \in \mathcal{F}$

$$P_N \mathcal{L}_f \geq \left(\frac{1}{N} \sum_{i=1}^N \ell'(\xi_i)(f - f^*)(X_i) - \mathbb{E}\ell'(\xi)(f - f^*) \right) + \frac{1}{2N} \sum_{i=1}^N \ell''(Z_i)(f - f^*)^2(X_i), \quad (7.1)$$

for midpoints $(Z_i)_{i=1}^N$ that belong to the intervals whose ends are ξ_i and $\xi_i + (f - f^*)(X_i)$; thus, the intervals depend only on f and on the sample $(X_i, Y_i)_{i=1}^N$.

Just as in the squared-loss case, if $P_N\mathcal{L}_f > 0$ then f cannot be an empirical minimizer. Therefore, using (7.1), one may obtain a lower bound on $P_N\mathcal{L}_f$ by identifying the levels $\bar{\alpha}_N$ and $\bar{\beta}_N$, for which, if $\|f-f^*\|_{L_2} \geq \bar{\beta}_N$ then $\frac{1}{2N}\sum_{i=1}^N \ell''(Z_i)(f-f^*)^2(X_i) \geq c\|f-f^*\|_{L_2}^2$, for some constant c, and if $\|f-f^*\|_{L_2} \geq \bar{\alpha}_N$ then $\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \ell'(\xi_i)(f-f^*)(X_i)\right| \leq \frac{c}{4}\|f-f^*\|_{L_2}^2$. On that event, $\|\hat{f}-f^*\|_{L_2} \leq \max\{\bar{\alpha}_N, \bar{\beta}_N\}$.

To compare this situation with Theorem 2.2, observe that for the squared loss $\ell(t) = t^2$, $\ell''(Z_i) = 2$ regardless of Z_i and $\ell'(\xi_i) = \xi_i$. Hence, $\bar{\alpha}_N$ and $\bar{\beta}_N$ in the squared loss case lead to the parameters α_N^* and β_N^*. Also, if ℓ happens to be strictly convex, that is, if $\ell'' \geq \kappa > 0$, the mid-points Z_i need not play a real role in the lower bound on the quadratic term. However, when ℓ is only convex (or when it has areas in which it is concave), the role of the midpoints becomes more significant.

Indeed, if ℓ is convex, one has to identify $\bar{\beta}_N$ for which, if $\|f-f^*\|_{L_2} \geq \bar{\beta}_N$, there is a subset of $\{1,...,N\}$ of cardinality proportional to N (that depends on f and on the sample) on which both $|f-f^*|(X_i) \gtrsim \|f-f^*\|_{L_2}$ *and* $\ell''(Z_i) \geq \kappa_1$. This requires a more careful analysis than the proof of Theorem 2.2 and will be presented in Mendelson (b). It is also the reason behind the formulation of Theorem 5.2, which leads to information on a proportional subset of coordinates, rather than the analogous lemma from Koltchinskii and Mendelson in which a lower bound on the empirical L_2 norm of $f-f^*$ is obtained.

Acknowledgments

Partially supported by the Mathematical Sciences Institute – The Australian National University and by ISF grant 900/10.

References

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

P.L. Bartlett, S. Mendelson, and J. Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probability Theory and Related Fields*, 154:193–224, 2012.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

D. Chafaï, O. Guédon, G. Lecué, and A. Pajor. *Interacrions between compressed sensing, random matrices and high dimensional geometry*, volume 37. SMF, 2012.

V. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer-Verlag, 1999.

A.W. Van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12(4): 929–989, 1984.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture notes in Matehmatics*. Springer, 2011.

V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *preprint, arXiv:1312.3580*.

G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. *preprint, arXiv:1305.4825*, a.

G. Lecué and S. Mendelson. Compressed sensing under weak moment assumptions. *preprint, arXiv:1401.2188*, b.

G. Lecué and S. Mendelson. Necessary moment conditions for exact reconstruction via basis pursuit. *preprint, arXiv:1404.3116*, c.

M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, 1991.

S. Mendelson. Learning without concentration, extended version. *preprint, arXiv:1401.0304*, a.

S. Mendelson. Learning without concentration for general loss functions. *preprint*, b.

S. Mendelson. A remark on the diameter of random sections of convex bodies. *preprint, arXiv:1312.3608*, c.

S. Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geometric and Functional Analysis*, 20(4):988–1027, 2010.

S. Mendelson and G. Paouris. On generic chaining and the smallest singular values of random matrices with heavy tails. *Journal of Functional Analysis*, 262(9):3775–3811, 2012.

S. Mendelson and G. Paouris. On the singular values of random matrices. *Journal of the European Mathematics Society*, 2014.

S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.