Bernoulli **19**(2), 2013, 646–675
DOI: 10.3150/11-BEJ408

On the optimality of the aggregate with exponential weights for low temperatures

GUILLAUME LECUÉ[1] and SHAHAR MENDELSON[2]

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

Given a finite class of functions F, the problem of aggregation is to construct a procedure with a risk as close as possible to the risk of the best element in the class. A classical procedure (PAC-Bayesian statistical learning theory (2004) Paris 6, *Statistical Learning Theory and Stochastic Optimization* (2001) Springer, *Ann. Statist.* **28** (2000) 75–87) is the aggregate with exponential weights (AEW), defined by

$$\tilde{f}^{\mathrm{AEW}} = \sum_{f \in F} \widehat{\theta}(f) f, \qquad \text{where } \widehat{\theta}(f) = \frac{\exp(-(n/T)R_n(f))}{\sum_{g \in F} \exp(-(n/T)R_n(g))},$$

where $T > 0$ is called the temperature parameter and $R_n(\cdot)$ is an empirical risk.

In this article, we study the optimality of the AEW in the regression model with random design and in the low-temperature regime. We prove three properties of AEW. First, we show that AEW is a suboptimal aggregation procedure in expectation with respect to the quadratic risk when $T \le c_1$, where c_1 is an absolute positive constant (the low-temperature regime), and that it is suboptimal in probability even for high temperatures. Second, we show that as the cardinality of the dictionary grows, the behavior of AEW might deteriorate, namely, that in the low-temperature regime it might concentrate with high probability around elements in the dictionary with risk greater than the risk of the best function in the dictionary by at least an order of $1/\sqrt{n}$. Third, we prove that if a geometric condition on the dictionary (the so-called "Bernstein condition") is assumed, then AEW is indeed optimal both in high probability and in expectation in the low-temperature regime. Moreover, under that assumption, the complexity term is essentially the logarithm of the cardinality of the set of "almost minimizers" rather than the logarithm of the cardinality of the entire dictionary. This result holds for small values of the temperature parameter, thus complementing an analogous result for high temperatures.

Keywords: aggregation; empirical process; Gaussian approximation; Gibbs estimators

1. Introduction and main results

In this note we study the problem concerning the optimality of the AEW in the regression model with random design. To formulate the problem, we need to introduce several definitions.

Let $\mathcal{Z}$ and $\mathcal{X}$ be two measure spaces, and set Z and $Z_1, \dots, Z_n$ to be $n+1$ i.i.d. random variables with values in $\mathcal{Z}$. From a statistical standpoint, $\mathcal{D} = (Z_1, \dots, Z_n)$ is the set of given data at our disposal. The *risk* of a measurable real-valued function f defined on $\mathcal{X}$ is given by

$$R(f) = \mathbb{E}Q(Z, f),$$

1350-7265

where $Q : \mathcal{Z} \times \mathcal{L}(\mathcal{X}) \mapsto \mathbb{R}$ is a non-negative function, called the *loss function* and $\mathcal{L}(\mathcal{X})$ is the set of all real-valued measurable functions defined on $\mathcal{X}$. If $\widehat{f}$ is a statistic constructed using the data $\mathcal{D}$, then the risk of $\widehat{f}$ is the random variable

$$R(\widehat{f}) = \mathbb{E}[Q(Z, \widehat{f})|\mathcal{D}].$$

Throughout this article, we restrict our attention to functions f, loss functions Q, and random variables Z for which $|Q(Z, f)| \leq b$ almost surely. (Note that some results have been obtained in the same setup for unbounded loss functions in [7,13,32], and [4].) The loss function on which we focus throughout most of the article is the quadratic loss function, defined when $Z = (X, Y)$ by $Q((X, Y), f) = (Y - f(X))^2$.

In the aggregation framework, one is given a finite set F of real-valued functions defined on $\mathcal{X}$, usually called a *dictionary*. The problem of *aggregation* (see, e.g., [7,10], and [31]) is to construct a procedure, usually called an *aggregation procedure*, that produces a function with a risk as close as possible to the risk of the best element in F. Keeping this in mind, one can define the *optimal rate of aggregation* [16,26], which is the smallest price, as a function of the cardinality of the dictionary M and the sample size n, that one has to pay to construct a function with a risk as close as possible to that of the best element in the dictionary. We recall the definition for the "expectation case;" a similar definition for the "probability case" can be formulated as well (see, e.g., [16]).

Definition 1.1 ([26]). *Let $b > 0$. We say that $(\psi_n(M))_{n,M\in\mathbb{N}^*}$ is an optimal rate of aggregation in expectation when there exist two positive constants, c_0 and c_1, depending only on b, for which the following holds for any $n \in \mathbb{N}^*$ and $M \in \mathbb{N}^*$:*

1. *There exists an aggregation procedure $\tilde{f}_n$ such that for any dictionary F of cardinality M and any random variable Z satisfying $|Q(Z, f)| \leq b$ almost surely for all $f \in F$, one has*

$$\mathbb{E}R(\tilde{f}_n) \leq \min_{f\in F} R(f) + c_0\psi_n(M); \tag{1.1}$$

2. *For any aggregation procedure $\bar{f}_n$, there exists a dictionary F of cardinality M and a random variable Z such that $|Q(Z, f)| \leq b$ almost surely for all $f \in F$ and*

$$\mathbb{E}R(\bar{f}_n) \geq \min_{f\in F} R(f) + c_1\psi_n(M).$$

In our setup, one can show (cf. [26]) that in general, an optimal rate of aggregation (in the sense of [26] [optimality in expectation] and of [16] [optimality in probability]) is lower-bounded by $(\log M)/n$. Thus, procedures satisfying an exact oracle inequality like (1.1)—that is, an oracle inequality with a factor of 1 in front of $\min_{f\in F} R(f)$—with a residual term of $\psi_n(M) = (\log M)/n$ are said to be optimal. Only a few aggregation procedures have been shown to achieve this optimal rate, including the exponential aggregating schemes of [2,3,7,13,31], the the "empirical star algorithm" in [3], and the "preselection/convexification algorithm" in [16]. For a survey on optimal aggregation procedures, see the HDR dissertation of J.-Y. Audibert.

Our main focus here is on the problem of the optimality of the aggregation procedure with exponential weights (AEW). This procedure originate from the thermodynamic standpoint of

learning theory (see [8] for the state of the art in this direction). AEW can be viewed as a relaxed version of the trivial aggregation scheme, which is to minimize the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(Z_i, f) \tag{1.2}$$

in the dictionary F.

A procedure that minimizes (1.2) is called *empirical risk minimization* (ERM). It is well known that ERM generally cannot achieve the optimal rate of $(\log M)/n$, unless one assumes that the given class F has certain geometric properties, which we discuss below (see also [13,18,21]). To have any chance of obtaining better rates, one has to consider aggregation procedures that take values in larger sets than F. The most natural set is the convex hull of F. AEW is a very popular candidate for the optimal procedure, and it was one of the first procedures to be studied in the context of the aggregation framework [2,4,7,9,13,15,20,31]. It is defined by the following convex sum:

$$\tilde{f}^{\mathrm{AEW}} = \sum_{j=1}^{M} \widehat{\theta}_j f_j, \qquad \text{where } \widehat{\theta}_j = \frac{\exp(-(n/T)R_n(f_j))}{\sum_{k=1}^{M}\exp(-(n/T)R_n(f_k))} \tag{1.3}$$

for the dictionary $F = \{f_1, \ldots, f_M\}$. The parameter $T > 0$ is called the *temperature*.[1]

Thus far, there have been three main results concerning the optimality of the AEW. The first of these is that the progressive mixture rule is optimal in expectation for T larger than some parameters of the model (see [4,7,13,30,32] and [3]), and under certain convexity assumption on the loss function Q. This procedure is defined by

$$\bar{f} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_k^{\mathrm{AEW}}, \tag{1.4}$$

where $\tilde{f}_k^{\mathrm{AEW}}$ is the function generated by AEW (with a common temperature parameter T) associated with the dictionary F and constructed using only the first k observations $Z_1, \ldots, Z_k$. (See [3] for more details and for other procedures related to the progressive mixture rule.)

Second, the optimality in expectation of AEW was obtained by [9] for the regression model $Y_i = f(x_i) + \varepsilon_i$ with a deterministic design $x_1, \ldots, x_n \in \mathcal{X}$ with respect to the risk $\|g - f\|_n^2 = n^{-1}\sum_{i=1}^{n}(g(x_i) - f(x_i))^2$ (with its empirical version being $R_n(g) = n^{-1}\sum_{i=1}^{n}(Y_i - g(x_i))^2$). That is, it was shown that for $T \geq c\max(b, \sigma^2)$, where σ^2 is the variance of the noise ε,

$$\mathbb{E}\|\tilde{f}^{\mathrm{AEW}} - f\|_n^2 \leq \min_{g\in F}\|g - f\|_n^2 + \frac{T\log M}{n+1}. \tag{1.5}$$

Finally, [1,2], and [8] proved that in the high-temperature regime, AEW can achieve the optimal rate $(\log M)/n$ under the Bernstein assumption, recalled below in Definition 1.3 in expectation and in high probability. This result is discussedin more detail later.

[1]This terminology comes from thermodynamics, since the weights $(\widehat{\theta}_1, \ldots, \widehat{\theta}_M)$ can be seen as a Gibbs measure with temperature T on the dictionary F.

Despite the long history of AEW, the literature contains no results on the optimality (or suboptimality) of AEW in the regression model with random design in the general case (when the dictionary does not necessarily satisfy the Bernstein condition). In this article, we address this issue and complement the results (assuming the Bernstein condition) of [1,2,8] for the low-temperature regime by proving the following:

- AEW is suboptimal for low temperatures $T \le c_1$ (where c_1 is an absolute positive constant), both in expectation and in probability, for the quadratic loss function and a dictionary of cardinality 2 (Theorem A).
- AEW is suboptimal in probability for some large dictionaries (of cardinality $M \sim \sqrt{n \log n}$) and small temperatures $T \le c_1$ (Theorem B).
- AEW achieves the optimal rate $(\log M)/n$ for low temperatures under the Bernstein condition on the dictionary (Theorem C). Together with the high-temperature results of [1, 2] and [8], this proves that the temperature parameter has almost no impact (as long as $T = \mathcal{O}(1)$) on the performance of the AEW under this condition, with a residual term of the order of $((T+1)\log M)/n$ for every $T > 0$.

Theorem A. *There exist absolute constants $c_0, \dots, c_5$ for which the following holds. For any integer $n \ge c_0$, there are random variables (X, Y) and a dictionary $F = \{f_1, f_2\}$ such that $(Y - f_i(X))^2 \le 1$ almost surely for $i = 1, 2$, for which the quadratic risk of the AEW satisfies the following*:

1. *if $T \le c_1$ and n is odd, then*

$$\mathbb{E}R(\tilde{f}^{\mathrm{AEW}}) \ge \min_{f \in F} R(f) + \frac{c_2}{\sqrt{n}};$$

2. *if $T \le c_3\sqrt{n}/\log n$, then, with probability greater than c_4,*

$$R(\tilde{f}^{\mathrm{AEW}}) \ge \min_{f \in F} R(f) + \frac{c_5}{\sqrt{n}}.$$

Theorem A proves that AEW is suboptimal in expectation in the low-temperature regime and suboptimal in probability in both the low- and high-temperature regimes, since it is possible to construct procedures that achieve the rate C/n with high probability [3,16] and in expectation [3,4,7,13,30,32] in the same setup as for Theorem A. It should be noted that the problem of the optimality in probability of the progressive mixture rule (and other related procedures) was studied by [3], who proved that, for a loss function Q satisfying some convexity and regularity assumption (e.g., the quadratic loss used in Theorem A), the progressive mixture rule $\bar{f}$ defined in (1.4) satisfies that for any temperature parameter, with probability greater than an absolute constant $c_0 > 0$, $R(\bar{f}) \ge \min_{f \in F} R(f) + c_1 n^{-1/2}$.

In addition, it is important to observe that the suboptimality in probability does not imply suboptimality in expectation for the aggregation problem, or vice versa. This property of the aggregation problem was first noted by [3], who found the progressive mixture rule (and other related aggregation procedures) to be suboptimal in probability for dictionaries of cardinality two but, on the other hand, to be optimal in expectation ([7,30,32] and [13]). This peculiar property of

the problem of aggregation comes from the fact that an aggregate $\widehat{f}$ is not restricted to the set F, which allows $R(\widehat{f}) - \min_{f\in F} R(f)$ to take negative values. [3] showed that for the progressive mixture rule $\bar{f}$, these negative values do compensate on average for larger values, but there is still an event of constant probability on which $R(\bar{f}) - \min_{f\in F} R(f)$ takes values greater than $C/\sqrt{n}$.

The proof of Theorem A shows that a dictionary consisting of two functions is sufficient to yield a lower bound in expectation in the low-temperature regime and in probability in both the small temperature regime, $0 \le T \le c_1$, and the large temperature regime, $c_1 \le T \le c_3\sqrt{n}/\log n$. In the following theorem, we study the behavior of AEW for larger dictionaries. To the best of our knowledge, negative results on the behavior of exponential weights based aggregation procedures are not known for dictionaries with more than two functions, and we show that the behavior of the AEW deteriorates in some sense as the cardinality of the dictionary increases.

Theorem B. *There exist an integer n_0 and absolute constants c_1 and c_2 for which the following holds. For every $n \ge n_0$, there are random variables (X, Y) and a dictionary $F = \{f_1, \dots, f_M\}$ of cardinality, $M = \lceil c_1\sqrt{n\log n}\rceil$, for which the quadratic loss function of any element in F is bounded by 2 almost surely, and for every $0 < \alpha \le 1/2$, if $T \le c_2\alpha$, then with probability at least $1 - c_3(\alpha)n^{\alpha-1/2}$,*

$$R(\tilde{f}^{\mathrm{AEW}}) \ge \min_{f\in F} R(f) + c_4(\alpha)\sqrt{\frac{\log M}{n}}.$$

Moreover, if $f_F^* \in F$ denotes the optimal function in F with respect to the quadratic loss (the oracle), then there exists $f_j \ne f_F^*$ with an excess risk greater than $c_5(\alpha)n^{-1/2}$ and for which the weight of f_j in the AEW procedure satisfies $\widehat{\theta}_j \ge 1 - n^{-c_6(\alpha)/T}$.

Theorem B implies that the AEW procedure might cause the weights to concentrate around a "bad" element in the dictionary (i.e., an element whose risk is larger than the best in the class by at least $\sim n^{-1/2}$) with high probability. In particular, Theorem B provides additional evidence that the AEW procedure is suboptimal for low temperatures.

The analysis of the behavior of AEW for a dictionary of cardinality larger than two is considerably harder than in the two-function case and requires some results on rearrangement of independent random variables that are almost Gaussian (see Proposition 5.2 below). Fortunately, not all is lost as far as optimality results for AEW go. Indeed, we show that under some geometric condition, AEW can be optimal and in fact can even adapt to the "real complexity" of the dictionary.

Intuitively, a good aggregation scheme should be able to ignore the elements in the dictionary whose risk is far from the optimal risk in F, or at least the impact of such elements on the function produced by the aggregation procedure should be small. Thus, a good procedure is one with a residual term of the order of ψ/n, where ψ is a complexity measure that is determined only by the richness of the set of "almost minimizers" in the dictionary. This leads to the following question:

Question 1.2. *Is it possible to construct an aggregation procedure that adapts to the real complexity of the dictionary?*

This question was first addressed by the PAC-Bayesian approach. [1,2] and [8] showed that in the high-temperature regime, AEW satisfies the requirements of Question 1.2, assuming that the class has a geometric property, called the Bernstein condition.

Definition 1.3 ([5]). *We say that a function class F is a (β, B)-Bernstein class* ($0 < \beta \leq 1$ *and* $B \geq 1$) *with respect to Z if every $f \in F$ satisfies $\mathbb{E}f \geq 0$ and*

$$\mathbb{E}(f^2(Z)) \leq B(\mathbb{E}f(Z))^{\beta}. \tag{1.6}$$

There are many natural situations in which the Bernstein condition is satisfied. For instance, when Q is the quadratic loss function and the regression function is assumed to belong to F, the excess loss function class $\mathcal{L}_F = \{Q(\cdot, f) - Q(\cdot, f_F^*)\colon\ f \in F\}$ satisfies the Bernstein condition with $\beta = 1$, where $f_F^* \in F$ is the minimizer of the risk in the class F. Another generic example is when the target function Y is far from the set of targets with "multiple minimizers" in F and $\mathcal{L}_F$ satisfies the Bernstein condition with $\beta = 1$. (See [21,22] for an exact formulation of this statement and related results.)

The Bernstein condition is very natural in the context of ERM because it has two consequences: that the empirical excess risk has better concentration properties around the excess risk, and that the complexity of the subset of F consisting of almost minimizers is smaller under this assumption. Consequently, if the class $\mathcal{L}_F$ is a (β, B)-Bernstein class for $0 < \beta \leq 1$, then the ERM algorithm can achieve fast rates (see, e.g., [5] and references therein). As the results below show, the same is true for AEW. Indeed, under a Bernstein assumption, [1,2] and [8] proved that if $R(\cdot)$ is a convex risk function and if F is such that $|Q(Z, f)| \leq b$ almost surely for any $f \in F$, then for every $T \geq c_1 \max\{b, B\}$ and $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$R(\tilde{f}^{\mathrm{AEW}}) \leq \min_{f \in F} R(f) + \frac{Tc_2}{n}\left(x + \log\left(\sum_{f \in F} \exp\bigl(-(n/2T)\bigl(R(f) - R(f_F^*)\bigr)\bigr)\right)\right). \tag{1.7}$$

Although the PAC-Bayesian approach cannot be used to obtain (1.7) in the low-temperature regime ($T \leq c_1 \max\{b, B\}$), such a result is not surprising. Indeed, because fast error rates for the ERM are expected when the underlying excess loss functions class satisfies the Bernstein condition, and because AEW converges to the ERM when the temperature T tends to 0, it is likely that for "small values" of T, AEW inherits some of the properties of ERM, such as fast rates under a Bernstein condition. We show this in Theorem C, proving that AEW answers Question 1.2 for low temperatures under the Bernstein condition.

Before formulating Theorem C, we introduce the following measure of complexity. For every $r > 0$, let

$$\begin{aligned}\psi(r) = {} & \log\bigl(|\{f \in F\colon\ R(f) - R(f_F^*) \leq r\}| + 1\bigr) \\ & + \sum_{j=1}^{\infty} 2^{-j} \log\bigl(|\{f \in F\colon\ 2^{j-1}r < R(f) - R(f_F^*) \leq 2^j r\}| + 1\bigr),\end{aligned}$$

where $|A|$ denotes the cardinality of the set A.

Observe that $\psi(r)$ is a weighted sum of the number of elements in F that assigns smaller and smaller weights to functions with a relatively large excess risk.

Theorem C. *There exist absolute constants c_0, c_1, c_2, and c_3 for which the following holds. Let F be a class of functions bounded by b such that the excess loss class $\mathcal{L}_F$ is a $(1, B)$-Bernstein class with respect to Z. If the risk function $R(\cdot)$ is convex and if $T \le c_0 \max\{b, B\}$, then for every $x > 0$, with probability at least $1 - 2\exp(-x)$, the function $\tilde{f}^{\mathrm{AEW}}$ produced by the AEW algorithm satisfies*

$$R(\tilde{f}^{\mathrm{AEW}}) \le R(f_F^*) + c_1(b + B)\frac{x + \psi(\theta)}{n},$$

where $\theta = c_2(b + B)(\log|F|)/n$.

In particular,

$$\mathbb{E}R(\tilde{f}^{\mathrm{AEW}}) \le R(f_F^*) + c_3(b + B)\frac{\psi(\theta)}{n}.$$

In other words, the scaling factor θ that we use is proportional to $(b + B)(\log|F|)/n$, and if the class is regular (in the sense that the complexity of F is well spread and not concentrated just around one point), then $\psi(\theta)$ is roughly the cardinality of the elements in F with risk at most $\sim(b + B)(\log|F|)/n$.

Observe that for every $r > 0$, $\psi(r) \le c \log|F|$ for a suitable absolute constant c. Thus, if T is reasonably small (below a level proportional to $\max\{B, b\}$), then the resulting aggregation rate is the optimal one, proportional to $(b + B)(x + \log M)/n$ with probability $1 - 2\exp(-x)$, and proportional to $(b + B)(\log M)/n$ in expectation. Thus, Theorem C indeed gives a positive answer to Question 1.2 in the presence of a Bernstein condition and for low temperatures.

Although the residual terms in Theorem C and in (1.7) are not the same, they are comparable. Indeed, the contribution of each element in F in the residual term depends exponentially on its excess risk.

Theorem C together with the results for high temperatures from [1,2] and [8] show that the AEW is an optimal aggregation procedure under the Bernstein condition as long as $T = \mathcal{O}(1)$ when M and n tend to infinity. In general, the residual term obtained is on the order of $((T + 1)\log M)/n$, and it can be proven that the optimal rate of aggregation under the Bernstein condition is proportional to $(\log M)/n$ using the classical tools in [28].

Finally, a word about the organization of the article. In the next section we present some comments about our results. The proofs of the three theorems follow in the subsequent sections. Throughout, we denote absolute constants or constants that depend on other parameters by c_1, c_2, etc. (Of course, we specify when a constant is absolute and when it depends on other parameters.) The values of constants may change from line to line. We write $a \sim b$ if there are absolute constants c and C such that $bc \le a \le Cb$, and write $a \lesssim b$ if $a \le Cb$.

2. Comments

Although from a theoretical standpoint, whether AEW is an optimal procedure in expectation and for high temperatures in the regression model with random design remains to be seen, from

a practical standpoint, we believe that exponential aggregating schemes simply should not be used in the setup of this article, because of the following reasons (see also the comments in [3]):

1. For any temperature $T \leq c_0\sqrt{n}/\log n$, there is an event of constant probability on which AEW performs poorly (this is the second part of Theorem A).
2. If the temperature parameter is chosen to be too small, then the AEW can perform poorly even in expectation (the first part of Theorem A).

Another consequence of the lower bounds stated in Theorem A is that AEW cannot be an optimal aggregation procedure both in expectation and in probability at low temperatures for two other aggregation problems: the problem of *convex aggregation*, in which one wants to mimic the best element in the convex hull of F, and the problem of *linear aggregation*, where one wishes to mimic the best linear combination of elements in F. Indeed, clearly

$$\min_{f\in F} R(f) \geq \min_{f\in \mathrm{conv}(F)} R(f) \geq \min_{f\in \mathrm{span}(F)} R(f).$$

Moreover, the optimal rates of aggregation for the convex and linear aggregation problems for dictionaries of cardinality two are of the order of n^{-1} (see [14,17,26]), whereas the residual terms obtained in Theorem A are on the order of $n^{-1/2}$ for such a dictionary. Thus AEW is suboptimal for these two other aggregation problems in the low-temperature regime.

We end this section by comparing two seemingly related assumptions, the margin assumption of [27] and the Bernstein condition of [5]. Note that in the proof of Theorem C, we have restricted ourselves to the case $\beta = 1$ simply to make the presentation as simple as possible. A very similar result, with the residual term $((x + \psi(\theta))/n)^{1/(2-\beta)}$ for the exact oracle inequality in probability and $(\psi(\theta)/n)^{1/(2-\beta)}$ for the exact oracle inequality in expectation, holds if one assumes a Bernstein condition for any $0 < \beta < 1$, and the proof is identical to that in the case where $\beta = 1$. This makes the discussion about β-Bernstein classes relevant here.

Recall the definition of the margin assumption:

Definition 2.1 ([27]). *We say that F has margin with parameters (β, B) $(0 < \beta \leq 1$ and $B \geq 1)$ if for every $f \in F$,*

$$\mathbb{E}\big(\big(Q(Z, f) - Q(Z, f^*)\big)^2\big) \leq B\big(R(f) - R(f^*)\big)^\beta,$$

where f^* is defined such that $R(f^*) = \min_f R(f)$, and the minimum is taken with respect to all measurable functions f on the given probability space.

Although the margin condition appears similar to the Bernstein condition, they are in fact very different, and have been introduced in the context of different types of problems. In the first of these, the "classical" statistical setup, one is given a function class F (the *model*) with an upper bound on its complexity and an unknown target function f^*, the minimizer of the risk over *all* measurable functions. One usually assumes that f^* belongs to F, and the aim is to construct an estimator $\widehat{f} = \widehat{f}(\cdot, \mathcal{D})$ for which the risk $R(\widehat{f})$ tends to 0 quickly as the sample size tends to infinity. In this setup, the margin assumption can improve this rate of convergence because of a better concentration of empirical means of $Q(\cdot, f) - Q(\cdot, f^*)$ around its mean [27]. The margin

assumption (MA) for $\beta = 1$ compares the performance of each $f \in F$ with the *best possible measurable function*, but it has nothing to do with the geometric structure of F. The margin is determined for every f separately, because f^* does not depend on the choice of F.

In the second type of problem, the "learning theory" setup, one does not assume that the target function f^* belongs to F. The aim is to construct a function $\widehat{f}$ with a risk as close as possible to that of the best element $f_F^* \in F$. Assuming that the excess loss class $\mathcal{L}_F$ satisfies the Bernstein condition (BC), the error rate can be improved (see, e.g., [5,22]).

At a first glance, MA and BC (for $\beta = 1$) share very strong similarities. Indeed, saying that $\mathcal{L}_F$ is a $(1, B)$-Bernstein class means that for every $f \in F$,

$$\mathbb{E}\big(\big(Q(Z, f) - Q(Z, f_F^*)\big)^2\big) \leq B\big(R(f) - R(f_F^*)\big),$$

but nevertheless they are different. Indeed, as mentioned earlier, MA is only a matter of concentration (and classical statistics questions are mostly a question of the trade-off between concentration and complexity). On the other hand, BC involves a lot of geometry of the function class F, because f_F^* might change significantly by adding a single function to F or by removing a function. In fact, the difficulty of learning theory problems is determined by the trade-off between concentration and complexity, *and* the geometry of the given class, since one measures the performance of the learning algorithm relative to the best *in the class*. Assuming that $f^* \in F$, as is usually done in classical statistics, exempts one from the need to consider the geometry of F, but one does not have that freedom in the aggregation framework. Indeed, since in the AEW algorithm the estimator is determined by the empirical means $R_n(f) - R_n(f_F^*)$, this is a learning problem rather than a problem in classical statistics, despite the fact that it has been used in statistical frameworks to construct adaptive estimators (see, e.g., [2,4,6,11,15,20,25,27,31]). Therefore, given their nature, aggregation procedures like the AEW are more natural under a BC assumption than under the MA. (A by-product of Theorem A is that the MA cannot improve the performance of AEW since in the setup of Theorem A, it is easy to check that MA is satisfied with the best possible margin parameter $\beta = 1$.)

3. Preliminary results on Gaussian approximation

Our starting point is the Berry–Esséen theorem on Gaussian approximation. Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d., mean-0 random variables with variance 1, set g to be a standard Gaussian variable, and write

$$\bar{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i.$$

Theorem 3.1 ([23]). *There exists an absolute constant $A > 0$ such that for every integer n,*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}[\bar{X}_n \leq x] - \mathbb{P}[g \leq x]| \leq \frac{A\mathbb{E}|W_1|^3}{\sqrt{n}}.$$

From here on, we let A denote the constant appearing in Theorem 3.1.

When the tail behavior of the W_i has a subexponential decay, the Gaussian approximation can be improved. Indeed, recall that a real-valued random variable W belongs to L_{ψ_α} for some $\alpha \geq 1$ if there exists $0 < c < \infty$ such that

$$\mathbb{E}\exp(|W|^\alpha / c^\alpha) \leq 2. \tag{3.1}$$

The infimum over all constants c for which (3.1) holds defines an Orlicz norm, which is called the ψ_α norm and is denoted by $\|\cdot\|_{\psi_\alpha}$. (For more information on Orlicz norms, see, e.g., [29] and [24].)

Proposition 3.2 (Chapter 5 in [23]). *For every $L > 0$, there exist constants B_0, c_1, and c_2 that depend only on L for which the following holds. If $\|W\|_{\psi_1} \leq L$, then for any $x \geq 0$, such that $x \leq B_0 n^{1/6}$,*

$$\mathbb{P}[\bar{X}_n \geq x] = \mathbb{P}[g \geq x]\exp\left(\frac{x^3\mathbb{E}W^3}{6\sqrt{n}}\right)\left[1 + \mathrm{O}\left(\frac{x+1}{\sqrt{n}}\right)\right]$$

and

$$\mathbb{P}[\bar{X}_n \leq -x] = \mathbb{P}[g \leq -x]\exp\left(-\frac{x^3\mathbb{E}W^3}{6\sqrt{n}}\right)\left[1 + \mathrm{O}\left(\frac{x+1}{\sqrt{n}}\right)\right],$$

where by $v = \mathrm{O}(u)$ we mean that $-c_1 u \leq v \leq c_1 u$.

In particular, if $|x| \leq B_0 n^{1/6}$ and $\mathbb{E}W^3 = 0$, then

$$|\mathbb{P}[\bar{X}_n \leq x] - \mathbb{P}[g \leq x]| \leq c_2\big(n^{-1/2}\exp(-x^2/2)\big).$$

From here on, we let B_0 denote the constant appearing in Proposition 3.2.

4. Proof of Theorem A

Before presenting the proof of Theorem A, we introduce the following notation. Given a probability measure ν and $(Z_i)_{i=1}^n$ selected independently according to ν, we set $P_n = n^{-1}\sum_{i=1}^n \delta_{Z_i}$ the empirical measure supported on $(Z_i)_{i=1}^n$. We let P denote the expectation $\mathbb{E}_\nu$. We assume that $T \leq 1$ and recall that n is an odd integer.

Let $Y = 0$ and define X by $\mathbb{P}[X = 1] = 1/2 - n^{-1/2}$ and $\mathbb{P}[X = -1] = 1/2 + n^{-1/2}$. Let $f_1 = \mathbb{1}_{[0,1]}$ and $f_2 = \mathbb{1}_{[-1,0]}$, and consider the dictionary $F = \{f_1, f_2\}$. It is easy to verify that the best function in F (the oracle) with respect to the quadratic risk is f_1, and that the excess loss function of f_2, $\mathcal{L}_2 = f_2^2 - f_1^2 = f_2 - f_1$, satisfies that

$$\mathcal{L}_2(X) = -X, \qquad \mathbb{E}\mathcal{L}_2(X) = 2n^{-1/2} \quad \text{and} \quad \sigma^2 = \mathbb{E}(\mathcal{L}_2(X) - \mathbb{E}\mathcal{L}_2(X))^2 = 1 - 4/n.$$

To simplify notation, set $P\mathcal{L}_2 = \mathbb{E}\mathcal{L}_2(X)$ and $P_n\mathcal{L}_2 = n^{-1}\sum_{i=1}^n \mathcal{L}_2(X_i)$.

An important parameter that lies at the heart of this counterexample is the Bernstein constant (which is very bad in this case),

$$\alpha = \frac{\mathbb{E}(f_1 - f_2)^2}{P\mathcal{L}_2} = \frac{\sqrt{n}}{2}. \tag{4.1}$$

Straightforward computation shows that AEW on F with temperature T is given by

$$\tilde{f}^{\mathrm{AEW}} = \widehat{\theta}_1 f_1 + (1 - \widehat{\theta}_1) f_2, \qquad \widehat{\theta}_1 = \frac{1}{1 + \exp(-(n/T) P_n \mathcal{L}_2)},$$

and that for $h(\theta) = \theta + \alpha\theta(1-\theta)$ defined for all $\theta \in [0,1]$,

$$\begin{aligned}
\mathbb{E}[R(\tilde{f}^{\mathrm{AEW}}) - R(f_1)] &= \mathbb{E}[1 - \widehat{\theta}_1 - \alpha\widehat{\theta}_1(1-\widehat{\theta}_1)] P\mathcal{L}_2 = \mathbb{E}[1 - h(\widehat{\theta}_1)] P\mathcal{L}_2 \\
&= \left[1 - \int_0^\infty h'(t)\mathbb{P}[\widehat{\theta}_1 \geq t]\,\mathrm{d}t\right] P\mathcal{L}_2 \\
&= \left[1 + \int_0^1 \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[\widehat{\theta}_1 \geq t]\,\mathrm{d}t\right] P\mathcal{L}_2 \\
&= \left[1 + \int_0^1 \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \geq \gamma(t)]\,\mathrm{d}t\right] P\mathcal{L}_2,
\end{aligned} \tag{4.2}$$

where $\gamma(t)$ is an increasing function defined for any $t \in (0,1)$ by

$$\gamma(t) = \frac{T}{n} \log\left(\frac{t}{1-t}\right).$$

In particular,

$$\mathbb{E}[R(\tilde{f}^{\mathrm{AEW}}) - R(f_1)] = [I_1 + I_2] P\mathcal{L}_2$$

for

$$I_1 = \int_0^{\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \geq \gamma(t)]\,\mathrm{d}t + 1$$

and

$$I_2 = \int_{\alpha^{-1}}^1 \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \geq \gamma(t)]\,\mathrm{d}t.$$

First, we bound I_1 from below. To that end, we note the following facts. First, for every $0 \leq t \leq \alpha^{-1}$, $1 + \alpha - 2\alpha t \geq 0$ and

$$\int_0^{\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\,\mathrm{d}t = -1.$$

Second, if we set $E = \exp(nP\mathcal{L}_2/T)$, then for $T \lesssim \sqrt{n}/\log n$, $0 < (1+E)^{-1} \leq \alpha^{-1}$. In particular, this holds under our assumption that $T \leq 1$. Moreover, because γ is increasing, for

$(1+E)^{-1} \le t \le \alpha^{-1}$, $\gamma(t) \ge \gamma((1+E)^{-1}) = -P\mathcal{L}_2$. Therefore,

$$\begin{aligned}
I_1 &= \int_0^{\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \ge \gamma(t)]\,\mathrm{d}t + 1 \\
&= \int_0^{\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\big(\mathbb{P}[P_n\mathcal{L}_2 \ge \gamma(t)] - 1\big)\,\mathrm{d}t \\
&\ge \int_{(1+E)^{-1}}^{\alpha^{-1}} (1+\alpha-2\alpha t)\mathbb{P}[P_n\mathcal{L}_2 < \gamma(t)]\,\mathrm{d}t \\
&\ge \int_{(1+E)^{-1}}^{\alpha^{-1}} (1+\alpha-2\alpha t)\,\mathrm{d}t \cdot \mathbb{P}\big[\big(\sqrt{n}/\sigma\big)(P_n\mathcal{L}_2 - P\mathcal{L}_2) < \big(\sqrt{n}/\sigma\big)(-2P\mathcal{L}_2)\big] \\
&\ge \int_{(1+E)^{-1}}^{\alpha^{-1}} (1+\alpha-2\alpha t)\,\mathrm{d}t\big(\mathbb{P}[g \le -8] - A/\sqrt{n}\big) \ge c_0 > 0,
\end{aligned}$$

where in the last step we used the Berry–Esséen theorem, with $|\mathcal{L}_2| \le 1$ and $n \ge 8 \vee (2A/\mathbb{P}[g \le -8])^2$, implying that $0 < c_0 < 1/2$.

We turn to a lower bound for I_2. Applying a change of variables $t \mapsto 1+\alpha^{-1}-u$ in the second term of I_2, it is evident that

$$\begin{aligned}
I_2 &= \int_{\alpha^{-1}}^{(\alpha+1)/(2\alpha)} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \ge \gamma(t)]\,\mathrm{d}t \\
&\quad + \int_{(\alpha+1)/(2\alpha)}^{1} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[P_n\mathcal{L}_2 \ge \gamma(t)]\,\mathrm{d}t \\
&= \int_{\alpha^{-1}}^{(\alpha+1)/(2\alpha)} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[\gamma(t) \le P_n\mathcal{L}_2 < \gamma(1+\alpha^{-1}-t)]\,\mathrm{d}t = I_3 + I_4
\end{aligned}$$

for

$$I_3 = \int_{\alpha^{-1}}^{(1+c_0/4)\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[\gamma(t) \le P_n\mathcal{L}_2 < \gamma(1+\alpha^{-1}-t)]\,\mathrm{d}t$$

and

$$I_4 = \int_{(1+c_0/4)\alpha^{-1}}^{(\alpha+1)/(2\alpha)} \big(2\alpha t - (1+\alpha)\big)\mathbb{P}[\gamma(t) \le P_n\mathcal{L}_2 < \gamma(1+\alpha^{-1}-t)]\,\mathrm{d}t.$$

To estimate I_3, note that $2\alpha t - (1+\alpha) \le 0$ for $t \in [\alpha^{-1}, (\alpha+1)/(2\alpha)]$, and thus

$$I_3 \ge \int_{\alpha^{-1}}^{(1+c_0/4)\alpha^{-1}} \big(2\alpha t - (1+\alpha)\big)\,\mathrm{d}t \ge \frac{-c_0}{4}\left(1+\frac{1}{\alpha}\right) \ge -\frac{c_0}{3}$$

for our choice of α.

The final step of the proof is to bound I_4 and in particular to show that for small values of T, $I_4 \geq -c_0/3$.

For any $0 < t \leq (\alpha+1)/(2\alpha)$, consider the intervals $I_T(t) = [n\gamma(t), n\gamma(1+\alpha^{-1}-t))$, and set $N_T(t) = |\{I_T(t) \cap \mathbb{Z}\}|$, which is the number of integers in $I_T(t)$. Because $\mathcal{L}_2(X) = -X$,

$$\mathbb{P}[\gamma(t) \leq P_n\mathcal{L}_2 < \gamma(1+\alpha^{-1}-t)] = \mathbb{P}\left[\sum_{i=1}^n -X_i \in I_T(t)\right] = \mathbb{P}_T(t).$$

Recall that $X \in \{-1,1\}$, and thus $\mathbb{P}[\sum_i -X_i \in I_T(t)] = \mathbb{P}[\sum_i -X_i \in I_T(t) \cap \mathbb{Z}]$. Because $n\gamma(t)$ is increasing and non-negative for $t > 1/2$, then if $1/2 < t \leq (\alpha+1)/(2\alpha)$, it follows that $0 < n\gamma(t) < n\gamma(1+1/\alpha - t) < 1$, provided that $T \leq 1$. Thus, for such values of t, $N_T(t) = 0$, implying that $\mathbb{P}_T(t) = 0$. On the other hand, if $t \leq 1/2$, then $\{0\} \subset I_T(t) \cap \mathbb{Z}$. In particular, if $N_T(t) = 1$, then $I_T(t) \cap \mathbb{Z} = \{0\}$, and since n is odd, then $\mathbb{P}_T(t) = \mathbb{P}[\sum_{i=1}^n -X_i = 0] = 0$. Otherwise, $N_T(t) \geq 2$, which implies that $N_T(t) \leq 2\Delta_T(t)$, where $\Delta_T(t)$ is the length of $I_T(t)$, given by

$$\Delta_T(t) = n\big(\gamma(1+\alpha^{-1}-t) - \gamma(t)\big) = T\log\left(\frac{(1-t)(\alpha+1-\alpha t)}{t(\alpha t-1)}\right).$$

Therefore, for every t in our range,

$$\mathbb{P}_T(t) \leq N_T(t) \max_{k\in I_T(t)} \mathbb{P}\left[\sum_{i=1}^n -X_i = k\right] \leq 2\Delta_T(t) \max_{k\in\mathbb{Z}} \mathbb{P}\left[\sum_{i=1}^n X_i = k\right].$$

Since $2\alpha t - (1+\alpha) \leq 0$ for every $0 < t \leq (\alpha+1)/(2\alpha)$, it is evident that

$$I_4 \geq 2T \max_{k\in\mathbb{Z}} \mathbb{P}\left[\sum_{i=1}^n X_i = k\right] \cdot \int_{(1+c_0/4)\alpha^{-1}}^{(\alpha+1)/(2\alpha)} \big(2\alpha t - (1+\alpha)\big) \log\left(\frac{(1-t)(\alpha+1-\alpha t)}{t(\alpha t-1)}\right) \mathrm{d}t.$$

It can be shown that $\max_{k\in\mathbb{Z}} \mathbb{P}[\sum_{i=1}^n X_i = k]$ is on the order of $n^{-1/2}$ either by a direct computation or by the Berry–Esséen theorem. Moreover, for any $(1+c_0/4)\alpha^{-1} \leq t \leq (\alpha+1)/(2\alpha)$, one has $\alpha t - 1 \geq c_0(4+c_0)^{-1}\alpha t$, and thus,

$$\log\left(\frac{(1-t)(\alpha+1-\alpha t)}{t(\alpha t-1)}\right) \leq \log\left(\frac{2(4+c_0)}{c_0 t^2}\right).$$

Therefore, combining the two observations with a change of variables $u = Ct$ for $C = (c_0/(2(4+c_0)))^{1/2}$, it is evident that there are absolute constants c_1, c_2 for which

$$I_4 \geq \frac{c_1 T}{\sqrt{n}} \int_{C(1+c_0/4)\alpha^{-1}}^{(C(\alpha+1)/(2\alpha))} (1+\alpha-2\alpha u/C)(\log u)\,\mathrm{d}u \geq -c_2\frac{T\alpha}{\sqrt{n}}.$$

Thus, there is an absolute constant c_3 such that if $T \leq c_3$, then $I_4 \geq -c_0/3$, implying that

$$\mathbb{E}[R(\tilde{f}^{\mathrm{AEW}}) - R(f_1)] \geq \frac{c_0}{3\sqrt{n}},$$

and proving the first part of Theorem A.

To prove the second part of the theorem, note that by the Berry–Esséen theorem, for every $x \in \mathbb{R}$, with probability greater than $\mathbb{P}[g \le x] - 2A/\sqrt{n}$,

$$\frac{\sqrt{n}}{\sigma(\mathcal{L}_2)}(P_n\mathcal{L}_2 - P\mathcal{L}_2) \le x.$$

Thus, if n is large enough to ensure that $\mathbb{P}[g \le -4] - 2A/\sqrt{n} \ge \mathbb{P}[g \le -4]/2 = c_4$, and taking $x = -4$, then with probability at least c_4, $P_n\mathcal{L}_2 \le -n^{-1/2}$. In that case, $\widehat{\theta}_1 \le \exp(-\sqrt{n}/T)$, which yields that

$$R(\tilde{f}^{\mathrm{AEW}}) - R(f_1) = \big(1 - \widehat{\theta}_1 - \alpha\widehat{\theta}_1(1-\widehat{\theta}_1)\big) \cdot P\mathcal{L}_2 \ge P\mathcal{L}_2/4 = n^{-1/2}/2,$$

provided that $T \lesssim \sqrt{n}/\log n$.

5. Proof of Theorem B

The first step in the proof of Theorem B involves a general statement regarding a monotone rearrangement of independent random variables that are close to being Gaussian. Let W be a mean 0, variance 1 random variable that is absolutely continuous with respect to the Lebesgue measure. Further assume that $|W|$ has a finite third moment (in fact, the random variables in which we are interested are bounded) and set $\beta(W) = A\mathbb{E}|W|^3$, where A is the constant appearing in the Berry–Esséen theorem (Theorem 3.1). Let $W_1, \dots, W_n$ be independent random variables distributed as W and set $\bar{X} = n^{-1/2}\sum_{i=1}^n W_i$. Let $(\bar{X}_j)_{j=1}^{\ell}$ be ℓ independent copies of $\bar{X}$, and put $\gamma_1 = \gamma_1(\ell) \in \mathbb{R}$ to satisfy that

$$\mathbb{P}\Big[\min_{1\le j\le \ell} \bar{X}_j \le \gamma_1(\ell)\Big] = 1 - \frac{1}{n}.$$

Note that such a γ_1 exists because W has a density with respect to the Lebesgue measure.

Throughout the proof of Theorem B, we require the following simple estimates on γ_1.

Lemma 5.1. *There exist absolute constants $c_0, \dots, c_3$ for which the following hold*:

1. *If $\ell \ge c_0 \log n$, then*
$$1 - \frac{\log n}{\ell} \le \mathbb{P}[\bar{X} > \gamma_1] \le 1 - c_1\frac{\log n}{\ell}.$$
2. *If ℓ and n are such that $(\beta(W)/\sqrt{n} + (\log n)/\ell) < \mathbb{P}[g < -2]$, then $\gamma_1 \le -2$.*
3. *If $\gamma_1 \le -2$ and $c_0 \log n \le \ell \le c_2\beta^{-1}(W)\sqrt{n}\log n$, then*
$$|\gamma_1| \sim \log^{1/2}\left(\frac{c_3\ell}{\log n}\right) \quad \text{and} \quad \exp(-\gamma_1^2/2) \sim \frac{\log n}{\ell}\log^{1/2}\left(\frac{c_3\ell}{\log n}\right).$$

Before we present the proof of Lemma 5.1, recall that for every $x \geq 2$,

$$\frac{3}{4\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x} \leq \mathbb{P}[g \geq x] \leq \frac{1}{\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x}. \tag{5.1}$$

Proof of Lemma 5.1. To prove the first part, note that by independence and because $\exp(-x) \geq 1 - x$,

$$\mathbb{P}[\bar{X} > \gamma_1] = \mathbb{P}\Big[\min_{1 \leq j \leq \ell} \bar{X}_j > \gamma_1\Big]^{1/\ell} = \left(\frac{1}{n}\right)^{1/\ell} \geq 1 - \frac{\log n}{\ell}. \tag{5.2}$$

The reverse inequality follows in an identical fashion, because $\exp(-x) \leq 1 - x/3$ if $0 \leq x \leq 1$.

Turning to the second part, if $\gamma_1 > -2$, then

$$1 - \frac{1}{n} = \mathbb{P}\Big[\min_{1 \leq j \leq \ell} \bar{X}_j \leq -\gamma_1\Big] \geq \mathbb{P}\Big[\min_{1 \leq j \leq \ell} \bar{X}_j \leq -2\Big] = 1 - (\mathbb{P}[\bar{X} > -2])^{\ell},$$

implying that $\mathbb{P}[\bar{X} \leq -2] \leq (\log n)/\ell$. On the other hand, by the Berry–Esséen theorem, $\mathbb{P}[\bar{X} \leq -2] \geq \mathbb{P}[g \leq -2] - \beta(W)/\sqrt{n}$, which is impossible under the assumptions of (2).

Finally, to prove (3), we use the Berry–Esséen theorem combined with the lower and upper estimates on the Gaussian tail (5.1) and (5.2). Thus,

$$\frac{3}{4\sqrt{2\pi}} \frac{1}{|\gamma_1|} \exp\left(-\frac{|\gamma_1|^2}{2}\right) \leq \mathbb{P}[g < \gamma_1] \leq \mathbb{P}[\bar{X} < \gamma_1] + \frac{\beta(W)}{\sqrt{n}} \leq \frac{\beta(W)}{\sqrt{n}} + c_1 \frac{\log n}{\ell},$$

and

$$\frac{1}{\sqrt{2\pi}} \frac{1}{|\gamma_1|} \exp\left(-\frac{|\gamma_1|^2}{2}\right) \geq \frac{\log n}{\ell} - \frac{\beta(W)}{\sqrt{n}},$$

from which both parts of the third claim follow. □

Proposition 5.2. *There exist constants c_1, c_2, c_3, and c_4 that depend only on $\|W\|_{\psi_2}$ for which the following holds. Let $2M^2 \exp(-c_1 n^{1/3}) < \delta \leq 1$, and assume that $\mathbb{E}W^3 = 0$ and that $\gamma_1 = \gamma_1(M-1) \leq -2$. Then*

$$\mathbb{P}[\exists j \in \{2, \ldots, M\}\colon \bar{X}_j \leq \gamma_1 \text{ and for every } k \in \{2, \ldots, M\} \setminus \{j\}, \bar{X}_k - \bar{X}_j \geq \delta]$$
$$\geq 1 - \frac{1}{n} - c_2\left(\frac{1}{\sqrt{n}} + \delta\right)(\log n)^2 \sqrt{\log M},$$

provided that $c_3 \log n \leq M \leq c_4 \sqrt{n}(\log n)$.

Proof. For every $2 \leq j \leq M$, let

$$\Omega_j = \big\{\bar{X}_j \leq \gamma_1 \text{ and } \bar{X}_k - \bar{X}_j \geq \delta \text{ for every } k \in \{2, \ldots, M\} \setminus \{j\}\big\}.$$

The events Ω_j for $2 \le j \le M$ are disjoint, and thus

$$\begin{aligned}&\mathbb{P}[\exists j \in \{2,\dots,M\}\colon\ \bar{X}_j \le \gamma_1 \text{ and } \bar{X}_k - \bar{X}_j \ge \delta \text{ for every } k \in \{2,\dots,M\}\setminus\{j\}]\\ &= \mathbb{P}\Bigg[\bigcup_{j=2}^{M} \Omega_j\Bigg] = (M-1)\mathbb{P}[\Omega_2].\end{aligned}$$

Since the variables $(\bar{X}_j)_{j=2}^M$ are independent, we have

$$\mathbb{P}[\Omega_2] = \int_{-\infty}^{\gamma_1} f_{\bar{X}}(z)\Bigg(\int_{z+\delta}^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)^{M-2}\mathrm{d}\mu(z),$$

where $f_{\bar{X}}$ is a density function of $\bar{X}$ with respect to the Lebesgue measure μ.

On the other hand, for any $z \le \gamma_1$, $\mathbb{P}[\bar{X} \ge z] > 0$ because of (5.2). Thus, for every $z \le \gamma_1$,

$$\int_{z+\delta}^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t) = \Bigg(1 - \frac{\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)}{\int_z^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)}\Bigg)\cdot\int_z^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t). \tag{5.3}$$

Note that for every $0 \le x \le 1$, $(1-x)^{M-2} \ge 1-(M-2)x$, and applied to (5.3),

$$\begin{aligned}\mathbb{P}[\Omega_2] &\ge \int_{-\infty}^{\gamma_1} f_{\bar{X}}(z)\Bigg(\int_z^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)^{M-2}\mathrm{d}\mu(z)\\ &\quad - (M-2)\int_{-\infty}^{\gamma_1} f_{\bar{X}}(z)\Bigg(\int_z^{\infty} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)^{M-3}\Bigg(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)\mathrm{d}\mu(z)\\ &\ge \mathbb{P}[\bar{X}_2 \le \gamma_1 \text{ and } \bar{X}_k \ge \bar{X}_2, \text{ for every} k \ge 3] - T_2\\ &= \frac{1}{M-1}\mathbb{P}\Big[\min_{2\le j\le M} \bar{X}_j \le \gamma_1\Big] - T_2,\end{aligned}$$

where

$$T_2 = (M-2)\int_{-\infty}^{\gamma_1} f_{\bar{X}}(z)\Bigg(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)\mathrm{d}\mu(z).$$

Recall the if (W_i) are independent mean-0 random variables and (a_i) are real numbers, then $\|\sum a_i W_i\|_{\psi_2} \le c(\sum a_i^2 \|W_i\|_{\psi_2}^2)^{1/2}$, where c is an absolute constant [29]. Thus, $\|\bar{X}\|_{\psi_2} \le c\|W\|_{\psi_2}$, and for any $t < 0$,

$$\int_{-\infty}^{t} f_{\bar{X}}(z)\Bigg(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\Bigg)\mathrm{d}\mu(z) \le \mathbb{P}[\bar{X} \le t] \le 2\exp(-t^2/c^2\|W\|_{\psi_2}^2).$$

Let $t_0 < 0$ be such that

$$2\exp(-t_0^2/c^2\|W\|_{\psi_2}^2) = \frac{\delta\sqrt{\log(M-1)}}{(M-1)(M-2)}.$$

Thus,

$$(M-2)\int_{-\infty}^{t_0} f_{\bar{X}}(z)\left(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\right)\mathrm{d}\mu(z) \le \frac{\delta\sqrt{\log(M-1)}}{M-1}.$$

Note that if $t_0 \ge \gamma_1$, then our claim follows. Indeed, because $\mathbb{P}[\min_{2\le j\le M} \bar{X}_j \le \gamma_1] = 1 - n^{-1}$, we have

$$\mathbb{P}[\Omega_2] \ge \frac{1}{M-1}\left(1-\frac{1}{n}\right) - \delta\frac{\sqrt{\log(M-1)}}{M-1}.$$

Otherwise, we split the interval $(-\infty, \gamma_1] = (-\infty, t_0) \cup [t_0, \gamma_1]$, and to upper bound T_2, it remains to control the integral on the second interval $[t_0, \gamma_1]$.

Recall that $W \in L_{\psi_1}$ and that $\mathbb{E}W^3 = 0$. Therefore, by Proposition 3.2, it is evident that if z and δ satisfy that $z \le z+\delta \le 0$ and $|z|, |z+\delta| \le B_0 n^{1/6}$, then

$$\begin{aligned}\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t) &= \mathbb{P}[z \le \bar{X} \le z+\delta] \\ &\le \mathbb{P}[z \le g \le z+\delta] + \frac{B_1}{\sqrt{n}}\exp(-z^2/2),\end{aligned} \tag{5.4}$$

where B_0 and B_1 are constants that depend only on $\|W\|_{\psi_1}$. In addition, for every $z \le 0$,

$$\mathbb{P}[z \le g \le z+\delta] \le \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\int_0^{\delta}\exp(-zt)\,\mathrm{d}t \le \frac{\delta}{\sqrt{2\pi}}\exp(-z^2/2). \tag{5.5}$$

If $2M^2\exp(-B_0^2 n^{1/3}/\|W\|_{\psi_2}^2) < \delta \le 1$, then $|t_0| \le B_0 n^{1/6}$. Combining (5.4) and (5.5) with the definition of T_2, we have

$$\begin{aligned}&(M-2)\int_{t_0}^{\gamma_1} f_{\bar{X}}(z)\left(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\right)\mathrm{d}\mu(z) \\ &\le (M-2)\left(\frac{B_1}{\sqrt{n}}+\frac{\delta}{\sqrt{2\pi}}\right)\int_{t_0}^{\gamma_1} f_{\bar{X}}(z)\exp(-z^2/2)\,\mathrm{d}\mu(z) \\ &\le (M-2)\left(\frac{B_1}{\sqrt{n}}+\frac{\delta}{\sqrt{2\pi}}\right)\exp(-\gamma_1^2/2)\mathbb{P}[\bar{X} \le \gamma_1] \\ &\le (M-2)\left(\frac{B_1}{\sqrt{n}}+\frac{\delta}{\sqrt{2\pi}}\right)\exp(-\gamma_1^2/2)\frac{\log n}{M-1},\end{aligned}$$

where the last inequality follows from (5.2). By Lemma 5.1, and since $M \lesssim \sqrt{n}\log n$,

$$\begin{aligned}&(M-2)\int_{t_0}^{\gamma_1} f_{\bar{X}}(z)\left(\int_z^{z+\delta} f_{\bar{X}}(t)\,\mathrm{d}\mu(t)\right)\mathrm{d}\mu(z) \\ &\le c\left(\frac{1}{\sqrt{n}}+\delta\right)\left(\frac{\log n}{M}\right)(\log n)\sqrt{\log M}\end{aligned}$$

for some constant $c = c(\beta)$, from which our claim follows. □

We next describe the construction needed for the proof of Theorem B. Let (X, Y) and $F = \{f_1, \dots, f_M\}$ be defined by

$$\begin{aligned} Y &= 0, \\ f_1(X) &= (12)^{1/4}\mathcal{U}_1, \\ f_j(X) &= (12)^{1/4}(\mathcal{U}_j + \lambda) \qquad \text{for every } 2 \le j \le M, \end{aligned}$$

where $\mathcal{U}_1, \dots, \mathcal{U}_M$ are M independent random variables with density $u \longmapsto 2(u+\lambda)\mathbb{1}_{[-\lambda,1-\lambda]}(u)$ for $0 < \lambda < 1/2$ to be fixed later. Note that for this choice of density function, $(\mathcal{U}_1 + \lambda)^2$ is uniformly distributed on $[0, 1]$, and the best element in F with respect to the quadratic risk is f_1.

Let $(\mathcal{U}_j^{(i)}\colon\ j = 1, \dots, M, i = 1, \dots, n)$ be a family of independent random variables distributed as $\mathcal{U}_1$. Thus, for every $1 \le i \le n$, $f_j(X_i) = (12)^{1/4}(\mathcal{U}_j^{(i)} + \lambda)$ for every $2 \le j \le M$ and $f_1(X_i) = (12)^{1/4}\mathcal{U}_1^{(i)}$. For every $1 \le j \le M$, set

$$\bar{R}_j = \sqrt{\frac{12}{n}}\left(\sum_{i=1}^{n}\big(\mathcal{U}_j^{(i)} + \lambda\big)^2 - \mathbb{E}\big(\mathcal{U}_j^{(i)} + \lambda\big)^2\right),$$

and observe that if $W = \sqrt{12}((\mathcal{U} + \lambda)^2 - \mathbb{E}(\mathcal{U} + \lambda)^2)$, then W is a mean 0, variance 1 random variable that is absolutely continuous with respect to the Lebesgue measure and $W \in L_{\psi_2}$ and satisfies that $\mathbb{E}W^3 = 0$. These properties allow us to apply Proposition 5.2 to the random variables $\bar{R}_1, \dots, \bar{R}_M$.

Let $0 < \rho < 1$ (to be named later), and set

$$\xi(\bar{R}_1) = \bar{R}_1 + \frac{T}{\sqrt{n}}\log\left[\frac{\rho}{2(1-\rho)}\right] - \sqrt{12}\lambda(2-\lambda)\sqrt{n},$$

and

$$\delta = \frac{-T}{\sqrt{n}}\log\left[\frac{\rho}{2(M-2)(1-\rho)}\right].$$

Consider the system of inequalities

$$\begin{cases} \bar{R}_j \le \xi(\bar{R}_1), \\ \bar{R}_k - \bar{R}_j \ge \delta \qquad \text{for every } k \ne 1, j, \end{cases} \tag{C_j}$$

and recall that for each $j = 1, \dots, M$ $\widehat{\theta}_j$ denotes the weight of f_j in the AEW procedure.

Proposition 5.3. *There exist absolute constants c_1 and c_2 for which the following holds. Let $0 < \rho < 1/2$ and $2 \le j \le M$. If the system (C_j) is satisfied, then*

$$\widehat{\theta}_j \ge 1 - \rho.$$

Moreover, if $\rho \le c_1\lambda$, then the quadratic risk of the function produced by the AEW procedure satisfies

$$R(\tilde{f}^{\mathrm{AEW}}) \ge \min_{f\in F} R(f) + c_2\lambda.$$

Proof. Let $2 \le j \le M$, and assume that (C_j) is satisfied. Recall that $R_n(f)$ is the empirical risk of f, and note that for any $k \in \{2,\dots,M\}\setminus\{j\}$,

$$\begin{aligned} R_n(f_k) - R_n(f_j) &= \frac{1}{n}\sum_{i=1}^n [f_k(X_i)^2 - f_j(X_i)^2] = \frac{\bar{R}_k - \bar{R}_j}{\sqrt{n}} \\ &\ge \frac{\delta}{\sqrt{n}} = \frac{-T}{n}\log\left[\frac{\rho}{2(M-2)(1-\rho)}\right]. \end{aligned} \tag{5.6}$$

In addition, since $\mathcal{U}_1^{(i)} \le 1-\lambda$ almost surely for any $1 \le i \le n$,

$$\begin{aligned} R_n(f_1) - R_n(f_j) &= \frac{1}{n}\sum_{i=1}^n [f_1(X_i)^2 - f_j(X_i)^2] \\ &= \frac{\bar{R}_1 - \bar{R}_j}{\sqrt{n}} - \sqrt{12}\left(\lambda^2 + \frac{2\lambda}{n}\sum_{i=1}^n \mathcal{U}_1^{(i)}\right) \\ &\ge \frac{\bar{R}_1 - \xi(\bar{R}_1)}{\sqrt{n}} - \sqrt{12}\lambda(2-\lambda) \ge \frac{-T}{n}\log\left[\frac{\rho}{2(1-\rho)}\right]. \end{aligned} \tag{5.7}$$

Combining (5.6) and (5.7), it is evident that

$$\begin{aligned} \widehat{\theta}_j &= \frac{1}{\sum_{k=1}^M \exp[(-n/T)(R_n(f_k) - R_n(f_j))]} \\ &\ge \frac{1}{1 + (M-2)\rho/(2(M-2)(1-\rho)) + \rho/(2(1-\rho))} = 1-\rho. \end{aligned}$$

Since the functions $f_1,\dots,f_M$ are independent in $L_2(X)$ and $\mathbb{E}f_j \ge 0$,

$$\begin{aligned} R(\tilde{f}^{\mathrm{AEW}}) &= \mathbb{E}\left(\sum_{j=1}^M \widehat{\theta}_j f_j(X)\right)^2 \\ &= (\widehat{\theta}_j)^2\mathbb{E}f_j^2 + \sum_{\ell\ne j}(\widehat{\theta}_\ell)^2\mathbb{E}f_\ell^2 + \sum_{\ell\ne j}\widehat{\theta}_j\widehat{\theta}_\ell\mathbb{E}f_j f_\ell \ge (\widehat{\theta}_j)^2\mathbb{E}f_j^2, \end{aligned}$$

and there is an absolute constant c_0 for which $\mathbb{E}f_j^2 \ge \mathbb{E}f_1^2 + c_0\lambda$. Thus,

$$(\widehat{\theta}_j)^2\mathbb{E}f_j^2 - \mathbb{E}f_1^2 \ge (1-\rho)(\mathbb{E}f_1^2 + c_0\lambda) - \mathbb{E}f_1^2 \ge c_2\lambda,$$

provided that $\rho \le c_1\lambda$, giving

$$R(\tilde{f}^{\mathrm{AEW}}) \ge \mathbb{E}f_1^2 + c_2\lambda = \min_{f\in F} R(f) + c_2\lambda,$$

as claimed. □

Next, we formulate a general statement, from which Theorem B follows immediately.

Theorem 5.4. *There exist absolute constants $c_i, i = 0,\dots,5$ and an integer n_0 for which the following holds. For any $n \ge n_0$, $1 \le \kappa \le c_0\sqrt{n\log n}$, $0 < T \le 1$, and $c_1 T/\sqrt{n\log n} < \varepsilon < 1/8$, let $M = \lceil c_2\sqrt{n\log n}\rceil$, $\lambda = c_3\varepsilon\sqrt{(\log n)/n}$, and $\rho = n^{-\varepsilon\kappa/T}$. Set F to be the class of functions defined above with those parameters. Then, with probability at least*

$$1 - c_4(\varepsilon\kappa + T + 1)\big((\log^3 n)/n\big)^{(1-2\varepsilon)^2/2},$$

there exists $j \ge 2$ such that

$$\widehat{\theta}_j \ge 1 - \frac{1}{n^{\varepsilon\kappa/T}}.$$

In particular, with the same probability and if $0 \le T < \min\{1, 2\varepsilon\kappa\}$,

$$R(\tilde{f}^{\mathrm{AEW}}) \ge \min_{f\in F} R(f) + c_5\varepsilon\sqrt{\frac{\log M}{n}}.$$

Proof. Set

$$\mathbb{P}_0 = \mathbb{P}[\exists j \in \{2,\dots,M\} \text{ such that } \widehat{\theta}_j \ge 1-\rho],$$

and, by Proposition 5.3,

$$\mathbb{P}_0 \ge \mathbb{P}[\exists j \in \{2,\dots,M\} \text{ for which } (C_j) \text{ is satisfied}] = \mathbb{P}_1.$$

Let $\gamma_1 = \gamma_1(M-1)$ be defined by $\mathbb{P}[\min_{2\le j\le M} \bar{R}_j \le \gamma_1] = 1 - n^{-1}$, and observe that γ_1 is well defined and satisfies all three parts of Lemma 5.1 for $\ell = M-1$. Set $\Omega_0 = \{\xi(\bar{R}_1) \ge \gamma_1\}$,

$$A = \big\{\exists j \in \{2,\dots,M\}\colon\ \bar{R}_j \le \xi(\bar{R}_1),\ \text{and } \bar{R}_k - \bar{R}_j \ge \delta \text{ for every } k \ne 1, j\big\}$$

and

$$B = \big\{\exists j \in \{2,\dots,M\}\colon\ \bar{R}_j \le \gamma_1 \text{ and } \bar{R}_k - \bar{R}_j \ge \delta \text{ for every } k \ne 1, j\big\}.$$

Since the functions $\bar{R}_j$, $j = 1,\dots,M$ are independent, we have

$$\mathbb{P}_1 \ge \mathbb{E}_{\bar{R}_1}[\mathbb{P}[A|\bar{R}_1]\mathbb{1}_{\Omega_0}] \ge \mathbb{P}[B]\mathbb{P}[\Omega_0].$$

Applying Proposition 5.2, we then have

$$\mathbb{P}[B] \ge 1 - \frac{1}{n} - c_2\left(\frac{1}{\sqrt{n}} + \delta\right)(\log n)^2\sqrt{\log M},$$

provided that $c_3 \log n \le M \le c_4 \sqrt{n}(\log n)$.

To lower bound $\mathbb{P}[\Omega_0]$, note that

$$\mathbb{P}[\Omega_0] = \mathbb{P}\left[\bar{R}_1 \ge \gamma_1 - \frac{T}{\sqrt{n}} \log\left(\frac{\rho}{2(1-\rho)}\right) + \sqrt{12}\lambda(2-\lambda)\sqrt{n}\right].$$

Fix $0 < \varepsilon < 1/8$ and assume that λ, ρ and T are such that

$$\sqrt{12}\lambda(2-\lambda)\sqrt{n} \le -\varepsilon\gamma_1 \quad \text{and} \quad -\frac{T}{\sqrt{n}} \log\left(\frac{\rho}{2(1-\rho)}\right) \le -\varepsilon\gamma_1. \tag{5.8}$$

By the Berry–Esséen theorem and (5.1),

$$\begin{aligned}
\mathbb{P}[\Omega_0] &\ge \mathbb{P}[\bar{R}_1 \ge (1-2\varepsilon)\gamma_1] = 1 - \mathbb{P}[\bar{R}_1 < (1-2\varepsilon)\gamma_1] \\
&\ge 1 - \mathbb{P}[g \le (1-2\varepsilon)\gamma_1] - \frac{2\beta(W)}{\sqrt{n}} \\
&\ge 1 - \frac{1}{\sqrt{2\pi}(1-2\varepsilon)|\gamma_1|} \exp\bigl(-(1-2\varepsilon)^2\gamma_1^2/2\bigr) - \frac{2A}{\sqrt{n}},
\end{aligned}$$

and by Lemma 5.1,

$$\exp\bigl(-(1-2\varepsilon)^2\gamma_1^2/2\bigr) \le c_5\left(\frac{\log n}{M-1}\log^{1/2}\left(\frac{c_5 M}{\log n}\right)\right)^{(1-2\varepsilon)^2}.$$

Therefore,

$$\mathbb{P}_0 \ge \left(1 - \frac{1}{n} - c_2\left(\frac{1}{\sqrt{n}} + \delta\right)(\log n)^2\sqrt{\log M}\right) \cdot \left(1 - c_5\left(\frac{\log^3 n}{M}\right)^{(1-2\varepsilon)^2}\right),$$

provided that $c_2 \log n \le M \le c_3\sqrt{n \log n}$.

To complete the proof, we need to chose λ and ρ for which (5.8) holds. By Lemma 5.1,

$$|\gamma_1| \gtrsim \log^{1/2}\left(\frac{M}{\log n}\right),$$

and thus (5.8) holds for λ and ρ for which

$$\lambda \le c_8\varepsilon\left[\frac{1}{n}\log\left(\frac{M}{\log n}\right)\right]^{1/2} \quad \text{and} \quad \rho \ge 2\exp\left[\frac{-c_9\varepsilon\sqrt{n}}{T}\log^{1/2}\left(\frac{M}{\log n}\right)\right].$$

In particular, when we take $M \sim \sqrt{n \log n}$, $\lambda \sim \varepsilon((\log M)/n)^{1/2}$, and $\rho = n^{-\varepsilon\kappa/T}$, ρ satisfies the required condition as long as $\varepsilon \gtrsim T/\sqrt{n \log n}$ and $\kappa \lesssim \sqrt{n/\log n}$, as assumed. Moreover,

$$\delta \lesssim (\varepsilon\kappa + T)\frac{\log n}{\sqrt{n}},$$

implying that

$$\mathbb{P}_0 \geq 1 - c_8(\varepsilon\kappa + T + 1)\left(\frac{\log^3 n}{n}\right)^{(1-2\varepsilon)^2/2}.$$

The lower bound on the risk of the AEW procedure now follows from Proposition 5.3. □

6. Proof of Theorem C

In this section we prove Theorem C, which we reformulate below. From here on, we assume that the dictionary F is finite, consisting of M functions, and that the functions are indexed according to their risk in an increasing order. Thus, $f_1 = f_F^*$. In addition, we denote $\mathcal{L}_f(\cdot) = Q(\cdot, f) - Q(\cdot, f_1)$, and thus $R(f) - R(f_1) = \mathbb{E}\mathcal{L}_f$.

For every $r > 0$, recall that

$$\psi(r) = \log(|\{f \in F\colon \mathbb{E}\mathcal{L}_f \leq r\}| + 1) + \sum_{j=1}^{\infty} 2^{-j} \log(|\{f \in F\colon 2^{j-1}r < \mathbb{E}\mathcal{L}_f \leq 2^j r\}| + 1),$$

which serves as a measure of complexity for the class F.

The first component needed in the proof of Theorem C is the level $\lambda(x)$ with the following property: with probability at least $1 - 2\exp(-x)$, $R_n(f_j) - R_n(f_1)$ is equivalent to $R(f_j) - R(f_1)$ if $R(f_j) - R(f_1) \geq \lambda(x)$. This "isomorphism" constant was introduced by [5]. To formulate the exact properties that we need, first recall the following definitions and notation.

If $G = \mathcal{L}_F$ is the excess loss functions class $\{\mathcal{L}_f\colon f \in F\}$, then let $\operatorname{star}(G, 0) = \{\theta g\colon 0 \leq \theta \leq 1, g \in G\}$ is the star-shaped hull of G and 0. Set $G_r = \operatorname{star}(G, 0) \cap \{g\colon \mathbb{E}g = r\}$, that is, the set of functions in the star-shaped hull of $\mathcal{L}_F$ and 0, with expectation r. Let

$$r^* = \inf\Big\{r\colon \mathbb{E}\sup_{g \in G_r} |P_n g - Pg| \leq r/2\Big\},$$

where, as always, P_n denotes the empirical mean and P is the mean according to the underlying probability measure of Z.

Theorem 6.1 ([5]). *There exists an absolute constant c for which the following holds. Let F be a class of functions bounded by b, such that $\mathcal{L}_F$ is a $(1, B)$-Bernstein class. For every $x > 0$ and an integer n, let*

$$\lambda(x) = c\max\Big\{r^*, (b + B)\frac{x}{n}\Big\}. \tag{6.1}$$

Then, with probability at least $1 - 2\exp(-x)$, for every $f \in F$ with $R(f) - R(f_F^*) \geq \lambda(x)$,

$$R_n(f) - R_n(f_F^*) \geq \frac{1}{2}\big(R(f) - R(f_F^*)\big).$$

Let $\rho = \kappa_1(B+b)/n$, where κ_1 is an absolute constant to be named later. Recall that functions in F are indexed according to their risk in an increasing order. Let $J_-(x) = \{j\colon R(f_j) - R(f_1) \le \lambda(x)\}$, and set $J_+(x)$ as its complement. Define the sets $J_{+,0} = \{j \in J_+(x)\colon R(f_j) - R(f_1) \le \rho\}$ and, for $k \ge 1$,

$$J_{+,k} = \{j \in J_+(x)\colon 2^{k-1}\rho < R(f_j) - R(f_1) \le 2^k\rho\}.$$

(Note that some of the sets $J_{+,k}$ may be empty.) Set

$$k_0 = \sup\{k \ge 0\colon 2^k \le \log(|J_{+,k}| + 1)\},$$

and let $I = J_- \cup \bigcup_{k \le k_0} J_{+,k}$.

From Theorem 6.1, it follows that for every $k \ge 0$ and every $j \in J_{+,k}$, $R_n(f_j) - R_n(f_F^*) \ge \frac{1}{2}(R(f_j) - R(f_F^*))$. This is because $R(f_j) - R(f_F^*) \ge \lambda(x)$ by the definition of $J_+(x)$, and $J_+(x) \supset J_{+,k}$.

The key factor in the proof of Theorem C is Theorem 6.2.

Theorem 6.2. *There exist absolute constants c_1 and c_2 for which the following holds. Let F be a class of functions bounded by b, such that $\mathcal{L}_F$ is a $(1, B)$-Bernstein class with respect to a convex risk function R. Then, with probability at least $1 - 2\exp(-x)$, if $\tilde{f}^{\mathrm{AEW}}$ is produced by the AEW algorithm and $T \le c_1(b+B)$, then*

$$R(\tilde{f}^{\mathrm{AEW}}) - R(f_F^*) \le c_2\left(\lambda(x) + (b+B)\frac{2^{k_0}}{n}\right), \tag{6.2}$$

where $\lambda(x)$ is as defined in (6.1).

Proof. Let $(\widehat{\theta}_j)_{j=1}^M$ be the weights of the AEW algorithm, and set $\tilde{f}^{\mathrm{AEW}} = \sum_{j=1}^M \widehat{\theta}_j f_j$ to be the aggregate function. Because R is a convex function,

$$R\left(\sum_{j=1}^M \widehat{\theta}_j f_j\right) - R(f_1) \le \sum_{j=1}^M \widehat{\theta}_j\big(R(f_j) - R(f_1)\big).$$

Note that for every $j \in I$, $R(f_j) - R(f_1) \le \lambda(x) + 2^{k_0}\rho = \lambda(x) + \kappa_1 2^{k_0}(b+B)/n$. In particular, because $\sum_{j=1}^M \widehat{\theta}_j = 1$,

$$\sum_{j \in I} \widehat{\theta}_j\big(R(f_j) - R(f_1)\big) \le \lambda(x) + \kappa_1 2^{k_0}(b+B)/n.$$

On the other hand, with probability at least $1 - 2\exp(-x)$, for every $k > k_0$ and every $j \in J_{+,k}$,

$$R_n(f_j) - R_n(f_1) \ge \big(R(f_j) - R(f_1)\big)/2.$$

Applying the definition of the weights in the AEW algorithm and given that $\widehat{\theta}_1 \le 1$,

$$\begin{aligned}\sum_{j\in I^c} \widehat{\theta}_j\big(R(f_j)-R(f_1)\big) &= \widehat{\theta}_1 \sum_{j\in I^c} \frac{\widehat{\theta}_j}{\widehat{\theta}_1}\big(R(f_j)-R(f_1)\big)\\ &\le \sum_{j\in I^c} \exp\bigg(-\frac{n}{T}\big(R_n(f_j)-R_n(f_1)\big)\bigg)\big(R(f_j)-R(f_1)\big)\\ &\le \sum_{k>k_0}\sum_{j\in J_{+,k}} \exp\bigg(-\frac{n}{2T}\big(R(f_j)-R(f_1)\big)\bigg)\big(R(f_j)-R(f_1)\big) = (\star).\end{aligned}$$

From the definition of k_0, it is evident that for every $k > k_0$, $2^k \ge \log|J_{+,k}|$, and thus if $T \le c_1 \max\{b, B\}$ and κ_1 is sufficiently large, then

$$(\star) \le \sum_{k>k_0} \exp\bigg(\log|J_{+,k}| - \frac{n}{2T}2^{k-1}\rho\bigg)2^k\rho \le \sum_{k>k_0} \exp\bigg(-c_2\frac{n}{T}2^k\rho\bigg)2^k\rho \le c_3\frac{T}{n}.$$

Indeed, this follows because for that choice of T, $(n/T)2^{k_0}\rho \ge c_4$, with c_4 an absolute constant.

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

$$R(\tilde{f}) - R(f_1) \le \lambda(x) + \kappa_1 2^{k_0}(b+B)/n + c_3\frac{T}{n} \le \lambda(x) + c_5 2^{k_0}\frac{b+B}{n},$$

as claimed. □

The next step in the proof of Theorem C requires several simple facts regarding the empirical process indexed by a localization of the star-shaped hull of a Bernstein class. First, it is simple to verify that the star-shaped hull of a $(1, B)$-Bernstein class is a $(1, B)$-Bernstein class as well. Second, if $G = \mathrm{star}(\mathcal{L}_F, 0)$ and $G_r = \{h \in G\colon \mathbb{E}h = r\}$, then

$$G_r = \bigcup_{j\ge 1}\bigg\{\frac{r\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f}\colon f\in F, 2^{j-1}r \le \mathbb{E}\mathcal{L}_f \le 2^j r\bigg\} \equiv \bigcup_{j\ge 1} H_{r,j}.$$

In particular,

$$\mathbb{E}\sup_{h\in G_r}\bigg|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\bigg| \le \sum_{i=1}^\infty \mathbb{E}\sup_{h\in H_{r,j}}\bigg|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\bigg|.$$

Lemma 6.3. *There exists an absolute constant c for which the following holds. If $\mathcal{L}_F$ is a $(1, B)$-Bernstein class with respect to Z, then for every r and $j \ge 1$,*

$$\mathbb{E}\sup_{h\in H_{r,j}}|P_n h - Ph| \le c\max\bigg\{\frac{b2^{-j}\log(|H_{r,j}|+1)}{n}, \sqrt{\frac{\log(|H_{r,j}|+1)}{n}}\sqrt{rB2^{-j}}\bigg\}.$$

Proof. Fix $r > 0$ and $j \geq 1$, and let

$$D = \sup_{h \in H_{r,j}} \left(\frac{1}{n} \sum_{i=1}^{n} h^2(Z_i) \right)^{1/2}.$$

Note that every $h \in H_{r,j}$ satisfies that $h = r\mathcal{L}_f / \mathbb{E}\mathcal{L}_f$ for some $f \in F$, and for which $\mathbb{E}\mathcal{L}_f \geq r2^{j-1}$. Therefore, using the Bernstein condition on $\mathcal{L}_F$,

$$\mathbb{E}h^2 = r^2 \frac{\mathbb{E}(\mathcal{L}_f)^2}{(\mathbb{E}\mathcal{L}_f)^2} \leq rB2^{-j+1}.$$

Moreover, $\|h\|_\infty \leq (r/\mathbb{E}\mathcal{L}_f)\|\mathcal{L}_f\|_\infty \leq b2^{-j+1}$. Thus, by the Giné–Zinn symmetrization theorem and a contraction argument (see, e.g., [12] and [19]),

$$\begin{aligned}
\mathbb{E}D^2 &\leq \mathbb{E} \sup_{h \in H_{r,j}} \left| \frac{1}{n} \sum_{i=1}^{n} h^2(Z_i) - \mathbb{E}h^2 \right| + rB2^{-j+1} \\
&\leq \frac{2}{\sqrt{n}} \mathbb{E}_Z \mathbb{E}_\varepsilon \sup_{h \in H_{r,j}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i h^2(Z_i) \right| + rB2^{-j+1} \\
&\leq \frac{b2^{-j+2}}{\sqrt{n}} \mathbb{E}_Z \mathbb{E}_\varepsilon \sup_{h \in H_{r,j}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i h(Z_i) \right| + rB2^{-j+1} \\
&\leq \frac{c_0 rb2^{-j+2}}{\sqrt{n}} \sqrt{\log(|H_{r,j}|+1)} \mathbb{E}D + rB2^{-j+1},
\end{aligned}$$

where the last inequality is evident by the sub-Gaussian properties of the Rademacher process (cf. [19]). Since $\mathbb{E}D \leq (\mathbb{E}D^2)^{1/2}$, it follows that

$$\mathbb{E}D^2 \leq c_0 b2^{-j+2} \sqrt{\frac{\log(|H_{r,j}|+1)}{n}} (\mathbb{E}D^2)^{1/2} + rB2^{-j+1},$$

implying that

$$\mathbb{E}D^2 \leq c_1 \max\left\{ b^2 2^{-2j} \frac{\log(|H_{r,j}|+1)}{n}, rB2^{-j} \right\}.$$

Thus, again using a symmetrization argument and the sub-Gaussian properties of the Rademacher process, we have

$$\begin{aligned}
\mathbb{E} \sup_{h \in H_{r,j}} \left| \frac{1}{n} \sum_{i=1}^{n} h(Z_i) - \mathbb{E}h \right| &\leq \frac{c_2}{\sqrt{n}} \sqrt{\log(|H_{r,j}|+1)} \mathbb{E}D \\
&\leq c_3 \max\left\{ \frac{b2^{-j} \log(|H_{r,j}|+1)}{n}, \sqrt{\frac{\log(|H_{r,j}|+1)}{n}} \sqrt{rB2^{-j}} \right\}. \qquad \square
\end{aligned}$$

Corollary 6.4. *There exist absolute constants c_1 and c_2 for which the following holds. Let F be a finite class consisting of M functions bounded by b, such that the excess loss class $\mathcal{L}_F$ is a $(1, B)$-Bernstein class. If we set $\theta = c_1(b+B)(\log M)/n$, then*

$$r^* \le c_2\left(\frac{b+B}{n}\right)\psi(\theta).$$

Proof. Observe that for every $r > 0$,

$$\begin{aligned}
&\mathbb{E}\sup_{h\in G_r}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\right| \\
&\quad\le \sum_{j\ge 1}\mathbb{E}\sup_{h\in H_{r,j}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\right| \\
&\quad\le c_1\max\left\{\frac{b}{n}\sum_{j\ge 1}2^{-j}\log(|H_{r,j}|+1), \sqrt{\frac{Br}{n}}\sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}\right\} \\
&\quad\le c_1\frac{b}{n}\left(\log(|H_{r,0}|+1) + \sum_{j\ge 1}2^{-j}\log(|H_{r,j}|+1)\right) \\
&\qquad + c_1\sqrt{\frac{Br}{n}}\left(\sqrt{\log(|H_{r,0}|+1)} + \sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}\right) \\
&\quad\equiv u(r),
\end{aligned}$$

where we define $H_{r,0} = \{(r\mathcal{L}_f)/(\mathbb{E}\mathcal{L}_f)\colon\ f \in F, \mathbb{E}\mathcal{L}_f \le r\}$. Let $\bar{r} = \inf\{r\colon\ u(r) \le r/2\}$. Since $|H_{r,j}| \le M$ for every $j \ge 0$, we have

$$u(r) \le c_2\max\left\{b\frac{\log M}{n}, \sqrt{\frac{rB\log M}{n}}\right\},$$

and thus

$$\bar{r} \le c_3(b+B)(\log M)/n = \theta.$$

Moreover, the functions of r,

$$\log(|H_{r,0}|+1) + \sum_{j\ge 1}2^{-j}\log(|H_{r,j}|+1),$$

and

$$\sqrt{\log(|H_{r,0}|+1)} + \sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)},$$

are increasing, and thus for any $r \le \theta$,

$$\frac{b}{n}\left(\log(|H_{r,0}|+1)+\sum_{j\ge 1}2^{-j}\log(|H_{r,j}|+1)\right)$$
$$\le \frac{b}{n}\left(\log(|H_{\theta,0}|+1)+\sum_{j\ge 1}2^{-j}\log(|H_{\theta,j}|+1)\right)$$

and

$$\sqrt{\frac{Br}{n}}\left(\sqrt{\log(|H_{r,0}|+1)}+\sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}\right)$$
$$\le \sqrt{\frac{Br}{n}}\left(\sqrt{\log(|H_{\theta,0}|+1)}+\sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{\theta,j}|+1)}\right).$$

Thus, if we consider

$$r = c_3\frac{b}{n}\left(\log(|H_{\theta,0}|+1)+\sum_{j\ge 1}2^{-j}\log(|H_{\theta,j}|+1)\right)$$
$$+ c_3\frac{B}{n}\left(\sqrt{\log(|H_{\theta,0}|+1)}+\sum_{j\ge 1}2^{-j/2}\sqrt{\log(|H_{\theta,j}|+1)}\right)^2$$
$$\le c_4\left(\frac{b+B}{n}\right)\psi(\theta)$$

for appropriate constants c_3 and c_4, then $r \le \theta$. Thus, $u(r) \le r/2$ and, therefore,

$$\bar{r} \le c_4\left(\frac{b+B}{n}\right)\psi(\theta).$$

Finally, because

$$\mathbb{E}\sup_{h\in G_r}|P_n h - Ph| \le u(r)$$

and $r^* = \inf\{r\colon \mathbb{E}\sup_{g\in G_r}|P_n g - Pg| \le r/2\}$, we have $r^* \le \bar{r}$. □

Proof of Theorem C. The proof of Theorem C follows from estimates of $\lambda(x)$ and 2^{k_0}. From Corollary 6.4, it is evident that

$$\lambda(x) \le c_1 \max\left\{\left(\frac{b+B}{n}\right)\psi\left(c_1(b+B)\frac{\log M}{n}\right), (b+B)\frac{x}{n}\right\},$$

where c_1 is an absolute constant to be identified later. (Note that ψ is an increasing function.)

Next, by the definition of k_0, $2^{k_0} \le \log M$. Therefore, using the notation of Theorem 6.2,

$$\bigcup_{k \le k_0} \{f_j\colon\ j \in J_{+,k}\} \subset \left\{ f_j\colon\ R(f_j) - R(f_1) \le \kappa_1 (b+B) \frac{\log M}{n} \right\}$$

and, in particular,

$$\begin{aligned} 2^{k_0} &\le \log\left(\left| \bigcup_{k \le k_0} \{f_j : j \in J_{+,k}\} \right| + 1 \right) \\ &\le \log\left(\left| \left\{ f_j\colon\ R(f_j) - R(f_1) \le \kappa_1 (b+B) \frac{\log M}{n} \right\} \right| + 1 \right) \le \log(|H_{\theta,0}| + 1), \end{aligned}$$

for an appropriate choice of constant c_1.

The second part of Theorem C follows from a standard integration argument. □

Acknowledgements

This article was written while G. Lecué was visiting the Department of Mathematics, Technion, and the Centre for Mathematics and Its Applications, Australian National University. The authors thank both of these institutions for their hospitality. They also thank Pierre Alquier and Olivier Catoni for useful discussions. G. Lecué was supported by French Agence Nationale de la Recherche ANR Grant "PROGNOSTIC" ANR-09-JCJC-0101-01. S. Mendelson was supported in part by the Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia, by an Australian Research Council Discovery Grant DP0559465, DP0986563 and by the European Community's Seventh Framework Programme (FP7/2007-2013), ERC grant agreement 203134.

References

[1] Alquier, P. (2006). Transductive and inductive adaptative inference for density and regression estimation. Ph.D. thesis, Paris 6.
[2] Audibert, J.-Y. (2004). PAC-Bayesian statistical learning theory. Ph.D. thesis, Paris 6.
[3] Audibert, J.-Y. (2007). No fast exponential deviation inequalities for the progressive mixture rule. Technical report, CERTIS.
[4] Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** 1591–1646. MR2533466
[5] Bartlett, P.L. and Mendelson, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689
[6] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101
[7] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Lecture Notes in Math.* **1851**. Berlin: Springer. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920

[8] Catoni, O. (2007). *Pac-Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. Beachwood, OH: IMS. MR2483528
[9] Dalalyan, A.S. and Tsybakov, A.B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 97–111. Berlin: Springer. MR2397581
[10] Emery, M., Nemirovski, A. and Voiculescu, D. (2000). *Lectures on Probability Theory and Statistics*. *Lecture Notes in Math*. **1738**. Berlin: Springer. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998, Edited by Pierre Bernard. MR1775638
[11] Gaïffas, S. and Lecué, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electron*. *J*. *Stat*. **1** 538–573. MR2369025
[12] Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes (with discussion). *Ann*. *Probab*. **12** 929–998. MR0757767
[13] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann*. *Statist*. **36** 2183–2206. MR2458184
[14] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann*. *Statist*. **34** 2593–2656. MR2329442
[15] Lecué, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann*. *Statist*. **35** 1698–1721. MR2351102
[16] Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab*. *Theory Related Fields* **145** 591–613. MR2529440
[17] Lecué, G. and Mendelson, S. (2010). On the optimality of the empirical risk minimization procedure for the convex aggregation problem. Unpublished manuscript.
[18] Lecué, G. and Mendelson, S. (2010). Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli* **16** 605–613. MR2730641
[19] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Berlin: Springer. MR1102015
[20] Leung, G. and Barron, A.R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans*. *Inform*. *Theory* **52** 3396–3410. MR2242356
[21] Mendelson, S. (2008). Lower bounds for the empirical minimization algorithm. *IEEE Trans*. *Inform*. *Theory* **54** 3797–3803. MR2451042
[22] Mendelson, S. (2008). Obtaining fast error rates in nonconvex situations. *J*. *Complexity* **24** 380–397. MR2426759
[23] Petrov, V.V. (1995). *Limit Theorems of Probability Theory*: *Sequences of Independent Random Variables*. *Oxford Studies in Probability* **4**. New York: Oxford Univ. Press. MR1353441
[24] Rao, M.M. and Ren, Z.D. (1991). *Theory of Orlicz Spaces*. *Monographs and Textbooks in Pure and Applied Mathematics* **146**. New York: Dekker. MR1113700
[25] Samarov, A. and Tsybakov, A. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference*. *Ser*. *Biostat*. **3** 233–251. Hackensack, NJ: World Sci. Publ. MR2416118
[26] Tsybakov, A. (2003). Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines* (*COLT*-2003). *Lecture Notes in Artificial Intelligence* **2777** 303–313. Heidelberg: Springer.
[27] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann*. *Statist*. **32** 135–166. MR2051002
[28] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359

[29] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671
[30] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal*. **74** 135–161. MR1790617
[31] Yang, Y. (2000). Mixing strategies for density estimation. *Ann*. *Statist*. **28** 75–87. MR1762904
[32] Yang, Y. (2001). Adaptive regression by mixing. *J*. *Amer. Statist*. *Assoc*. **96** 574–588. MR1946426

Received November 2010 and revised July 2011