Empirical risk minimization is optimal for the convex aggregation problem

Guillaume Lecué

Abstract

Let $F$ be a finite model of cardinality $M$ and denote by $\text{conv}(F)$ its convex hull. The problem of convex aggregation is to construct a procedure having a risk as close as possible to the minimal risk over $\text{conv}(F)$. Consider the bounded regression model with respect to the squared risk denoted by $R(\cdot)$. If $f_{\text{ERM}}$ denotes the empirical risk minimization procedure over $\text{conv}(F)$ then we prove that for any $x > 0$, with probability greater than $1 - 4 \exp(-x)$,

$$R(f_{\text{ERM}} - C_n) \leq \min_{f \in \text{conv}(F)} R(f) + c_0 \max \left( \psi_n^{(C)}(M), \frac{x}{n} \right)$$

where $c_0 > 0$ is an absolute constant and $\psi_n^{(C)}(M)$ is the optimal rate of convex aggregation defined in [37] by $\psi_n^{(C)}(M) = M/n$ when $M \leq \sqrt{n}$ and $\psi_n^{(C)}(M) = \sqrt{\log (eM/\sqrt{n})/n}$ when $M > \sqrt{n}$.

1 Introduction and main results

Let $\mathcal{X}$ be a probability space and let $(X, Y)$ and $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $n + 1$ i.i.d. random variables with values in $\mathcal{X} \times \mathbb{R}$. From the statistical point of view, the set $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is the set of given data where the $X_i$’s are usually considered as input data taking their values in some space $\mathcal{X}$ and the $Y_i$’s are some outputs or labels associated with these inputs. We are interested in the prediction of $Y$ associated with a new observation $X$. The data $D$ are thus used to construct functions $f: \mathcal{X} \rightarrow \mathbb{R}$ such that $f(X)$ provides a good guess of $Y$. We measure the quality of this prediction by means of the squared risk

$$R(f) = \mathbb{E}(Y - f(X))^2,$$

when $f$ is a real-valued function defined on $\mathcal{X}$ and by

$$R(\hat{f}) = \mathbb{E} \left[ (Y - \hat{f}(X))^2 | D \right]$$

when $\hat{f}$ is a function constructed using the data $D$. For the sake of simplicity, throughout this article we restrict ourselves to functions $f$ and random variables $(X, Y)$ for which

---

1CNRS, LAMA, Université Paris-Est Marne-la- Vallée, 77454 France. Supported by French Agence Nationale de la Recherche ANR Grant “PROGNOSTIC” ANR-09-JCJC-0101-01. Email: guillaume.lecue@univ-mlv.fr
\(|Y| \leq b\) and \(|f(X)| \leq b\) almost surely, for some fixed \(b \geq 0\). Note that \(b\) does not have to be known from the statistician for the construction of the procedures we are studying in this note.

Given a finite set \(F\) of real-valued measurable functions defined on \(X\) (usually called a dictionary), there are three main types of aggregation problems:

1. \textit{model selection aggregation}: construct a procedure whose risk is as close as possible to the risk of the best element in \(F\) (cf. \([2, 3, 12, 14, 15, 30, 17, 18, 22, 37, 38, 41, 42]\)).

2. \textit{convex aggregation}: construct a procedure whose risk is as close as possible to the risk of the best function in the convex hull of \(F\) (cf. \([1, 9, 11, 12, 17, 37, 43, 26]\)).

3. \textit{linear aggregation}: construct a procedure whose risk is as close as possible to the risk of the best function in the linear span of \(F\) (cf. \([12, 30, 20, 37, 4]\)).

In this note, we focus on the convex aggregation problem. We want to construct a procedure \(\hat{f}\) for which, with high probability,

\[
R(\hat{f}) \leq \min_{f \in \text{conv}(F)} R(f) + \psi_n(M) \tag{1.1}
\]

where \(\psi_n(M)\) is called the residual term. The residual term is the quantity that we want as small as possible. Results in expectation are also of interest: construct a procedure \(\bar{f}\) such that \(\mathbb{E}R(\bar{f}) \leq \min_{f \in \text{conv}(F)} R(f) + \psi_n(M)\).

In \([37]\) the author defined the optimal rates of the convex aggregation, by the smallest price in the minimax sense that one has to pay to solve the convex aggregation problem. The definition of \([37]\) is given in expectation, as a function of the cardinality \(M\) of the dictionary \(F\) and of the sample size \(n\). It has been proved in \([37]\) (see also \([17]\) and \([43]\)) that the optimal rate of convex aggregation is

\[
\psi_n^{(C)}(M) = \begin{cases} 
\frac{M}{\sqrt{n}} & \text{if } M \leq \sqrt{n} \\
\sqrt{\frac{1}{n} \log \left( \frac{eM}{\sqrt{n}} \right)} & \text{if } M > \sqrt{n}.
\end{cases}
\]

This rate is defined up to some multiplying constant. Note that the rate \(\psi_n^{(C)}(M)\) was achieved in \([37]\) in expectation for the Gaussian regression model with a known variance and a known marginal distribution of the design. In \([11]\), the authors were able to remove these assumptions at a price of an extra \(\log n\) factor for \(1 \leq M \leq \sqrt{n}\) (results are still in expectation). Last year, there has been some striking results on different problems of aggregation including the convex aggregation problem. To mention few of them, we refer the reader to \([33, 32]\) and \([40]\). Finally, we also refer the reader to \([7, 43]\) for non-exact oracle inequalities (inequalities like (1.1) where \(\min_{f \in \text{conv}(F)} R(f)\) is multiplied by a constant strictly larger than 1) in the context of convex aggregation.

A lower bound in deviation for the convex aggregation problem follows from the arguments of \([37]\): there exist absolute positive constants \(c_0, c_1\) and \(c_2\) such that for any sample cardinality \(n \geq 1\), any cardinality of dictionary \(M \geq 1\) such that \(\log M \leq c_0 n\), there exists a dictionary \(F\) of size \(M\) such that for any aggregation procedure \(\hat{f}_n\), there
exists a random couple \((X,Y)\) such that \(|Y| \leq b\) and \(\max_{f \in F} |f(X)| \leq b\) a.s. and with probability larger than \(c_1\),

\[
R(\hat{f}_n) \geq \min_{f \in \text{conv}(F)} R(f) + c_2 b^2 \psi_n(C_n) (M) \tag{1.2}
\]

This means that, from a minimax point of view, one cannot do better than the rate \(\psi_n(C_n) (M)\) for the convex aggregation problem. Therefore, any procedure achieving the rate \(\psi_n(C_n) (M)\) for any dictionary \(F\) and couple \((X,Y)\) such that \(|Y| \leq b\) and \(\max_{f \in F} |f(X)| \leq b\) a.s. in an oracle inequality like (1.1) is called an optimal procedure in deviation for the convex aggregation problem.

The procedure constructed in [37] achieves the rate \(\psi_n(C_n) (M)\) in expectation (i.e. a procedure satisfying (1.1) in expectation with the optimal residual term \(\psi_n(C_n) (M)\)). An optimal procedure in deviation has been constructed in Theorem 2.8.1 in [21]. In both cases, the construction of these optimal aggregation procedures require the aggregation of an exponential number in \(M\) of functions in \(\text{conv}(F)\) and thus cannot be used in practice. On the other side, it would be much simpler and natural to consider the classical procedure of empirical risk minimization (cf. [39]) over the convex hull of \(F\) to solve the convex aggregation problem:

\[
\hat{f}_{\text{ERM}} \in \arg\min_{f \in \text{conv}(F)} R_n(f) \text{ where } R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2. \tag{1.3}
\]

In [17, 30, 24], the authors prove that, for every \(x > 0\), with probability greater than \(1 - 4 \exp(-x)\)

\[
R(\hat{f}_{\text{ERM}}) \leq \min_{f \in \text{conv}(F)} R(f) + c_0 \max \left( \phi_n(M), \frac{x}{n} \right) \text{ where } \phi_n(M) = \min \left( \frac{M}{n}, \sqrt{\log M / n} \right).
\]

The rate \(\phi_n(M)\) behaves like the optimal rate \(\psi_n(C_n) (M)\) except for values of \(M\) such that \(n^{1/2} < M \leq c(\epsilon)n^{1/2+\epsilon}\) for \(\epsilon > 0\) for which there is a logarithmic gap. In this note, we were able to remove this logarithmic loss proving that \(\hat{f}_{\text{ERM}}\) is indeed optimal for the convex aggregation problem. Finally, note that in [24], the authors show that the rate \(\psi_n(C_n) (M)\) can be achieved by \(\hat{f}_{\text{ERM}}\) for any orthogonal dictionary (i.e. such that \(\forall f \neq g \in F, E f(X) g(X) = 0\)). The performance of ERM in the convex hull has been studied for an infinite dictionary in [9]. The resulting upper bounds, in the case of a finite dictionary, is of the order of \(M/n\) for every \(n\) and \(M\).

Another motivation for this work comes from what is known about ERM in the context of the three aggregation schemes mentioned above. It is well-known that ERM in \(F\) is, in general, a suboptimal aggregation procedure for the model selection aggregation problem (see [18], [29] or [23]). It is also known that ERM in the linear span of \(F\) is an optimal procedure for the linear aggregation problem [20] (cf. Theorem 13 and Example 1) or [4]. Therefore, studying the performances of ERM in the convex hull of \(F\) in the context of convex aggregation can be seen as an “intermediate” problem which remained open. In fact, a lot of effort has been invested in finding any procedure that would be optimal.
for the convex aggregation problem. For example, many boosting algorithms (see [34] or [10] for recent results on this topic) are based on finding the best convex combination in a large dictionary (for instance, dictionaries consisting of “decision stumps”), while random forest algorithms can be seen as procedures that try finding the best convex combination of decision trees. Thus, finding an optimal procedure for the problem of convex aggregation for a general dictionary is of high practical importance. In the following result, we prove that empirical risk minimization is an optimal procedure for the convex aggregation problem.

**Theorem A** There exists absolute constants $c_0$ and $c_1$ such that the following holds. Let $F$ be a finite dictionary of cardinality $M$ and $(X,Y)$ be a random couple of $\mathcal{X} \times \mathbb{R}$ such that $|Y| \leq b$ and $\max_{f \in F} |f(X)| \leq b$ a.s. for some $b > 0$. Then, for any $x > 0$, with probability greater than $1 - 4 \exp(-x)$

$$R(f_n^{ERM-C}) \leq \min_{f \in \text{conv}(F)} R(f) + c_0 b^2 \max \left[ \psi_n^{(C)}(M), \frac{x}{n} \right].$$

The optimality also holds in expectation:

$$\mathbb{E} R(f_n^{ERM-C}) \leq \min_{f \in \text{conv}(F)} R(f) + c_1 b^2 \psi_n^{(C)}(M).$$

## 2 Preliminaries on isomorphic properties of functions classes

We recall the machinery developed in [5] to prove isomorphic results between the empirical and actual structures of functions classes.

Let $(Z,\sigma)$ be a measurable space, $Z, Z_1, \ldots, Z_n$ be $n + 1$ i.i.d. random variables with values in $Z$ distributed according to $P_Z$ and $G$ be a class of real-valued measurable functions defined on $Z$. We consider the star shaped hull of $G$ in zero and its localized set at some level $\lambda > 0$:

$$V(G) = \{\alpha g : 0 \leq \alpha \leq 1, g \in G\} \text{ and } V(G)_\lambda = \{h \in V(G) : Ph \leq \lambda\}.$$  

For any functions class $H$ (in particular for $H$ being $G$, $V(G)$ or $V(G)_\lambda$ for some $\lambda$), we denote $\|P - P_n\|_H = \sup_{h \in H} \|P - P_n\|_h$, where $Ph = E h(Z)$ and $P_n h = n^{-1} \sum_{i=1}^n h(Z_i)$, $\sigma(H) = \sup_{h \in H} \sqrt{Ph^2}$ and $\|H\|_\infty = \sup_{h \in H} \|h\|_{L_\infty(P_Z)}$. We also recall the separability condition of [28] (cf. Condition (M)) for which Talagrand’s concentration inequality holds:

(M) There exists $G_0 \subset G$ such that $G_0$ is countable and for any $g \in G$, there exists a sequence $(g_k)_k$ in $G_0$ such that for any $z \in Z$, $(g_k(z))_k$ tends to $g(z)$ when $k$ tends to infinity.

**Theorem 2.1 ([5])** There exists an absolute constant $c_0 > 0$ such that the following holds. Let $G$ be a class of real-valued measurable functions defined on $Z$ satisfying condition (M) and such that $Pg^2 \leq BPg, \forall g \in G$ for some constant $B > 0$. Let $\lambda^* > 0$ be such that

$$\mathbb{E} \|P - P_n\|_{V(G)_{\lambda^*}} \leq (1/8)\lambda^*.$$

(2.1)

4
For every $x > 0$, with probability greater than $1 - 4 \exp(-x)$, for every $g \in G$, 

$$|P_g - P_n g| \leq (1/2) \max(P_g, \rho_n(x)) \text{ where } \rho_n(x) = \max\left(\lambda^*, \frac{c_0(B + \|G\|_\infty x)}{n}\right).$$

For the reader convenience, we recall the short proof of [5].

**Proof.** Without loss of generality, we can assume that $G$ is countable. From a limit argument the result holds for classes of functions satisfying condition (M).

Fix $\lambda > 0$ and $x > 0$, and note that by Talagrand’s concentration inequality (cf. [35, 36, 27, 19, 8]), with probability larger than $1 - 4 \exp(-x)$,

$$\|P - P_n\|_{V(G)_\lambda} \leq 2\mathbb{E} \|P - P_n\|_{V(G)_\lambda} + K\sigma(V(G)_\lambda) \sqrt{\frac{x}{n}} + K\|V(G)_\lambda\|_\infty \frac{x}{n} \quad (2.2)$$

where $K$ is an absolute constant. Clearly, we have $\|V(G)_\lambda\|_\infty \leq \|G\|_\infty$ and

$$\sigma^2(V(G)_\lambda) = \sup\left(P(\alpha g)^2 : 0 \leq \alpha \leq 1, g \in G, P(\alpha g) \leq \lambda\right) \leq B\lambda.$$

Moreover, since $V(G)$ is star-shaped, $\lambda > 0 \rightarrow \phi(\lambda) = \mathbb{E} \|P - P_n\|_{V(G)_\lambda} / \lambda$ is non-increasing, and since $\phi(\lambda^*) \leq 1/8$ and $\rho_n(x) \geq \lambda^*$ then

$$\mathbb{E} \|P - P_n\|_{V(G)_{\rho_n(x)}} \leq (1/8)\rho_n(x).$$

Combined with (2.2), there exists an event $\Omega_0(x)$ of probability greater than $1 - 4 \exp(-x)$, and on $\Omega_0(x)$,

$$\|P - P_n\|_{V(G)_{\rho_n(x)}} \leq (1/4)\rho_n(x) + K\sqrt{\frac{B\rho_n(x)x}{n}} + K\|G\|_\infty \frac{x}{n} \leq (1/2)\rho_n(x)$$

as long as $c_0 \geq 64(K^2 + K)$. Hence, on $\Omega_0(x)$, if $g \in V(G)$ satisfies that $P_g \leq \rho_n(x)$, then $|P_g - P_n g| \leq (1/2)\rho_n(x)$. Moreover, if $g \in V(G)$ is such that $P_g > \rho_n(x)$, then $h = \rho_n(x)/P_g \in V(G)_{\rho_n(x)}$; hence $|P_h - P_n h| \leq (1/2)\rho_n(x)$, and so in both cases $|P_g - P_n g| \leq (1/2)\max(P_g, \rho_n(x))$.

Therefore, if one applies Theorem 2.1 to obtain isomorphic properties between the empirical and actual structures, one has to check the condition $Pg^2 \leq B Pg, \forall g \in G$, called the Bernstein condition in [5], and to find a point $\lambda^*$ satisfying (2.1).

A point $\lambda^*$ such that (2.1) holds can be found thanks to the peeling argument of [6]: for any $\lambda > 0$,

$$V(G)_\lambda \subset \bigcup_{i=0}^{\infty} \{\alpha g : 0 \leq \alpha \leq 2^{-i}, g \in G, P_g \leq 2^{i+1}\lambda\}$$

which implies

$$\mathbb{E} \|P - P_n\|_{V(G)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E} \|P - P_n\|_{G_{2^{i+1}\lambda}} \quad (2.3)$$

where, for any $\mu > 0$, $G_\mu = \{g \in G : Pg \leq \mu\}$. Then if $\lambda^* > 0$ is such that $\lambda^*/8$ upper bounds the RHS in (2.3) this point also satisfies (2.1).
The Bernstein condition usually follows from some convexity argument. For instance, it is now standard to check the Bernstein condition for the excess loss functions class \( L_F = \{ L_f : f \in F \} \) associated with a convex model \( F \) with respect to the squared loss function \( \ell_f(x, y) = (y - f(x))^2 \), \( \forall x \in \mathcal{X}, y \in \mathbb{R} \), where \( f^*_F \in \text{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2 \) and \( \mathcal{L}_f = \ell_f - \ell_{f^*_F} \). Indeed, if \( F \) is a convex set of functions and \( (X, Y) \) is a random couple on \( \mathcal{X} \times \mathbb{R} \) such that \( |Y| \leq b \) and \( \sup_{f \in F} |f(X)| \leq b \) a.s. then it follows from convexity and definition of \( f^*_F \) that for any \( f \in F \), \( \mathbb{E}[(f^*_F(X) - Y)(f^*_F(X) - f(X))] \leq 0 \) and so

\[
\mathbb{E}L_f = 2\mathbb{E}(f^*_F(X) - f(X))(Y - f^*_F(X)) + \mathbb{E}(f^*_F(X) - f(X))^2 \geq \mathbb{E}(f^*_F(X) - f(X))^2. \tag{2.4}
\]

Moreover, since \( |Y| \leq b \) and \( \sup_{f \in F} |f(X)| \leq b \) a.s. then

\[
\mathbb{E}L^2_f = \mathbb{E}(2Y - f^*_F(X) - f(X))^2(f(X) - f^*_F(X))^2 \leq (4b)^2\mathbb{E}(f(X) - f^*_F(X))^2. \tag{2.5}
\]

Therefore, any \( f \) in \( F \) is such that \( \mathbb{E}L^2_f \leq (4b)^2\mathbb{E}L_f \).

### 3 Proof of Theorem A

The proof of Theorem A for the case \( M \leq \sqrt{n} \) is now very classical and can be found in [20] (cf. Theorem 13 and Example 1). Nevertheless, we reproduce here this short proof in order to provide a self-contained note. The proof for the case \( M > \sqrt{n} \) is much trickier and relies on isomorphic properties of an exponential number of segments in \( \text{conv}(F) \) together with Maurey’s empirical method (cf. [31, 13]) which was first used in the context of convex aggregation in [30] and [37]. Note that segments are models of particular interest in Learning theory because they are convex models (in particular, they satisfy the Bernstein condition) and they are of small complexity (essentially the same complexity as a model of cardinality two). On the contrary to the classical entropy based approach which essentially consists in approximating a set by finite sets, approaching models by union of segments may be of particular interest in Learning theory beyond the convex aggregation problem. Note that finite models have no particular geometrical structure and therefore are somehow “bad models” as far as ERM procedures are concerned.

Proofs are given for the deviation result of Theorem A. The result in expectation of Theorem A follows from a direct integration argument.

#### 3.1 The case \( M > \sqrt{n} \)

We apply Theorem 2.1 to excess loss functions classes indexed by segments. First note that segments of bounded functions are functions classes satisfying condition (M). We consider a set \( \mathcal{C} = \{ g_1, \ldots, g_N \} \) of real-valued measurable functions defined on \( \mathcal{X} \) such that \( \max_{g \in \mathcal{C}} |g(X)| \leq b \) a.s. For every \( i, j \in \{ 1, \ldots, N \} \) we consider the segment \( \{ g_i, g_j \} = \{ \theta g_i + (1 - \theta)g_j : 0 \leq \theta \leq 1 \} \) and take \( g_{ij}^* \in \text{argmin}_{g \in [g_i, g_j]} R(g) \) where \( R(\cdot) \) is the squared risk. We consider the excess loss functions class

\[
\mathcal{L}^{ij} = \{ L_g^{ij} : g \in [g_i, g_j] \} \text{ where } L_g^{ij} = \ell_g - \ell_{g_{ij}^*},
\]
for \( \ell_g(x, y) = (y - g(x))^2 \), \( \forall x \in \mathcal{X}, y \in \mathbb{R} \).

As a consequence of convexity of segments, we have for any \( g \in [g_i, g_j] \), \( \mathbb{E}(\mathcal{L}^{ij}_g)^2 \leq (4b)^2 \mathcal{L}^{ij}_g \) (cf. (2.4) and (2.5) in Section 2). This implies that the functions class \( \mathcal{L}^{ij}_g \) satisfies the Bernstein condition of Theorem 2.1. Now, it remains to find \( \lambda \) together with a union bound to obtain the following result.

Let \( C \) be a random couple with values in \( \mathcal{X} \times \mathbb{R} \), \( \mathbb{E} \) satisfies the Bernstein condition of Theorem 2.1. Now, it remains to find \( \lambda \) (48) be a random couple with values in \( \mathcal{X} \times \mathbb{R} \), \( \mathbb{E} \) satisfies the Bernstein condition of Theorem 2.1. Now, it remains to find \( \lambda \) together with a union bound to obtain the following result.

Let \( C \) be a random couple with values in \( \mathcal{X} \times \mathbb{R} \), \( \mathbb{E} \) satisfies the Bernstein condition of Theorem 2.1. Now, it remains to find \( \lambda \) together with a union bound to obtain the following result.

Note that for any \( g \in [g_i, g_j] \), \( P\mathcal{L}^{ij}_g \geq P(g - g^{*}_{ij})^2 = \mathbb{E}(g(X) - g^{*}_{ij}(X))^2 \) (cf. (2.4)). It follows from the symmetrization argument and the contraction principle (cf. [25] page 95) that if \( g_i \neq g_j \) then

\[
\mathbb{E} \|P - P_n\|_{\mathcal{V}^{(ij)}_\mu} \leq 2 \mathbb{E} \sup_{g \in [g_i, g_j]: \mathcal{P}_{\mathcal{L}^{ij}_g} \leq \mu} \left| \frac{1}{n} \sum_{k=1}^{n} \epsilon_k \mathcal{L}^{ij}_g (X_k, Y_k) \right|
\]

\[
\leq 8b \mathbb{E} \sup_{g \in [g_i, g_j]: \mathcal{P}(g - g^{*}_{ij})^2 \leq \mu} \left| \frac{1}{n} \sum_{k=1}^{n} \epsilon_k (g(X_k) - g^{*}_{ij}(X_k)) \right|
\]

\[
\leq 8b \mathbb{E} \sup_{g \in [g_i, g_j]: \mathcal{P}(g - g^{*}_{ij})^2 \leq \mu} \left| \frac{1}{n} \sum_{k=1}^{n} \epsilon_k g(X_k) \right| \leq 8b \mathbb{E} \sup_{g \in \mathcal{P}(g - g^{*}_{ij})^2 \leq \mu} \left| \frac{1}{n} \sum_{k=1}^{n} \epsilon_k g(X_k) \right|
\]

\[
= \frac{8b \sqrt{\mu}}{P(g_i - g_j)^2} \mathbb{E} \left( \frac{1}{n} \sum_{k=1}^{n} \epsilon_k (g_i - g_j)(X_k) \right) \leq \frac{8b \sqrt{\mu}}{P(g_i - g_j)^2} \left( \mathbb{E} \left( \frac{1}{n} \sum_{k=1}^{n} \epsilon_k (g_i - g_j)(X_k) \right)^2 \right)^{1/2}
\]

\[
= 8b \sqrt{\frac{\mu}{n}}.
\]

Note that when \( g_i = g_j \) the result is also true. Now, we use the peeling argument of (2.3) to obtain

\[
\mathbb{E} \|P - P_n\|_{\mathcal{V}(\mathcal{L}^{ij})_{\lambda}} \leq \sum_{k=1}^{\infty} 2^{-k} \mathbb{E} \|P - P_n\|_{\mathcal{V}(\mathcal{L}^{ij})_{\lambda+1}} \leq \sum_{k=0}^{\infty} 2^{-k} 8b \sqrt{\frac{9k+1}{n}} \lambda \leq c_0 b \sqrt{\lambda/n}.
\]

Therefore, for \( \lambda^* = (8c_0 b)^2 / n \), we have \( \mathbb{E} \|P - P_n\|_{\mathcal{V}(\mathcal{L}^{ij})_{\lambda}} \leq (1/8) \lambda^* \).

Now, we can apply Theorem 2.1 to the family of excess loss functions classes \( \{\mathcal{L}^{ij}_g\}_{1 \leq i, j \leq N} \) together with a union bound to obtain the following result.

**Proposition 3.1** There exists an absolute constant \( c_0 > 0 \) such that the following holds.

Let \( \mathcal{C}' = \{g_1, \ldots, g_N\} \) be a set of measurable real-valued functions defined on \( \mathcal{X} \). Let \( (X, Y) \) be a random couple with values in \( \mathcal{X} \times \mathbb{R} \) such that \( |Y| \leq b \) and \( \max_{g \in \mathcal{C}'} |g(X)| \leq b \) a.s.. For any \( x > 0 \), with probability greater than \( 1 - 4 \exp(-x) \), for any \( i, j \in \{1, \ldots, N\} \) and any \( g \in [g_i, g_j] \),

\[
|P\mathcal{L}^{ij}_g - P_n\mathcal{L}^{ij}_g| \leq (1/2) \max \left( P\mathcal{L}^{ij}_g, \gamma(x) \right) \text{ where } \gamma(x) = \frac{c_0 b^2 (x + 2 \log N)}{n}.
\]
Now, we want to apply the isomorphic result of Proposition 3.1 to a wisely chosen subset \( C' \) of \( C = \text{conv}(F) \). For that we consider the integer
\[
m = \left\lceil \frac{n}{\log(eM/\sqrt{n})} \right\rceil
\]
and the set \( C' \) is defined by
\[
C' = \left\{ \frac{1}{m} \sum_{i=1}^{m} h_i : h_1, \ldots, h_m \in F \right\}.
\]
The set \( C' \) is an approximating set of the convex hull \( \text{conv}(F) \). We will for instance use the following approximation property:
\[
\min_{f \in C'} R(f) \leq \min_{f \in C} R(f) + \frac{4b^2}{m}.
\] (3.1)
Indeed, to obtain such a result, we use Maurey’s empirical method. Let \( f^*_C \in \text{argmin}_{f \in C} R(f) \) and denote \( f^*_C = \sum_{j=1}^{M} \lambda_j f_j \) where \( \lambda_j \geq 0, \forall j = 1, \ldots, M \) and \( \sum_{j=1}^{M} \lambda_j = 1 \). Consider a random variable \( \Theta : \Omega \rightarrow F \) such that \( P[\Theta = f_j] = \lambda_j, \forall j = 1, \ldots, M \) and let \( \Theta_1, \ldots, \Theta_m \) be \( m \) i.i.d. random variables distributed according to \( \Theta \) and independent of \( (X,Y) \). Denote by \( \mathbb{E}_\Theta \) the expectation with respect to \( \Theta_1, \ldots, \Theta_m \). Since \( \mathbb{E}_\Theta \Theta_j = f^*_C \) for any \( j = 1, \ldots, m \), we have
\[
\min_{f \in C'} R(f) \leq \mathbb{E}_\Theta R\left( \frac{1}{m} \sum_{j=1}^{m} \Theta_j \right) = \mathbb{E}_\Theta \mathbb{E}\left( \frac{1}{m} \sum_{j=1}^{m} \Theta_j(X) - Y \right)^2
\]
\[
= \mathbb{E}\left( \frac{1}{m^2} \sum_{j,k=1}^{m} \mathbb{E}_\Theta(Y - \Theta_j(X))(Y - \Theta_k(X)) \right) = R(f^*_C) + \frac{\mathbb{E}V_\Theta(Y - \Theta(X))}{m}
\]
where \( V_\Theta \) stands for the variance symbol with respect to \( \Theta \). Equation (3.1) follows since \( |Y| \leq b \) and \( \max_{f \in F} |f(X)| \leq b \) a.s..

Denote by \( N = |C'| \) the cardinality of \( C' \) and by \( g_1, \ldots, g_N \) the functions in \( C' \). For simplicity assume that \( R(g_1) = \min_{g \in C'} R(g) \). Thanks to [13] for the first inequality and [16], page 218, or [27] proposition 2, for the second inequality, we know that
\[
|C'| = N \leq \left( \frac{M + m - 1}{m} \right) \leq \left( \frac{2eM}{m} \right)^m.
\] (3.2)

Let \( x > 0 \). Consider the event \( \Omega(x) \subset \Omega \) such that the following isomorphic property holds for all the segments \( [g_1, g_j], j = 1, \ldots, N \):
\[
|P_n L_g^{ij} - P L_g^{ij}| \leq (1/2) \max (P L_g^{ij}, \gamma(x)), \quad \forall g \in [g_1, g_j]
\] (3.3)
where we recall that \( L_g^{ij} = \ell_g - \ell_{g_{ij}} \) is the excess loss function of \( g \in [g_1, g_j] \) for the model \( [g_1, g_j] \) and
\[
\gamma(x) = \frac{c_0 b^2 (x + 2 \log N)}{n}.
\]
Thanks to Proposition 3.1, we know that $\mathbb{P}[\Omega(x)] \geq 1 - 4 \exp(-x)$.

We are going to work on the event $\Omega(x)$ but for the moment, we use a second time Maurey’s empirical method. Fix $X_1, \ldots, X_n$ and write $\tilde{f}_n^{ERM-C} = \sum_{j=1}^{M} \beta_j f_j$. Consider a random variable $\Theta : \Omega' \rightarrow F$ defined on another probability space $(\Omega', \mathcal{A}', \mathbb{P}')$ such that $\mathbb{P}'[\Theta = f_j] = \beta_j, \forall j = 1, \ldots, M$ and let $\Theta_1, \ldots, \Theta_m$ be $m$ i.i.d. random variables having the same probability distribution as $\Theta$. Once again, denote by $\mathbb{E}_\Theta$ the expectation with respect to $\Theta_1, \ldots, \Theta_m$ and by $\mathbb{V}_\Theta$ the variance with respect to $\Theta$. Since $\mathbb{E}_\Theta \Theta_j = \tilde{f}_n^{ERM-C}$ for any $j = 1, \ldots, m$, it follows from the same method used to obtain (3.1) that

$$
\mathbb{E}_\Theta R \left( \frac{1}{m} \sum_{j=1}^{m} \Theta_j \right) = R(\tilde{f}_n^{ERM-C}) + \frac{\mathbb{E} \mathbb{V}_\Theta(Y - \Theta(X))}{m}
$$

and the same holds for the empirical risk:

$$
\mathbb{E}_\Theta R_n \left( \frac{1}{m} \sum_{j=1}^{m} \Theta_j \right) = R_n(\tilde{f}_n^{ERM-C}) + \frac{1}{m} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{V}'_\Theta(Y_i - \Theta(X_i)) \right).
$$

Consider the following notation:

$$
ge_\Theta = \frac{1}{m} \sum_{j=1}^{m} \Theta_j \text{ and } i_\Theta \in \{1, \ldots, N\} \text{ such that } g_{i_\Theta} = g_\Theta.
$$

Note that $g_\Theta$ is a random point in $\mathcal{C}'$ (as a measurable function from $\Omega'$ to $\mathcal{C}'$) and that, on the event $\Omega(x)$, the following isomorphic property on the segment $[g_1, g_\Theta]$ holds:

$$
|P_n L_{g_\Theta} - P L_{g_\Theta}| \leq (1/2) \max \left( P L_{g_{i_\Theta}} \gamma(x) \right), \quad \forall g \in [g_1, g_{i_\Theta}].
$$

First note that for every $\Theta_1, \ldots, \Theta_m$, we have

$$
R(\tilde{f}_n^{ERM-C}) = R(g_{i_\Theta}^*) + R(g_\Theta) - R(g_{1_{i_\Theta}}^*) + R(\tilde{f}_n^{ERM-C}) - R(g_\Theta).
$$

By definition of $g_{i_\Theta}^* \in \text{argmin}_{g \in [g_{i_\Theta}, g_1]} R(g)$, we have $R(g_{i_\Theta}^*) \leq R(g_1) = \min_{g \in \mathcal{C}} R(g)$ and according to (3.1) we have $\min_{f \in \mathcal{C}} R(f) \leq \min_{f \in \mathcal{C}} R(f) + (4b^2)/m$. Therefore, it follows from (3.7) that

$$
R(\tilde{f}_n^{ERM-C}) \leq \min_{f \in \mathcal{C}} R(f) + \frac{4b^2}{m} + P L_{g_\Theta} + R(\tilde{f}_n^{ERM-C}) - R(g_\Theta).
$$

On the event $\Omega(x)$, we use (3.6) to obtain for every $\Theta_1, \ldots, \Theta_m$

$$
R(\tilde{f}_n^{ERM-C}) \leq \min_{f \in \mathcal{C}} R(f) + \frac{4b^2}{m} + 2P_n L_{g_\Theta} + \gamma(x) + R(\tilde{f}_n^{ERM-C}) - R(g_\Theta).
$$

Moreover, by definition of $\tilde{f}_n^{ERM-C}$, we have

$$
P_n L_{g_\Theta} = R_n(g_\Theta) - R_n(g_{i_\Theta}^*) \leq R_n(g_\Theta) - R_n(\tilde{f}_n^{ERM-C}).
$$
Therefore, on the event $\Omega(x)$, we have for every $\Theta_1, \ldots, \Theta_m$

\[
R(\hat{f}_{n}^{\text{ERM}-C}) \leq \min_{f \in C} R(f) + \frac{4b^2}{m} + \gamma(x) + 2(R_n(g_0) - R_n(\hat{f}_{n}^{\text{ERM}-C})) + R(\hat{f}_{n}^{\text{ERM}-C}) - R(g_0).
\]

In particular, one can take the expectation with respect to $\Theta_1, \ldots, \Theta_m$ (defined on $\Omega'$) in the last inequality. We have on $\Omega(x)$,

\[
R(\hat{f}_{n}^{\text{ERM}-C}) \leq \min_{f \in C} R(f) + \frac{4b^2}{m} + \gamma(x) + 2\mathbb{E}_\Theta'(R_n(g_0) - R_n(\hat{f}_{n}^{\text{ERM}-C})) + \mathbb{E}_\Theta'(R(\hat{f}_{n}^{\text{ERM}-C}) - R(g_0)).
\]

Thanks to (3.4), we have $\mathbb{E}_\Theta'(R(\hat{f}_{n}^{\text{ERM}-C}) - R(g_0)) \leq 0$ and it follows from (3.5) that $\mathbb{E}_\Theta'(R_n(g_0) - R_n(\hat{f}_{n}^{\text{ERM}-C})) \leq (2b)^2/m$. Therefore, on the event $\Omega(x)$, we have

\[
R(\hat{f}_{n}^{\text{ERM}-C}) \leq \min_{f \in C} R(f) + \frac{8b^2}{m} + \gamma(x) \leq \min_{f \in C} R(f) + c_1b^2 \max(\psi(C)(M), x/n)
\]

where the last inequality follows from (3.2) and the definition of $m$.

### 3.2 The case $M \leq \sqrt{n}$

We use the strategy developed in [5] together with the one of [20] (cf. example 1) to prove Theorem A in the case $M \leq \sqrt{n}$. Define $C = \text{conv}(F)$ and $\mathcal{L}_C = \{L_f : f \in C\}$ the excess loss class associated with $C$ where $L_f = \ell_f - \ell_f^*, \forall f \in C$ and $f^*_C \in \text{argmin}_{f \in C} R(f)$.

Let $x > 0$. Assume that we can find some $\rho_n(x) > 0$ such that with probability greater than $1 - 4 \exp(-x)$, for any $f \in C$,

\[
|P_n(L_f - P\mathcal{L}_f)| \leq (1/2) \max (P\mathcal{L}_f, \rho_n(x)).
\]

Then, the ERM over $\text{conv}(F)$ would satisfy with probability greater than $1 - 4 \exp(-x)$,

\[
R(\hat{f}_{n}^{\text{ERM}-C}) - \min_{f \in \text{conv}(F)} R(f) = P\mathcal{L}_f^2 \leq 2P_nL_{\hat{f}_{n}^{\text{ERM}-C}} + \rho_n(x) \leq \rho_n(x).
\]

This means that if we can prove some isomorphic properties between the empirical and the actual structures of the functions class $\mathcal{L}_C$ like in (3.9), then we can derive oracle inequalities for $\hat{f}_{n}^{\text{ERM}-C}$. This is the strategy used in [5] that we follow here.

According to Theorem 2.1 in Section 2, a function $\rho_n(x)$ satisfying (3.9) can be constructed if we prove that $\mathcal{L}_C$ satisfies some Bernstein condition and if we find some fixed point $\lambda^* > 0$ such that $\mathbb{E}\|P - P_n\|_{\text{V}(\mathcal{L}_C)^{\lambda^*}} \leq (1/8)\lambda^*$. The Bernstein condition follows from the convexity of $\text{conv}(F)$ and the strategy used in Section 2: for any $f \in C, P\mathcal{L}_f^2 \leq (4b)^2P\mathcal{L}_f$.

We use the peeling argument of Section 2 together with the following observations due to [20] (cf. example 1) to find a fixed point $\lambda^*$. Let $S$ be the linear subspace of
$L^2(P_X)$ spanned by the dictionary $F$ and consider an orthonormal basis $(e_1, \ldots, e_{M'})$ of $S$ in $L^2(P_X)$ (where $M' = \dim(S) \leq M$). For any $\mu > 0$, it follows from the symmetrization argument and the contraction principle (cf. Chapter 4 in [25]) that

$$
E \|P - P_n\|_{(L^2)_{\mu}} \leq 8b \mathbb{E} \sup_{f \in S; \|f\|_2^2 \leq \mu} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|
$$

\begin{align*}
&\leq 8b \mathbb{E} \sup_{\beta \in \mathbb{R}^{M'}; \|\beta\|_2 \leq \sqrt{\mu}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left( \sum_{j=1}^{M'} \beta_j e_j(X_i) \right) \right| \\
&\leq 8b \sqrt{\mu} \mathbb{E} \left( \sum_{j=1}^{M'} \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i e_j(X_i) \right)^2 \right)^{1/2} \leq 8b \sqrt{\frac{M'\mu}{n}}.
\end{align*}

We use the peeling argument of (2.3) to prove that for $\lambda^* = c_0 b^2 M/n$ and $c_0$ an absolute constant large enough, we have indeed $E \|P - P_n\|_{V(L^2)} \leq (1/8)\lambda^*$. Now, it follows from Theorem 2.1 that for any $x > 0$, with probability greater than $1 - 4 \exp(-x)$,

$$
R(\tilde{f}_n^{\text{ERM}-C}) \leq \min_{f \in \mathcal{C}} R(f) + c_1 b^2 \max\left( \frac{M}{n}, \frac{x}{n} \right).
$$

This concludes the proof for the case $M \leq \sqrt{n}$.

**Remark 3.2** We did not use the condition $M \leq \sqrt{n}$ in the last proof. In fact, the result holds in the following more general framework. Let $\Lambda$ be any closed convex subset of $\mathbb{R}^M$ and for any dictionary $F = \{f_1, \ldots, f_M\}$ denote by $\Lambda(F)$ the set of all functions $\sum_{j=1}^{M} \lambda_j f_j$ when $(\lambda_1, \ldots, \lambda_M)^\top \in \Lambda$. Let $(X,Y)$ be a random couple with values in $X \times \mathbb{R}$ such that $|Y| \leq b$ and $\max_{f \in F} |f(X)| \leq b$ a.s.. Consider the ERM procedure

$$
\tilde{f}_n \in \arg\min_{f \in \Lambda(F)} R_n(f).
$$

Then, it follows from Theorem 2.1 and the argument used previously in this section that for any $x > 0$, with probability greater than $1 - 4 \exp(-x)$,

$$
R(\tilde{f}_n) \leq \min_{f \in \Lambda(F)} R(f) + c_1 b^2 \max\left( \frac{M}{n}, \frac{x}{n} \right).
$$

The same result can be found in [4] under very weak moment assumptions.

**Acknowledgements:** We would like to thank Alexandre Tsybakov for helping us for the presentation of this result.

**References**


