Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2013, Vol. 49, No. 1, 288–306
DOI: 10.1214/11-AIHP458

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

On the optimality of the empirical risk minimization procedure for the Convex aggregation problem[1]

Guillaume Lecué[a] and Shahar Mendelson[b]

[a]*CNRS, LAMA, Université Paris-Est Marne-la-Vallée, 77454 Marne-la-Vallée, France. E-mail: guillaume.lecue@univ-mlv.fr*
[b]*Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel. E-mail: shahar@tx.technion.ac.il*

Received 24 September 2010; revised 12 October 2011; accepted 14 October 2011

Abstract. We study the performance of *empirical risk minimization* (ERM), with respect to the quadratic risk, in the context of *convex aggregation*, in which one wants to construct a procedure whose risk is as close as possible to the best function in the convex hull of an arbitrary finite class F. We show that ERM performed in the convex hull of F is an optimal aggregation procedure for the convex aggregation problem. We also show that if this procedure is used for the problem of model selection aggregation, in which one wants to mimic the performance of the best function in F itself, then its rate is the same as the one achieved for the convex aggregation problem, and thus is far from optimal. These results are obtained in deviation and are sharp up to logarithmic factors.

Résumé. Nous étudions les performances de la procédure de minimisation du risque empirique, par rapport au risque quadratique, pour le problème d'agrégation convexe. Dans ce problème, on souhaite construire des procédures dont le risque est aussi proche que possible du risque du meilleur élément dans l'enveloppe convexe d'une classe finie F de fonctions. Nous prouvons que la procédure obtenue par minimisation du risque empirique sur la coque convexe de F est une procédure optimale pour le problème d'aggrégation convexe. Nous prouvons aussi que si cette procédure est utilisée pour le problème d'agrégation en sélection de modèle, pour lequel on souhaite imiter le meilleur dans F, alors le résidu d'agrégation est le même que celui obtenue pour le problème d'agrégation convexe. Cette procédure est donc loin d'être optimale pour le problème d'agrégation en sélection de modèle. Ces résultats sont obtenus en déviation et sont optimaux à des facteurs logarithmiques prés.

MSC: 62G08

Keywords: Learning theory; Aggregation theory; Empirical process theory

1. Introduction and main results

In this note, we study the optimality of the empirical risk minimization procedure in the aggregation framework.

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, $\mathcal{D} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is the family of given data.

The *quadratic risk* of a real-valued function f defined on $\mathcal{X}$ is given by

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

[1]Part of this research was supported by the Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia, by an Australian Research Council Discovery Grant DP0559465 and by the European Community's Seventh Framework Programme (FP7/2007-2013), ERC Grant Agreement 203134.

If $\widehat{f}$ is a function constructed using the data $\mathcal{D}$, the quadratic risk of $\widehat{f}$ is the random variable

$$R(\widehat{f}) = \mathbb{E}\big[\big(Y - \widehat{f}(X)\big)^2 | \mathcal{D}\big].$$

For the sake of simplicity, throughout this article we will restrict ourselves to functions f and random variables (X, Y) for which $|Y|, |f(X)| \leq b$ almost surely, for some fixed $b \geq 1$. One should note, though, that it is possible to extend the results beyond this case, to functions with well behaved tail – though at a high technical price (cf. the chaining arguments in [20] and [21]).

In the aggregation framework, one is given a finite set F of real-valued functions defined on $\mathcal{X}$ (usually called a *dictionary*) of cardinality M. There are three main types of aggregation problems:

1. In the *Model Selection (MS) aggregation* problem, one has to construct a procedure that produces a function whose risk is as close as possible to the risk of the best element in the given class F (cf. [2,3,9–12,16,24,25,27]).
2. In the *Convex (C) aggregation* problem (cf. [1,7–9,12,24,28]) one wants to construct a procedure whose risk is as close as possible to the risk of the best function in the convex hull of F (later denoted by $\mathrm{conv}(F)$).
3. In the *Linear (L) aggregation* problem (cf. [9,11,15,24]), one wants to construct a procedure whose risk is as close as possible to the risk of the best function in the linear span of F (later denoted by $\mathrm{span}(F)$).

The aim in the aggregation framework is to construct a procedure $\tilde{f}$ for which, with high probability

$$R(\tilde{f}) \leq C \min_{f \in \Delta(F)} R(f) + \psi_n^{\Delta(F)}(M) \tag{1.1}$$

with $C = 1$ and $\Delta(F)$ is either F, or $\mathrm{conv}(F)$ or $\mathrm{span}(F)$. It is worth mentioning that it is desirable for the constant C in (1.1) to be one in the aggregation setup for at least two reasons. First, there are some obvious mathematical differences in the analysis leading to exact oracle inequalities ($C = 1$) and non-exact oracle inequalities ($C > 1$). In particular, the geometry of the set $\Delta(F)$ has a key role in an attempt to obtain exact oracle inequalities, whereas non-exact oracle inequalities are mainly based on complexity and concentration argument (cf. [17]). Second, an exact oracle inequality for the prediction risk $R(\cdot)$ leads to an exact oracle inequality for the estimation risk; namely, with high probability

$$\mathbb{E}\big[\big(\tilde{f}(X) - f^*(X)\big)^2 | \mathcal{D}\big] \leq \min_{f \in \Delta(F)} \mathbb{E}\big[\big(f(X) - f^*(X)\big)^2\big] + \psi_n^{\Delta(F)}(M),$$

where f^* denotes the regression function of Y given X. Such an estimate on the regression function cannot follow from a non-exact oracle inequality, and thus, exact oracle inequalities can provide prediction and estimation results whereas non-exact oracle inequalities only lead to prediction results.

One can define the *optimal rates of the (MS), (C) and (L) aggregation* problems, respectively denoted by $\psi_n^{(\mathrm{MS})}(M)$, $\psi_n^{(\mathrm{C})}(M)$ and $\psi_n^{(\mathrm{L})}(M)$ (see, for example, [24]). The optimal rates are the smallest prices in the minimax sense that one has to pay to solve the (MS), (C) or (L) aggregation problems in expectation, as a function of the cardinality M of the dictionary and of the sample size n. It has been proved in [24] (see also [12] and [28] for the (C) aggregation problem) that

$$\psi_n^{(\mathrm{MS})}(M) \sim \frac{\log M}{n}, \qquad \psi_n^{(\mathrm{C})}(M) \sim \begin{cases} \frac{M}{n} & \text{if } M \leq \sqrt{n}, \\ \sqrt{\frac{1}{n}\log\big(\frac{eM}{\sqrt{n}}\big)} & \text{if } M > \sqrt{n} \end{cases} \quad \text{and} \quad \psi_n^{(\mathrm{L})}(M) \sim \frac{M}{n},$$

where we denote $a \sim b$ if there are absolute positive constants c and C such that $cb \leq a \leq Cb$. Note that the rates obtained in [24] hold in expectation and in particular, the rate $\psi_n^{(\mathrm{C})}(M)$ was achieved in the Gaussian regression model with a known variance and a known marginal distribution of the design. In [8], 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). We also refer the reader to [6,28] for non-exact oracle inequalities in the (C) aggregation context.

Lower bounds in deviation follow from the arguments of [24] for the three aggregation problems with the same rates $\psi_n^{(\mathrm{MS})}(M)$, $\psi_n^{(\mathrm{C})}(M)$ and $\psi_n^{(\mathrm{L})}(M)$. In other words, there exist two absolute constants $c_0, c_1 > 0$ such that for

any sample cardinality $n \geq 1$, any cardinality of a dictionary $M \geq 1$ and any aggregation procedure $\bar{f}_n$, there exists a dictionary F of size M such that with probability larger than c_0,

$$R(\bar{f}_n) \geq \min_{f \in \Delta(F)} R(f) + c_1 \psi_n^{\Delta(F)}(M), \tag{1.2}$$

where the residual term $\psi_n^{\Delta(F)}(M)$ is $\psi_n^{(\mathrm{MS})}(M)$ (resp. $\psi_n^{(\mathrm{C})}(M)$ or $\psi_n^{(\mathrm{L})}(M)$) when $\Delta(F) = F$ (resp. $\Delta(F) = \operatorname{conv}(F)$ or $\Delta(F) = \operatorname{span}(F)$). Procedures achieving these rates in deviation have been constructed for the (MS) aggregation problem ([2] and [16]) and the (L) aggregation problem ([15]). So far, there was no example of a procedure that achieves the rate of aggregation $\psi_n^{(\mathrm{C})}(M)$ with high probability for the (C) aggregation problem and the aim of this note is to prove that the most natural procedure, empirical risk minimization over the convex hull of F, achieves the rate of $\psi_n^{(\mathrm{C})}(M)$ in deviation (up to a $\log n$ factor for values of M close to $\sqrt{n}$).

Indeed, we will show that the procedure $\tilde{f}^{\mathrm{ERM\text{-}C}}$ minimizing the empirical risk functional

$$f \longmapsto R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2, \tag{1.3}$$

in $\operatorname{conv}(F)$ achieves, with high probability, the rate $\min(\frac{M}{n}, \sqrt{\frac{\log M}{n}})$ for the (C) aggregation problem (see the exact formulation in Theorem A.1 in the Appendix). Moreover, we will show that the rate $\psi_n^{(\mathrm{C})}(M)$ can be achieved by $\tilde{f}^{\mathrm{ERM\text{-}C}}$ for any orthogonal dictionary (formulated in Theorem B). On the other hand, it turns out that the same algorithm is far from the conjectured optimal rate $\psi_n^{(\mathrm{MS})}(M)$ for the (MS) aggregation problem (see Theorem A and [16] for the conjecture).

Our first main result is to prove a lower bound on the performance of $\tilde{f}^{\mathrm{ERM\text{-}C}}$ (ERM in the convex hull) in the context of the (MS) aggregation problem. In [16], it was proved that this procedure is suboptimal for the problem of (MS) aggregation when the size of the dictionary is of the order of $\sqrt{n}$. Here we complement the result by providing a lower bound for almost all values of M and n.

Theorem A. *There exist two absolute positive constants c_0 and c_1 for which the following holds. For any integer n and M such that $\log M \leq c_0 n^{1/3}$, there exists a dictionary F of cardinality M such that, with probability greater than $9/12$*

$$R\big(\tilde{f}^{\mathrm{ERM\text{-}C}}\big) \geq \min_{f \in F} R(f) + c_2 \psi_n(M),$$

where $\psi_n(M) = M/n$ when $M \leq \sqrt{n}$ and $(n \log(\mathrm{e}M/\sqrt{n}))^{-1/2}$ when $M > \sqrt{n}$. Moreover, for the same class F, if $M \geq \sqrt{n}$, then with probability larger than $7/12$,

$$R\big(\tilde{f}^{\mathrm{ERM\text{-}C}}\big) \leq \min_{f \in F} R(f) + c_3 \psi_n(M).$$

Note that the residual term $\psi_n(M)$ of Theorem A is much larger than the optimal rate $\psi_n^{(\mathrm{MS})}(M) = (\log M)/n$ for the (MS) aggregation problem. It shows that ERM in the convex hull satisfies a much stronger lower bound than the one mentioned in (1.2) that holds for any algorithm. This result is of particular importance since optimal aggregation procedures for the (MS) aggregation problem take their values in $\operatorname{conv}(F)$, and it was thus conjectured that $\tilde{f}^{\mathrm{ERM\text{-}C}}$ could be an optimal aggregation procedure for the (MS) aggregation problem (cf. [16] for more details on this problem). In [16] it was proved that this not the case for $M = \sqrt{n}$; Theorem A shows that this is not the case for all the values of M and n in the significant range (when M is sub-exponential in n).

The proof of Theorem A requires two separate arguments (as in the proofs of the lower bounds in [24] and [28]). The case $M \leq \sqrt{n}$ is easier, and follows an identical path to the one used in [16] for $M = \sqrt{n}$. Its proof is presented for the sake of completeness, and to allow the reader a comparison with the situation in the other case, when $M > \sqrt{n}$. In the "large M" range things are very different and we present a more intuitive description of the idea behind the construction in Section 2.

The performance of ERM in the convex hull has been studied for an infinite dictionary in [7], in which estimates on its performance have been obtained in terms of the metric entropy of F. The resulting upper bounds were conjectured to be suboptimal in the case of a finite dictionary, since they provide an upper bound of M/n for every n and M whereas it is possible to achieve the rate $\sqrt{(\log M)/n}$ when $M \geq \sqrt{n}$. Although this result is probably known to experts and relies on standard machinery (see for instance [15,14]), we present its proof in the Appendix.

The residual term $\min(\frac{M}{n}, \sqrt{\frac{\log M}{n}})$ of Theorem A.1 behaves like $\psi_n^{(\mathrm{C})}(M)$ except for values of M for which $n^{1/2} < M \leq c(\varepsilon)n^{1/2+\varepsilon}$ for $\varepsilon > 0$. And, although there is a gap in this range in the general case, under the additional assumption that the dictionary is orthogonal, this gap can be removed.

Theorem B. *For every $b > 0$ there is a constant $c_1(b)$ and an absolute constant c_2 for which the following holds. Let n and M be integers which satisfy that $\log M \leq c_1(b)\sqrt{n}$. Let F be a finite dictionary F of cardinality M and (X, Y) such that $|Y|, \sup_{f\in F} |f(X)| \leq b$. If $F = \{f_1, \ldots, f_M\}$ satisfies that $\mathbb{E} f_i(X) f_j(X) = 0$ for any $i \neq j \in \{1, \ldots, M\}$, then $\tilde{f}^{\text{ERM-C}}$ achieves the rate $\psi_n^{(C)}(M)$: for any $u > 0$, with probability greater than $1 - \exp(-u)$*

$$R\big(\tilde{f}^{\text{ERM-C}}\big) \leq \min_{f\in \operatorname{conv}(F)} R(f) + c_2 b^2 \max\left[\psi_n^{(C)}(M), \frac{u}{n}\right].$$

Removing the gap in the general case is likely to be a much harder problem, although we believe that the orthogonal case should be the "worst" one.

Finally, a word about notation. Throughout, we denote absolute constants or constants that depend on other parameters by c, C, c_1, c_2, etc. (and, of course, we will specify when a constant is absolute and when it depends on other parameters). The values of constants may change from line to line. The notation $x \sim y$ (resp. $x \lesssim y$) means that there exist absolute constants $0 < c < C$ such that $cy \leq x \leq Cy$ (resp. $x \leq Cy$). If $b > 0$ is a parameter then $x \lesssim_b y$ means that $x \leq C(b)y$ for some constant $C(b)$ depending only on b. We denote by ℓ_p^M the space $\mathbb{R}^M$ endowed with the ℓ_p norm. The unit ball there is denoted by B_p^M. We also denote the unit Euclidean sphere in $\mathbb{R}^M$ by $\mathcal{S}^{M-1}$.

If F is a class of functions, let f^* be a minimizer in F of the true risk; in our case, f^* is the minimizer of $\mathbb{E}(f(X) - Y)^2$. For every $f \in F$ set $\mathcal{L}_f = (Y - f(X))^2 - (Y - f^*(X))^2$, and let $\mathcal{L}_F = \{\mathcal{L}_f\colon\ f \in F\}$ be the excess loss class associated with F, the target Y and the quadratic risk.

2. On the complexity of B_1^M with respect to ℓ_2^M

The aim of this section is to give some of the ideas needed in the proof of Theorem A in the case $M \geq \sqrt{n}$. It is also presented to explain why the seemingly unlikely fact that the rate

$$\frac{1}{\sqrt{n \log(\mathrm{e}M/\sqrt{n})}} \tag{2.1}$$

actually improves as the size of the dictionary M increases in our construction is true.

The example used for this result is a class $F_M = \{0, \pm\phi_1, \ldots, \pm\phi_M\}$ where $(\phi_i)_{i=1}^M$ is a bounded orthonormal family of $L_2(P^X)$ and $Y = \phi_{M+1}(X)$ is orthogonal to this family. We also assume that $\Phi(X) = (\phi_1(X), \ldots, \phi_M(X))$ is isotropic, that is, for every $\lambda \in \mathbb{R}^n$, $\mathbb{E}\langle\Phi(X), \lambda\rangle^2 = \|\lambda\|_2^2$.

An element in $\operatorname{conv}(F_M)$ is of the form $f_\lambda = \langle\Phi, \lambda\rangle$ for some $\lambda \in B_1^M$, its excess loss is $\mathcal{L}_{f_\lambda} = \langle\Phi, \lambda\rangle^2 - 2\langle\Phi, \lambda\rangle\phi_{M+1}$ and the process one has to minimize is indexed by B_1^M and given by

$$P_n\mathcal{L}_{f_\lambda} = \frac{1}{n}\sum_{i=1}^n \big\langle\Phi(X_i), \lambda\big\rangle^2 - \frac{2}{n}\sum_{i=1}^n \big\langle\Phi(X_i), \lambda\big\rangle\phi_{M+1}(X_i). \tag{2.2}$$

It follows from [21] that the oscillations of the quadratic term $\lambda \in B_1^M \to |(P_n - P)(\langle\Phi, \lambda\rangle^2)|$ are of lower order, and that the empirical process (2.2) behaves like $\lambda \in B_1^M \to \|\lambda\|_2^2 - 2n^{-1/2}\langle V, \lambda\rangle$ where $V =$

$n^{-1/2}\sum_{i=1}^n \phi_{M+1}\Phi(X_i)$, while a Gaussian approximation shows that V essentially behaves like a standard Gaussian vector G in $\mathbb{R}^M$. Hence, the excess risk $P\mathcal{L}_{\widehat{f}} = \|\widehat{\lambda}\|_2^2$ of the empirical risk minimization procedure $\widehat{f} = f_{\widehat{\lambda}}$ will be located around

$$\arg\min_{0\le r\le 1}\min_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}}\left(r - 2\frac{\langle G,\lambda\rangle}{\sqrt{n}}\right) = \arg\min_{0\le r\le 1}\left(r - 2n^{-1/2}\sup_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}}\langle G,\lambda\rangle\right).$$

Observe that for every radius $0 < r \le 1$, $\sup_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}}\langle\cdot,\lambda\rangle$, is an interpolation norm, which will be denoted by $\|\cdot\|_{A_r^\circ}$. The problem arises because in the range $1/M \le r \le 1$ (which is the range we are interested in), a proportional change in the radius r only results in a logarithmic change in the value of $\mathbb{E}\|G\|_{A_r^\circ}$, which is why one has to obtain a sharp estimate on $\mathbb{E}\|G\|_{A_r^\circ}$ for every r.

It turns out that a rather accurate estimate on the complexity of $B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}$ comes from vectors of "short" support. Namely, for every $I\subset\{1,\dots,M\}$, let $\mathcal{S}^I$ be the set of vectors in $\mathcal{S}^{M-1}$ supported in I. Set

$$C_k = \bigcup_{|I|=k}\frac{1}{\sqrt{k}}\mathcal{S}^I \subset B_1^M\cap\frac{1}{\sqrt{k}}\mathcal{S}^{M-1}.$$

If one replaces $B_1^M\cap\frac{1}{\sqrt{k}}\mathcal{S}^{M-1}$ by C_k, it is much easier to analyze ERM over that set. Indeed, it is straightforward to verify that ERM is likely to choose a vector in C_k, where k minimizes the functional

$$k\to\frac{1}{\sqrt{k}} - \frac{2}{\sqrt{n}}\cdot\mathbb{E}\sup_{v\in C_k}\langle G,v\rangle = \frac{1}{\sqrt{k}} - \frac{2}{\sqrt{n}}\mathbb{E}\left(\sum_{i=1}^k (g_i^2)^*\right)^{1/2}, \tag{2.3}$$

where (x_i^*) is a non-increasing ordering of the vector $(|x_i|)$.

A sharp estimate on the Gaussian quantity reveals that the gap between the "level" k and the "level" ℓ decrease with the dimension M. Thus, the minimum of (2.3) – which is proportional to (2.1) – decreases as M increases.

The proof of Theorem A will be a combination of two approximation arguments – first, of the measure $n^{-1/2}\sum_{i=1}^n X_i$ by a Gaussian, and second, an approximation of B_1^M by the sets C_k, reducing the problem to the one described above.

One should comment that it is possible to approximate $B_1^M\cap\frac{1}{\sqrt{k}}\mathcal{S}^{M-1}$ using a completely combinatorial set $\bigcup_{|I|=k}k^{-1}\{-1,1\}^I$, and the way the complexities change between the levels k and ℓ as M increases gives a more geometric explanation to why the minimizer moves closer to 0.

3. Proof of the lower bound for the (MS) aggregation problem (Theorem A)

The proof of Theorem A consists of two parts. The first, simpler part, is when $M\le\sqrt{n}$. This is due to the fact that if $0<\theta<1$ and $\rho=\theta r\sim M/n$, the set $B_1^M\cap\sqrt{r}S^{M-1}$ is much "larger" than the set $B_1^M\cap\sqrt{\rho}B_2^M$. This results in much larger "oscillations" of the appropriate empirical process on the former set than on the latter one, leading to very negative values of the empirical excess risk functional for functions whose excess risk larger than ρ. The case $M\ge\sqrt{n}$ is much harder because when considering the required values of r and ρ, the complexity of the two sets is very close, and comparing the two oscillations accurately involves a far more delicate analysis.

3.1. *The case $M\le\sqrt{n}$*

We will follow the method used in [16]. Let $(\phi_i)_{i\in\mathbb{N}}$ be a sequence of functions defined on $[0,1]$ and set μ to be a probability measure on $[0,1]$ such that $(\phi_i\colon i\in\mathbb{N})$ is a sequence of independent Rademacher variables in $L_2([0,1],\mu)$.

Let $M\le\sqrt{n}$ be fixed and put (X,Y) to be a couple of random variables; X is distributed according to μ and $Y=\phi_{M+1}(X)$. Let $F=\{0,\pm\phi_1,\dots,\pm\phi_M\}$ be the dictionary, and note that any function in the convex hull of F

can be written as $f_\lambda = \sum_{j=1}^M \lambda_j \phi_j$ for $\lambda \in B_1^M$. Since relative to $\mathrm{conv}(F)$, $f^* = 0$, the excess quadratic loss function is

$$\mathcal{L}_\lambda(X,Y) = -2\phi_{M+1}(X)\big\langle \lambda, \Phi(X)\big\rangle + \big\langle \lambda, \Phi(X)\big\rangle^2,$$

where we set $\Phi(\cdot) = (\phi_1(\cdot), \dots, \phi_M(\cdot))$.

The following is a reformulation of Lemma 5.4 in [16].

Lemma 3.1. *There exist absolute constants c_0, c_1 and c_2 for which the following holds. Let $(X_i, Y_i)_{i=1,\dots,n}$ be n independent copies of (X,Y). Then, for every $r > 0$, with probability greater than $1 - 8\exp(-c_0 M)$, for any $\lambda \in \mathbb{R}^M$,*

$$\left| \|\lambda\|_2^2 - \frac{1}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle^2 \right| \le \frac{1}{2}\|\lambda\|_2^2 \tag{3.1}$$

and

$$c_1\sqrt{\frac{rM}{n}} \le \sup_{\lambda \in \sqrt{r}B_2^M} \frac{1}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle \phi_{M+1}(X_i) \le c_2\sqrt{\frac{rM}{n}}. \tag{3.2}$$

Set $r = \beta M/n$ for some $0 < \beta \le 1$ to be named later, and observe that $B_1^M \cap \sqrt{r}S^{M-1} = \sqrt{r}S^{M-1}$ because $r \le 1/M$. For any $\lambda \in \sqrt{r}\mathcal{S}^{M-1}$, $P\mathcal{L}_\lambda = \|\lambda\|_2^2 = r$, and thus applying (3.1) and (3.2), it is evident that with probability greater than $1 - 8\exp(-c_0 M)$,

$$\begin{aligned}
\inf_{\lambda \in B_1^M \cap \sqrt{r}S^{M-1}} P_n\mathcal{L}_\lambda &= r - \sup_{\lambda \in \sqrt{r}S^{M-1}} (P - P_n)\mathcal{L}_\lambda \\
&\le r + \sup_{\lambda \in \sqrt{r}S^{M-1}} \left| \|\lambda\|_2^2 - \frac{1}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle^2 \right| - \sup_{\lambda \in \sqrt{r}S^{M-1}} \frac{2}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle \phi_{M+1}(X_i) \\
&\le \frac{3r}{2} - 2c_1\sqrt{\frac{rM}{n}} = \left(\frac{3\beta}{2} - 2c_1\sqrt{\beta}\right)\frac{M}{n} \le -c_1\sqrt{\beta}\frac{M}{n},
\end{aligned}$$

provided that $\beta \le (2c_1/3)^2$.

On the other hand, let $\rho = \alpha M/n$ for some α to be chosen later. Using (3.1) and (3.2) again, it follows that with probability at least $1 - 8\exp(-c_0 M)$, for any $\lambda \in B_1^M \cap \sqrt{\rho}B_2^M$

$$\begin{aligned}
|P_n\mathcal{L}_\lambda| &\le P\mathcal{L}_\lambda + \left| \|\lambda\|_2^2 - \frac{1}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle^2 \right| + \left| \frac{2}{n}\sum_{i=1}^n \big\langle \lambda, \Phi(X_i)\big\rangle \phi_{M+1}(X_i) \right| \\
&\le \frac{3\rho}{2} + 2c_2\sqrt{\frac{\rho M}{n}} = \left(\frac{3\alpha}{2} + 2c_2\sqrt{\alpha}\right)\frac{M}{n}.
\end{aligned}$$

Therefore, if $0 < \alpha < \beta$ satisfies that $3\alpha/2 + 2c_2\sqrt{\alpha} < c_1\sqrt{\beta}$ for some $0 < \beta \le (2c_1/3)^2$ then with probability greater than $1 - 16\exp(-c_0 M)$, the empirical risk function $\lambda \longmapsto R_n(f_\lambda)$ achieves smaller values on $B_1^M \cap \sqrt{r}S^{M-1}$ than on $B_1^M \cap \sqrt{\rho}B_2^M$. Hence, with the same probability, $R(\tilde{f}^{\text{ERM-C}}) \ge \rho = \alpha M/n$.

3.2. *The case $M \ge \sqrt{n}$*

Let us reformulate the second part of Theorem A.

Theorem 3.2. *There exist absolute constants c_0, c_1, c_2 and n_0 for which the following holds. For every integers $n \ge n_0$ and M, if $\sqrt{n} \le M \le \exp(c_0 n^{1/3})$, there is a function class F_M of cardinality M consisting of functions that are*

bounded by 1, *and a couple* (X,Y) *distributed according to a probability measure* μ, *such that with* $\mu^{\otimes n}$*-probability at least* $9/12$,

$$R(\widehat{f}) \geq \min_{f\in F_M} R(f) + \frac{c_1}{\sqrt{n\log(\mathrm{e}M/\sqrt{n})}},$$

where $\widehat{f}$ *is the empirical minimizer in* $\operatorname{conv}(F_M)$. *Moreover, with* $\mu^{\otimes n}$*-probability greater than* $7/12$,

$$R(\widehat{f}) \leq \min_{f\in F_M} R(f) + \frac{c_2}{\sqrt{n\log(\mathrm{e}M/\sqrt{n})}}.$$

The proof will require accurate information on a monotone rearrangement of almost Gaussian random variables.

Lemma 3.3. *There exists an absolute constant* C *for which the following holds. Let* g *be a standard Gaussian random variable, set* $H(x)=\mathbb{P}(|g|>x)$ *and put* $W(p)=H^{-1}(p)$ *(the inverse function of* H*). Then for every* $0<p<1$,

$$\bigl|W^2(p) - \log\bigl(2/\bigl(\pi p^2\bigr)\bigr) + \log\log\bigl(2/\bigl(\pi p^2\bigr)\bigr)\bigr| \leq C\frac{\log\log(2/(\pi p^2))}{\log(2/(\pi p^2))}.$$

Moreover, for every $0<\varepsilon<1/2$ *and* $0<p<1/(1+\varepsilon)$,

$$\bigl|W^2(p) - W^2\bigl((1+\varepsilon)p\bigr)\bigr| \leq C\varepsilon, \qquad \bigl|W^2(p) - W^2\bigl((1-\varepsilon)p\bigr)\bigr| \leq C\varepsilon.$$

Proof. The proof of the first part follows from the observation that for every $x>0$,

$$\frac{\sqrt{2}}{x\sqrt{\pi}}\exp\bigl(-x^2/2\bigr)\biggl(1-\frac{1}{x^2}\biggr) \leq \mathbb{P}\bigl(|g|>x\bigr) \leq \frac{\sqrt{2}}{x\sqrt{\pi}}\exp\bigl(-x^2/2\bigr), \tag{3.3}$$

where c is a suitable absolute constant (see, e.g. [22]), combined with a straightforward (yet tedious) computation. The second part of the claim follows from the first one, and is omitted. □

The next step is a Gaussian approximation of a variable $Y=n^{-1/2}\sum_{i=1}^n X_i$, where $X_1,\dots,X_n$ are i.i.d. random variables, with mean zero, variance 1, under the additional assumption that X has well behaved tails.

Definition 3.4 ([18,26]). *Let* $1\leq\alpha\leq 2$. *We say that a random variable* X *belongs to* L_{ψ_α} *if there exists a constant* C *such that*

$$\mathbb{E}\exp\bigl(|X|^\alpha/C^\alpha\bigr) \leq 2. \tag{3.4}$$

The infimum over all constants C *for which* (3.4) *holds defines a norm called the* ψ_α *norm of* X, *and we denote it by* $\|X\|_{\psi_\alpha}$.

Proposition 3.5 ([22], p. 183). *For every* L *there exist constants* c_1 *and* c_2 *that depend only on* L *and for which the following holds. Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of i.i.d., mean zero random variables with variance* 1, *and* $\|X\|_{\psi_1}\leq L$. *If* $Y=\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ *then for any* $0<x\leq c_1 n^{1/6}$,

$$\mathbb{P}[Y\geq x] = \mathbb{P}[g\geq x]\exp\biggl(\frac{\mathbb{E}X_1^3x^3}{6\sqrt{n}}\biggr)\biggl[1+c_2\frac{x+1}{\sqrt{n}}\biggr]$$

and

$$\mathbb{P}[Y\leq -x] = \mathbb{P}[g\leq -x]\exp\biggl(\frac{-\mathbb{E}X_1^3x^3}{6\sqrt{n}}\biggr)\biggl[1+c_2\frac{x+1}{\sqrt{n}}\biggr].$$

In particular, if $0 < x \le c_1 n^{1/6}$ *and* $\mathbb{E}X_1^3 = 0$ *then*

$$\big|\mathbb{P}\big[|Y| \ge x\big] - \mathbb{P}\big[|g| \ge x\big]\big| = c_2 \mathbb{P}\big[|g| \ge x\big]\frac{x+1}{\sqrt{n}}.$$

Since Proposition 3.5 implies a better Gaussian approximation than the standard Berry–Esséen bounds, one may consider the following family of random variables that will be used in the construction.

Definition 3.6. *We say that a random variable* Y *is* (L,n)*-almost Gaussian for* $L > 0$ *and* $n \in \mathbb{N}$*, if* $Y = n^{-1/2}\sum_{i=1}^n X_i$*, where* $X_1, \dots, X_n$ *are independent copies of* X*, which is a non-atomic random variable with mean* 0*, variance* 1*, and satisfies that* $\mathbb{E}X^3 = 0$ *and* $\|X\|_{\psi_1} \le L$*.*

Let $X_1, \dots, X_n$ and Y be such that $Y = n^{-1/2}\sum_{i=1}^n X_i$ is (L,n)-almost Gaussian. For $0 < p < 1$ set

$$U(p) = \big\{x > 0\colon \mathbb{P}\big(|Y| > x\big) = p\big\}.$$

Since X is non-atomic then $U(p)$ is non-empty and let

$$u^+(p) = \sup U(p) \quad \text{and} \quad u^-(p) = \inf U(p).$$

We shall apply Lemma 3.3 and Proposition 3.5 in the following case to bound $u^+(i/M)$ and $u^-(i/M)$ for every i, as long as M is not too large (i.e. $\log M \le c_1 n^{1/3}$). To that end, set $\varepsilon_{M,n} = [(\log M)/n]^{1/2}$, and for fixed values of M and n, and $1 \le i \le M$ let

$$u_i^+ = u^+(i/M) \quad \text{and} \quad u_i^- = u^-(i/M).$$

Corollary 3.7. *For every* $L > 0$ *there exist a constant* C_0 *that depends on* L *and an absolute constant* C_1 *for which the following holds. Assume that* Y *is* (L,n)*-almost Gaussian and that* $\log M \le C_0 n^{1/3}$*. Then, for every* $1 \le i \le M/2$*,*

$$(u_i^+)^2 \le \log\left(\frac{2M^2}{\pi i^2}\right) - \log\left(\log\left(\frac{2M^2}{\pi i^2}\right)\right) + C_1 \max\left\{\frac{\log(\log(2M^2/(\pi i^2)))}{\log(2M^2/(\pi i^2))}, \varepsilon_{M,n}\right\}$$

and

$$(u_i^-)^2 \ge \log\left(\frac{2M^2}{\pi i^2}\right) - \log\left(\log\left(\frac{2M^2}{\pi i^2}\right)\right) - C_1 \max\left\{\frac{\log(\log(2M^2/(\pi i^2)))}{\log(2M^2/(\pi i^2))}, \varepsilon_{M,n}\right\}.$$

Proof. Since $\sqrt{\log M} \le C_0 n^{1/6}$, one may use the Gaussian approximation from Proposition 3.5 to obtain

$$\begin{aligned}\mathbb{P}\big[|Y| \ge \sqrt{4\log M}\big] &\le \mathbb{P}\big[|g| \ge \sqrt{4\log M}\big]\left(1 + c_1\left(\frac{\sqrt{4\log M}+1}{\sqrt{n}}\right)\right)\\ &\le \sqrt{\frac{2}{4\pi\log M}}\exp(-2\log M)\left(1 + c_1\left(\frac{\sqrt{4\log M}+1}{\sqrt{n}}\right)\right) \le \frac{1}{M^2}.\end{aligned}$$

Thus, for every $1 \le i \le M$, if $x \in U(i/M)$ then $x \le \sqrt{4\log M}$.

Let $1 \le i \le M/2$ and $x \in U(i/M)$. Since $x \le 2C_0 n^{1/6}$ (because $x \le \sqrt{4\log M} \le 2C_0 n^{1/6}$), it follows from Proposition 3.5 that

$$\big|i/M - H(x)\big| \le c_3 H(x)\frac{x+1}{\sqrt{n}} \le c_4 H(x)\varepsilon_{M,n}, \tag{3.5}$$

where $H(x) = \mathbb{P}[|g| \ge x]$. Observe that if $W(p) = H^{-1}(p)$, then

$$\big|W^2(i/M) - x^2\big| \le c_5 \varepsilon_{M,n}.$$

Indeed, since $H(x)(1-c_4\varepsilon_{M,n}) \le i/M \le H(x)(1+c_4\varepsilon_{M,n})$, then by the monotonicity of W and the second part of Lemma 3.3, setting $p = H(x)$,

$$W^2(i/M) \le W^2\big((1+c_4\varepsilon_{M,n})H(x)\big) \le W^2\big(H(x)\big) + c_6\varepsilon_{M,n} = x^2 + c_6\varepsilon_{M,n}.$$

One obtains the lower bound in a similar way. The claim follows by using the approximate value of $W^2(i/M)$ provided in the first part of Lemma 3.3. □

The parameters u_i^+ and u_i^- can be used to estimate the distribution of a non-increasing rearrangement $(Y_i^*)_{i=1}^M$ of the absolute values of M independent copies of Y.

Lemma 3.8. *There exists constants $c > 0$ and $j_0 \in \mathbb{N}$ for which the following holds. Let $Y_1, \dots, Y_M$ be i.i.d. non-atomic random variables. For every $1 \le s \le M$, with probability at least $1 - 2\exp(-cs)$,*

$$\big|\big\{i\colon |Y_i| \ge u_s^-\big\}\big| \ge s/2 \quad \text{and} \quad \big|\big\{i\colon |Y_i| \ge u_s^+\big\}\big| \le 3s/2.$$

In particular, with probability at least $11/12$, for every $j_0 \le j \le M/2$,

$$u_{2j}^- \le Y_j^* \le u_{\lceil 2(j-1)/3\rceil}^+,$$

where $\lceil x\rceil = \min\{n \in \mathbb{N}\colon x \le n\}$.

Proof. Fix $0 < p < 1$ to be named later and let $(\delta_i)_{i=1}^M$ be independent $\{0,1\}$-valued random variables with $\mathbb{E}\delta_i = p$. A straightforward application of Bernstein's inequality [26] shows that

$$\mathbb{P}\left(\left|\frac{1}{M}\sum_{i=1}^M \delta_i - p\right| \ge t\right) \le 2\exp\big(-cM\min\big\{t^2/p, t\big\}\big).$$

In particular, with probability at least $1 - 2\exp(-c_1 Mp)$,

$$(1/2)Mp \le \sum_{i=1}^M \delta_i \le (3/2)Mp.$$

We will apply this observation to the independent random variables $\delta_i = \mathbb{1}_{\{|Y_i|>a\}}$, $1 \le i \le M$, for an appropriate choice of a. Indeed, if we take a for which $\mathbb{P}(|Y_1| > a) = s/M$ (such an a exists because Y_1 is non-atomic), then with probability at least $1 - 2\exp(-c_1 s)$, at least $s/2$ of the $|Y_i|$ will be larger than a, and at most $3s/2$ will be larger than a. Since this result holds for any $a \in U(s/M)$ the first part of the claim follows.

Now take s_0 to be the smallest integer such that $1 - 2\sum_{s=s_0}^M \exp(-cs) \ge 11/12$ (in particular $c^{-1}\log 24 \le s_0 \le c^{-1}(\log 48 + 1)$). Applying the union bound and a change of variables, it is evident that with probability at least $5/6$, for every $\lfloor (3s_0)/2\rfloor + 1 \le j \le M/2$,

$$\big|\big\{i\colon |Y_i| \ge u_{2j}^-\big\}\big| \ge j \quad \text{and} \quad \big|\big\{i\colon |Y_i| \ge u_{\lceil (2(j-1))/3\rceil}^+\big\}\big| \le j-1,$$

and thus $u_{2j}^- \le Y_j^* \le u_{\lceil (2(j-1))/3\rceil}^+$. □

With Lemma 3.8 and Corollary 3.7 in hand, one can bound the following functional of the random variables $(Y_i^*)_{i=1}^M$.

Lemma 3.9. *For every $L > 0$ there exist constants $c_1, \dots, c_4$, j_0 and $\alpha < 1$ that depend only on L for which the following holds. Let Y be (L,n)-almost Gaussian and let $Y_1, \dots, Y_M$ be independent copies of Y. Then, with probability at least $11/12$, for every $j_0 \le \ell \le k \le \alpha M$,*

$$c_1\left(\frac{\log(\mathrm{e}k/\ell) - \varepsilon_{M,n}}{\sqrt{\log(\mathrm{e}M/\ell)}}\right) \le Y_\ell^* - Y_k^* \le c_2\left(\frac{\log(\mathrm{e}k/\ell) + \varepsilon_{M,n}}{\sqrt{\log(\mathrm{e}M/\ell)}}\right).$$

Moreover, with probability at least $10/12$, *for every* $j_0 \le \ell \le k \le \alpha M$

$$Y^*_\ell - Y^*_k - \left(\frac{1}{k}\sum_{i=1}^k \big(Y^*_i - Y^*_k\big)^2\right)^{1/2} \ge c_3 \frac{\log(\mathrm{e}k/\ell)}{\sqrt{\log(\mathrm{e}M/\ell)}} - \frac{c_4}{\sqrt{\log(\mathrm{e}M/k)}},$$

and if $j_0 \le k \le \alpha M$, *then* $u^-_{2k} \le Y^*_k \le u^+_{\lceil 2(k-1)/3\rceil}$ *and*

$$\left(\frac{1}{k}\sum_{i=1}^k \big(Y^*_i - Y^*_k\big)^2\right)^{1/2} \le \frac{c_4}{\sqrt{\log(\mathrm{e}M/k)}},$$

provided that $\log^2 M \lesssim_L k$ *and that* $\varepsilon_{M,n} = \sqrt{(\log M)/n} \le 1$.

Proof. The first part of the claim follows from Lemma 3.8 and Corollary 3.7, combined with a straightforward computation. For the second part, observe that, for some well chosen constant $c_1(L)$ depending only on L, with probability at least $11/12$, $Y^*_1 \le c_1(L)\sqrt{\log M}$. Hence, applying the first part of the claim, with probability at least $10/12$,

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^k \big(Y^*_i - Y^*_k\big)^2 &\le c_1(L)\frac{j_0 \log M}{k} + \frac{1}{k}\sum_{i=j_0}^k \big(Y^*_i - Y^*_k\big)^2 \\ &\le c_1(L)\frac{j_0 \log M}{k} + \frac{c_2}{k}\sum_{i=j_0}^k \left(\frac{\log^2(\mathrm{e}k/i)}{\log(\mathrm{e}M/i)} + \frac{\varepsilon^2_{M,n}}{\log(\mathrm{e}M/i)}\right) \\ &\le c_1(L)\frac{j_0 \log M}{k} + c_3\frac{1+\varepsilon^2_{M,n}}{\log(\mathrm{e}M/k)} \le \frac{c_4}{\log(\mathrm{e}M/k)},\end{aligned}$$

provided that $\log^2 M \lesssim_L k$ and that $\varepsilon_{M,n} \le 1$. Note that to estimate the sum we have used that

$$\frac{1}{k}\sum_{i=j_0}^k \frac{\log^2(\mathrm{e}k/i)}{\log(\mathrm{e}M/i)} \le \frac{1}{\log(\mathrm{e}M/k)}\frac{1}{k}\sum_{i=j_0}^k \log^2(\mathrm{e}k/i) \le \frac{c_3}{\log(\mathrm{e}M/k)}.$$

Now the second and the third parts follow from the first one. □

The next preliminary step we need is a simple bound on the dual norm to the one whose unit ball is $A_r = B_1^M \cap \sqrt{r}B_2^M$. Recall that for a convex body $C \subset \mathbb{R}^M$, the polar body of C is $C^\circ = \{x \in \mathbb{R}^M\colon \sup_{y\in C}\langle x, y\rangle \le 1\}$, and in our case, $A_r^\circ = \mathrm{conv}(B_\infty^M \cup r^{-1/2}B_2^M)$ (see, for example, [23]). From here on, given $v \in \mathbb{R}^M$, set

$$\|v\|_{A_r^\circ} = \sup_{w\in A_r}\langle v, w\rangle,$$

and, as always, $(v_i^*)_{i=1}^M$ is the monotone rearrangement of $(|v_i|)_{i=1}^M$.

Lemma 3.10. *For every* $v \in \mathbb{R}^M$ *and any* $0 < \rho < r \le 1$ *such that* $1/r$ *and* $1/\rho$ *are integers*,

$$\|v\|_{A_r^\circ} - \|v\|_{A_\rho^\circ} \ge v^*_{1/r} - v^*_{1/\rho} - \left(\rho\sum_{i=1}^{1/\rho}\big(v^*_i - v^*_{1/\rho}\big)^2\right)^{1/2}$$

and in general for any $0 < r \le 1$,

$$v^*_{\lfloor 1/r\rfloor} \le \|v\|_{A_r^\circ} \le v^*_{\lceil 1/r\rceil} + \sqrt{\lceil 1/r\rceil}\left(\sum_{i=1}^{\lceil 1/r\rceil}\big(v^*_i - v^*_{\lceil 1/r\rceil}\big)^2\right)^{1/2}.$$

Proof. First, observe that for every $v \in \mathbb{R}^M$,

$$\|v\|_{A_r^\circ} = \min_{1 \le j \le M} \left(\sqrt{r} \left(\sum_{i=1}^{j} \left(v_i^* - v_j^*\right)^2 \right)^{1/2} + v_j^* \right). \tag{3.6}$$

Indeed, since $A_r^\circ = \operatorname{conv}(B_\infty^M \cup (1/\sqrt{r}) B_2^M)$, it is evident that $\|v\|_{A_r^\circ} = \inf\{\|u\|_\infty + \sqrt{r}\|w\|_2, v = u + w\}$. One may verify that if $v = u + w$ is an optimal decomposition then $\operatorname{supp}(w) \subset \{i\colon |u_i| = \|u\|_\infty\}$. Hence, if $\|u\|_\infty = K$ then for every $1 \le i \le M$, $u_i = K \operatorname{sgn}(v_i) \mathbb{1}_{\{|v_i| \ge K\}} + v_i \mathbb{1}_{\{|v_i| < K\}}$, and thus, $w_i = \mathbb{1}_{\{|v_i| \ge K\}}(v_i - \operatorname{sgn}(v_i)K)$. Therefore,

$$\|v\|_{A_r^\circ} = \inf_{K>0} \left\{ K + \sqrt{r} \left(\sum_{\{i:|v_i| \ge K\}} \left(|v_i| - K\right)^2 \right)^{1/2} \right\}.$$

Moreover, since it is enough to consider only values of K in $\{v_j^*\colon 1 \le j \le M\}$, (3.6) is verified. In particular, if $1/r$ is an integer then

$$\|v\|_{A_r^\circ} \le \sqrt{r} \left(\sum_{i=1}^{1/r} \left(v_i^* - v_{1/r}^*\right)^2 \right)^{1/2} + v_{1/r}^*.$$

On the other hand, if $T_r = \{u \in \mathbb{R}^M\colon \|u\|_2 \le \sqrt{r}, |\operatorname{supp}(u)| \le 1/r\}$ then $T_r \subset B_1^M \cap \sqrt{r} B_2^M$. Hence,

$$\|v\|_{A_r^\circ} \ge \sup_{w \in T_r} \langle v, w \rangle = \sqrt{r} \left(\sum_{i=1}^{1/r} \left(v_i^*\right)^2 \right)^{1/2}.$$

Therefore, if $1/r$ and $1/\rho$ are integers, it follows that

$$\begin{aligned} \|v\|_{A_r^\circ} - \|v\|_{A_\rho^\circ} &\ge \sqrt{r} \left(\sum_{i=1}^{1/r} \left(v_i^*\right)^2 \right)^{1/2} - \left(\sqrt{\rho} \left(\sum_{i=1}^{1/\rho} \left(v_i^* - v_{1/\rho}^*\right)^2 \right)^{1/2} + v_{1/\rho}^* \right) \\ &\ge v_{1/r}^* - v_{1/\rho}^* - \sqrt{\rho} \left(\sum_{i=1}^{1/\rho} \left(v_i^* - v_{1/\rho}^*\right)^2 \right)^{1/2}, \end{aligned}$$

because $(r \sum_{i=1}^{1/r} (v_i^*)^2)^{1/2} \ge v_{1/r}^*$.

The second part follows in a similar fashion and it omitted. □

Proof of the lower bound of Theorem 3.2. Let $\phi_1, \dots, \phi_M$, X and $a > 0$ be such that $\phi_1(X), \dots, \phi_M(X)$ are uniformly distributed on $[-a, a]$ and have variance 1 (in particular $a = \sqrt{3}$). Set $T(X) = \phi_{M+1}(X) = Y$ to be a Rademacher variable. Assume further that $(\phi_i)_{i=1}^{M+1}$ are independent in $L_2(P^X)$ and let $F_M = \{0, \pm\phi_1, \dots, \pm\phi_M\}$. Note that the functions in $\operatorname{conv}(F_M)$ are given by $f_\lambda = \langle \Phi, \lambda \rangle$ where $\Phi = (\phi_1, \dots, \phi_M)$ and $\lambda \in B_1^M$.

It is straightforward to verify that the excess loss function of f_λ relative to $\operatorname{conv}(F_M)$ is

$$\mathcal{L}_{f_\lambda} = (f_\lambda - \phi_{M+1})^2 - (0 - \phi_{M+1})^2 = \langle \Phi, \lambda \rangle^2 - 2\langle \Phi, \lambda \rangle \phi_{M+1}$$

(since $f^* = 0$), implying that $\mathbb{E}\mathcal{L}_{f_\lambda} = \|\lambda\|_2^2$.

Let us consider the problem of empirical minimization in $\operatorname{conv}(F_M) = \{\langle \lambda, \Phi \rangle\colon \lambda \in B_1^M\}$. Recall that $A_r = B_1^M \cap \sqrt{r} B_2^M$ and, for an independent sample $(\Phi(X_i), \phi_{M+1}(X_i))_{i=1}^n$, define the functional

$$\psi(r, \rho) = n \Big(\inf_{\lambda \in A_r} R_n(f_\lambda) - \inf_{\mu \in A_\rho} R_n(f_\mu) \Big).$$

If we show that for some $r \ge \rho$, $\psi(r, \rho) < 0$, then for that sample, $\mathbb{E}\mathcal{L}_{\widehat{f}} \ge \rho$.

Note that, for any $r, \rho > 0$,

$$\psi(r,\rho) \le \sup_{\lambda\in A_r}\sum_{i=1}^n \big\langle \Phi(X_i),\lambda\big\rangle^2 - 2\sup_{\lambda\in A_r}\sum_{i=1}^n \big\langle \Phi(X_i),\lambda\big\rangle \phi_{M+1}(X_i)$$
$$+ 2\sup_{\mu\in A_\rho}\sum_{i=1}^n \big\langle \Phi(X_i),\lambda\big\rangle \phi_{M+1}(X_i),$$

and let us estimate the supremum of the process

$$\lambda\in A_r \to \sum_{i=1}^n \big\langle \Phi(X_i),\lambda\big\rangle^2 = n\big((P_n-P)\big(\langle\Phi,\lambda\rangle^2\big) + \|\lambda\|_2^2\big).$$

Observe that $\Phi(X)$ is isotropic (that is, for every $\lambda\in\mathbb{R}^M$, $\mathbb{E}\langle\lambda,\Phi(X)\rangle^2 = \|\lambda\|_2^2$), and sub-Gaussian – since $\|\langle\lambda,\Phi(X)\rangle\|_{\psi_2}\le 4a\|\lambda\|_2$. Hence, applying the results from [21], it is evident that with probability at least $11/12$,

$$\sup_{\lambda\in A_r}\big|(P_n-P)\big(\langle\lambda,\Phi\rangle^2\big)\big| \le c(a)\max\left\{\operatorname{diam}\big(A_r,\|\cdot\|_2\big)\frac{\gamma_2(A_r,\|\cdot\|_2)}{\sqrt{n}}, \frac{\gamma_2^2(A_r,\|\cdot\|_2)}{n}\right\}. \tag{3.7}$$

Recall that for $r\ge 1/M$, $\gamma_2(A_r,\|\cdot\|_2)\sim\sqrt{\log(\mathrm{e}Mr)}$ (see, for instance, [21]), and thus, if $r\ge\max(1/M,1/n)$, then with probability at least $11/12$,

$$n\sup_{\lambda\in A_r}\big((P_n-P)\big(\langle\Phi,\lambda\rangle^2\big)+\|\lambda\|_2^2\big) \le nr + c_1\sqrt{nr\log(\mathrm{e}Mr)},$$

where c_1 is a constant that depends only on a.

Next, to estimate the first two terms, let

$$Y_j = n^{-1/2}\sum_{i=1}^n \phi_{M+1}(X_i)\big\langle\Phi(X_i),e_j\big\rangle$$

and observe that $(Y_j)_{j=1}^M$ are independent copies of a $(2,n)$-almost Gaussian variable. If we set $V=(Y_i)_{i=1}^M$ then

$$\sup_{\lambda\in A_r}\sum_{i=1}^n\big\langle\lambda,\Phi(X_i)\big\rangle\phi_{M+1}(X_i) - \sup_{\theta\in A_\rho}\sum_{i=1}^n\big\langle\theta,\Phi(X_i)\big\rangle\phi_{M+1}(X_i)$$
$$= \sqrt{n}\Big(\sup_{\lambda\in A_r}\langle\lambda,V\rangle - \sup_{\theta\in A_\rho}\langle\theta,V\rangle\Big) = \sqrt{n}\big(\|V\|_{A_r^\circ}-\|V\|_{A_\rho^\circ}\big) = (*).$$

By Lemma 3.10, if $1/r=\ell$ and $1/\rho=k$ are integers, then

$$(*)\ge\sqrt{n}\left[Y_\ell^* - Y_k^* - \left(\frac{1}{k}\sum_{i=1}^k\big(Y_i^*-Y_k^*\big)^2\right)^{1/2}\right]$$

and thus, if ℓ, k, M and n are as in Lemma 3.9, then with probability at least $9/12$,

$$(*)\ge\sqrt{n}\left(\frac{c_2\log(\mathrm{e}k/\ell)}{\sqrt{\log(\mathrm{e}M/\ell)}} - \frac{c_3}{\sqrt{\log(\mathrm{e}M/k)}}\right)$$
$$\ge c_4\sqrt{n}\frac{\log(\mathrm{e}k/\ell)}{\sqrt{\log(\mathrm{e}M/\ell)}},$$

provided that $k\ge c_5\ell$ for c_5 large enough.

Hence, with probability at least $9/12$,

$$\psi(r,\rho) \le -2c_4\sqrt{n}\frac{\log(\mathrm{e}k/\ell)}{\sqrt{\log(\mathrm{e}M/\ell)}} + nr + c_1\sqrt{rn\log(\mathrm{e}Mr)}.$$

It follows that if we select $r \sim 1/\sqrt{n\log(\mathrm{e}M/\sqrt{n})}$ and $\rho \sim r$ with $\rho < r$ so that the conditions of Lemma 3.9 are satisfied, then with probability at least $9/12$, $\psi(r,\rho) < 0$. Hence, with the same probability,

$$R(\widehat{f}) - \min_{f\in F_M} R(f) = \mathbb{E}\mathcal{L}_{\widehat{f}} \ge \frac{c_6}{\sqrt{n\log(\mathrm{e}M/\sqrt{n})}}. \qquad \square$$

Proof of the upper bound in Theorem 3.2. We will show that with constant probability,

$$\inf_{0\le r\le r_0}\inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}} R_n(f_\lambda) < \inf_{r_0\le r\le 1}\inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}} R_n(f_\lambda) \tag{3.8}$$

for $r_0 \sim 1/\sqrt{n\log(\mathrm{e}M/\sqrt{n})}$, and thus, on that event, $R(\widehat{f}) \le r_0$. To that end, one has to show that

$$\inf_{0\le r\le r_0}\inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}} P_n\mathcal{L}_{f_\lambda} < \inf_{r_0\le r\le 1}\inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}} P_n\mathcal{L}_{f_\lambda}.$$

Let

$$Q(r) = \sup_{\lambda\in B_1^M\cap\sqrt{r}B_2^M} \big|(P_n - P)\big(\langle\Phi,\lambda\rangle^2\big)\big|$$

and set $r^* = \inf\{r > 0\colon \mathbb{E}Q(r) \le r/2\}$. Applying (3.7) and since $\gamma_2(A_r, \|\cdot\|_2) \sim \sqrt{\log(\mathrm{e}Mr)}$, then $r^* \le c_0\sqrt{\log(\mathrm{e}M/\sqrt{n})/n}$. Hence, by a standard fixed point argument (see for instance, [4]), it follows that with probability greater than $11/12$, if $\lambda \in B_1^M$ and $\|\lambda\|_2^2 \ge r^*$, then

$$\frac{\|\lambda\|_2^2}{2} \le \frac{1}{n}\sum_{i=1}^n \big\langle\Phi(X_i),\lambda\big\rangle^2 \le \frac{3\|\lambda\|_2^2}{2}.$$

In particular, by Lemmas 3.9, 3.10 and Corollary 3.7, with probability larger than $9/12$, for every $r \ge r^*$,

$$\begin{aligned}
\inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}} P_n\mathcal{L}_{f_\lambda} &= \inf_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}}\left(\frac{1}{n}\sum_{i=1}^n\big\langle\Phi(X_i),\lambda\big\rangle^2 - \frac{2}{n}\sum_{i=1}^n\big\langle\Phi(X_i),\lambda\big\rangle\phi_{M+1}(X_i)\right)\\
&\ge \frac{r}{2} - \frac{2}{\sqrt{n}}\sup_{\lambda\in B_1^M\cap\sqrt{r}\mathcal{S}^{M-1}}\langle\lambda,V\rangle \ge \frac{r}{2} - \frac{2}{\sqrt{n}}\sup_{\lambda\in B_1^M\cap\sqrt{r}B_2^M}\langle\lambda,V\rangle = \frac{r}{2} - \frac{2\|V\|_{A_r^\circ}}{\sqrt{n}}\\
&\ge \frac{r}{2} - \frac{2}{\sqrt{n}}\left(Y^*_{\lceil 1/r\rceil} + \sqrt{\lceil 1/r\rceil}\left(\sum_{i=1}^{\lceil 1/r\rceil}\big(Y_i^* - Y^*_{\lceil 1/r\rceil}\big)^2\right)^{1/2}\right)\\
&\ge \frac{r}{2} - \frac{2}{\sqrt{n}}\left(u^+_{\lceil 2(\lceil 1/r\rceil-1)/3\rceil} + \frac{c_1}{\sqrt{\log(c_2Mr)}}\right)\\
&\ge \frac{r}{2} - \frac{2}{\sqrt{n}}\left(c_3\sqrt{\log(c_4Mr)} + \frac{c_1}{\sqrt{\log(c_5Mr)}}\right) > 0
\end{aligned} \tag{3.9}$$

provided that $r \ge c_6\sqrt{\log(\mathrm{e}M/\sqrt{n})/n}$ for some constant c_6 large enough. Therefore, on that event, if $\|\lambda\|_2^2 \ge c_7\sqrt{\log(\mathrm{e}M/\sqrt{n})/n}$ then $P_n\mathcal{L}_{f_\lambda} > 0$. On the other hand, $P_n\mathcal{L}_{f_0} = 0$, and thus $\|\widehat{\lambda}\|_2^2 \le c_7\sqrt{\log(\mathrm{e}M/\sqrt{n})/n}$ (where $\widehat{f} = f_{\widehat{\lambda}}$) with probability at most $9/12$.

It remains to show that with sufficiently high constant probability

$$\inf_{0\le r\le r_0}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda} < \inf_{r_0\le r\le r_1}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda}$$

for $r_0\sim 1/\sqrt{n\log(eM/\sqrt{n})}$ and $r_1=c_7\sqrt{\log(eM/\sqrt{n})/n}$.

Using the same argument as in (3.9) and applying Lemmas 3.9, 3.10 and Corollary 3.7, it is evident that with probability at least 10/12,

$$\begin{aligned}&\inf_{r_0\le r\le r_1}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda}\\ &\quad\ge \inf_{r_0\le r\le r_1}\left(r-\frac{2}{\sqrt{n}}\left(\log(C_0Mr)-c_8\log\log(C_0Mr_0)\right)^{1/2}\right)-\frac{c_1}{\sqrt{n\log(c_5Mr_0)}}-Q(r_1)\end{aligned}\tag{3.10}$$

and for some $r_2\le r_0$ to be named later,

$$\begin{aligned}\inf_{0\le r\le r_0}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda} &\le \inf_{0\le r\le r_2}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda}\\ &\le \inf_{0\le r\le r_2}\left(r-\frac{2}{\sqrt{n}}\big(\log(C_1Mr)-c_9\log\log(C_1Mr_2)\big)^{1/2}\right)+Q(r_2).\end{aligned}\tag{3.11}$$

Moreover, thanks to (3.7), with probability greater than 10/12,

$$Q(r_1)+Q(r_2)\le c_{10}\sqrt{r_1\log(eMr_1)/n}.$$

Fix $0<\beta_2<\beta_0$ to be named later and set

$$r_0=\frac{\beta_0}{\sqrt{n\log(eM/\sqrt{n})}}\quad\text{and}\quad r_2=\frac{\beta_2}{\sqrt{n\log(eM/\sqrt{n})}}.$$

For β_0 large enough (resp. β_2 small enough), the infimum in (3.10) (resp. (3.11)) is achieved in r_0 (resp. r_2). Therefore, with probability greater than 8/12

$$\begin{aligned}&\inf_{0\le r\le r_0}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda}-\inf_{r_0\le r\le r_1}\inf_{\lambda\in B_1^M\cap\sqrt{r}S^{M-1}} P_n\mathcal{L}_{f_\lambda}\\ &\le \frac{c_{11}\log((C_0r_0)/(C_1r_2))}{\sqrt{\log(C_1Mr_2)}}+r_2-r_0+\frac{c_1}{\sqrt{n\log(c_5Mr_0)}}+c_{12}\sqrt{r_1\log\left(\frac{eMr_1}{\sqrt{n}}\right)}\\ &\le \frac{c_{13}\log((C_0\beta_0)/(C_1\beta_2))}{\sqrt{n\log(eM/\sqrt{n})}}+\frac{\beta_2-\beta_0}{\sqrt{n\log(eM/\sqrt{n})}}+\frac{c_{14}}{\sqrt{n\log(eM/\sqrt{n})}}+\frac{c_{14}\log(eM/\sqrt{n})}{n}.\end{aligned}$$

Therefore, there exists some β_0 for which the latter quantity is negative and thus (3.8) holds for $r_0=\beta_0/\sqrt{n\log(C_0\beta_0M/\sqrt{n})}$. □

4. Proof of Theorem B

Our starting point is to describe the machinery developed in [4], leading to the desired estimates on the performance of ERM in a general class of functions. Let G be a class of functions and denote by $\mathcal{L}_G=\{(x,y)\longmapsto(y-g(x))^2-(y-g_G^*(x))^2\colon\ g\in G\}$ the associated class of quadratic excess loss functions, where g_G^* is the minimizer of the quadratic risk in G. Let $V=\operatorname{star}(\mathcal{L}_G,0)=\{\theta\mathcal{L}\colon\ 0\le\theta\le1,\mathcal{L}\in\mathcal{L}_G\}$ and for every $\lambda>0$ set $V_\lambda=\{h\in V\colon\ \mathbb{E}h\le\lambda\}$.

Theorem 4.1 ([4]). *For every positive B and b there exists a constant $c_0 = c_0(B,b)$ for which the following holds. Let G be a class of functions for which $\mathcal{L}_G$ consists of functions that are bounded by b almost surely. Assume further that for any $\mathcal{L} \in \mathcal{L}_G$, $\mathbb{E}\mathcal{L}^2 \le B\mathbb{E}\mathcal{L}$. If $x > 0$, $\lambda^* > 0$ satisfies that $\mathbb{E}\|P - P_n\|_{V_{\lambda^*}} \le \lambda^*/8$ and*

$$\lambda^*(x) = c_0 \max\left(\lambda^*, \frac{x}{n}\right),$$

then with probability greater than $1 - \exp(-x)$, the empirical risk minimization procedure $\widehat{g}$ in G satisfies

$$R(\widehat{g}) \le \inf_{g\in G} R(g) + \lambda^*(x).$$

Let F be the given dictionary and set $G = \operatorname{conv}(F)$. Using the notation of Theorem 4.1, put $\mathcal{L}_{\operatorname{conv}(F)} = \{\mathcal{L}_f\colon\ f \in \operatorname{conv}(F)\}$, consider the star-shaped hull $V = \operatorname{star}(\mathcal{L}_{\operatorname{conv}(F)}, 0)$ and its localizations $V_\lambda = \{g \in V\colon\ \mathbb{E}g \le \lambda\}$ for any $\lambda > 0$. Thanks to convexity, the following observation holds in our case (see [19] for the proof).

Proposition 4.2. *If $f \in \operatorname{conv}(F)$ then $\mathbb{E}\mathcal{L}_f \ge \|f - f^*\|^2_{L_2(P^X)}$ where f^* is the minimizer of the quadratic risk in $\operatorname{conv}(F)$. In particular,*

1. *$\mathbb{E}\mathcal{L}^2 \le 4b^2\mathbb{E}\mathcal{L}$ for any $\mathcal{L} \in \mathcal{L}_{\operatorname{conv}(F)}$;*
2. *for $\mu > 0$, if $f \in \operatorname{conv}(F)$ satisfies that $\mathbb{E}\mathcal{L}_f \le \mu$, then $f \in f^* + K_\mu$, where*

$$K_\mu = 2\big[\operatorname{conv}\{\pm f_1, \dots, \pm f_M\} \cap \sqrt{\mu}\mathcal{B}\big(L_2\big(P^X\big)\big)\big].$$

The first part of Proposition 4.2 shows that $\mathcal{L}_{\operatorname{conv}(F)}$ satisfies the assumptions of Theorem 4.1 with $B = 4b^2$. To apply Theorem 4.1 one has to find $\lambda^* > 0$ for which $\mathbb{E}\|P - P_n\|_{V_{\lambda^*}} \le \lambda^*/8$, and to that end we will use the second part of Proposition 4.2. First, observe that it was shown in [5] that

$$\mathbb{E}\|P - P_n\|_{V_\lambda} \le \sum_{i\ge 0} 2^{-i}\mathbb{E}\|P - P_n\|_{\mathcal{L}_{2^{i+1}\lambda}}, \tag{4.1}$$

where from here on we set $\mathcal{L}_\mu = \{\mathcal{L} \in \mathcal{L}_{\operatorname{conv}(F)}\colon\ \mathbb{E}\mathcal{L} \le \mu\}$. Applying the second part of Proposition 4.2 it is evident that $\{f \in \operatorname{conv}(F)\colon\ \mathbb{E}\mathcal{L}_f \le \mu\} \subset f^* + K_\mu$.

Proof of Theorem B. By the Giné–Zinn symmetrization Theorem [26],

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_\mu} \le 2\mathbb{E}\mathbb{E}_\varepsilon \sup_{\mathcal{L}\in\mathcal{L}_\mu}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{L}(X_i, Y_i)\right|. \tag{4.2}$$

Note that if $\mathcal{L} \in \mathcal{L}_\mu$ and $f \in \operatorname{conv}(F)$ satisfies that $\mathcal{L} = \mathcal{L}_f$, then for any (x, y),

$$\begin{aligned}\big|\mathcal{L}(x,y)\big| &= \big|\big(y - f(x)\big)^2 - \big(y - f^*(x)\big)^2\big| \\ &= \big|\big(f^*(x) - f(x)\big)\big(2y - f(x) - f^*(x)\big)\big| \le 4b\big|f(x) - f^*(x)\big|.\end{aligned}$$

Thus, by the contraction principle (see, e.g. [18]) and Proposition 4.2,

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_\mu} \le \frac{8b}{\sqrt{n}}\mathbb{E}\mathbb{E}_\varepsilon \sup_{f\in K_\mu}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right|.$$

Observe that since the dictionary consists of an orthogonal family, if $(e_1, \dots, e_M)$ is the standard basis in ℓ_2^M and $F(\cdot) = (f_1(\cdot), \dots, f_M(\cdot))$, then

$$K_\mu = \big\{2\langle\lambda, F\rangle\colon\ \lambda \in B_1^M \cap \sqrt{\mu}\mathcal{E}\big\},$$

where $\mathcal{E}$ is an ellipsoid with principal axes $(\|f_i\|_{L_2} e_i)_{i=1}^M$. From here on we will assume that $(\|f_i\|_{L_2})_{i=1}^M$ is a non-increasing sequence.

Now, we want to bound

$$\mathbb{E} \sup_{f\in K_\mu} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| = \mathbb{E} \sup_{\lambda \in B_1^M \cap \sqrt{\mu}\mathcal{E}} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i \langle \lambda, F(X_i) \rangle \right| = \frac{2}{\sqrt{n}} \mathbb{E} \left\| \sum_{i=1}^n \frac{1}{\sqrt{n}} \varepsilon_i F(X_i) \right\|_{(B_1^M \cap \sqrt{\mu}\mathcal{E})^\circ},$$

where $\|\cdot\|_{(B_1^M \cap \sqrt{\mu}\mathcal{E})^\circ}$ denotes the dual norm to the one whose unit ball is $B_1^M \cap \sqrt{\mu}\mathcal{E}$. We will use two different strategies to bound this process depending on $M \le \sqrt{n}$ or $M > \sqrt{n}$. First start with the case $M \ge \sqrt{n}$. Since both B_1^M and $\mathcal{E}$ are unconditional with respect to the coordinate structure given by $(e_i)_{i=1}^M$, it follows that

$$\|v\|_{(B_1^M \cap \sqrt{\mu}\mathcal{E})^\circ} \sim \inf_{I \subset \{1,\dots,M\}} \left[\sqrt{\mu} \left(\sum_{i\in I} \left(\frac{v_i}{\|f_i\|_{L_2}} \right)^2 \right)^{1/2} + \max_{i \in I^c} |v_i| \right], \tag{4.3}$$

and in our case, $v = (v_j)_{j=1}^M = ((1/\sqrt{n}) \cdot \sum_{i=1}^n \varepsilon_i f_j(X_i))_{j=1}^M$.

Let

$$J_0 = \{ j\colon \|f_j\|_{L_2} \ge c_0 b \sqrt{\log M}/\sqrt{n} \},$$

where c_0 is a constant to be named later. A straightforward application of Bernstein inequality [26] shows that, for $t \ge c_1$,

$$\mathbb{P}\big(\exists j \in J_0\colon\ P_n f_j^2 \ge (t+1)\|f_j\|_{L_2}^2\big) \le \sum_{j\in J_0} \exp\big(-c_2 n \big(\|f_j\|_{L_2}^2/b^2\big) \min\big(t^2, t\big)\big) \le M \exp(-c_3 t \log M) \le \exp(-c_4 t \log M)$$

and

$$\mathbb{P}\big(\exists j \in J_0^c\colon\ P_n f_j^2 \ge (t+1) b^2 n^{-1} \log M\big) \le \exp(-c_4 t \log M).$$

For every integer $\ell \ge c_1$, let

$$\mathcal{A}_\ell = \big\{ \forall j \in J_0\colon\ P_n f_j^2 \le (\ell+1)\|f_j\|_{L_2}^2 \big\} \cap \big\{ \forall j \in J_0^c\colon\ P_n f_j^2 \le (\ell+1) b^2 n^{-1} \log M \big\}.$$

Set $\mathcal{B}_\ell = \mathcal{A}_{\ell+1} \cap \mathcal{A}_\ell^c$ and note that $\mathbb{P}(\mathcal{B}_\ell) \le \mathbb{P}(\mathcal{A}_\ell^c) \le 2\exp(-c_4 \ell \log M)$ for any $\ell \ge c_1$.

For every $\ell \ge c_1$, consider the random variables conditioned on $\mathcal{B}_\ell$,

$$U_{j,\ell} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j(X_i)/\|f_j\|_{L_2} \Big| \mathcal{B}_\ell \quad \forall j \in J_0$$

and

$$U_{j,\ell} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j(X_i) \Big| \mathcal{B}_\ell \quad \forall j \in J_0^c.$$

Hence, by Hoeffding's inequality (cf. [26]), there exists an absolute constant c_5 such that, for any $j \in J_0$,

$$\|U_{j,\ell}\|_{\psi_2(\varepsilon)}^2 \le c_5 \frac{n^{-1} \sum_{i=1}^n f_j^2(X_i)}{\|f_j\|_{L_2}^2} \le c_5(\ell+1),$$

and for any $j \in J_0^c$,

$$\|U_{j,\ell}\|_{\psi_2(\varepsilon)}^2 \le c_5(\ell+1)b^2(\log M)/n.$$

By a result due to Klartag [13], it follows that for every such ℓ and any $1 \le j \le |J_0|$,

$$\mathbb{E}_\varepsilon\left(\sum_{i=1}^{j}\big(U_{i,\ell}^2\big)^*\right)^{1/2} \le c_6\sqrt{\ell}\sqrt{j\log\big(\mathrm{e}|J_0|/j\big)},$$

where $(U_{j,\ell}^*)_{j=1}^{|J_0|}$ is a decreasing rearrangement of $(|U_{j,\ell}|)_{j\in J_0}$. Moreover, by a standard maximal inequality (see, e.g. [26])

$$\mathbb{E}_\varepsilon \max_{j\in J_0^c} U_{j,\ell} \le c_7\sqrt{\log\big|J_0^c\big|}\max_{j\in J_0^c}\|U_{j,\ell}\|_{\psi_2(\varepsilon)} \le c_8\sqrt{\ell}b\frac{\log M}{\sqrt{n}}.$$

For every $1 \le j \le |J_0|$, let I be the set of the j largest coordinates of $(|U_{j,\ell}|)_{j\in J_0}$. Hence, by (4.3) and since $\|f_j\|_{L_2} \le b$,

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left(\left\|\sum_{i=1}^n \frac{1}{\sqrt{n}}\varepsilon_i F(X_i)\right\|_{(B_1^M\cap\sqrt{\mu}\mathcal{E})^\circ}\Big|(X_i)_{i=1}^n \in \mathcal{B}_\ell\right)\\
&\quad\lesssim \mathbb{E}_\varepsilon\sqrt{\mu}\left(\sum_{i=1}^{j}\big(U_{i,\ell}^2\big)^*\right)^{1/2} + \mathbb{E}_\varepsilon\max\Big(\|f_j\|_{L_2}U_{j,\ell}^*, \max_{j\in J_0^c}|U_{j,\ell}|\Big)\\
&\quad\lesssim \sqrt{\ell}\Big(\sqrt{\mu}\sqrt{j\log\big(\mathrm{e}|J_0|/j\big)} + b\sqrt{\log\big(\mathrm{e}|J_0|/j\big)}\Big) + \mathbb{E}_\varepsilon\max_{j\in J_0^c}|U_{j,\ell}|\\
&\quad\le \sqrt{\ell}\Big(\sqrt{\mu}\sqrt{j\log\big(\mathrm{e}|J_0|/j\big)} + b\sqrt{\log\big(\mathrm{e}|J_0|/j\big)}\Big) + \sqrt{\ell}b\frac{\log M}{\sqrt{n}}.
\end{aligned}$$

Therefore, if we take $j = \min\{[1/\mu], |J_0|\}$ it is evident that

$$\mathbb{E}_\varepsilon\left(\left\|\sum_{i=1}^n \frac{1}{\sqrt{n}}\varepsilon_i F(X_i)\right\|_{(B_1^M\cap\sqrt{\mu}\mathcal{E})^\circ}\Big|(X_i)_{i=1}^n \in \mathcal{B}_\ell\right) \lesssim b\sqrt{\ell}\left(\sqrt{\log(\mathrm{e}M\mu)} + \frac{\log M}{\sqrt{n}}\right).$$

Thus, integration with respect to $X_1,\dots,X_n$ and applying the estimates on the measure of $\mathcal{B}_\ell$,

$$\frac{2}{\sqrt{n}}\mathbb{E}\left\|\sum_{i=1}^n \frac{1}{\sqrt{n}}\varepsilon_i F(X_i)\right\|_{(B_1^M\cap\sqrt{\mu}\mathcal{E})^\circ} \lesssim b\sqrt{\frac{\log(\mathrm{e}M\mu)}{n}}.$$

Finally, by (4.1), for any $\lambda > 1/M$,

$$\begin{aligned}
\mathbb{E}\|P - P_n\|_{V_\lambda} &\le \sum_{i\ge 0} 2^{-i}\mathbb{E}\|P - P_n\|_{\mathcal{L}_{2^{i+1}\lambda}}\\
&\lesssim b^2\sum_{i\ge 0}2^{-i}\sqrt{\frac{\log(\mathrm{e}M2^{i+1}\lambda)}{n}} \lesssim b^2\sqrt{\frac{\log(\mathrm{e}M\lambda)}{n}},
\end{aligned}$$

and, if

$$\lambda^* \sim b^2\sqrt{\frac{1}{n}\log\left(\frac{\mathrm{e}Mb^2}{\sqrt{n}}\right)},$$

then $\mathbb{E}\|P - P_n\|_{V_{\lambda^*}} \le \lambda^*/8$, as required.

When $M \le \sqrt{n}$, we use the strategy developed in [15]. Let S be the linear subspace of $L^2(P^X)$ spanned by F and take $(e_1, \dots, e_{M'})$ to be an orthonormal basis of S (where $M' = \dim(S) \le M$). Since $K_\mu \subset S \cap 2\sqrt{\mu}\mathcal{B}(L_2(P^X))$, then

$$\mathbb{E}\sup_{f\in K_\mu}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \le 2\mathbb{E}\sup_{\|\lambda\|_{\ell_2^{M'}}\le 2\sqrt{\lambda}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(\sum_{j=1}^{M'}\lambda_j e_j(X_i)\right)\right|$$

$$\lesssim \sqrt{\mu}\mathbb{E}\left(\sum_{j=1}^{M'}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i e_j(X_i)\right)^2\right)^{1/2} \lesssim \sqrt{\frac{M'\mu}{n}}.$$

The rate obtained in the case $M \le \sqrt{n}$ follows now from (4.1). □

Appendix

We establish the following upper bound on the risk of $\tilde{f}^{\text{ERM-C}}$ as a (C)-aggregation procedure in the general case. Its proof follows the same path as in Section 4. But, rather than studying the empirical process indexed by the interpolation body $B_1^M \cap \sqrt{\mu}\mathcal{E}$, in the case $M \ge \sqrt{n}$, one simply uses the approximation $B_1^M \cap \sqrt{\mu}\mathcal{E} \subset B_1^M$ to get, conditionally on $X_1, \dots, X_n$,

$$\mathbb{E}_\varepsilon \sup_{f\in K_\mu}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \le \mathbb{E}_\varepsilon \sup_{\lambda\in B_1^M}\left|\left\langle \lambda, \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i F(X_i)\right\rangle\right| = \mathbb{E}_\varepsilon \max_{1\le j\le M}|\gamma_j|,$$

where, for all $j = 1, \dots, M$, γ_j is the sub-Gaussian random variable $n^{-1/2}\sum_{i=1}^n \varepsilon_i f_j(X_i)$ with ψ_2-norm bounded by $n^{-1}\sum_{i=1}^n f_j(X_i)^2 \le c_0 b^2$ and thus by a maximal inequality [18],

$$\mathbb{E}_g \max_{1\le j\le M}|\gamma_j| \le c_1 b\sqrt{\log M}.$$

The result below follows from this upper bound and (4.1) for the case $M > \sqrt{n}$, and the case $M \le \sqrt{n}$ follows the same path as the proof of Theorem B, and thus its proof is omitted.

Theorem A.1. *For every $b > 0$ there is a constant $c_1(b)$ and an absolute constant c_2 for which the following holds. Let n and M be integers which satisfy that $\log M \le c_1(b)\sqrt{n}$. For any couple (X, Y) and any finite dictionary F of cardinality M such that $|Y|, \sup_{f\in F}|f(X)| \le b$, and for any $u > 0$, with probability greater than $1 - \exp(-u)$,*

$$R\big(\tilde{f}^{\text{ERM-C}}\big) \le \min_{f\in\text{conv}(F)} R(f) + c_2 b^2 \max\left[\min\left(\frac{M}{n}, \sqrt{\frac{\log M}{n}}\right), \frac{u}{n}\right].$$

Acknowledgment

The authors would like to thank Vladimir Koltchinskii for a helpful discussion.

References

[1] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré Probab. Stat.* **40** (2004) 685–736. MR2096215

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

[3] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** (2009) 1591–1646. MR2533466

[4] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probab. Theory Related Fields* **135** (2006) 311–334. MR2240689

[5] P. L. Bartlett, S. Mendelson and J. Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields* **154** (2012) 193–224. MR2981422
[6] L. Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** (2006) 273–325. MR2219712
[7] O. Bousquet, V. Koltchinskii and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *Computational Learning Theory (Sydney, 2002)* 59–73. *Lecture Notes in Comput. Sci.* **2375**. Springer, Berlin, 2002. MR2040405
[8] F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory* **54** (2008) 1725–1735. MR2450298
[9] F. Bunea, A. B. Tsybakov and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.* **35** (2007) 1674–1697. MR2351101
[10] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory* 97–111. *Lecture Notes in Comput. Sci.* **4539**. Springer, Berlin, 2007. MR2397581
[11] M. Emery, A. Nemirovski and D. Voiculescu. *Lectures on Probability Theory and Statistics*. *Lecture Notes in Math.* **1738**. Springer, Berlin, 2000. MR1775638
[12] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.* **28** (2000) 681–712. MR1792783
[13] B. Klartag. 5*n* Minkowski symmetrizations suffice to arrive at an approximate Euclidean ball. *Ann. of Math. (2)* **156** (2002) 947–960. MR1954241
[14] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Été de Probabilités de Saint-Flour XXXVIII-2008*. *Lecture Notes in Math.* Springer, Berlin, 2011. MR2829871
[15] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** (2006) 2593–2656. MR2329442
[16] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** (2009) 591–613. MR2529440
[17] G. Lecué and S. Mendelson. General non-exact oracle inequalities in the unbounded case. Preprint.
[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer, Berlin, 1991. MR1102015
[19] S. Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity* **24** (2008) 380–397. MR2426759
[20] S. Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.* **20** (2010) 988–1027. MR2729283
[21] S. Mendelson, A. Pajor and N. Tomczak-Jaegermann. Reconstruction and sub-Gaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17** (2007) 1248–1282. MR2373017
[22] V. V. Petrov. *Limit Theorems of Probability Theory*. *Oxford Studies in Probability* **4**. Oxford Univ. Press, New York, 1995. MR1353441
[23] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. *Cambridge Tracts in Math.* **94**. Cambridge Univ. Press, Cambridge, 1989. MR1036275
[24] A. B. Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)* 303–313. *Lecture Notes in Artificial Intelligence* **2777**. Springer, Heidelberg, 2003.
[25] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** (2004) 135–166. MR2051002
[26] A. W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Ser. Statist.* Springer, New York, 1996. MR1385671
[27] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.* **28** (2000) 75–87. MR1762904
[28] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli* **10** (2004) 25–47. MR2044592