Bernoulli **22**(3), 2016, 1520–1534
DOI: 10.3150/15-BEJ701

Performance of empirical risk minimization in linear aggregation

GUILLAUME LECUÉ[1] and SHAHAR MENDELSON[2]

[1]*CNRS, CMAP, Ecole Polytechnique, 91120 Palaiseau, France.*
E-mail: guillaume.lecue@cmap.polytechnique.fr
[2]*Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel. E-mail: shahar@tx.technion.ac.il*

We study conditions under which, given a dictionary $F = \{f_1, \ldots, f_M\}$ and an i.i.d. sample $(X_i, Y_i)_{i=1}^N$, the empirical minimizer in $\operatorname{span}(F)$ relative to the squared loss, satisfies that with high probability

$$R\big(\tilde{f}^{\mathrm{ERM}}\big) \leq \inf_{f \in \operatorname{span}(F)} R(f) + r_N(M),$$

where $R(\cdot)$ is the squared risk and $r_N(M)$ is of the order of M/N.

Among other results, we prove that a uniform small-ball estimate for functions in $\operatorname{span}(F)$ is enough to achieve that goal when the noise is independent of the design.

Keywords: aggregation theory; empirical processes; empirical risk minimization; learning theory

1. Introduction and main results

Let $(\mathcal{X}, \mu)$ be a probability space, set X to be distributed according to μ and put Y to be an unknown target random variable.

In the usual setup in learning theory, one observes N independent couples $(X_i, Y_i)_{i=1}^N$ in $\mathcal{X} \times \mathbb{R}$, distributed according to the joint distribution of X and Y. The goal is to construct a real-valued function f which is a good guess/prediction of Y. A standard way of measuring the prediction capability of f is via the risk $R(f) = \mathbb{E}(Y - f(X))^2$. The conditional expectation

$$R(\hat{f}) = \mathbb{E}\big(\big(Y - \hat{f}(X)\big)^2 | (X_i, Y_i)_{i=1}^N\big)$$

is the risk of the function $\hat{f}$ that is chosen by the procedure, using the observations $(X_i, Y_i)_{i=1}^N$.

There are many different ways in which one may construct learning procedures (see, e.g., the books [1,5,10,12,29,31] for numerous examples), but in general, there is no 'universal' choice of an optimal learning procedure.

The variety of learning algorithms motivated the introduction of aggregation or ensemble methods, in which one combines a batch or *dictionary*, created by learning procedures, in the hope of obtaining a function with 'better' prediction capabilities than individual members of the dictionary.

Aggregation procedures have been studied extensively (see, e.g., [7,9,13,14,26,30,33–35] and references therein), and among the more well-known aggregation procedures are boosting [28] and bagging [5].

1350-7265

Our aim is to explore the problem of *linear aggregation*: given a dictionary $F = \{f_1, \dots, f_M\}$, one wishes to construct a procedure $\tilde{f}$ whose risk is almost as small as the risk of the best element in the linear span of the dictionary, denoted by $\operatorname{span}(F)$; namely, a procedure which ensures that with high probability

$$R(\tilde{f}) \le \inf_{f \in \operatorname{span}(F)} R(f) + r_N(M). \tag{1.1}$$

This type of inequality is called an *oracle inequality* and the function f^* for which $R(f^*) = \inf_{f \in \operatorname{span}(F)} R(f)$ is called the *oracle*.

Of course, in (1.1) one is looking for the smallest possible residual term $r_N(M)$, that holds uniformly for all choices of couples (X, Y) and dictionaries F that satisfy certain assumptions.

The linear aggregation problem has been studied in [26] in the Gaussian white noise model; in [6,30] for the Gaussian model with random design; in [27] for the density estimation problem and in [3] in the learning theory setup, under moment conditions. And, based on these cases, it appears that the best possible residual term $r_N(M)$ that one may hope for is of the order of M/N.

This rate is usually called the *optimal rate of linear aggregation* and, in fact, its optimality holds in some minimax sense, introduced in [30].

The only procedure we will focus on here is empirical risk minimization (ERM) performed in the span of the dictionary:

$$\hat{f}^{\mathrm{ERM}} \in \operatorname*{arg\,min}_{f \in \operatorname{span}(F)} R_N(f) \qquad \text{where } R_N(f) = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - f(X_i)\big)^2.$$

We do not claim that ERM is always the best procedure for the linear aggregation problem, but rather, our aim is to identify conditions under which it achieves the optimal rate of M/N.

The benchmark result on the performance of ERM in linear aggregation is Theorem 2.2 in [3]. To formulate it, let F be a dictionary of cardinality M and set f^* to be the oracle in $\operatorname{span}(F)$ (i.e., $R(f^*) = \inf_{f \in \operatorname{span}(F)} R(f)$). We also denote by L_p for $1 \le p \le \infty$ the Banach spaces $L_p(\mathcal{X}, \mu)$, and in particular, $\|f\|_{L_2} = (\mathbb{E} f(X)^2)^{1/2}$.

Theorem 1.1 [3]. *Assume that* $\mathbb{E}(Y - f^*(X))^4 < \infty$ *and that for every* $f \in \operatorname{span}(F)$,

$$\|f\|_{L_\infty} \le \sqrt{B}\, \|f\|_{L_2}. \tag{1.2}$$

If $x > 0$ *satisfies that* $2/N \le 2\exp(-x) \le 1$ *and*

$$N \ge 1280 B^2 \left[3BM + x + \frac{16 B^2 M^2}{N} \right],$$

then with probability at least $1 - 2\exp(-x)$,

$$R\big(\hat{f}^{\mathrm{ERM}}\big) - R\big(f^*\big) \le 1920 B \sqrt{\mathbb{E}\big(Y - f^*(X)\big)^4} \left[\frac{3BM + x}{N} + \frac{16 B^2 M^2}{N^2} \right].$$

It follows from Theorem 1.1 that under an L_4 assumption on $Y - f^*(X)$ and the equivalence between the L_2 and L_∞ norms on the span of F, ERM achieves a rate of convergence of order B^2M/N when $N \geq cB^3M$ for an absolute constant c.

However, it should be noted that the best probability estimate one may obtain in Theorem 1.1 is $1 - 2/N$; also, it is possible to show that the constant B defined in (1.2) is *necessarily larger than the dimension M of* $\mathrm{span}(F)$. For the sake of completeness, we shall provide a proof of that fact in the Appendix. Therefore, the rate that Theorem 1.1 guarantees is, at best, of the order of M^3/N, to achieve that rate, at least $N \geq cM^4$ observations are needed, and even with that sample size, the probability estimate is, at best, $1 - 2/N$. This estimate is far from the anticipated rate of M/N, which should be achieved when $N \geq cM$ and preferably, with significantly higher probability.

Nevertheless, the optimal rate of M/N can be obtained by relaxing assumption (1.2) and using a different method of proof. Recall that the ψ_2 norm of a function f is

$$\|f\|_{\psi_2} = \inf\bigl\{C > 0\colon \mathbb{E}\exp\bigl(f^2(X)/C^2\bigr) \leq 2\bigr\}.$$

One may show that $\|f\|_{\psi_2} \leq c\|f\|_{L_\infty}$ for a suitable absolute constant c (see, e.g., Section 1 in [8]). Therefore, assuming that the ψ_2-norm and the L_2-norm are equivalent in $\mathrm{span}(F)$ is a weaker requirement than the one in (1.2). The assumption that for every $f \in \mathrm{span}(F)$,

$$\|f\|_{\psi_2} \leq \sqrt{C}\|f\|_{L_2}, \tag{1.3}$$

means that $\mathrm{span}(F)$ is a *sub-Gaussian class*, following the definition from [18]. To put this assumption in some perspective, there are numerous examples of sub-Gaussian classes (the simplest of which are classes of linear functionals on $\mathbb{R}^M$ endowed with a sub-Gaussian design) for which the equivalence constant C is an absolute constant, unlike the constant B in (1.2), which is at least M.

Naturally, the analysis of ERM under a sub-Gaussian assumption requires a more sophisticated technical machinery than in situations in which the L_2/L_∞ equivalence assumption used in Theorem 1.1 holds. Invoking the main result from [18], one can show that if $Y - f^*(X)$ is sub-Gaussian and $\mathrm{span}(F)$ is a *sub-Gaussian class*, then for every $x > 0$, ERM achieves a rate $r_N(M) = c_1xM/N$ with probability at least $1 - \exp(-c_2xM)$.

Although the sub-Gaussian case is interesting, the goal of this note is the study of ERM as a linear aggregation procedure under much weaker assumptions.

Theorem A. *Let $F = \{f_1, \ldots, f_M\}$ and assume that there are constants κ_0 and β_0 for which*

$$P\bigl\{\bigl|f(X)\bigr| \geq \kappa_0\|f\|_{L_2}\bigr\} \geq \beta_0 \tag{1.4}$$

for every $f \in \mathrm{span}(F)$. Let $N \geq (400)^2M/\beta_0^2$ and set $\zeta = Y - f^*(X)$. Assume further that one of the following two conditions holds:

1. *ζ is independent of X and $\mathbb{E}\zeta^2 \leq \sigma^2$, or*
2. *$|\zeta| \leq \sigma$ almost surely.*

Then, for every $x > 0$, *with probability at least* $1 - \exp(-\beta_0^2 N/4) - (1/x)$,

$$\big\| \hat{f}^{\mathrm{ERM}} - f^* \big\|_{L_2}^2 = R\big(\hat{f}^{\mathrm{ERM}}\big) - \min_{f \in \mathrm{span}(F)} R(f) \le \left(\frac{16}{\beta_0 \kappa_0^2}\right)^2 \frac{\sigma^2 M x}{N}.$$

Since the loss is the squared one, one has to assume that Y and functions in $\mathrm{span}(F)$ have a second moment. It follows from Theorem A that in some cases, this is (almost) all that is needed for an optimal rate. Indeed, if $\zeta = Y - f^*(X)$ is independent of the design X – as is the case in any regression model with independent noise $Y = f^*(X) + \zeta$, and if (1.4) holds, ERM achieves the optimal rate M/N.

Corollary 1.2. *Consider the regression model* $Y = f^*(X) + \zeta$ *where* ζ *is a mean-zero noise that is independent of* X. *Assume that* $\zeta \in L_2$ *and that* $f^* \in \mathrm{span}(F)$. *If* $\mathrm{span}(F)$ *satisfies* (1.4) *and* $N \ge (400)^2 M/\beta_0^2$, *then for every* $x > 0$, *with probability at least* $1 - \exp(-\beta_0^2 N/4) - 1/x$,

$$\big\| \hat{f}^{\mathrm{ERM}} - f^* \big\|_{L_2}^2 \le \left(\frac{16}{\beta_0 \kappa_0^2}\right)^2 \frac{\sigma^2 M x}{N}.$$

From a statistical point of view, (1.4), which is a *small-ball assumption* on $\mathrm{span}(F)$, is a quantified version of *identifiability*. Indeed, consider the statistical model $\mathcal{M} = \{\mathbb{P}_f \colon f \in \mathrm{span}(F)\}$ where $\mathbb{P}_f$ is the probability distribution of the couple (X, Y), $Y = f(X) + \zeta$ and ζ is, for instance, a Gaussian noise that is independent of X. Assuming that $\mathcal{M}$ is identifiable is equivalent to having $P(|f(X) - g(X)| > 0) > 0$ for every $f, g \in \mathrm{span}(F)$, which, by linearity, is equivalent to $P(|f(X)| > 0) > 0$ for every $f \in \mathrm{span}(F)$. Comparing this with the small-ball condition in (1.4) shows that the latter is just a 'robust' version of identifiability.

It is possible to slightly modify the assumptions of Theorem A and still obtain the same type of estimate. For example, it is straightforward to verify that the small-ball condition (1.4) holds when the L_2 and L_p norms are equivalent on $\mathrm{span}(F)$ for some $p > 2$. This type of L_p/L_2 equivalence assumption on $\mathrm{span}(F)$ is weaker than the equivalence between the L_{ψ_2} and the L_2 norms in (1.3) because for every $p \ge 1$, $\|f\|_{L_p} \le c\sqrt{p}\|f\|_{\psi_2}$ for a suitable absolute constant c. And, it is clearly weaker than the L_∞/L_2 equivalence assumption (1.2) used in Theorem 1.1.

It turns out that if the L_2 and L_4 norms are equivalent on $\mathrm{span}(F)$, one may obtain the optimal rate for an arbitrary target Y, as long as $\zeta = Y - f^*(X)$ has a fourth moment. The difference between such a result and Theorem A is that ζ need not be independent of X, nor must it be bounded.

Theorem 1.3. *There exist absolute constants* c_0, c_1 *and* c_2 *for which the following holds. Assume that there exists* θ_0 *for which*

$$\|f\|_{L_4} \le \theta_0 \|f\|_{L_2} \tag{1.5}$$

for every $f \in \mathrm{span}(F)$, *and let* $N \ge (c_0 \theta_0^4)^2 M$. *Set* $\zeta = Y - f^*(X)$ *and put* $\sigma = (\mathbb{E}\zeta^4)^{1/4}$. *Then, for every* $x > 0$, *with probability at least* $1 - \exp(-N/(c_1\theta_0^8)) - (1/x)$,

$$\big\| \hat{f} - f^* \big\|_{L_2}^2 = R(\hat{f}) - \min_{f \in \mathrm{span}(F)} R(f) \le c_2 \theta_0^{12} \cdot \frac{\sigma^2 M x}{N}.$$

Remark 1.4. One may show that a possible choice of constants in Theorem 1.3 is $c_0 = 1600$, $c_1 = 64$ and $c_2 = (256)^2$, but since we have not made any real attempt of optimizing the choice of constants – because identifying the correct rate is the main focus of this note – we will not keep track of the values of constants in what follows.

One example in which Theorem 1.3 may be used is the regression problem with a misspecified model: $Y = f_0(X) + W$ where the regression function f_0 may not be in the model $\operatorname{span}(F)$ and $\zeta = (f_0 - f^*)(X) + W$ has a fourth moment. If $\operatorname{span}(F)$ satisfies (1.4), then with high probability,

$$\|\hat{f} - f^*\|_{L_2}^2 = \|\hat{f} - f_0\|_{L_2}^2 - \|f_0 - f^*\|_{L_2}^2 \leq c(\theta_0)\big(\mathbb{E}\zeta^4\big)^{1/2}\frac{M}{N}, \tag{1.6}$$

for a constant $c(\theta_0)$ that only depends on θ_0. Hence, one may select M as the solution of an optimal trade-off between the variance term $(\mathbb{E}\zeta^4)^{1/2}M/N$ and the bias; we refer the reader to Chapter 1 in [31] for techniques of a similar flavour.

The standard way of analyzing the performance of ERM is via certain trade-offs between concentration and complexity. However, in the case we study here, the functions involved may have 'heavy tails', and empirical means do not exhibit strong, two-sided concentration around their true means – which is a crucial component in the standard method of analysis. Therefore, a completely different path must be taken if one is to obtain the results formulated above.

The method we shall employ here has been introduced in [22,23] for problems in Learning Theory; in [24] in the context of the geometry of convex bodies; in [25] for applications in random matrix theory; and in [20] for Compressed Sensing.

Obviously, and regardless of the method of analysis, the (seemingly) unsatisfactory probability estimate is the price one pays for the moment assumptions on the 'noise' $Y - f^*(X)$. The next result shows that without stronger moment assumptions, only weak polynomial probability estimates are true.

Proposition 1.5. *Let $x \geq 1$, assume that $N \geq c_0 M$ for a suitable absolute constant c_0 and that X is the standard Gaussian vector in $\mathbb{R}^M$. There exists a mean-zero, variance one random variable ζ, that is independent of X and for which the following holds.*

Fix $t^* \in \mathbb{R}^M$ and consider the model $Y = \langle X, t^*\rangle + \zeta$. With probability at least c_1/x, ERM produces $\hat{t} \in \arg\min_{t\in\mathbb{R}^M} \sum_{i=1}^N (Y_i - \langle X_i, t\rangle)^2$ that satisfies

$$\|\hat{t} - t^*\|_2^2 = R(\hat{t}) - R\big(t^*\big) \geq \frac{c_2 x M}{N},$$

where c_1 and c_2 are absolute constants and $R(t) = \mathbb{E}(Y - \langle X, t\rangle)^2$ is the squared risk of t.

Note that the class of linear functional $\{\langle\cdot, t\rangle\colon\ t \in \mathbb{R}^M\}$ is a linear space of dimension M and it satisfies the small-ball condition when X is the standard Gaussian vector (actually, this class is sub-Gaussian). It follows from Proposition 1.5 that there is no hope of obtaining an exponential probability bound on the excess risk of ERM under an L_2-moment assumption on the noise – only polynomial bounds are possible. In particular, the probability estimate obtained in Theorem A under the L_2-assumption on the noise cannot be improved.

Finally, we would like to address the problem of linear aggregation under the classical boundedness assumptions: that $|Y| \leq 1$ and $|f(X)| \leq 1$ almost surely for every $f \in F$.

These are the standard assumptions that have been considered for the three problems of aggregation with a random design. For instance, optimal rates of aggregation have been obtained under these assumptions for the model selection aggregation problem in [2,16,21] and for the convex aggregation problem in [15]. And, it has been established that while ERM is suboptimal for the model selection aggregation problem (see, e.g., Section 3.5 in [7] or [17]), it is optimal for the convex aggregation problem. However, the optimality of ERM in the linear aggregation problem under the boundedness assumption was left open. The final result of this article addresses that problem – and it turns out that the answer is negative in a very strong way.

Proposition 1.6. *For every $0 < \eta < 1$ and integers N and M, there exists a couple (X, Y) and a dictionary $F = \{f_1, \ldots, f_M\}$ with the following properties*:

1. $|Y| \leq 1$ *almost surely and* $|f(X)| \leq 1$ *almost surely for every* $f \in F$.
2. *With probability at least η, for every $\kappa > 0$ there is some*

$$\hat{f}^{\mathrm{ERM}} \in \operatorname*{arg\,min}_{f \in \mathrm{span}(F)} \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - f(X_i)\big)^2$$

for which

$$R\big(\hat{f}^{\mathrm{ERM}}\big) \geq \inf_{f \in \mathrm{span}(F)} R(f) + \kappa.$$

Proposition 1.6 shows that even if one assumes that $|Y| \leq 1$ and $|f(X)| \leq 1$ almost surely for every function in the dictionary, and despite the convexity of $\mathrm{span}(F)$, the empirical risk minimization procedure performs poorly. This illustrates the major difference between assuming that the class is well bounded in L_∞ and assuming that the L_2 and L_p norms are equivalent on its span: while the latter suffices for an optimal bound, the former is rather useless.

An obvious outcome of Proposition 1.6 is that ERM should not be used to solve the linear aggregation problem under the boundedness assumption and one has to look for different procedures in the bounded setup. It should also be noted that since Proposition 1.6 is a non-asymptotic lower bound and X may depend on N and M, the asymptotic result appearing in Theorem 2.1 in [3] does not apply here.

Notation. For every function f, let $\|f\|_{L_p} = (\mathbb{E}|f(X)|^p)^{1/p}$. The excess loss of a function $f \in \mathrm{span}(F)$ is defined for every $x \in \mathcal{X}$ and $y \in \mathbb{R}$ by

$$\mathcal{L}_f(x, y) = \big(y - f(x)\big)^2 - \big(y - f^*(x)\big)^2;$$

thus, $R(f) - R(f^*) = P\mathcal{L}(X, Y) \geq 0$. The empirical measure over the data is denoted by P_N and

$$P_N \mathcal{L}_f = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - f(X_i)\big)^2 - \big(Y_i - f^*(X_i)\big)^2.$$

For every vector $x \in \mathbb{R}^M$, let $\|x\|_{\ell_p^M} = (\sum_{j=1}^M |x_j|^p)^{1/p}$ be its ℓ_p^M-norm.

Finally, all absolute constants are denoted by c_1, c_2, etc. Their value may change from line to line. We write $A \lesssim B$ if there is an absolute constant c for which $A \leq cB$, and $A \lesssim_\alpha B$ if $A \leq c(\alpha)B$ for a constant c that depends only on α.

2. Proofs of Theorem A and Theorem 1.3

The starting point of the proof of Theorem A is the same as in [18,19,22,23]: a decomposition of the excess loss function

$$\mathcal{L}_f(x,y) = \big(f^*(x) - f(x)\big)^2 + 2\big(y - f^*(x)\big)\big(f^*(x) - f(x)\big) \tag{2.1}$$

to a sum of quadratic and linear terms in $(f - f^*)(X)$. The idea of the proof is to control the quadratic term from below using a 'small-ball' argument, and the linear term from above using standard methods from empirical processes theory. A combination of these two bounds suffices to show that if $\|f - f^*\|_{L_2} \geq r_N^*$ for an appropriate choice of r_N^*, the quadratic term dominates the linear one, and in particular, for such functions $P_N \mathcal{L}_f > 0$. Since the empirical excess loss of the empirical minimizer is non-positive, it follows that $\|\hat{f} - f^*\|_{L_2} < r_N^*$.

Lemma 2.1. *There exists an absolute constant c_0 for which the following holds. Assume that there are κ_0 and β_0 for which*

$$P\big(|f(X)| \geq \kappa_0 \|f\|_{L_2}\big) \geq \beta_0$$

for every $f \in \operatorname{span}(F)$. If $N \geq c_0 M/\beta_0^2$, then with probability at least $1 - \exp(-\beta_0^2 N/4)$, for every $f \in \operatorname{span}(F)$,

$$\big|\big\{i \in \{1,\dots,N\}\colon \big|f(X_i)\big| \geq \kappa_0 \|f\|_{L_2}\big\}\big| \geq \frac{\beta_0 N}{2}.$$

Proof. Let $x > 0$ and set

$$H = \sup_{f \in \operatorname{span}(F)} \left| \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{|f(X_i)| \geq \kappa_0 \|f\|_{L_2}\}} - P\big(|f(X)| \geq \kappa_0 \|f\|_{L_2}\big) \right|.$$

Set $W = (f_1(X), \dots, f_M(X))$ – a random vector endowed on $\mathbb{R}^M$ by the dictionary F and the random variable X. Note that $\operatorname{span}(F) = \{\sum_{j=1}^M t_j f_j\colon (t_1,\dots,t_M) \in \mathbb{R}^M\}$ and set $\|t\|_{L_2} = \|\sum_{j=1}^M t_j f_j\|_{L_2}$.

Since N independent copies of X, $X_1, \dots, X_N$, endow N independent copies of W, denoted by $W_1, \dots, W_N$, it follows that

$$H = \sup_{t \in \mathbb{R}^M} \left| \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{|\langle t, \cdot\rangle| \geq \kappa_0 \|t\|_{L_2}\}}(W_i) - P\big(|\langle t, W\rangle| \geq \kappa_0 \|t\|_{L_2}\big) \right|.$$

By the bounded differences inequality (see, e.g., Theorem 6.2 in [4]), with probability at least $1-\exp(-x^2/2)$,

$$H \le \mathbb{E}H + \frac{1}{2}\sqrt{\frac{x}{N}}, \tag{2.2}$$

and a standard argument based on the VC-dimension of half-spaces in $\mathbb{R}^M$ shows that

$$\mathbb{E}H = \mathbb{E}H(X_1,\dots,X_N) \le c_1\sqrt{\frac{M}{N}}$$

(one may show the $c_1 \le 100$ using a rough estimate on Dudley's entropy integral combined with Exercise 2.6.4 in [32]). Therefore, if $c_1\sqrt{M/N} \le \beta_0/4$ and $(1/2)\sqrt{x/N} = \beta_0/4$, then with probability at least $1-\exp(-\beta_0^2 N/4)$, $H \le \beta_0/2$.

Finally, since

$$\inf_{f\in\operatorname{span}(F)} P\big(|f(X)| \ge \kappa_0\|f\|_{L_2}\big) \ge \beta_0$$

it follows that on the event $\{H \le \beta_0/2\}$,

$$\inf_{f\in\operatorname{span}(F)} \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{|f(X_i)|\ge\kappa_0\|f\|_{L_2}\}}(X_i) \ge \frac{\beta_0}{2}. \tag{2.3}$$

Therefore, (2.3) holds with probability at least $1-\exp(-\beta_0^2 N/4)$. □

Lemma 2.2. *Let $\zeta = Y - f^*(X)$ and assume that one of the following two conditions hold*:

1. *ζ is independent of X and $\mathbb{E}\zeta^2 \le \sigma^2$, or*
2. *$|\zeta| \le \sigma$ almost surely.*

Then, for every $x>0$, with probability larger than $1-(1/x)$,

$$\left|\frac{1}{N}\sum_{i=1}^N \big(Y_i - f^*(X_i)\big)\big(f^*(X_i) - f(X_i)\big)\right| \le 2\sigma\sqrt{\frac{Mx}{N}}\|f^* - f\|_{L_2}$$

for every $f \in \operatorname{span}(F)$.

Proof. Recall that $f^*(X)$ is the best L_2-approximation of Y in the linear space $\operatorname{span}(F)$; hence, $\mathbb{E}(Y-f^*(X))(f^*(X)-f(X)) = 0$ for every $f \in \operatorname{span}(F)$.

Let $\varepsilon_1,\dots,\varepsilon_N$ be independent Rademacher variables that are also independent of the couples $(X_i,Y_i)_{i=1}^N$. A standard symmetrization argument shows that

$$\mathbb{E}\sup_{f\in\operatorname{span}(F)\setminus\{f^*\}}\left|\frac{1}{N}\sum_{i=1}^N \big(Y_i - f^*(X_i)\big)\frac{f^*(X_i)-f(X_i)}{\|f^*-f\|_{L_2}}\right|^2$$

$$\le 4\mathbb{E}\sup_{f\in\operatorname{span}(F)\setminus\{f^*\}}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\big(Y_i - f^*(X_i)\big)\frac{f^*(X_i)-f(X_i)}{\|f^*-f\|_{L_2}}\right|^2.$$

Let $T = \{t \in \mathbb{R}^M\colon \| \sum_{j=1}^M t_j f_j \|_{L_2} = 1\}$ and observe that if $\zeta_1, \dots, \zeta_N$ are independent copies of ζ, then

$$\mathbb{E} \sup_{f \in \operatorname{span}(F)\setminus\{f^*\}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \big(Y_i - f^*(X_i)\big) \frac{f^*(X_i) - f(X_i)}{\|f^* - f\|_{L_2}} \right|^2$$
$$= \mathbb{E} \sup_{t \in T} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \zeta_i \left(\sum_{j=1}^M t_j f_j(X_i) \right) \right|^2 = (*).$$

Recall that $W = (f_1(X), \dots, f_M(X))$ and set Σ to be the covariance matrix associated with W. Let $\Sigma^{-1/2}$ be the pseudo-inverse of the squared-root of Σ, set $Z = \Sigma^{-1/2} W$ and note that $\mathbb{E}\|Z\|_{\ell_2^M}^2 \le M$.

If $Z_1, \dots, Z_N$ are independent copies of Z, it follows that

$$(*) = \mathbb{E} \sup_{\|t\|_{\ell_2^M}=1} \left| \left\langle t, \frac{1}{N} \sum_{i=1}^N \varepsilon_i \zeta_i Z_i \right\rangle \right|^2 = \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \zeta_i Z_i \right\|_{\ell_2^M}^2$$
$$= \mathbb{E}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_N} \left\| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \zeta_i Z_i \right\|_{\ell_2^M}^2 = \mathbb{E}\left(\frac{1}{N^2} \sum_{i=1}^N \zeta_i^2 \|Z_i\|_{\ell_2^M}^2 \right) = \frac{\mathbb{E}\zeta^2 \|Z\|_{\ell_2^M}^2}{N}$$
$$\le \frac{\sigma^2 \mathbb{E}\|Z\|_{\ell_2^M}^2}{N},$$

implying that

$$\mathbb{E} \sup_{f \in \operatorname{span}(F)\setminus\{f^*\}} \left| \frac{1}{N} \sum_{i=1}^N \big(Y_i - f^*(X_i)\big) \frac{f^*(X_i) - f(X_i)}{\|f^* - f\|_{L_2}} \right|^2 \le \frac{4\sigma^2 M}{N}.$$

The claim now follows from Markov's inequality. □

Proof of Theorem A. Combining Lemma 2.1 and Lemma 2.2 when $N \ge c_0 M/\beta_0^2$, it follows that with probability at least $1 - \exp(-\beta_0^2 N/4) - (1/x)$, if $f \in \operatorname{span}(F)$ and

$$\|\hat{f} - f^*\|_{L_2} > \frac{16\sigma}{\beta_0 \kappa_0^2} \sqrt{\frac{Mx}{N}}, \tag{2.4}$$

one has

$$\frac{1}{N} \sum_{i=1}^N \big(f^*(X_i) - f(X_i)\big)^2$$
$$\ge \kappa_0^2 \|f - f^*\|_{L_2}^2 \big|\{i\colon |f^*(X_i) - f(X_i)| \ge \kappa_0 \|f - f^*\|_{L_2}\}\big|/N$$

$$\geq \frac{\beta_0\kappa_0^2}{2}\big\|f-f^*\big\|_{L_2}^2 > 8\sigma\sqrt{\frac{Mx}{N}}\big\|f^*-f\big\|_{L_2}$$
$$> \frac{2}{N}\sum_{i=1}^N \big(Y_i - f^*(X_i)\big)\big(f^*(X_i)-f(X_i)\big).$$

Hence, on the same event, if $f \in \operatorname{span}(F)$ and (2.4) is satisfied then $P_N\mathcal{L}_f > 0$. Since $P_N\mathcal{L}_{\hat{f}^{\mathrm{ERM}}} \leq 0$, it follows that

$$\big\|\hat{f}^{\mathrm{ERM}} - f^*\big\|_{L_2}^2 \leq \left(\frac{16\sigma}{\beta_0\kappa_0^2}\right)^2 \frac{Mx}{N}. \qquad \square$$

Proof of Theorem 1.3. The proof of Theorem 1.3 is almost identical to the proof of Theorem A, and we will only outline the minor differences.

The small-ball condition (1.4) follows from the Paley–Zygmund inequality (see, for instance, Proposition 3.3.1 in [11]): if V is a real-valued random variable then

$$P\big(|V| \geq \kappa_0\big(\mathbb{E}V^2\big)^{1/2}\big) \geq (1-\kappa_0)^2\frac{(\mathbb{E}V^2)^2}{\mathbb{E}|V|^4}.$$

In particular, if $(\mathbb{E}|V|^4)^{1/4} \leq \theta_0(\mathbb{E}|V|^2)^{1/2}$ then

$$P\big(|V| \geq (1/2)\big(\mathbb{E}V^2\big)^{1/2}\big) \geq \big(4\theta_0^4\big)^{-1}$$

and thus the assertion of Lemma 2.1 holds for $\kappa_0 = 1/2$ and $\beta_0 = (4\theta_0^4)^{-1}$.

As for the analogous version of Lemma 2.2, the one change in its proof is that

$$\mathbb{E}\zeta^2\|Z\|_{\ell_2^M}^2 \leq \big(\mathbb{E}\zeta^4\big)^{1/2}\big(\mathbb{E}\|Z\|_{\ell_2^M}^4\big)^{1/2}$$

and

$$\mathbb{E}\|Z\|_{\ell_2^M}^4 = \mathbb{E}\left(\sum_{j=1}^M \langle e_j, Z\rangle^2\right)^2 = \mathbb{E}\sum_{p,q=1}^M \langle e_p, Z\rangle^2\langle e_q, Z\rangle^2$$
$$\leq \sum_{p,q=1}^M \big(\mathbb{E}\langle e_p, Z\rangle^4\mathbb{E}\langle e_q, Z\rangle^4\big)^{1/2} \leq \theta_0^4\sum_{p,q=1}^M \mathbb{E}\langle e_p, Z\rangle^2\mathbb{E}\langle e_q, Z\rangle^2 = \theta_0^4 M^2. \qquad \square$$

3. Proof of Proposition 1.6

Fix $Y = 1$ as the target and let $\mathcal{X} = \bigcup_{j=0}^M \mathcal{X}_j$ be some partition of $\mathcal{X}$. Consider a random variable X which is distributed as follows: fix $k \geq M$ to be chosen later; for $1 \leq j \leq M$, set $P(X \in \mathcal{X}_j) = \frac{1}{k}$ and put $P(X \in \mathcal{X}_0) = 1 - \frac{M}{k}$.

Finally, set

$$f_j(x) = \begin{cases} 1, & \text{if } x \in \mathcal{X}_j, \\ 0, & \text{otherwise} \end{cases}$$

and put $F = \{f_1, \ldots, f_M\}$.

Note that $|Y| \leq 1$ almost surely and that for every $f \in F$, $|f(X)| \leq 1$ almost surely. It is straightforward to verify that the oracle in $\operatorname{span}(F)$ is $f^* = \sum_{j=1}^M f_j(\cdot)$, and thus

$$\inf_{f \in \operatorname{span}(F)} R(f) = R\big(f^*\big) = \mathbb{E}\big(Y - f^*(X)\big)^2 = P(X \in \mathcal{X}_0) = 1 - \frac{M}{k}.$$

Let $X_1, \ldots, X_N$ be independent copies of X. Given $0 < \eta < 1$ and k large enough (for instance, $k \geq c(\eta)N/\log M$ for a sufficiently large constant $c(\eta)$ would suffice), there exists an event Ω_0 of probability at least η on which the following holds: there exists $j_0 \in \{1, \ldots, M\}$ for which $X_i \notin \mathcal{X}_{j_0}$ for every $1 \leq i \leq N$ (this is a slight modification of the coupon-collector problem).

For every $j = 1, \ldots, M$, let $N_j = |\{i \in \{1, \ldots, N\}\colon\ X_i \in \mathcal{X}_j\}|$. Hence, for $t \in \mathbb{R}^M$, the empirical risk of $\sum_{i=1}^M t_j f_j$ is

$$R_N\left(\sum_{j=1}^M t_j f_j\right) = \frac{1}{N}\sum_{i=1}^N \left(Y_i - \sum_{j=1}^M t_j f_j(X_i)\right)^2 = \sum_{j=1}^M \frac{N_j}{N}(1 - t_j)^2.$$

For $\xi > 0$ define $\hat{t}(\xi) \in \mathbb{R}^M$ by setting

$$\hat{t}(\xi)_j = \begin{cases} 1, & \text{if there exists } i \in \{1, \ldots, N\} \text{ s.t. } X_i \in \mathcal{X}_j, \\ \xi, & \text{if there is no } i \in \{1, \ldots, N\} \text{ s.t. } X_i \in \mathcal{X}_j. \end{cases}$$

Hence, $\hat{t}(\xi) \in \arg\min_{t \in \mathbb{R}^M} R_N(\sum_{j=1}^M t_j f_j)$ and $\hat{h}_\xi = \sum_{j=1}^M \hat{t}(\xi)_j f_j$ is an empirical minimizer in $\operatorname{span}(F)$.

For every sample in Ω_0, let $j_0 \in \{1, \ldots, N\}$ be the index for which $X_i \notin \mathcal{X}_{j_0}$ for every $1 \leq i \leq N$. Therefore,

$$R(\hat{h}_\xi) = \mathbb{E}\big(Y - \hat{h}_\xi(X)\big)^2 \geq (\xi - 1)^2 P(X \in \mathcal{X}_{j_0}) = \frac{(\xi - 1)^2}{k}$$

and the claim follows by selecting ξ large enough.

Appendix

We begin by presenting a proof of the well-known fact that if the L_∞ and L_2 norms are $\sqrt{B}$-equivalent on the span of M linearly-independent functions, then $B \geq M$.

Let $F = \{f_1, \ldots, f_M\} \subset L_2$ be a dictionary whose span is of dimension M, and recall that

$$\sqrt{B} = \sup_{f \in \operatorname{span}(F) \setminus \{0\}} \frac{\|f\|_{L_\infty}}{\|f\|_{L_2}}. \tag{A.1}$$

For every $u \in \mathbb{R}^M$ set $f_u = \sum_{j=1}^M u_j f_j$ and define an inner-product on $\mathbb{R}^M$ by

$$\langle u, v \rangle_F = \mathbb{E} f_u(X) f_v(X).$$

Let $(v_1, \ldots, v_M)$ be an orthonormal basis of $\mathbb{R}^M$ relative to $\langle \cdot, \cdot \rangle_F$ and for every $1 \le j \le M$, set $\phi_j = f_{v_j}$. Observe that $(\phi_1, \ldots, \phi_M)$ is an orthonormal basis of $\mathrm{span}(F)$ in L_2.

For μ-almost every $x \in \mathcal{X}$,

$$\sum_{j=1}^M \phi_j^2(x) \le \operatorname*{ess\,sup}_{z \in \mathcal{X}} \sum_{j=1}^M \phi_j(x)\phi_j(z) = \left\| \sum_{j=1}^M \phi_j(x)\phi_j \right\|_{L_\infty},$$

and by the definition of B in (A.1),

$$\left\| \sum_{j=1}^M \phi_j(x)\phi_j \right\|_{L_\infty} \le \sqrt{B} \left\| \sum_{j=1}^M \phi_j(x)\phi_j \right\|_{L_2} = \sqrt{B} \left(\sum_{j=1}^M \phi_j^2(x) \right)^{1/2}.$$

Hence, for μ-almost every $x \in \mathcal{X}$,

$$\sum_{j=1}^M \phi_j^2(x) \le B,$$

and by integrating this inequality with respect to μ and recalling that $\mathbb{E}\phi_j^2(X) = 1$, it follows that $M \le B$.

Proof of Proposition 1.5. Consider the model $Y = \langle X, t^* \rangle + \zeta$ where $t^* \in \mathbb{R}^M$, X is a standard Gaussian vector in $\mathbb{R}^M$ and ζ is a mean-zero noise that is independent of X. To make the presentation simpler, assume that $t^* = 0$, and thus one only observes the noise $Y = \zeta$. The aim here is to estimate the distance between $\hat{t}$ and $t^* = 0$ when the noise ζ is only assumed to be in L_2.

Let us begin by showing that, conditionally on $\zeta_1, \ldots, \zeta_N$, and if $\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \zeta_i^2$, then with probability at least $1 - 2\exp(-c_0 N)$,

$$R(\hat{t}) - R\big(t^*\big) = \|\hat{t}\|_2^2 \ge \frac{c\hat{\sigma}_N^2 M}{N}, \tag{A.2}$$

for a suitable absolute constant c.

To that end, observe that the excess empirical risk for every $v \in \mathbb{R}^M$ is

$$P_N \mathcal{L}_v = R_N(v) - R_N(0) = \frac{1}{N} \sum_{i=1}^N \langle X_i, v \rangle^2 - \frac{2}{N} \sum_{i=1}^N \zeta_i \langle X_i, v \rangle, \tag{A.3}$$

and that for every sample, if $r_1 < r_2$ and

$$\inf_{0 \le r < r_1} \inf_{\|v\|_2 = r} P_N \mathcal{L}_v > \inf_{r \ge r_2} \inf_{\|v\|_2 = r} P_N \mathcal{L}_v,$$

one has $\|\hat{t}\|_2 \ge r_1$.

Using a standard ε-net argument together with Gaussian concentration, one may show that if $N \geq c_0 M$, then with μ^N-probability at least $1 - 2\exp(-c_1 N)$, for every $x \in \mathbb{R}^M$,

$$\frac{1}{2}\|x\|_2^2 \leq \frac{1}{N}\sum_{i=1}^N \langle X_i, x\rangle^2 \leq \frac{3}{2}\|x\|_2^2. \tag{A.4}$$

Moreover, on that event, setting

$$I = \sup_{\{x\in\mathbb{R}^M:\ \|x\|_2=1\}} \left| \frac{1}{N}\sum_{i=1}^N \zeta_i \langle X_i, x\rangle \right|,$$

one has that for any $\zeta_1, \ldots, \zeta_N$

$$c_1 \hat{\sigma}_N \sqrt{\frac{M}{N}} \leq I \leq c_2 \hat{\sigma}_N \sqrt{\frac{M}{N}}$$

for suitable absolute constants c_1 and c_2. We refer the reader to Lemma 2.6.4 and Theorem 2.6.5 in [8] for more details on the techniques used to obtain these observations.

Clearly, for every $r > 0$,

$$\inf_{\{x\in\mathbb{R}^M:\ \|x\|_2=r\}} \frac{1}{N}\sum_{i=1}^N \zeta_i \langle X_i, x\rangle = -rI. \tag{A.5}$$

Hence, by (A.3), it follows that for $N \geq c_0 M$ and conditioned on $\zeta_1, \ldots, \zeta_N$, with probability at least $1 - 2\exp(-c_3 N)$,

$$\begin{aligned}\inf_{0\leq r<I/6} \inf_{\|v\|_2=r} P_N \mathcal{L}_v &\geq \inf_{0\leq r<I/6} \left(\frac{r^2}{2} - rI\right) \\ &> \inf_{r\geq I/3} \left(\frac{3r^2}{2} - rI\right) \geq \inf_{r\geq I/3} \inf_{\|v\|_2=r} P_N \mathcal{L}_v.\end{aligned}$$

Therefore, on that event

$$\|\hat{t}\|_2 \geq I/6 \geq c_4 \hat{\sigma}_N \sqrt{\frac{M}{N}}.$$

Now, all that remains is to show that $P(\hat{\sigma}_N^2 \geq x) \geq c_5/x$. □

Lemma A.1. *For every $N \geq 2$ and $x \geq 1$, there exists a mean-zero, variance one random variable ζ for which*

$$P\left(\hat{\sigma}_N^2 \geq x\right) \geq \frac{c_1}{x}.$$

Proof. Fix $x \geq 1$, let ε be a symmetric, $\{-1,1\}$-valued random variable, set $\delta = 1/(xN)$ and put η to be a $\{0,1\}$-valued random variable with mean δ that is independent of ε. Finally, let $R = 1/\sqrt{\delta}$ and set $\zeta = R\varepsilon\eta$. Thus, $\mathbb{E}\zeta = 0$ and $\|\zeta\|_{L_2} = R\delta^{1/2} = 1$.

Let $\zeta_i = R\varepsilon_i\eta_i$, $i = 1,\dots,N$ be independent copies of ζ. Recall that $NR^{-2}x = 1$ and that $\delta N \leq 1$. Therefore,

$$P\big(\hat{\sigma}_N^2 \geq x\big) = P\left(\frac{1}{N}\sum_{i=1}^N \zeta_i^2 \geq x\right) = P\left(\sum_{i=1}^N \eta_i \geq 1\right)$$
$$= P\big(\exists i \in \{1,\dots,N\}, \eta_i = 1\big) = 1 - (1-\delta)^N \geq c_1 N\delta = c_1/x,$$

as claimed. □

Acknowledgment

Shahar Mendelson was supported by the Mathematical Sciences Institute – The Australian National University and by ISF Grant 900/10.

References

[1] Anthony, M. and Bartlett, P.L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge: Cambridge Univ. Press. MR1741038

[2] Audibert, J.-Y. (2007). Progressive mixture rules are deviation suboptimal. In *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, *Vancouver*, *British Columbia*, *Canada*, *December* 3–6, 2007. *Advances in Neural Information Processing Systems* (*NIPS*) **20** 41–48. Vancouver: MIT Press.

[3] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.* **39** 2766–2794. MR2906886

[4] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193

[5] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Belmont, CA: Wadsworth Advanced Books and Software. MR0726392

[6] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[7] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math.* **1851**. Berlin: Springer. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920

[8] Chafaï, D., Guédon, O., Lecué, G. and Pajor, A. (2012). *Interactions Between Compressed Sensing Random Matrices and High Dimensional Geometry*. *Panoramas et Synthèses* [*Panoramas and Syntheses*] **37**. Paris: Société Mathématique de France. MR3113826

[9] Dalalyan, A.S. and Tsybakov, A.B. (2008). Aggregation by exponential weighting, sharp pac-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.

[10] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. New York: Springer. MR1383093

[11] de la Peña, V.H. and Giné, E. (1999). *Decoupling*: *From Dependence to Independence*. *Probability and Its Applications* (*New York*). *Randomly Stopped Processes. U-Statistics and Processes. Martingales and Beyond*. New York: Springer. MR1666908

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2722294
[13] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. MR1792783
[14] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184
[15] Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166. MR3160549
[16] Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** 591–613. MR2529440
[17] Lecué, G. and Mendelson, S. (2010). Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli* **16** 605–613. MR2730641
[18] Lecué, G. and Mendelson, S. (2013). Learning subGaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion.
[19] Lecué, G. and Mendelson, S. (2013). Minimax rate of convergence and the performance of ERM in phase recovery. Technical report. *Electron. J. Probab*. To appear.
[20] Lecué, G. and Mendelson, S. (2014). Sparse recovery under weak moment assumptions. Technical report. *J. Eur. Math. Soc.* To appear.
[21] Lecué, G. and Rigollet, P. (2014). Optimal learning with Q-aggregation. *Ann. Statist.* **42** 211–224. MR3178462
[22] Mendelson, S. (2013). Learning without concentration. Technical report. *J. ACM*. Available at arXiv:1401.0304. To appear.
[23] Mendelson, S. (2014). Learning without concentration for general loss functions. Technical report, Technion, Israel and ANU, Australia. Available at arXiv:1410.3192.
[24] Mendelson, S. (2014). A remark on the diameter of random sections of convex bodies. In *Geometric Aspects of Functional Analysis* (*GAFA Seminar Notes*). *Lecture Notes in Math*. **2116** 395–404.
[25] Mendelson, S. and Koltchinskii, V. (2013). Bounding the smallest singular value of a random matrix without concentration. Technical report, Technion and Georgia Tech. Available at arXiv:1312.3580.
[26] Nemirovski, A. (2000). *Lectures on Probability Theory and Statistics*. *Lecture Notes in Math*. **1738**. Berlin: Springer. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998, edited by Pierre Bernard.
[27] Rigollet, Ph. and Tsybakov, A.B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821
[28] Schapire, R.E. and Freund, Y. (2012). *Boosting*: *Foundations and Algorithms*. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. MR2920188
[29] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. *Information Science and Statistics*. New York: Springer. MR2450103
[30] Tsybakov, A.B. (2003). Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines* (*COLT*-2003). *Lecture Notes in Artificial Intelligence* **2777** 303–313. Heidelberg: Springer.
[31] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359
[32] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671
[33] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28** 75–87. MR1762904
[34] Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96** 574–588. MR1946426
[35] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592

Received March 2014 and revised February 2015