# Performance of empirical risk minimization in linear aggregation

Guillaume Lecué[1,3] Shahar Mendelson[2,4,5]

February 24, 2014

**Abstract**

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

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

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

We show that if one assumes that $|Y| \leq 1$ and $|f(X)| \leq 1$ almost surely for every function in the dictionary, the empirical risk minimization procedure may still perform poorly, and in particular, its performance is far from the rate $M/N$.

On the other hand, under mild assumptions on $F$ (a uniform small-ball estimates for functions in span$(F)$), ERM in span$(F)$ does achieve the rate of $M/N$.

## 1 Introduction and main results

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

In the learning theory setup, 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

[1]CNRS, CMAP, Ecole Polytechnique, 91120 Palaiseau, France.

[2]Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.

[3]Email: guillaume.lecue@cmap.polytechnique.fr

[4]Email: shahar@tx.technion.ac.il

[5]Supported by the Mathematical Sciences Institute – The Australian National University and by ISF grant 900/10.

of $Y$. A standard way of measuring the prediction capability of $f$ is via the square risk: $R(f) = \mathbb{E}\big(Y - f(X)\big)^2$. The conditional expectation

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

is the square risk of the function $\hat{f}$ that was 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 the books [9], [1], [24], [5], [10], [26] and references therein for numerous examples), but in general, there is no canonical choice of a prediction 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 to obtain a function with 'better' prediction capabilities than individual members of the dictionary.

Aggregation procedures have been studied extensively (see, e.g. [11, 21, 7, 29, 28, 27, 12, 8, 25] and references therein), and among the more well-known aggregation procedures are boosting (see, for example, [23]) and bagging [5].

Here, our aim is to study the problem of linear aggregation: given a dictionary $F = \{f_1, \ldots, f_M\}$, 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 $\mathrm{span}(F)$; namely, a procedure ensuring that with high probability

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

This type of inequality is called an *oracle inequality* and the function $f^*$ for which $R(f^*) = \inf_{f \in \mathrm{span}(F)} R(f)$ is 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 [21] in the gaussian white noise model; in [25, 6] for the gaussian model with random design; in [22] for the density estimation problem and in [3] in the learning theory setup under moment conditions.

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 its optimality holds in some minimax sense, introduced in [25].

The procedure that will be studied in this note is empirical risk minimization (ERM), performed in the span of the dictionary:

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

This procedure has been studied extensively and we refer the reader to [19] and [13] for results on ERM in the bounded or the gaussian setup.

The benchmark result for ERM performed in the span of a dictionary as an aggregation procedure is Theorem 2.2 in [3]. To formulate it, here, and throughout this note we will denote by $F$ a dictionary of cardinality $M$ and by $f^*$ the oracle in $\mathrm{span}(F)$ (i.e. $R(f^*) = \inf_{f \in \mathrm{span}(F)} R(f)$).

**Theorem 1.1** *There exist absolute constants $c_0$ and $c_1$ for which the following holds. Assume that $\mathbb{E}(Y - f^*(X))^4 < \infty$ and that*

$$\sup_{f \in \mathrm{span}(F) - \{0\}} \frac{\|f(X)\|_{L_\infty}}{\|f(X)\|_{L_2}} = \sqrt{B} < \infty. \tag{1.2}$$

*If $x > 0$ satisfies that $2/N \leq 2\exp(-x) \leq 1$ and $N \geq c_0 B^2(BM + x)$, then with probability at least $1 - 2\exp(-x)$,*

$$R(\hat{f}^{ERM}) - R(f^*) \leq c_1 \sqrt{\mathbb{E}(Y - f^*(X))^4}\left(\frac{BM + x}{N}\right).$$

It follows from Theorem 1.1 that under an $L_4$ assumption on $Y - f^*(X)$ and the equivalence of the $L_2$ and $L_\infty$ norms on the span of $F$, ERM achieves the optimal rate of convergence with exponential probability.

Recall that the $\psi_2$ norm of a random variable $Z$ is defined by $\|Z\|_{\psi_2} = \inf\big\{c > 0 : \mathbb{E}\exp(-Z^2/c^2) \leq 2\big\}$. Since (1.2) implies that for any $f, h \in \mathrm{span}(F)$, $\|(f-g)(X)\|_{\psi_2} \leq \sqrt{B}\,\|(f-g)(X)\|_{L_2}$, then $\mathrm{span}(F)$ is a subgaussian class. Applying the results of [16], one may show a similar result to Theorem 1.1, if one assumes that $\|Y - f^*(X)\|_{\psi_2} < \infty$. However, as such an assumption appear rather restrictive, one may wonder whether weaker moment conditions still suffice to guarantee the optimal rate of convergence. Theorem A and Theorem B show that this is indeed the case.

Throughout, $N$ and $M$ denote integers, $F$ is a dictionary consisting of $M$ functions, $(X, Y)$ is a couple of random variables and $\zeta = Y - f^*(X)$.

**Theorem A.** There exist absolute constants $c_0$ and $c_1$ for which the following holds. Assume that there are $\kappa_0$ and $\beta_0$ such that $M \leq c_0\beta_0 N$ and

$$P\big\{|f(X)| \geq \kappa_0 \|f(X)\|_{L_2}\big\} \geq \beta_0 \tag{1.3}$$

for every $f \in \mathrm{span}(F)$. Assume further that one of the two conditions hold:

1. $\zeta$ 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 $1 - \exp(-c_0\beta_0^2 N) - (1/x)$,

$$\left\| \hat{f}^{ERM} - f^* \right\|_{L_2}^2 = R(\hat{f}^{ERM}) - \min_{f \in \mathrm{span}(F)} R(f) \leq \frac{c_2 \sigma^2 M x}{N}.$$

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 to obtain an optimal rate. Indeed, if the noise $Y - f^*(X)$ is independent of the design $X$ – as is the case in any regression model with independent noise – ERM achieves the rate $M/N$ under Assumption (1.3). And, one obvious case in which Assumption (1.3) may be verified is when the $L_2$ and $L_p$ norms are equivalent on $\mathrm{span}(F)$.

A result of a similar flavour is the following:

**Theorem B.** There are absolute constants $c_0$, $c_1$ and $\theta_0$ for which the following holds. Assume that $\theta_0^8 M \leq c_0 N$, that

$$\left(\mathbb{E}f^4(X)\right)^{1/4} \leq \theta_0 \left(\mathbb{E}f^2(X)\right)^{1/2} \tag{1.4}$$

for every $f \in \mathrm{span}(F)$ and that $\left(\mathbb{E}\zeta^4\right)^{1/4} \leq \sigma$.

Then, for every $x > 0$, with probability at least than $1 - \exp(-c_0 N/\theta_0^8) - (1/x)$,

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

In both results, the price that has to be paid for the very weak moment assumptions is probability estimate – but that can not be helped, and it seems that without stronger assumptions, the exponential probability estimate is not realistic.

Finally, we would like to address another related problem. Since optimal rates of aggregation have been obtained in the learning theory framework under the assumptions that $|Y| \leq 1$ and $|f(X)| \leq 1$ almost surely for every $f \in F$, both for the Model Selection [15, 2, 18] and for convex aggregation

[14], it is natural to study the problem of linear aggregation in this learning theory setup and the performance of ERM under the same assumptions.

The next result shows that the behaviour of ERM as a linear aggregation procedure in that case is very far from optimal.

**Theorem C.** 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 $\eta$, for every $\kappa > 0$ there is some

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

for which

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

One should note that since Theorem C is a non-asymptotic lower bound and the probability distribution of $X$ may depend on $N$ and $M$, the asymptotic result appearing in Theorem 2.1 in [3] does not apply here. What is also surprising is that even though the model $\mathrm{span}(F)$ is convex, ERM is still a suboptimal procedure.

## 2 Proofs of Theorem A and Theorem B

The starting point of the proofs of Theorem A and Theorem B is the same as in [16, 20, 17]: a decomposition of the excess loss function

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

to a sum of a quadratic term and a linear term. The idea of the two proofs 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 $r_N^*$, then 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 exist absolute constants $c_0, c_1, c_2, c_3$ and $c_4$ for which the following holds. If there are $\kappa_0$ and $\beta_0$ for which*

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

*for every $f \in \mathrm{span}(F)$, and if $M \leq c_0\beta_0^2 N$, then with probability at least $1 - c_1\exp(-c_2\beta_0^2 N)$, for every $f \in \mathrm{span}(F)$,*

$$\left|\{i \in \{1,\ldots,N\} : |f(X_i)| \geq \kappa_0 \left\|f(X)\right\|_{L_2}\}\right| \geq c_3\beta_0 N.$$

**Proof.** Let $x > 0$ and set

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

Set $W = (f_1(X), ..., f_M(X))$, and thus the dictionary $F$ and the random variable $X$ define a random vector on $\mathbb{R}^M$. Each $f \in \mathrm{span}(F)$ is associated with $t \in \mathbb{R}^M$, since $\sum_{i=1}^M t_j f_j(X) = \langle t, W\rangle$. Set $\|t\|_{L_2} = \|\sum_{j=1}^M t_j f_j\|_{L_2}$, and observe that $N$ independent copies of $X$, $X_1, ..., X_N$, endow $N$ independent copies of $W$. Thus,

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

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

$$H \leq \mathbb{E}H + c_0\sqrt{\frac{x}{N}}, \tag{2.2}$$

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

$$\mathbb{E}H(X_1,\ldots,X_N) \leq c_1\sqrt{\frac{M}{N}}.$$

Therefore, if $c_1\sqrt{M/N} \leq \beta_0/4$ and $c_0\sqrt{x/N} = \beta_0/4$, then with probability at least $1 - c_0\exp(-c_1\beta_0^2 N)$, $H \leq \beta_0/2$, and since

$$\inf_{f\in\mathrm{span}(F)} P\left(\left\{|f(X)| \geq \kappa_0\left\|f(X)\right\|_{L_2}\right\}\right) \geq \beta_0$$

then on the same event

$$\inf_{f\in\mathrm{span}(F)} \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left\{|f(X_i)|\geq\kappa_0\left\|f(X)\right\|_{L_2}\right\}}(X_i) \geq \frac{\beta_0}{2}.$$

■

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

1. *$\zeta$ 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 larger than $1 - (1/x)$,*

$$\left| \frac{1}{N} \sum_{i=1}^{N} (Y_i - f^*(X_i)) (f^*(X_i) - f(X_i)) \right| \leq c_1 \sigma \sqrt{\frac{Mx}{N}} \|f^*(X) - f(X)\|_{L_2}$$

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

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

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

$$\begin{aligned} &\mathbb{E} \sup_{f \in \mathrm{span}(F)} \left| \frac{1}{N} \sum_{i=1}^{N} (Y_i - f^*(X_i)) \frac{f^*(X_i) - f(X_i)}{\|f^*(X) - f(X)\|_{L_2}} \right|^2 \\ &\leq c_1 \mathbb{E} \sup_{f \in \mathrm{span}(F)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (Y_i - f^*(X_i)) \frac{f^*(X_i) - f(X_i)}{\|f^*(X) - f(X)\|_{L_2}} \right|^2 \end{aligned}$$

for a suitable absolute constant $c_1$.

Let $T = \{t \in \mathbb{R}^M : \| \sum_{j=1}^M t_j f_j \|_{L_2} = 1\}$. If $\zeta_1, ..., \zeta_N$ are independent copies of $\zeta$, then

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

Recall that $W = (f_1(X), ..., f_M(X))$, set $\Sigma$ to be the covariance matrix associate with $W$ and let $\Sigma^{-1/2}$ be the pseudo-inverse of the squared-root of $\Sigma$. Set $Z = \Sigma^{-1/2} W$ and observe that $Z$ is an isotropic random vector

on its image and that $\mathbb{E}\left\|Z\right\|_{\ell_2^M}^2 \leq M$. Hence, if $Z_1, ..., Z_N$ are independent copies of $Z$,

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

implying that

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

The claim follows from Markov's inequality. ∎

**Proof of Theorem A:** Combining Lemma 2.1 and Lemma 2.2, it is evident that with probability at least $1-\exp(-c_0 N)-(1/x)$, if $f\in\mathrm{span}(F)$ and $\left\|f(X)-f^*(X)\right\|_{L_2} \geq c_0\sigma^2 Mx/n$ for a sufficiently large constant $c_0$, then

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N (f^*(X_i)-f(X_i))^2 \\ &\geq \kappa_0^2\left\|f-f^*\right\|_{L_2}^2 \left|\{i : |f^*(X_i)-f(X_i)| \geq \kappa_0\left\|f-f^*\right\|_{L_2}\}\right|/N \\ &\geq c_3\beta_0\kappa_0^2\left\|f-f^*\right\|_{L_2}^2 > 2c_1\sigma\sqrt{\frac{Mx}{N}}\left\|f^*-f\right\|_{L_2} \\ &> \frac{2}{N}\sum_{i=1}^N (Y_i - f^*(X_i))\left(f^*(X_i)-f(X_i)\right).\end{aligned}$$

Hence, on the same event, if $f\in\mathrm{span}(F)$ and $\left\|f-f^*\right\|_{L_2} \geq c_0\sigma^2 Mx/n$ then $P_N\mathcal{L}_f > 0$. Since $P_N\mathcal{L}_{\hat{f}^{ERM}} \leq 0$, then

$$\left\|\hat{f}^{ERM}-f^*\right\|_{L_2}^2 < c_0\frac{\sigma^2 Mx}{n}.$$

∎

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

The small-ball property (1.3) follows from the Paley-Zygmund inequality: if $Z$ is a real valued random variable then

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

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

$$P\left(|Z| \geq (1/2)\left(\mathbb{E}Z^2\right)^{1/2}\right) \geq \left(4\theta_0^4\right)^{-1}$$

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

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

and

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

■

## 3 Proof of Theorem C

Fix $Y = 1$ as the target and let $\mathcal{X} = \cup_{i=0}^M \mathcal{X}_M$ be some partition of $\mathcal{X}$. Define the distribution $X$ 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 $F = \{f_1, ..., f_M\}$.

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 $\mathrm{span}(F)$ is $f^* = \sum_{j=1}^{M} f_j(\cdot)$, and thus

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

Let $X_1, \ldots, X_N$ be independent copies of $X$. Given $0 < \eta < 1$ and $k$ large enough (for instance $k \gtrsim_\eta N/\log M$ would suffice), there exists an event $\Omega_0$ of probability at least $\eta$ on which the following holds: if $X_1, ..., X_N \in \Omega_0$, then there exists some $j_0 \in \{1, \ldots, M\}$ and $X_i \notin \mathcal{X}_{j_0}$ for every $1 \leq i \leq N$ (this is a slight modification of the coupon-collector problem).

For every $X_1, ..., X_N \in \Omega_0$, let $N_j = |\{i \in \{1, \ldots, N\} : X_i \in \mathcal{X}_j\}|$. Hence, if $t \in \mathbb{R}^M$ then the empirical risk of $\sum_{i=1}^{M} t_j f_j$ is

$$R_N\left(\sum_{j=1}^{M} t_j f_j\right) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{M} t_j f_j(X_i)\right)^2 = \frac{N_0}{N} + \sum_{j=1}^{M}\left(\frac{N_j}{N}\left(1 - t_j\right)^2\right).$$

In particular, for $\kappa > 0$ define $\hat{t}(\kappa) \in \mathbb{R}^M$ by setting

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

Hence, $\hat{t}(\kappa) \in \mathrm{argmin}_{t \in \mathbb{R}^M} R_N(\sum_{j=1}^{M} t_j f_j)$, and $\hat{h}_\kappa = \sum_{j=1}^{M} \hat{t}(\kappa)_j f_j$ minimizes the empirical risk in $\mathrm{span}(F)$. Thus, if $(X_1, ..., X_N) \in \Omega_0$ and $X_i \notin \mathcal{X}_{j_0}$ for every $1 \leq i \leq N$, then

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

The result follows by selecting $\kappa$ large enough. ∎

# References


[1] Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.

[2] Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[3] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[5] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[6] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[7] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[8] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1–2):39–61, 2008.

[9] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[11] Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

[12] Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[13] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. 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].

[14] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013.

[15] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009.

[16] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[17] Guillaume Lecué and Shahar Mendelson. Minimax rate of convergence and the performance of erm in phase recovery. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[18] Guillaume Lecué and Philippe Rigollet. Optimal learning with $q$-aggregation. Technical report, CNRS, Ecole Polytechnique and Princeton University, 2013.

[19] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[20] Shahar Mendelson. Learning without concentration. Technical report, Technion, 2013. arXiv:1401.0304.

[21] Arkadii Nemirovski. *Lectures on probability theory and statistics*, volume 1738 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998, Edited by Pierre Bernard.

[22] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

[23] Robert E. Schapire and Yoav Freund. *Boosting*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012. Foundations and algorithms.

[24] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[25] Alexandre B. Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.

[26] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[27] Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.

[28] Yuhong Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001.

[29] Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.