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

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

Abstract

In this article, we study the optimality of the aggregate with exponential weights (AEW) in the regression model with random design, and in the low temperature regime. We prove three properties of AEW. First, that AEW is a suboptimal aggregation procedure in expectation with respect to the quadratic risk when $T \leq c_1$ 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 whose risk is larger than the risk of the best function in the dictionary by at least order of $1/\sqrt{n}$. On the other hand, we prove that if one assumes a geometric condition on the dictionary (the so-called Bernstein condition), 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 completing an analogous result for high temperatures.

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 $Z_1, \ldots, Z_n$ to be $n+1$ i.i.d. random variables with values in $\mathcal{Z}$. From the statistical point of view, $\mathcal{D} = (Z_1, \ldots, Z_n)$

[1]CNRS, LAMA, Marne-la-valle, 77454 France.

[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.

[3]Email: guillaume.lecue@univ-mlv.fr

[4]Email: shahar.mendelson@anu.edu.au

is the set of given data at our disposal. The *risk* of a real-valued function f defined on $\mathcal{X}$ is given by

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

where $Q : \mathcal{Z} \times \mathcal{X} \mapsto \mathbb{R}$ is a nonnegative function, called the *loss function*. If $\widehat{f}$ is a random function constructed using the data $\mathcal{D}$, the risk of $\widehat{f}$ is the random variable

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

Throughout this article we will restrict ourselves 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], [29], [13] or [4]). The loss function we will focus on through most of this article is the quadratic loss function, defined 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, for example, [10], [7] and [28]) is to construct a procedure that produces a function whose risk is as close as possible to the risk of the best element in F. Having this in mind, one can define the *optimal rate of aggregation* [24, 15], 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 whose risk is as close as possible to that of the best element in the dictionary. We recall here the definition for the "expectation case". A similar definition for the "probability case" can also be formulated (see, for example, [15]).

Definition 1.1 ([24]) *Let $b > 0$. We say that $(\psi_n(M))_{n,M\in\mathbb{N}^*}$ is the 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 size M and a random variable Z such that $|Q(Z, f)| \leq b$ a.s. 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. [24]) that, in general, the optimal rate of aggregation (in the sense of [24] – optimality in expectation and of [15] – optimality in probability) is lower bounded by $(\log M)/n$. Hence, procedures satisfying an exact oracle inequality like (1.1) with the residual term $\psi_n(M) = (\log M)/n$ are said to be optimal. There are only a few aggregation procedures that have been proved to achieve this optimal rate. The first results dealing with optimal aggregation procedures can be found in [7] and [27] (and for a survey on optimal aggregation procedures we refer the reader to the HDR dissertation of J.-Y. Audibert).

Our main focus here is the problem of the optimality of the aggregation procedure with exponential weights (AEW). The origin of this procedure comes from the thermodynamical point of view of learning theory (see [8] for the state of the art in this direction). AEW can be seen 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), and it is well known that ERM cannot, in general, achieve the optimal rate of $(\log M)/n$, unless one assumes that the given class F has certain geometric properties which will be discussed below (see also [16, 19, 13]). To have any chance of obtaining better rates, one has to consider aggregation procedures that are taking values in larger subsets than F, and the most natural set is the convex hull of F. AEW has been a very popular candidate for an optimal procedure, and it was one of the first procedures to be studied in the context of the aggregation framework [13, 4, 14, 18, 7, 2, 28, 9]. It is defined by the following convex sum

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

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

Despite its long history, the optimality of AEW remained open. In this work, we study the following question:

Question 1.2 *Is the AEW an optimal aggregation procedure in expectation or in probability in the regression model with random design?*

We will show that the answer to Question 1.2 is:

[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.

- negative for low temperatures $T \leq c_1$ (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);

- negative in probability for some large dictionaries and small temperatures $T \leq c_1$ (Theorem B);

- positive for low temperatures under a geometric condition on the dictionary (Theorem C); Together with the high temperature result of [1], [2] and [8], this proves that the temperature parameter has almost no impact 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.

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

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

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

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

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

Theorem A proves that AEW is suboptimal in expectation in the low temperature regime, and suboptimal in probability in both low and high temperature regimes.

It should be mentioned that suboptimality in probability does not imply suboptimality in expectation for the aggregation problem, nor vice-versa. This property of the aggregation problem was first noticed in [3] where an aggregation procedure called the *progressive mixture rule* was proved to be suboptimal in probability for dictionaries of cardinality two, whereas it was known to be optimal in expectation (cf. [7], [27], [29] or [13]).

The proof of Theorem A shows that a dictionary consisting of two functions is enough to give the lower bound in expectation in the low temperature regime and in probability in both regimes. And, although the second part is to be expected, the first one is surprising, as we will explain below.

In Theorem B we study the behavior of AEW for larger dictionaries. To our knowledge, negative results on the behavior of exponential weights based aggregation procedures are not known for dictionaries with more than two functions, and what we show is that the behavior of the AEW deteriorates, in some sense, as the cardinality of the dictionary grows.

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

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

Moreover, if $f_F^* \in F$ denotes the optimal function in F with respect to the quadratic loss (the oracle), then there exists $f_j \neq f_F^*$ whose excess risk larger than $c_5(\alpha)n^{-1/2}$ and for which the weight of f_j in the AEW procedure satisfies

$$\widehat{\theta}_j \geq 1 - \frac{1}{n^{c_6(\alpha)/T}}.$$

Theorem B implies that the AEW procedure might cause the weights to concentrate around a "bad" element in the dictionary (that is, 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 gives additional evidence that the AEW procedure is suboptimal for low temperatures.

The analysis of the behavior of AEW for dictionary of cardinality larger than two is considerably harder than the two-function case, and it requires some results on rearrangement of independent random variables which are almost Gaussian (see Proposition 5.2 below).

Fortunately, not all is lost as far as optimality results for AEW go. Indeed, we will show that under some geometric condition, AEW can be optimal; in fact, it 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. Hence, a good procedure is one whose residual term is of the order of ψ/n, where ψ is a complexity measure that is determined only by the complexity of the set of "almost minimizers" in the dictionary.

Question 1.3 *Is it possible to construct an aggregation procedure that adapts to the real complexity of the dictionary?*

This question was first answered by the PAC-Bayesian approach. It was shown in [1], [2] and [8] that in the high temperature regime, AEW satisfies the requirements of Question 1.3, assuming that the class has a geometric property, called the Bernstein condition.

Definition 1.4 ([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*

$$\mathbb{E}\left(f^2(Z)\right) \leq B\left(\mathbb{E}f(Z)\right)^{\beta}. \tag{1.4}$$

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 then the excess loss functions class $\mathcal{L}_F = \{Q(\cdot, f) - Q(\cdot, f_F^*) : f \in F\}$ 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 targets with "multiple minimizers" in F, and, in which case as well, $\mathcal{L}_F$ satisfies the Bernstein condition with $\beta = 1$ (see [19, 20] 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. Firstly, the empirical excess risk has better concentration properties around the excess risk, and secondly, the complexity of the subset of F consisting of almost minimizers is smaller under this assumption. As a consequence, 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, for example [5], and references therein). As the results below show, the same is true for AEW. Indeed, under a Bernstein assumption it was proved in ([1], [2] or [8]) 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 $1 - 2\exp(-x)$,

$$R(\tilde{f}^{AEW}) \leq \min_{f \in F} R(f) + \frac{Tc_2}{n}\Big(x + \log\Big(\sum_{f \in F} \exp\big(-(n/2T)(R(f) - R(f_F^*))\big)\Big)\Big). \tag{1.5}$$

Although the PAC-Bayesian approach can not be used to obtain (1.5) in the low temperature regime ($T \leq c_1 \max\{b, B\}$), such a result is not surprising. Indeed, since fast error rates for the ERM are to be expected when the underlying excess loss functions class satisfies the Bernstein condition and since AEW converges to the ERM when the temperature T tends to zero, it is likely that for "small values" of T, AEW inherits some of the properties of ERM, for example, fast rates under a

Bernstein condition. This is what we show in Theorem C, proving that AEW does answer Question 1.3 for low temperatures under the Bernstein condition.

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

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

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

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

Theorem C. *There exists 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 $\tilde{f}^{AEW}$ produced by the AEW algorithm satisfies*

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

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

In particular,

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

In other words, the scaling factor θ we use is proportional to $(b + B)(\log |F|)/n$, and if the class is reasonably regular, $\psi(\theta)$ is roughly the cardinality of the elements in F whose risk is 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. Therefore, if T is reasonably small – below a level proportional to $\max\{B, b\}$, the resulting aggregation rate is the optimal one, proportional to $(b + B)(x + \log M)/n$ with probability of $1 - 2\exp(-x)$, and proportional to $(b + B)(\log M)/n$ in expectation. Therefore, Theorem C gives a positive answer to Question 1.3 in the presence of a Bernstein condition and for a low temperature.

Although the residual terms in Theorem C and in (1.5) 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 result for high temperatures from [1], [2] and [8] shows 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 one obtains is of the order of $((T+1)\log M)/n$.

Finally, a word about the organization of the article. In the next section we provide some comments about our results. The proofs of the three theorems follow in the other sections. Throughout, we denote absolute constants or constants that depend on other parameters by 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. We write $a \sim b$ if there are absolute constants c and C such that $bc \leq a \leq Cb$, and $a \lesssim b$ if $a \leq Cb$.

Acknowledgments

This article was written while G. Lecué was visiting the Department of Mathematics, Technion, and the Centre for Mathematics and its Applications, The Australian National University. The authors would like to thank both these institutions for their hospitality. We also would like to thank Pierre Alquier and Olivier Catoni for useful discussions.

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n^o [203134] and from the Australian Research Council Discovery Project DP0986563.

2 Comments

The suboptimality in expectation of the AEW, obtained in Theorem A, is rather surprising for two reasons. First of all, it is known that the progressive mixture rule is optimal in expectation for T larger than some parameters of the model (see [7], [27], [29], [13] or [4]). This procedure is defined by

$$\bar{f} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_k^{AEW}, \tag{2.1}$$

where $\tilde{f}_k^{AEW}$ is the function generated by AEW associated with the dictionary F and constructed using the first k observations $Z_1, \ldots, Z_k$. Thus, this aggregate is the mean of $\tilde{f}_k^{AEW}$ for $1 \leq k \leq n$, where, for every $k < n$, $\tilde{f}_k^{AEW}$ is constructed using only the first k observations. In particular, $\bar{f}$ is the mean of aggregates that are (or should be) less "efficient" than $\tilde{f}_n^{AEW}$, since the latter is constructed using all the

observations $Z_1, \ldots, Z_n$, rather than a subset of the given observations. That is why one expects the AEW to be an optimal aggregation procedure in expectation – at least in the high temperature regime. Theorem A shows that, even for temperature of the order of a constant, $\tilde{f}_n^{AEW}$ might have a very bad behavior, of the order of $(1/\sqrt{n})$.

Second, the optimality in expectation of AEW was obtained in [9] for the regression model $Y_i = f(x_i) + \epsilon_i$ 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),

$$\mathbb{E} \left\| \tilde{f}^{AEW} - f \right\|_n^2 \leq \min_{g \in F} \|g - f\|_n^2 + \frac{T \log M}{n+1}. \tag{2.2}$$

Theorem A shows that the behavior of the AEW is very different, at least in the low temperature regime. The fact that the same procedure (although in different models) can exhibit such two extreme behaviors - and for roughly the same temperature parameter is rather striking. The $1/\sqrt{n}$ lower bound of Theorem A vs. the $1/n$ upper bound derived from the oracle inequality (2.2) can have one of the two following explanations. Either that the two seemingly similar scenarios are, in fact very different, or that AEW exhibits a sharp phase transition at $T \sim c$. And, if the latter is true, then an important outcome of Theorem A is that the temperature parameter is of the highest importance with regard to the optimality of the AEW in expectation.

An indication that a phase transition is the likely explanation to the phenomenon observed in Theorem A, is that most of the optimal upper bounds on AEW or on the progressive mixture need T to be larger than some unknown parameters of the model (the variance of the noise for instance). This means that in practice, AEW is likely to be a very "risky" aggregation procedure because of its sensitivity to the temperature parameter. Moreover, and to make things even worse, even for large values of T, AEW is suboptimal with a constant probability for small dictionaries (Part 2 of Theorem A) and with probability that tends to 1 for larger dictionaries (Theorem B). Hence, given a set of data and a dictionary, AEW *is likely* to behave very poorly regardless of what T is. In contrast, Theorem C shows that the choice of the temperature parameter has no significant effect on the performance of the AEW (residual term of the order of $T(\log M)/n$) under the Bernstein condition. To conclude, although from a theoretical point of view it remains to be seen whether AEW displays a phase transition at constant temperature, and is indeed an optimal procedure in expectation for high temperatures $T \geq c_2 \max(b, \sigma^2)$ as one may conjecture based on the results from [7], [27], [29], [13] [4] and [9], from a practical point of view, we believe that exponential aggregating schemes simply should not be used in the setup of this article. The choice of T is simply too "risky", as indicated by the lower bounds in probability

of [3], Part 1 of Theorem A and Theorem B.

Another consequence of the lower bounds stated in Theorem A is that AEW cannot be an optimal aggregation procedure both in expectation and probability for 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

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

Also, 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 [24]), while the residual terms obtained in Theorem A are of the order of $n^{-1/2}$ for such a dictionary. Hence, AEW is suboptimal for these two other aggregation problems for low temperatures.

We end this section by comparing two seemingly related assumptions: the margin assumption of [25] and the Bernstein condition of [5]. Let us mention 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 holds if one assumes a Bernstein condition for any $0 < \beta < 1$, and the proof is identical to the one in the case $\beta = 1$. This makes the discussion about β-Bernstein classes relevant here.

Recall the definition of the margin assumption:

Definition 2.1 ([25]) *We say that F has margin with parameters (β, B) ($0 < \beta \leq 1$ and $B \geq 1$) if for every $f \in F$,*

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

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, "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^*, which is 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 zero quickly as the sample size tends to infinity. In this setup, the margin assumption can improve this rate of convergence thanks to a better concentration of empirical means of $Q(\cdot, f) - Q(\cdot, f^*)$ around its mean [25]. The margin assumption (MA for short) for $\beta = 1$ compares the performance of each

$f \in F$ to 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 at all.

In the second, "learning theory" setup, we do not assume that the target function f^* belongs to F. The aim is to construct a function $\widehat{f}$ whose risk is as close as possible to that of the best element $f_F^* \in F$. And, assuming that the excess loss class $\mathcal{L}_F$ satisfies the Bernstein condition (BC for short) one can improve the error rate (see, e.g., [20, 5]).

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 $f \in F$

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

but nevertheless, they are different. Indeed, as we mentioned, MA is only a matter of concentration (and classical statistics questions are mostly a question of the tradeoff 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 one. In fact, the difficulty of "learning theory" problems is determined by the tradeoff 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 we do not have that freedom in the aggregation framework. Indeed, since in the AEW algorithm the estimator is determined by the empirical means $R_n(f) - R_n(f_F^*)$, it 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, for example, [4, 11, 14, 23, 6, 18, 25, 2, 28]). Therefore, because of its nature, aggregation procedures like the AEW are more natural under a BC assumption and not the MA one (a by-product of Theorem A is that the MA cannot improve the performance of AEW since in Theorem A's setup 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 zero random variables with variance 1, set g to be a standard Gaussian variable and put

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

Theorem 3.1 *([21])There exists an absolute constant $A > 0$ such that for every integer n,*

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

From here on we will denote by A the constant appearing in Theorem 3.1.

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

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

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 facts on Orlicz norms see, for instance, [26] and [22].

Proposition 3.2 (Chapter 5 in [21]) *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 $\|W\|_{\psi_1} \le L$ then for any $x \ge 0$ such that $x \le B_0 n^{1/6}$,*

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

and

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

where by $v = O(u)$ we mean that $-c_1 u \le v \le c_1 u$.

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

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

From here on we will denote by B_0 the constant appearing in Proposition 3.2.

4 Proof of Theorem A

Before presenting the proof of Theorem A, let us 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 denote by P the expectation $\mathbb{E}_\nu$. From here on, we will assume that $T \le 1$ 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

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

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

An important parameter which is at the heart of this counter-example is the Bernstein constant (which is very bad in this case):

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

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

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

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

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

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,

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

for

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

and

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

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

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

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

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

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

Let us 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

$$\begin{aligned}
I_2 &= \int_{\alpha^{-1}}^{\frac{\alpha+1}{2\alpha}} (2\alpha t - (1+\alpha))\, \mathbb{P}\left[P_n\mathcal{L}_2 \ge \gamma(t)\right] dt + \int_{\frac{\alpha+1}{2\alpha}}^{1} (2\alpha t - (1+\alpha))\, \mathbb{P}\left[P_n\mathcal{L}_2 \ge \gamma(t)\right] dt \\
&= \int_{\alpha^{-1}}^{\frac{\alpha+1}{2\alpha}} (2\alpha t - (1+\alpha))\, \mathbb{P}\left[\gamma(t) \le P_n\mathcal{L}_2 < \gamma\left(1+\alpha^{-1}-t\right)\right] dt = I_3 + I_4
\end{aligned}$$

for

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

and

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

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

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

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 $I_T(t) = \big[n\gamma(t), n\gamma(1+\alpha^{-1}-t)\big)$, and set $N_T(t) = |\{I_T(t) \cap \mathbb{Z}\}|$, which is the number of integers in $I_T(t)$. Since $\mathcal{L}_2(X) = -X$ then

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

Recall that $X \in \{-1,1\}$ and thus $\mathbb{P}\left[\sum_i -X_i \in I_T(t)\right] = \mathbb{P}\left[\sum_i -X_i \in I_T(t) \cap \mathbb{Z}\right]$. Since $n\gamma(t)$ is increasing and nonnegative 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

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

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

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

One may show that $\max_{k\in\mathbb{Z}} \mathbb{P}\left[\sum_{i=1}^n X_i = k\right]$ is of 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

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

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

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

and proving the first part of Theorem A.

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

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

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

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

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

5 Proof of Theorem B

The first step in the proof of Theorem B is a general statement about a monotone rearrangement of independent random variables that are close to being gaussian.

Let W be a mean zero, variance one random variable, that is absolutely continuous with respect to the Lebesgue measure. Assume further that $|W|$ has a finite third moment (in fact, the random variables we will be interested in will be 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 $W_1, ..., W_n$ 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

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

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 will require the following simple estimates on γ_1.

Lemma 5.1 *There exist absolute constants $c_0, ..., c_3$ for which the following holds.*

1. *If* $\ell \geq c_0 \log n$ *then*

$$1 - c_1 \frac{\log n}{\ell} \leq \mathbb{P}[\bar{X} > \gamma_1] \leq 1 - \frac{\log n}{\ell}.$$

2. *If* ℓ *and* n *are such that* $(\beta(W)/\sqrt{n} + (\log n)/\ell) < \mathbb{P}[g < -2]$, *then* $\gamma_1 \leq -2$.

3. *If* $\gamma_1 \leq -2$ *and* $c_0 \log n \leq \ell \leq c_2 \beta^{-1}(W)\sqrt{n} \log n$ *then*

$$|\gamma_1| \sim \log^{1/2}\left(\frac{c_3 \ell}{\log n}\right) \text{ and } \exp(-\gamma_1^2/2) \sim \frac{\log n}{\ell} \log^{1/2}\left(\frac{c_3 \ell}{\log n}\right).$$

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

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{P}\left[\exists j \in \{2,\dots,M\} : \bar{X}_j \leq \gamma_1 \text{ and for every } k \in \{2,\dots,M\}\backslash\{j\},\ \bar{X}_k - \bar{X}_j \geq \delta\right] \\ &\quad \geq 1 - \frac{1}{n} - c_2\left(\frac{1}{\sqrt{n}} + \delta\right)(\log n)^2\sqrt{\log M},\end{aligned}$$

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

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

$$\Omega_j = \left\{\bar{X}_j \leq \gamma_1 \text{ and } \bar{X}_k - \bar{X}_j \geq \delta \text{ for every } k \in \{2,\dots,M\}\backslash\{j\}\right\}.$$

The events Ω_j for $2 \leq j \leq M$ are disjoints and thus

$$\begin{aligned}&\mathbb{P}\left[\exists j \in \{2,\dots,M\} : \bar{X}_j \leq \gamma_1 \text{ and } \bar{X}_k - \bar{X}_j \geq \delta \text{ for every } k \in \{2,\dots,M\}\backslash\{j\}\right] \\ &= \mathbb{P}[\cup_{j=2}^M \Omega_j] = (M-1)\mathbb{P}[\Omega_2].\end{aligned}$$

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

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

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

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

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

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

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

where

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

Recall the if (W_i) are independent, mean zero random variables then $\|\sum a_i W_i\|_{\psi_2} \leq c(\sum a_i^2 \|W_i\|_{\psi_2}^2)^{1/2}$ where c is an absolute constant [26]. Hence, $\|\bar{X}\|_{\psi_2} \leq c\,\|W\|_{\psi_2}$, and for any $t < 0$,

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

Let $t_0 < 0$ be such that

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

Hence,

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

Note that if $t_0 \geq \gamma_1$ then our claim follows. Indeed, since $\mathbb{P}\left[\min_{2\leq j\leq M}\bar{X}_j \leq \gamma_1\right] \leq 1 - n^{-1}$, then

$$\mathbb{P}[\Omega_0] \geq \frac{1}{M-1}\left(1-\frac{1}{n}\right) - \delta\frac{\sqrt{\log(M-1)}}{M-1}.$$

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 $z \leq z+\delta \leq 0$ and $|z|, |z+\delta| \leq B_0 n^{1/6}$, then

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

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

$$\mathbb{P}[z \leq g \leq z+\delta] \leq \frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right)\int_0^{\delta}\exp(-zt)dt \leq \frac{\delta}{\sqrt{2\pi}}\exp\left(-z^2/2\right). \tag{5.5}$$

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

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

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

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

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

Next, let us describe the construction we need for the proof of Theorem B. Let (X,Y) and $F = \{f_1,\ldots,f_M\}$ be defined by

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

where $\mathcal{U}_1,\ldots,\mathcal{U}_M$ are M independent random variables with the 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 that the best element in F with respect to the quadratic risk is f_1.

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

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

and observe that if $W = \sqrt{12}\left((\mathcal{U}+\lambda)^2 - \mathbb{E}(\mathcal{U}+\lambda)^2\right)$ then W is a mean zero, variance 1 random variable that is absolutely continuous with respect to the Lebesgue measure;

also, $W \in L_{\psi_2}$ 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

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

and recall that for each $j = 1, \dots, M$ we denote by $\widehat{\theta}_j$ 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

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

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 $k \in \{2, \dots, M\} \backslash \{j\}$,

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

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

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

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

$$\widehat{\theta}_j = \frac{1}{\sum_{k=1}^{M} \exp\left[\frac{-n}{T}\left(R_n(f_k) - R_n(f_j)\right)\right]} \geq \frac{1}{1 + (M-2)\frac{\rho}{2(M-2)(1-\rho)} + \frac{\rho}{2(1-\rho)}} = 1 - \rho.$$

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

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

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

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

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

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

as claimed.

■

Let us formulate a general statement from which Theorem B follows immediately.

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

$$1 - c_4(\epsilon\kappa + T + 1)\left((\log^3 n)/n\right)^{(1-2\epsilon)^2/2},$$

there exists $j \geq 2$ such that

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

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

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

Proof. Set

$$\mathbb{P}_0 = \mathbb{P}\left[\exists j \in \{2,\dots,M\} \text{ such that } \widehat{\theta}_{\mathrm{j}} \geq 1-\rho\right],$$

and by Proposition 5.3,

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

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

$$A = \left\{\exists j \in \{2,...,M\} : \bar{R}_j \leq \xi(\bar{R}_1), \text{ and } \bar{R}_k - \bar{R}_j \geq \delta \text{ for every } k \neq 1, j\right\},$$

and

$$B = \left\{\exists j \in \{2,...,M\} \ : \ \bar{R}_j \leq \gamma_1 \text{ and } \bar{R}_k - \bar{R}_j \geq \delta \text{ for every } k \neq 1, j\right\}.$$

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

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

Applying Proposition 5.2,

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

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

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

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

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

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

By the Berry-Esséen Theorem and (5.1),

$$\mathbb{P}[\Omega_0] \geq \mathbb{P}[\bar{R}_1 \geq (1-2\epsilon)\gamma_1] = 1 - \mathbb{P}[\bar{R}_1 < (1-2\epsilon)\gamma_1] \geq 1 - \mathbb{P}[g \leq (1-2\epsilon)\gamma_1] - \frac{2\beta(W)}{\sqrt{n}}$$
$$\geq 1 - \frac{1}{\sqrt{2\pi}(1-2\epsilon)|\gamma_1|}\exp\left(-(1-2\epsilon)^2\gamma_1^2/2\right) - \frac{2A}{\sqrt{n}},$$

and by Lemma 5.1,

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

Therefore,

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

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

To complete the proof, one has 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

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

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

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

implying that

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

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 will prove Theorem C, which is re-formulated below. From here on we will 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, $f_1 = f_F^*$. Also, we will 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

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

which will serve as a measure of complexity for the class F.

The first component that is needed in the proof of Theorem C is to find 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 $R(f_j) - R(f_1) \ge \lambda(x)$. This "isomorphism" constant was introduced in [5] and to formulate the exact properties we need, recall the following definitions and notation.

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

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

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*

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

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

$$R_n(f) - R_n(f_F^*) \ge \frac{1}{2}\left(R(f) - R(f_F^*)\right).$$

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 $J_-(x) = \{j : R(f_j) - R(f_1) \le \lambda(x)\}$ and set $J_+(x)$ to be its complement. Define the sets $J_{+,0} = \{j \in J_+(x) : R(f_j) - R(f_1) \le \rho\}$ and for $k \ge 1$,

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

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

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

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 $j \in J_{+,k}$, $R_n(f_j) - R_n(f_F^*) \ge \frac{1}{2}\left(R(f_j) - R(f_F^*)\right)$ (because $R(f_j) - R(f_F^*) \ge \lambda(x)$ by the definition of $J_+(x)$, and since $J_+(x) \supset J_{+,k}$).

The key ingredient 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 $\tilde{f}^{AEW}$ is produced by the AEW algorithm and $T \le c_1(b+B)$, then*

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

where $\lambda(x)$ has been defined in (6.1).

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

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

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, since $\sum_{j=1}^M \widehat{\theta}_j = 1$ then

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

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

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

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

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

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

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

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

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

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

as claimed.

■

The next step towards 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 of all, 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 $G_r = \{h \in G : \mathbb{E}h = r\}$ then

$$G_r = \bigcup_{j \geq 1} \left\{ \frac{r\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : f \in F, 2^{j-1} r \leq \mathbb{E}\mathcal{L}_f \leq 2^j r \right\} \equiv \bigcup_{j \geq 1} H_{r,j},$$

In particular,

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

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

$$\mathbb{E} \sup_{h \in H_{r,j}} |P_n h - Ph| \leq c \max\left\{ \frac{b 2^{-j} \log(|H_{r,j}| + 1)}{n}, \sqrt{\frac{\log(|H_{r,j}| + 1)}{n}} \sqrt{r B 2^{-j}} \right\}.$$

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, for example, [12] and [17]),

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

where the last inequality is evident by the subgaussian properties of the Rademacher process (cf. [17]). 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

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

Hence, using a symmetrization argument and the subgaussian properties of the Rademacher process once again,

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

■

Corollary 6.4 *There exists 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

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

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

$$\begin{aligned}
&\mathbb{E}\sup_{h\in G_r}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\right| \le \sum_{j\ge 1}\mathbb{E}\sup_{h\in H_{r,j}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h\right| \\
\le& c_1 \max\left\{\frac{b}{n}\sum_{j\ge1}2^{-j}\log(|H_{r,j}|+1), \sqrt{\frac{Br}{n}}\sum_{j\ge1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}\right\} \\
\le& c_1\frac{b}{n}\left(\log(|H_{r,0}|+1) + \sum_{j\ge1}2^{-j}\log(|H_{r,j}|+1)\right) \\
&+ c_1\sqrt{\frac{Br}{n}}\left(\sqrt{\log(|H_{r,0}|+1)} + \sum_{j\ge1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}\right) \\
\equiv& u(r),
\end{aligned}$$

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

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

and thus

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

Moreover, the functions of r

$$\log(|H_{r,0}|+1) + \sum_{j\ge1}2^{-j}\log(|H_{r,j}|+1)$$

and

$$\sqrt{\log(|H_{r,0}|+1)} + \sum_{j\ge1}2^{-j/2}\sqrt{\log(|H_{r,j}|+1)}$$

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

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

and

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

Hence, if we consider

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

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

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

Finally, since

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

and since $r^* = \inf\{r : \mathbb{E}\sup_{g\in G_r}|P_ng - Pg| \le r/2\}$, then $r^* \le \bar{r}$. ∎

Proof of Theorem C. The proof of Theorem C follows from estimates on $\lambda(x)$ and on 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 named later (note that ψ is an increasing function).

Next, if follows from the definition of k_0 that $2^{k_0} \leq \log M$. Therefore, using the notation of Theorem 6.2,

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

and in particular

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

for an appropriate choice of the constant c_1.

The second part of Theorem C is evident from a standard integration argument. ∎

References

[1] Pierre Alquier. *Transductive and Inductive Adaptative Inference for Density and Regression Estimation*. PhD thesis, Paris 6, Dec 2006.

[2] Jean-Yves Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Paris 6, Jul 2004.

[3] Jean-Yves Audibert. No fast exponential deviation inequalities for the progressive mixture rule. Technical report, CERTIS, 2007.

[4] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, Jun 2008.

[5] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

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

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

[8] Olivier Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

[9] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

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

[11] Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.*, 1:538–573, 2007.

[12] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion.

[13] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[14] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 2007.

[15] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related fields*, 145(3–4):591–613, 2004.

[16] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *to appear in Bernoulli*, 2009.

[17] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[18] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

[19] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54(8):3797–3803, 2008.

[20] Shahar Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity*, 24(3):380–397, 2008.

[21] Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[22] M. M. Rao and Z. D. Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 1991.

[23] Alexander Samarov and Alexandre Tsybakov. Aggregation of density estimators and dimension reduction. In *Advances in statistical modeling and inference*, volume 3 of *Ser. Biostat.*, pages 233–251. World Sci. Publ., Hackensack, NJ, 2007.

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

[25] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[26] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[27] Yuhong Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000.

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

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