On the optimality of the empirical risk minimization procedure for the Convex Aggregation problem

Guillaume Lecué\textsuperscript{1,3} Shahar Mendelson\textsuperscript{2,4}

September 24, 2010

Abstract

We study the performances of the empirical risk minimization procedure (ERM for short), relative to the quadratic risk, in the context of \((C)\) aggregation (the “C” is for “convex”), 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 \((C)\) aggregation problem. We also show that if this procedure is used for the problem of \((MS)\) 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 \((C)\) aggregation problem, and thus it is far from optimal. These results are obtained in deviation and are sharp up to logarithmic factors.

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 set of given data.

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

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

\textsuperscript{1}CNRS, LAMA, Université Paris-Est Marne-la-vallée, 77454 France and Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia.

\textsuperscript{2}Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel, and Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia.

\textsuperscript{3}Email: guillaume.lecue@univ-mlv.fr

\textsuperscript{4}Email: shahar@tx.technion.ac.il
If $\hat{f}$ is a random function constructed using the data $\mathcal{D}$, the quadratic risk of $\hat{f}$ is the random variable
\[ R(\hat{f}) = \mathbb{E} \left[ (Y - \hat{f}(X))^2 | \mathcal{D} \right]. \]

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.

In the aggregation framework, one is given a finite set $\mathcal{F}$ 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 $\mathcal{F}$ (cf. [2, 8, 9, 10, 11, 15, 24, 25, 27]).

2. In the Convex (C) aggregation problem (cf. [1, 5, 7, 8, 11, 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 $\mathcal{F}$.

3. In the linear (L) aggregation problem (cf. [8, 10, 14, 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 $\mathcal{F}$.

One can define the optimal rates of the (MS), (C) and (L) aggregation problems, respectively denoted by $\psi_n^{(MS)}(M)$, $\psi_n^{(C)}(M)$ and $\psi_n^{(L)}(M)$ (see, for example, [24]). The optimal rates are the smallest prices that one has to pay to solve the (MS), (C) or (L) aggregation problems in expectation, as a function of the cardinality of the dictionary $M$ and of the sample size $n$. It has been proved in [24] that
\[ \psi_n^{(MS)}(M) \sim \frac{\log M}{n}, \psi_n^{(L)}(M) \sim \frac{M}{n} \quad \text{and} \quad \psi_n^{(C)}(M) \sim \min \left( \frac{M}{n}, \sqrt{\frac{\log \left( eM/\sqrt{n} \right)}{n}} \right), \]
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 their optimality holds only in that context. Nevertheless, lower bounds in deviation follow from the argument of [24] for the three aggregation problems with the same rates $\psi_n^{(MS)}(M)$, $\psi_n^{(C)}(M)$ and $\psi_n^{(L)}(M)$. 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 dictionary $M \geq 1$ and any aggregation procedure $\tilde{f}_n$, there exists a dictionary $\mathcal{F}$ of size $M$ such that with probability larger than $c_0$,
\[ R(\tilde{f}_n) \geq \min_{f \in \Delta(\mathcal{F})} R(f) + c_1 \psi_n^{(MS)}(M), \]
where the residual term $\psi_n^{\Delta(F)}(M)$ is $\psi_n^{(MS)}(M)$ (resp. $\psi_n^{(C)}(M)$ or $\psi_n^{(L)}(M)$ ) when $\Delta(F) = F$ (resp. $\Delta(F) = \text{conv}(F)$ or $\Delta(F) = \text{span}(F)$).

Procedures achieving these rates in deviation have been constructed for the (MS) aggregation problem ([15]) and the (L) aggregation problem ([14]). So far, there was no example of a procedure that achieves the rate of aggregation $\psi_n^{(C)}(M)$ with high probability for the (C) aggregation problem. Moreover the rate $\psi_n^{(C)}(M)$ was achieved in [24] (in expectation) in the Gaussian regression model with a known variance and a known marginal distribution of the design. In [7], the authors were able to remove these assumptions at a price of an extra $\log n$ factor for $1 \leq M \leq \sqrt{n}$. Moreover, the procedures of [24] and [7] depend on a parameter $T$, called the temperature parameter, which has to be chosen larger than certain parameters of the model (depending on the tail of the output and on the parameter $b$). On one hand the residue is monotone in $T$, but on the other, when $T$ is too small, then these procedures are far from optimal in general (cf. [16]). This makes the use of these aggregates “risky” since there is no natural (deterministic or data-dependent) way of choosing $T$.

Despite that no procedure has been shown to achieve the rate $\psi_n^{(C)}(M)$ in deviation, we will call this rate the optimal rate of (C) aggregation, 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 $\psi_n^{(C)}(M)$ in deviation (up to a $\log n$ factor for values of $M$ close to $\sqrt{n}$) without requiring the choice of some “model dependent” parameter. Indeed, we will show that the procedure $\tilde{f}^{\text{ERM}-C}$ minimizing the empirical risk functional
\[
  f \mapsto R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,
\]  
(1.2)
in $\text{conv}(F)$ achieves, with high probability, the rate $\min \left( \frac{M}{n}, \sqrt{\frac{\log M}{n}} \right)$ for the (C) aggregation problem (see the exact formulation in Theorem B below). Moreover, we will show that the optimal rate $\psi_n^{(C)}(M)$ can be achieved by $\tilde{f}^{\text{ERM}-C}$ for any orthogonal dictionary (formulated in Theorem C). On the other hand, it turns out that the same algorithm is far from the conjectured optimal rate $\psi_n^{(MS)}(M)$ for the (MS) aggregation problem (see Theorem A).

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 (MS) aggregation problem (see [12], [19] or [17]). It is also known that performing ERM in the convex hull of $F$ is a suboptimal aggregation procedure for the (MS) aggregation problem for $M = \sqrt{n}$ (see [15]). Concerning the (L) aggregation problem, ERM in the linear span of $F$ is an optimal procedure ([14]). It is straightforward to see that ERM in
the convex hull of $F$ in the context of (L) aggregation or ERM in $F$ in the context of (C) or (L) aggregation are suboptimal. Therefore, studying the performances of ERM in the convex hull of $F$ in the context of (C) 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 achieve the (conjectured) optimal rate of the (C) aggregation problem. For example, many boosting algorithms (see [23] or [6] 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 (C) aggregation for a general dictionary is of high practical importance.

Our first main result is to prove a lower bound on the performance of $\tilde{f}_{\text{ERM}}$ (ERM in the convex hull) in the context of the (MS) aggregation problem. In [15], 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 all values of $M$ and $n$.

**Theorem A** There exist absolute positive constants $c_1, c_2$ for which the following holds. For any integer $n$ and $M$ there exists a dictionary $F$ of cardinality $M$ such that, with probability greater than $c_1$

$$R(\tilde{f}_{\text{ERM}}) \geq \min_{f \in F} R(f) + c_2 \psi_n(M),$$

where $\psi_n(M) = M/n$ when $M \leq \sqrt{n}$ and $(n \log(eM/\sqrt{n}))^{-1/2}$ when $M > \sqrt{n}$.

Note that the residual term $\psi_n(M)$ of Theorem A is much larger than the optimal rate $\psi_n^{(MS)}(M) = (\log M)/n$ for the (MS) aggregation problem. Theorem A shows that ERM in the convex hull satisfies a much stronger lower bound than the one mentioned in (1.1) that holds for any algorithm. This result is of particular importance since optimal aggregation procedures for the (MS) aggregation problem take their values in $\text{conv}(F)$, and it was thus conjectured that $\tilde{f}_{\text{ERM}}$ could be an optimal aggregation procedure for the (MS) aggregation problem. In [15] it was proved that this not the case for $M = \sqrt{n}$; Theorem A shows that this is not the case for any $M$ and $n$.

The proof of Theorem A requires two separate arguments. The case $M \leq \sqrt{n}$ is easier, and follows an identical path to the one used in [15] 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. The difficulty here stems from the fact that if $M > \sqrt{n}$,
the “complexity” of the sets of functions in the convex hull whose excess risk is at most \( r \approx \sqrt{\log(cM/\sqrt{n})/\sqrt{n}} \) does not change much with a proportional change in \( r \). Thus, accurate estimates on the risk of the ERM in that range require a far more delicate analysis than in the “small dictionary” case. Let us mention that for \( M > \sqrt{n} \) the result is optimal up to a logarithmic factor, and the conjectured rate is \( \sqrt{\log(cM/\sqrt{n})/\sqrt{n}} \).

The performance of ERM in the convex hull has been studied for an infinite dictionary in [5], 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 \( M/n \) for every \( n \) and \( M \). And indeed, we establish the following upper bound on the risk of \( \tilde{f}_{\text{ERM}} - C \) as a (C)-aggregation procedure:

**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 \( \log M \leq c_1(b)\sqrt{n} \). For any couple \((X,Y)\) and any finite dictionary \( F \) of cardinality \( M \) such that \(|Y|, \sup_{f \in F} |f(X)| \leq b \), and for any \( u > 0 \), with probability greater than \( 1 - \exp(-u) \),

\[
R(\tilde{f}_{\text{ERM}} - C) \leq \min_{f \in \text{conv}(F)} R(f) + c_2b^2 \max \left[ \min \left( \frac{M}{n}, \sqrt{\frac{\log M}{n}} \right), \frac{u}{n} \right].
\]

Although Theorem B is new, it is probably known to experts, and its proof is based on what is now, rather standard machinery.

Note that Theorem B is optimal except for values of \( M \) for which \( n^{1/2} < M \leq c(\epsilon)n^{1/2+\epsilon} \) for \( \epsilon > 0 \). 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 C** Under the assumptions of Theorem B, if \( F = \{f_1, \ldots, f_M\} \) is such 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 optimal rate of (C) aggregation: for any \( u > 0 \), with probability greater than \( 1 - \exp(-u) \)

\[
R(\tilde{f}_{\text{ERM}} - C) \leq \min_{f \in \text{conv}(F)} R(f) + c_2b^2 \max \left[ \min \left( \frac{M}{n}, \sqrt{\frac{1}{n} \log \left( \frac{eMb^2}{\sqrt{n}} \right)} \right), \frac{u}{n} \right].
\]

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

Combining Theorem A with Theorem B, it follows that up to some logarithmic terms, the rate \( \psi_n^{(C)}(M) \) is the rate of aggregation of \( \tilde{f}_{\text{ERM}} - C \) for the (MS) and (C) aggregation problems. In particular, it achieves the conjectured optimal rate for
the (C) aggregation problem, up to a logarithmic factor that appears when \( n^{1/2+\epsilon} < M \leq c_1(\epsilon)n^{1/2+\epsilon} \). However, it is far from the conjectured optimal rate for the (MS) aggregation problem.

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. \( x \preceq y \)) 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 \( x \preceq b y \) means that \( x \leq C(b)y \) for some constant \( C(b) \) depending only on \( b \). We denote by \( \ell_p^M \) the space \( \mathbb{R}^M \) endowed with the \( \ell_p \) norm. The unit ball there is denoted by \( B^M_p \). We also denote the unit Euclidean sphere in \( \mathbb{R}^M \) by \( S^{M-1} \).

If \( F \) is a class of functions, then let \( f^* \) be a minimizer in \( F \) of the true risk; in our case, \( f^* \) is the minimizer of \( E( f(X) - Y )^2 \). For every \( f \in F \) set \( L_f = (Y - f(X))^2 - (Y - f^*(X))^2 \), and let \( L_F = \{ L_f : f \in F \} \) be the excess loss class associated with \( F \), the target \( Y \) and the quadratic risk.

## 2 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 \( M \leq \sqrt{n} \). This is due to the fact that if \( 0 < \theta < 1 \) and \( \rho = \theta r \sim M/n \), the set \( B^M_1 \cap \sqrt{\rho}S^{M-1} \) is much “larger” than the set \( B^M_1 \cap \sqrt{r}B^M_2 \). 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 \( \rho \). The case \( M \geq \sqrt{n} \) is much harder because when considering the required values of \( r \) and \( \rho \), the complexity of the two sets is very close, and comparing the two oscillations accurately involves a far more delicate analysis.

### 2.1 The case \( M \leq \sqrt{n} \)

We will follow the method used in [15]. Let \( (\phi_i)_{i \in \mathbb{N}} \) be a sequence of functions defined on \( [0,1] \) and set \( \mu \) to be a probability measure on \( [0,1] \) such that \( (\phi_i : i \in \mathbb{N}) \) is a sequence of independent Rademacher variables in \( L_2([0,1], \mu) \).

Let \( M \leq \sqrt{n} \) be fixed and put \( (X,Y) \) to be a couple of random variables; \( X \) is distributed according to \( \mu \) and \( Y = \phi_{M+1}(X) \). Let \( F = \{ 0, \pm \phi_1, \ldots, \pm \phi_M \} \) 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 \( \lambda \in B^M_1 \). Since relative to \( \text{conv}(F) \), \( f^* = 0 \), the excess quadratic
loss function is
\[ \mathcal{L}_\lambda(X, Y) = 1 - 2\phi_{M+1}(X)\langle \lambda, \Phi(X) \rangle + \langle \lambda, \Phi(X) \rangle^2 \]
where we set \( \Phi(\cdot) = (\phi_1(\cdot), \ldots, \phi_M(\cdot)) \).

The following is a reformulation of Lemma 5.4 in [15].

**Lemma 2.1** There exist absolute constants \( c_0, c_1 \) and \( c_2 \) for which the following holds.
Let \( (X_i, Y_i)_{i=1,\ldots,n} \) be \( n \) independent copies of \((X, Y)\). Then, for every \( r > 0 \), with probability greater than 1 – 8 exp(−\( c_0 M \)), for any \( \lambda \in \mathbb{R}^M \),
\[
\left| \|\lambda\|^2 - \frac{1}{n} \sum_{i=1}^{n} \langle \lambda, \Phi(X_i) \rangle^2 \right| \leq \frac{1}{2} \|\lambda\|^2
\] (2.1)
and
\[
c_1 \sqrt{\frac{rM}{n}} \leq \sup_{\lambda \in \sqrt{r}B_2^M} \frac{1}{n} \sum_{i=1}^{n} \langle \lambda, \Phi(X_i) \rangle \phi_{M+1}(X_i) \leq c_2 \sqrt{\frac{rM}{n}}.
\] (2.2)

Set \( r = \beta M/n \) for some \( 0 < \beta \leq 1 \) to be named later, and observe that \( B_1^M \cap \sqrt{r}S_1^M = \sqrt{r}S_1^M \) because \( r \leq 1/M \). Applying (2.1) and (2.2), it is evident that with probability greater than 1 – 8 exp(−\( c_0 M \))
\[
\inf_{\lambda \in B_1^M \cap \sqrt{r}S_1^M} P_n \mathcal{L}_\lambda = r - \sup_{\lambda \in \sqrt{r}S_1^M} (P - P_n) \mathcal{L}_\lambda
\]
\[
\leq r + \sup_{\lambda \in \sqrt{r}S_1^M} \left| \|\lambda\|^2 - \frac{1}{n} \sum_{i=1}^{n} \langle \lambda, X_i \rangle^2 \right| - \sup_{\lambda \in \sqrt{r}S_1^M} \frac{2}{n} \sum_{i=1}^{n} \langle \lambda, \Phi(X_i) \rangle \phi_{M+1}(X_i)
\]
\[
\leq \frac{3r}{2} - 2c_1 \sqrt{\frac{rM}{n}} = \left( \frac{3\beta}{2} - 2c_1 \beta \right) \frac{M}{n} \leq -c_1 \sqrt{\frac{M}{n}},
\]
provided that \( \beta \leq \left( \frac{2c_1}{3} \right)^2 \).

On the other hand, let \( \rho = \alpha M/n \) for some \( \alpha \) to be chosen later. Using (2.1) and (2.2) again, it follows that with probability at least 1 – 8 exp(−\( c_0 M \)), for any \( \lambda \in B_1^M \cap \sqrt{\rho}B_2^M \)
\[
|P_n \mathcal{L}_\lambda| \leq P \mathcal{L}_\lambda + \left| \|\lambda\|^2 - \frac{1}{n} \sum_{i=1}^{n} \langle \lambda, \Phi(X_i) \rangle^2 \right| + \left| \frac{2}{n} \sum_{i=1}^{n} \langle \lambda, \Phi(X_i) \rangle \phi_{M+1}(X_i) \right|
\]
\[
\leq \frac{3\rho}{2} + 2c_2 \sqrt{\frac{\rho M}{n}} = \left( \frac{3\alpha}{2} + 2c_2 \alpha \right) \frac{M}{n}.
\]
Therefore, if \( 0 < \alpha < \beta \) satisfies that \( 3\alpha/2 + 2c_2 \sqrt{\alpha} < c_1 \sqrt{\beta} \) for some \( 0 < \beta \leq \left( \frac{2c_1}{3} \right)^2 \) then, with probability greater than 1 – 16 exp(−\( c_0 M \)), \( \lambda \mapsto R_n(f_\lambda) \) will be “more negative” on \( B_1^M \cap \sqrt{r}S_1^M \) than on \( B_1^M \cap \sqrt{\rho}B_2^M \). Hence, with that probability, \( R(f^{ERM-C}) \geq \rho = \alpha M/n \).
2.2 The case $M \geq \sqrt{n}$

Let us reformulate the second part of Theorem A.

**Theorem 2.2** There exist absolute constants $c$ and $n_0$ for which the following holds. For every integers $n \geq n_0$ and $M$, if $M \geq \sqrt{n}$, 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 $\mu$, such that with $\mu \otimes_n$-probability at least $1/2$,

$$R(\hat{f}) \geq \min_{f \in F_M} R(f) + \frac{c}{\sqrt{n \log(eM/\sqrt{n})}},$$

where $\hat{f}$ is the empirical minimizer in $\text{conv}(F_M)$.

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

**Lemma 2.3** 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 $0 < p < 1$,

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

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

$$|W^2(p) - W^2((1 + \epsilon)p)| \leq C\epsilon, \quad |W^2(p) - W^2((1 - \epsilon)p)| \leq C\epsilon.$$

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

$$\frac{\sqrt{2}}{x\sqrt{\pi}} \exp(-x^2/2) \left(1 - \frac{1}{x^2}\right) \leq \mathbb{P}(|g| > x) \leq \frac{\sqrt{2}}{x\sqrt{\pi}} \exp(-x^2/2), \quad (2.3)$$

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 $X_1, \ldots, X_n$ are i.i.d random variables, with mean zero, variance 1, under the additional assumption that $X$ has well behaved tails.

**Definition 2.4** [18, 26] Let $1 \leq \alpha \leq 2$. We say that a random variable $X$ belongs to $L_{\psi_\alpha}$ if there exists a constant $C$ such that

$$\mathbb{E} \exp(|X|^\alpha/C^\alpha) \leq 2. \quad (2.4)$$

The infimum over all constants $C$ for which (2.4) holds defines a norm called the $\psi_\alpha$ norm of $X$, and we denote it by $\|X\|_{\psi_\alpha}$. 

8
Proposition 2.5 ([22], pg. 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, for which $\|X\|_{\psi_1} \leq L$. If $Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ then for any $0 < x \leq c_1 n^{1/6}$,

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

and

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

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

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

Since Proposition 2.5 implies a better gaussian approximation than the standard Berry-Esseen bounds, one may consider the following family of random variables that will be used in the construction.

Definition 2.6 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 $X_1, \ldots, X_n$ are independent copies of $X$, which is a non-atomic random variable with mean 0, variance 1, and satisfies that $\mathbb{E} X_1^3 = 0$ and $\|X\|_{\psi_1} \leq L$.

Let $X_1, \ldots, X_n$ 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

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

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

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

We shall apply Lemma 2.3 and Proposition 2.5 in the following case to bound $\nu^+(i/M)$ and $\nu^- (i/M)$ for every $i$, as long as $M$ is not too large (i.e. $\log M \leq c_1 n^{1/3}$). To that end, set $\epsilon_{M,n} = [(\log M)/n]^{1/2}$, and for fixed values of $M$ and $n$, and $1 \leq i \leq M$ let

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

Corollary 2.7 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 $\log M \leq C_0 n^{1/3}$. Then, for every $1 \leq i \leq M/2$,

$$(\nu^+_i)^2 \leq \log \left( \frac{2M^2}{\pi i^2} \right) - \log \left( \log \left( \frac{2M^2}{\pi i^2} \right) \right) + C_1 \max \left\{ \log \left( \frac{\log \left( \frac{2M^2}{\pi i^2} \right)}{\log \left( \frac{2M^2}{\pi i^2} \right)} \right), \epsilon_{M,n} \right\},$$

9
and

\[(u_i^-)^2 \geq \log \left( \frac{2M^2}{\pi i^2} \right) - \log \left( \log \left( \frac{2M^2}{\pi i^2} \right) \right) - C_1 \max \left\{ \log \left( \frac{\log \left( \frac{2M^2}{\pi i^2} \right)}{\log \left( \frac{2M^2}{\pi i^2} \right)} \right), \epsilon_{M,n} \right\} \, .\]

**Proof.** Since \(\sqrt{\log M} \leq C_0 n^{1/6}\), one may use the gaussian approximation from Proposition 2.5 to obtain

\[
P(|Y| \geq \sqrt{4 \log M}) \leq P(|g| \geq \sqrt{4 \log M}) \left( 1 + c_1 \left( \sqrt{\log M} \right) \right) \leq c_2 M^2.
\]

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

Let \(1 \leq i \leq M/2\) and \(x \in U(i/M)\). Since \(x \leq 2C_0 n^{1/6}\) (because \(x \leq \sqrt{4 \log M} \leq 2C_0 n^{1/6}\)), it follows from Proposition 2.5 that

\[
|i/M - H(x)| \leq c_3 H(x) \frac{x + 1}{\sqrt{n}} \leq c_4 H(x) \epsilon_{M,n},
\]

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

\[
|W^2(i/M) - x^2| \leq c_5 \epsilon_{M,n}.
\]

Indeed, since \(H(x)(1 - c_4 \epsilon_{M,n}) \leq i/M \leq H(x)(1 + c_4 \epsilon_{M,n})\), then by the monotonicity of \(W\) and the second part of Lemma 2.3, setting \(p = H(x)\),

\[
W^2(i/M) \leq W^2((1 + c_4 \epsilon_{M,n}) H(x)) \leq W^2(H(x)) + c_6 \epsilon_{M,n} = x^2 + c_6 \epsilon_{M,n}.
\]

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 2.3.

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

\[
|\{i : |Y_i| \geq u_s^+\}| \geq s/2 \quad \text{and} \quad |\{i : |Y_i| \geq u_s^-\}| \leq 3s/2.
\]

In particular, with probability at least \(5/6\), for every \(j_0 \leq j \leq M/2\),

\[
u_{2j} \leq y_j^* \leq u_{[2(j-1)/3]}^+,
\]

where \([x] = \min\{n \in \mathbb{N} : x \leq n\}\).
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

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

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

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

We will apply this observation to the independent random variables $\delta_i = \mathbb{I}_{(|Y_i| > a)}$, $1 \leq i \leq M$ 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 $1 - 2 \sum_{a=s_0}^M \exp(-c s) \geq 5/6$ and apply an union bound and a change of variable. Hence, with probability at least $5/6$, for every $\lfloor (3s_0)/2 \rfloor + 1 \leq j \leq \lceil M/2 \rceil$,

$$|\{i : |Y_i| \geq u_{2j}\}| \geq j \quad \text{and} \quad |\{i : |Y_i| \geq u_{(2j-1)/3}\}| \leq j - 1$$

and thus $u_{2j} \leq Y^* \leq u_{(2j-1)/3}$. \(\blacksquare\)

With Lemma 2.8 and Corollary 2.7 in hand, we can bound the following functional of the random variables $(Y_i^*)_{i=1}^M$.

**Lemma 2.9** For every $L > 0$ there exist constants $c_1, \ldots, c_4, j_0$ and $\alpha < 1$ that depend only on $L$ for which the following holds. Let $Y$ be $(L,n)$-almost gaussian and let $Y_1, \ldots, Y_M$ be independent copies of $Y$. Then, with probability at least $5/6$, for every $\lfloor (3s_0)/2 \rfloor + 1 \leq j \leq \lceil M/2 \rceil$,

$$c_1 \frac{\log(ek/\ell) - \epsilon_{M,n}}{\sqrt{\log(eM/\ell)}} \leq Y^*_j - Y^*_k \leq c_2 \frac{\log(ek/\ell) + \epsilon_{M,n}}{\sqrt{\log(eM/\ell)}}.$$ 

Moreover, with probability at least $2/3$,

$$Y^*_j - Y^*_k - \left( \frac{1}{k} \sum_{i=1}^k (Y^*_i - Y^*_k) \right)^{1/2} \geq c_3 \sqrt{\frac{\log(ek/\ell)}{\log(eM/\ell)}} - c_4 \frac{1}{\sqrt{\log(eM/k)}}.$$ 

provided that $\log^2 M \lesssim_L k$ and that $\epsilon_{M,n} = \sqrt{\log M}/n \leq 1$. 

11
Proof. The first part of the claim follows from Lemma 2.8 and Corollary 2.7, 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 $5/6$, $Y_1^* \leq c_1(L)\sqrt{\log M}$. Hence, applying the first part of the claim, with probability at least $2/3$,

$$
\frac{1}{k} \sum_{i=1}^{k} (Y_i^* - Y_k^*)^2 \leq c_1(L) \frac{j_0 \log M}{k} + \frac{1}{k} \sum_{i=j_0}^{k} (Y_i^* - Y_k^*)^2
$$

$$
\leq c_1(L) \frac{j_0 \log M}{k} + \frac{c_2}{k} \sum_{i=j_0}^{k} \left( \frac{\log^2(ek/i)}{\log(eM/i)} + \frac{\epsilon_{M,n}^2}{\log(eM/i)} \right)
$$

$$
\leq c_1(L) \frac{j_0 \log M}{k} + \frac{c_3}{k} \log(eM/k) \leq c_4 \log(eM/k),
$$

provided that $\log^2 M \lesssim L$ and that $\epsilon_{M,n} \leq 1$. Note that to estimate the sum we used that

$$
\frac{1}{k} \sum_{i=j_0}^{k} \log^2(ek/i) = \frac{1}{\log(eM/k)} \sum_{i=j_0}^{k} \log^2(ek/i) \leq \frac{c_3}{\log(eM/k)}.
$$

Now the second preliminary part follows 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 $A_r = B_1^M \cap \sqrt{r}B_2^M$. From here on, given $v \in \mathbb{R}^M$ we shall set

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

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

**Lemma 2.10** For every $v \in \mathbb{R}^M$ and any $0 < \rho < r \leq 1$ such that $1/r$ and $1/\rho$ are integers,

$$
\|v\|_{A_r^*} - \|v\|_{A_\rho^*} \geq v_{1/r}^* - v_{1/\rho}^* - \left( \frac{1}{\rho} \sum_{i=1}^{1/\rho} (v_i^* - v_{1/\rho}^*)^2 \right)^{1/2}.
$$

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

$$
\|v\|_{A_r^*} = \inf_{1 \leq j \leq M} \left( \frac{1}{r} \left( \sum_{i=1}^{j} (v_i^* - v_j^*)^2 \right)^{1/2} + v_j^* \right). \quad (2.6)
$$
Indeed, since $A_r^o = \text{conv}(B_M^r \cup (1/\sqrt{r})B_2^M)$, it is evident that $\|v\|_{A_r^o} = \inf\{\|u\|_\infty + \sqrt{r}\|w\|_2, \ v = u + w\}$. One may verify that if $v = u + w$ is an optimal decomposition then $\text{supp}(w) \subset \{i : |u_i| = \|u\|_\infty\}$. Hence, if $\|u\|_\infty = K$ then $u_i = K\text{sgn}(v_i)\mathds{1}_{\{|v_i| \geq K\}} + v_i\mathds{1}_{\{|v_i| < K\}}$ for every $1 \leq i \leq M$, implying that $w_i = \mathds{1}_{\{|v_i| \geq K\}}(v_i - \text{sgn}(v_i)K)$ for every $1 \leq i \leq M$. Therefore,

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

Moreover, since it is enough to consider only values of $K$ in $\{v_j^* : 1 \leq j \leq M\}$, (2.6) is verified. In particular, if $1/r$ is an integer then

$$\|v\|_{A_r^o} \leq \sqrt{r} \left( \sum_{i=1}^{1/r} (v_i^* - v_1^*)^2 \right)^{1/2} + v_1^*/r.$$ 

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

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

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

$$\|v\|_{A_r^o} - \|v\|_{A_r^o} \geq \sqrt{r} \left( \sum_{i=1}^{1/r} (v_i^*)^2 \right)^{1/2} - \left( \sqrt{\rho} \left( \sum_{i=1}^{1/\rho} (v_i^* - v_1^*)^2 \right)^{1/2} + v_1^*/\rho \right) \geq v_1^*/r - v_1^*/\rho - \sqrt{\rho} \left( \sum_{i=1}^{1/\rho} (v_i^* - v_1^*)^2 \right)^{1/2},$$

because $(r \sum_{i=1}^{1/r} (v_i^*)^2)^{1/2} \geq v_1^*/r$. 

**Proof of Theorem 2.2.** Let $\Phi_1, \ldots, \Phi_M$, $X$ and $\alpha > 0$ be such that $\Phi_1(X), \ldots, \Phi_M(X)$ are uniformly distributed on $[-\alpha, \alpha]$ and have variance 1. 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 $F_M = \{0, \pm \Phi_1, \ldots, \pm \Phi_M\}$. Note that the functions in $\text{conv}(F_M)$ are given by $f_\lambda = \langle \Phi, \lambda \rangle$ where $\Phi = (\Phi_1, \ldots, \Phi_M)$ and $\lambda \in B_1^M$.

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

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

13
(since $f^* = 0$), implying that $\mathbb{E}L_{f^*} = \|\lambda\|^2_2$.

Let us consider the problem of empirical minimization in $\text{conv}(F_M) = \{\langle \lambda, \Phi \rangle : \lambda \in B^*_M \}$. Recall that $A_r = B^*_M \cap \sqrt{r}B^*_2$ and, for an independent sample $(\Phi(X_i), \Phi_{M+1}(X_i))_{i=1}^n$, define the functional

$$
\psi(r, \rho) = n \left( \inf_{\lambda \in A_r} R_n(f_\lambda) - \inf_{\mu \in A_\rho} R_n(f_\mu) \right)
= \inf_{\lambda \in A_r} \sum_{i=1}^n \langle \Phi(X_i), \lambda \rangle^2 - 2 \langle \Phi(X_i), \lambda \rangle \Phi_{M+1}(X_i)
- \inf_{\mu \in A_\rho} \sum_{i=1}^n \langle \Phi(X_i), \mu \rangle^2 - 2 \langle \Phi(X_i), \mu \rangle \Phi_{M+1}(X_i).
$$

If we show that for some $r \geq \rho$, $\psi(r, \rho) < 0$, then for that sample, $\mathbb{E}L_{\hat{f}} \geq \rho$.

Note that

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

and let us estimate the supremum of the process

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

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 subgaussian – since $\|\langle \lambda, \Phi(X) \rangle\|_{\psi_2} \leq 4a \|\lambda\|_2$. Hence, applying the results from [21], it is evident that with probability at least $3/4$,

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

Recall that for $r \geq 1/M$, $\gamma_2(A_r, \|\cdot\|_2) \sim \sqrt{\log(eMr)}$ (see, for instance, [21], and thus, if $r \geq \max(1/M, 1/n)$, then with probability at least $3/4$,

$$
n \sup_{\lambda \in A_r} \left( (P_n - P)(\langle \Phi, \lambda \rangle^2) + \|\lambda\|^2_2 \right) \leq nr + c_1 \sqrt{nr \log(eMr)},
$$
where $c_1$ is a constant that depends only on $a$.

Next, to estimate the first two terms, let

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

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

$$\sup_{\lambda \in A_r} \sum_{i=1}^n \langle \lambda, \Phi(X_i) \rangle \Phi_{M+1}(X_i) - \sup_{\theta \in A_\rho} \sum_{i=1}^n \langle \theta, \Phi(X_i) \rangle \Phi_{M+1}(X_i)$$

$$= \sqrt{n} \left( \sup_{\lambda \in A_r} \langle \lambda, V \rangle - \sup_{\theta \in A_\rho} \langle \theta, V \rangle \right) = \sqrt{n} (\|V\|_{A_r}^2 - \|V\|_{A_\rho}^2) = (*)$$

By Lemma 2.10, if $1/r = \ell$ and $1/\rho = k$ are integers, then

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

and thus, if $\ell, k, M$ and $n$ are as in Lemma 2.9, then with probability at least $2/3$,

$$(*) \geq \sqrt{n} \left( \frac{c_2 \log(ek/\ell)}{\sqrt{\log(eM/\ell)}} - \frac{c_3}{\sqrt{\log(eM/k)}} \right)$$

$$\geq c_4 \sqrt{n} \frac{\log(ek/\ell)}{\sqrt{\log(eM/\ell)}},$$

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

Hence, with probability at least $1/2$,

$$\psi(r, \rho) \leq -2c_4 \sqrt{n} \frac{\log(ek/\ell)}{\sqrt{\log(eM/\ell)}} + nr + c_1 \sqrt{rn \log(eMr)}.$$ 

It follows that if we select $r \sim 1/\sqrt{n \log(eM/\sqrt{n})}$ and $\rho \sim r$ with $\rho < r$ so that the conditions of Lemma 2.9 are satisfied, then with probability at least $1/2$, $\psi(r, \rho) < 0$. Hence, with the same probability,

$$R(\hat{f}) - \min_{f \in F_M} R(f) = \mathbb{E} L_{\hat{f}} \geq \frac{c_6}{\sqrt{n \log(eM/\sqrt{n})}}.$$ 

\[\blacksquare\]
3 Proof of the upper bound for the (C) aggregation problem (Theorem B)

Our starting point is to describe the machinery developed in [3], 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 $L_G = \{(x,y) \mapsto (y - g(x))^2 - (y - g^*_G(x))^2 : g \in G\}$ the associated class of quadratic excess loss functions, where $g^*_G$ is the minimizer of the quadratic risk in $G$. Let $V = \text{star}(L_G, 0) = \{\theta L : 0 \leq \theta \leq 1, L \in L_G\}$ and for every $\lambda > 0$ set $V_\lambda = \{h \in V : \mathbb{E} h \leq \lambda\}$.

**Theorem 3.1 ([3])** 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 $L_G$ consists of functions that are bounded by $b$ almost surely. Assume further that for any $L \in L_G$, $\mathbb{E} L^2 \leq B \mathbb{E} L$. If $x > 0$, $\lambda > 0$ satisfies that $\mathbb{E} \|P - P_n\| V_\lambda \leq \lambda^*/8$ 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 $\hat{g}$ in $G$ satisfies

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

Let $F$ be the given dictionary and set $G = \text{conv}(F)$. Using the notation of Theorem 3.1, put

$$
L_{\text{conv}(F)} = \{L_f : f \in \text{conv}(F)\},
$$

consider the star-shaped hull $V = \text{star}(L_{\text{conv}(F)}, 0)$ and its localizations $V_\lambda = \{g \in V : \mathbb{E} g \leq \lambda\}$ for any $\lambda > 0$.

Thanks to convexity, the following observation holds in our case (see [20] for the proof).

**Proposition 3.2** If $f \in \text{conv}(F)$ then $\mathbb{E} L_f \geq \|f - f^*\|_{L_2(P_X)}^2$ where $f^*$ is the minimizer of the quadratic risk in $\text{conv}(F)$. In particular,

1. $\mathbb{E} L^2 \leq 4b^2 \mathbb{E} L$ for any $L \in L_{\text{conv}(F)}$;

2. For $\mu > 0$, if $f \in \text{conv}(F)$ satisfies that $\mathbb{E} L_f \leq \mu$, then $f \in f^* + K_\mu$, where

$$
K_\mu = 2[\text{conv}\{\pm f_1, \ldots, \pm f_M\} \cap \sqrt{\mu} B(L_2(P_X))].
$$

The first part of Proposition 3.2 shows that $L_{\text{conv}(F)}$ satisfies the assumptions of Theorem 3.1 with $B = 4b^2$. 

16
To apply Theorem 3.1 one has to find $\lambda^* > 0$ for which $E \|P - P_n\|_{V_{\lambda^*}} \leq \lambda^*/8$, and to that end we will use the second part of Proposition 3.2. Observe that

$$V_{\lambda} \subset \bigcup_{i=0}^{\infty} \{ \theta \mathcal{L}_f : 0 \leq \theta \leq 2^{-i}, E \mathcal{L}_f \leq 2^{i+1} \lambda \},$$

and it was shown in [4] that

$$E \|P - P_n\|_{V_{\lambda}} \leq \sum_{i \geq 0} 2^{-i} E \|P - P_n\|_{L^{2i+1}_{\lambda}},$$

where from here on we set $L_{\mu} = \{ \mathcal{L} \in L_{\text{conv}(F)} : E \mathcal{L} \leq \mu \}$. Applying the second part of Proposition 3.2 it is evident that

$$\{ f \in \text{conv}(F) : E \mathcal{L}_f \leq \mu \} \subset f^* + K_{\mu}.$$  

There are two trivial ways of approximating $K_{\mu}$, both of which will be used. First, an “$L_1$” approximation, that $K_{\mu} \subset 2\text{conv}\{ \pm f_1, \ldots, \pm f_M \}$. Second, an “$L_2$” approximation, that $K_{\mu} \subset 2\sqrt{\mu B(L_2(P^X))} \cap \text{span}(F)$.

**Proposition 3.3** There exists an absolute constant $c_1 > 0$ for which the following holds. For any $\mu > 0$,

$$E \|P - P_n\|_{L_{\mu}} \leq c_1 \min \left\{ b^2 \sqrt{\frac{\log M}{n}}, b \sqrt{\frac{M \mu}{n}} \right\}.$$

**Proof.** By the Giné-Zinn symmetrization Theorem [26],

$$E \|P - P_n\|_{L_{\mu}} \leq 2E E_\epsilon \sup_{\mathcal{L} \in L_{\mu}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathcal{L}(X_i, Y_i) \right|. \quad (3.2)$$

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

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

$$= |(f^*(x) - f(x))(2y - f(x) - f^*(x))| \leq 4b |f(x) - f^*(x)|.$$

Thus, by the contraction principle (see, e.g. [18]) and Proposition 3.2,

$$E \|P - P_n\|_{L_{\mu}} \leq \frac{8b}{\sqrt{n}} E E_\epsilon \sup_{f \in K_{\mu}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|.$$

17
Set $F(\cdot) = (f_1(\cdot), \ldots, f_M(\cdot))$. By the comparison theorem between Rademacher and Gaussian processes [18] and the “$\ell_1$” approximation for $K_\mu$, it is evident that conditionally on $X_1, \ldots, X_n$,

$$
\mathbb{E}_e \sup_{f \in K_\mu} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i f(X_i) \right| \leq c_1 \mathbb{E}_g \sup_{f \in K_\mu} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i f(X_i) \right|
$$

$$
\leq c_2 \mathbb{E}_g \sup_{\lambda \in B^M_1} \left| \lambda, \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i F(X_i) \right| = c_2 \mathbb{E}_g \max_{1 \leq j \leq M} |\gamma_j|,
$$

where $\gamma_j$ is the Gaussian random variable $n^{-1/2} \sum_{i=1}^n g_i f_j(X_i)$ defined for all $j = 1, \ldots, M$. The variance of $\gamma_j$ is $n^{-1} \sum_{i=1}^n f_j(X_i)^2 \leq b^2$. Hence, by the Gaussian maximal inequality [18],

$$
\mathbb{E}_g \max_{1 \leq j \leq M} |\gamma_j| \leq c_3 b \sqrt{\log M},
$$

which completes the first part of the proof.

Turning to the second part, when $M \leq \sqrt{n}$, we will use the strategy developed in [14]. Let $S$ be the linear subspace of $L^2(P^X)$ spanned by $F$ and take $(e_1, \ldots, e_{M'})$ to be an orthonormal basis of $S$ (where $M' = \dim(S) \leq M$). Since $K_\mu \subset S \cap 2\sqrt{M} B(L^2(P^X))$, then

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

$$
\leq c_1 \sqrt{\mathbb{E} \left( \sum_{j=1}^{M'} \left( \frac{1}{n} \sum_{i=1}^n \epsilon_i e_j(X_i) \right)^2 \right)^{1/2}} \leq c_1 \sqrt{\frac{M'\mu}{n}}.
$$

**Proof of Theorem B:** By Proposition 3.3 it follows that for any $\mu > 0$,

$$
\mathbb{E} \|P - P_n\|_{L_\mu} \leq \min \left( c_1 \left( b^2 \sqrt{(\log M)/n}, b\sqrt{M\mu/n} \right) \right).
$$

Then, by (3.1), for every $\lambda > 0$,

$$
\mathbb{E} \|P - P_n\|_{V_\lambda} \leq \sum_{i \geq 0} 2^{-i} \mathbb{E} \|P - P_n\|_{V_{2^{i+1}\lambda}}
$$

$$
\leq c_1 \sum_{i \geq 0} 2^{-i} \min \left( b^2 \sqrt{\frac{\log M}{n}}, b\sqrt{\frac{M2^{i+1}\lambda}{n}} \right)
$$

$$
\leq c_2 \min \left( b^2 \sqrt{\frac{\log M}{n}}, b\sqrt{\frac{M\lambda}{n}} \right).
$$
Hence, $E \|P - P_n\|_{V^*} \leq \lambda^*/8$ if

$$\lambda^* \sim b^2 \min \left(\sqrt{\frac{\log M}{n}}, \frac{M}{n} \right),$$

and Theorem B follows from Theorem 3.1.

Next, we turn to the proof of Theorem C, in which the “trivial” approximations of $K_\mu$ are too weak to give the desired upper bound.

**Proof of Theorem C.** Observe that since the dictionary consists of an orthogonal family, if $F(x) = (f_1(x), \ldots, f_M(x))$ and if $(e_1, \ldots, e_M)$ is the standard basis in $\ell_2^M$, then

$$K_\mu = \left\{ 2\langle \lambda, F \rangle : \lambda \in B_1^M \cap \sqrt{\mu} E \right\},$$

where $E$ is an ellipsoid with principal axes $\parallel f_i \parallel_{L^2} e_i$ where $i = 1, \ldots, M$. From here on we will assume that $\parallel f_i \parallel_{L^2}$ is a non-increasing sequence.

Following the path used in the proof of Theorem B, it is enough to bound

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

where $\parallel \cdot \parallel_{(B_1^M \cap \sqrt{\mu} E)^\circ}$ denotes the dual norm to the one whose unit ball is $B_1^M \cap \sqrt{\mu} E$. Since both $B_1^M$ and $E$ are unconditional with respect to the coordinate structure given by $(e_i)_{i=1}^M$, it follows that

$$\parallel v \parallel_{(B_1^M \cap \sqrt{\mu} E)^\circ} \sim \inf_{I \subset \{1, \ldots, M\}} \left[ \sqrt{\mu} \left( \sum_{i \in I} \left( \frac{v_i}{\parallel f_i \parallel_{L^2}} \right)^2 \right)^{1/2} + \max_{i \in F^c} |v_i| \right], \quad (3.3)$$

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

Let

$$J_0 = \{ j : \parallel f_j \parallel_{L^2} \geq c_0 b \sqrt{\log M / \sqrt{n}} \},$$

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

$$\mathbb{P}(\exists j \in J_0 : P_n f_j^2 \geq (t + 1) \parallel f_j \parallel_{L^2}^2) \leq \sum_{j \in J_0} \exp \left( -c_2 n (\parallel f_j \parallel_{L^2}^2 / b^2 \min(t^2, t) \right) \leq M \exp(-c_3 t \log M) \leq \exp(-c_4 t \log M),$$
and
\[ P(\exists j \in J^c_0 : P_n f_j^2 \geq (t + 1)b^2n^{-1}\log M) \leq \exp(-c_4t \log M). \]

For every integer \( \ell \geq c_1 \), let
\[ A_{\ell} = \{ \forall j \in J_0 : P_n f_j^2 \leq (\ell + 1)\|f_j\|^2_{L_2} \} \bigcap \{ \forall j \in J^c_0 : P_n f_j^2 \leq (\ell + 1)b^2n^{-1}\log M \}. \]

Set \( B_{\ell} = A_{\ell + 1} \cap A_{\ell}^c \) and note that \( P(B_{\ell}) \leq P(A_{\ell}^c) \leq 2 \exp(-c_4\ell \log M) \) for any \( \ell \geq c_1 \).

For every \( \ell \geq c_1 \), consider the random variables conditioned on \( B_{\ell} \),
\[ U_{j,\ell} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f_j(X_i) \|f_j\|_{L_2} \big| B_{\ell} \quad \forall j \in J_0 \]
and
\[ U_{j,\ell} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f_j(X_i) \big| B_{\ell} \quad \forall j \in J^c_0. \]

Hence, by Höffding inequality (cf. [26]), there exists an absolute constant \( c_5 \) such that, for any \( j \in J_0 \),
\[ \|U_{j,\ell}\|_{\psi_2(\epsilon)}^2 \leq c_5 \frac{n^{-1} \sum_{i=1}^{n} f_j^2(X_i)}{\|f_j\|^2_{L_2}} \leq c_5(\ell + 1), \]
and for any \( j \in J^c_0 \),
\[ \|U_{j,\ell}\|_{\psi_2(\epsilon)}^2 \leq c_5(\ell + 1)b^2(\log M)/n. \]

By a result due to Klartag [13], it follows that for every such \( \ell \) and any \( 1 \leq j \leq |J_0| \),
\[ \mathbb{E}_e \left( \sum_{i=1}^{j} (U_{i,\ell}^2)^* \right)^{1/2} \leq c_6 \ell \sqrt{j \log(e |J_0|/j)}, \]
where \( (U^*_j)_{j=1}^{|J_0|} \) is a monotone rearrangement of \( (|U_{j,\ell}|)_{j=1}^{|J_0|} \). Moreover, by a standard maximal inequality (see, e.g. [26])
\[ \mathbb{E}_e \max_{j \in J^c_0} U_{j,\ell} \leq c_7 \sqrt{\log |J^c_0|} \max_{j \in J^c_0} \|U_{j,\ell}\|_{\psi_2(\epsilon)} \leq c_8 \sqrt{b \log M}/\sqrt{n}. \]
For every $1 \leq j \leq |J_0|$, let $I$ be the set of the $j$ largest coordinates of $(|U_j,\ell|)_{j \in J_0}$. Hence, by (3.3) and since $\|f_j\|_{L^2} \leq b$, 

$$\mathbb{E}_\epsilon \left( \left\| \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \epsilon_i F(X_i) \right\|_{(B_1^M \cap \sqrt{n} \mathcal{E})^c} |(X_i)_{i=1}^{n} \in \mathcal{B}_\ell \right) \leq c_0 \sqrt{\ell} \left( \mu \sqrt{j \log(e|J_0|/j)} + b \sqrt{\log(e|J_0|/j)} \right) + \mathbb{E}_\epsilon \max_{j \in J_0} |U_{j,\ell}|$$

$$\leq c_0 \sqrt{\ell} \left( \sqrt{\mu \sqrt{j \log(e|J_0|/j)} + b \sqrt{\log(e|J_0|/j)}} \right) + c_{10} \sqrt{\ell} b \log M \sqrt{n}.$$ 

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

$$\mathbb{E}_\epsilon \left( \left\| \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \epsilon_i F(X_i) \right\|_{(B_1^M \cap \sqrt{n} \mathcal{E})^c} |(X_i)_{i=1}^{n} \in \mathcal{B}_\ell \right) \lesssim b \sqrt{\ell} \left( \sqrt{\log(eM \mu) + \log M \sqrt{n}} \right).$$

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

$$\frac{2}{\sqrt{n}} \mathbb{E} \left( \left\| \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \epsilon_i F(X_i) \right\|_{(B_1^M \cap \sqrt{n} \mathcal{E})^c} |(X_i)_{i=1}^{n} \in \mathcal{B}_\ell \right) \lesssim b \sqrt{\frac{\log(eM \mu)}{n}}.$$ 

Finally, by (3.1), for any $\lambda > 1/M$, 

$$\mathbb{E} \|P - P_n\|_{V_\lambda} \leq \sum_{i \geq 0} 2^{-i} \mathbb{E} \|P - P_n\|_{L_{2^{-i+1}}^\lambda} \lesssim b^2 \sum_{i \geq 0} 2^{-i} \sqrt{\frac{\log(eM2^{i+1}\lambda)}{n}} \lesssim b^2 \sqrt{\frac{\log(eM\lambda)}{n}},$$

and, if 

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

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

Note that here, the localization is truly needed to obtain the correct rate of aggregation. To generalize Theorem C beyond the orthogonal case, one must study the geometry of the intersection of $B_1^M$ with an ellipsoid, but now the two bodies need not have the same unconditional structure, making the analysis considerably harder.
4 Concluding Remarks

The best rate of (MS) aggregation for the ERM in $F$ is $\sqrt{\log M}/n$ (which is larger than the rate achieved in Theorem B). Therefore, if one decides to use the ERM procedure, it is in general better to use it in $\text{conv}(F)$ than in $F$ itself – even for the problem of (MS) aggregation. In particular, for dictionaries of small cardinality (smaller than $\sqrt{n}$), ERM in the convex hull will outperform ERM in $F$ (since $M/n \leq \sqrt{\log M}/n$). Moreover, note that when $M \leq c_0$ for some absolute constant $c_0$ the rate $M/n$ achieved by $\tilde{f}_{\text{ERM}}$ is of the same order of magnitude as the optimal residue $1/n$, and so ERM in $\text{conv}(F)$ is an optimal aggregation procedure in deviation even for the (MS) aggregation problem for a very small dictionary. Finally, from a computational point of view, minimizing in the convex set $\text{conv}(F)$ is reasonable, and thus for a finite function class $F$, ERM in $\text{conv}(F)$ should be preferred to ERM in $F$.

Our final comment has to do with the logarithmic gaps that both upper and lower bounds have. We believe that the correct rate in both cases is $\sqrt{\log(eM/\sqrt{n})}/\sqrt{n}$ when $M > \sqrt{n}$, and the gap in both cases has a similar source.

To obtain the correct upper bound one must study the empirical process indexed by the intersection body of $B_1^M$ and an ellipsoid coming from the localization. The proof of Theorem C requires that $B_1^M$ and an ellipsoid $E$ share the same unconditional structure, and to obtain the general result, one must remove that assumption. Although it seems reasonable that the worst case is when the two share the same unconditional basis, proving it is nontrivial.

For the lower bound, one has to study the structure of $B_1^M \cap \sqrt{r}B_2^M$ in a range where a proportional change of the radius has a very small impact on the complexity of the sets and thus, on the oscillation of the process indexed by the set. Although we believe that the example we gave here should give the desired bound, we have not been able to improve our estimate. The main difficulty is that for the optimal bound one has to compare two very close levels - $r$ and $\rho$ - where $\sqrt{n}(r-\rho) \sim 1/\sqrt{\log(eM/\sqrt{n})}$, rather than $r \sim \rho$ which is what we have been able to do here.

References


