Известия НАН Армении, Математика, том 58, н. 6, 2023, стр. 3 – 14.

SIMPLE PROOF OF THE RISK BOUND FOR DENOISING BY EXPONENTIAL WEIGHTS FOR ASYMMETRIC NOISE DISTRIBUTIONS

A. S. DALALYAN

Institut Polytechnique de Paris, France[1]
E-mail: *arnak.dalalyan@ensae.fr*

Abstract. In this note, we consider the problem of aggregation of estimators in order to denoise a signal. The main contribution is a short proof of the fact that the exponentially weighted aggregate satisfies a sharp oracle inequality. While this result was already known for a wide class of symmetric noise distributions, the extension to asymmetric distributions presented in this note is new.

MSC2020 numbers: 62J05; 62H12.

Keywords: signal denoising; exponential weights; PAC-Bayesian bounds.

1. Introduction

Let us consider the problem of denoising an n dimensional noisy signal $\boldsymbol{Y}$ using a family of candidates $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m$. More precisely, we assume that

$$\boldsymbol{Y} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^n$ is the n dimensional true signal and $\boldsymbol{\xi}$ is random noise. Only the noisy vector $\boldsymbol{Y}$ is observed and the goal is to construct an estimator $\widehat{\boldsymbol{\theta}}$ such that the expected error $\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2]$ is as small as possible, where $\|\boldsymbol{v}\|$ stands for the Euclidean norm of $\boldsymbol{v} \in \mathbb{R}^n$. We consider the framework in which to achieve the aforementioned goal we are given a set of vectors $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m\}$. An estimator $\widehat{\boldsymbol{\theta}}$ is considered a good estimator, if the regret

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] - \min_{j=1,\dots,m} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*\|^2 \tag{1.1}$$

is as small as possible. This problem has been coined model-selection aggregation in (17), where it is also proved that the optimal rate of the difference in (1.1) is $\log m$. The problem of aggregation has been extensively studied in the literature, see for instance (3; 20; 22; 21; 13; 2; 16; 18; 1; 14; 4). In this note, we consider the

[1]The work of the author was supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047), the FAST Advance grant and the center Hi! PARIS.

exponentially weighted aggregate (EWA) defined as follows. Let $\pi_0(1), \dots, \pi_0(m)$ be some nonnegative weights summing to one. Each $\pi_0(j)$ represents our prior confidence in the approximation of $\boldsymbol{\theta}^*$ by $\boldsymbol{\theta}_j$. Based on these prior weights and the observed vector $\boldsymbol{Y}$, we define

$$\widehat{\boldsymbol{\theta}} = \sum_{j=1}^m \boldsymbol{\theta}_j \widehat{\pi}(j), \qquad \text{with} \qquad \widehat{\pi}(j) = \frac{\exp\{-\|\boldsymbol{Y} - \boldsymbol{\theta}_j\|^2/\beta\}\pi_0(j)}{\sum_{\ell=1}^m \exp\{-\|\boldsymbol{Y} - \boldsymbol{\theta}_\ell\|^2/\beta\}\pi_0(\ell)}.$$

In this expression, $\beta > 0$ is a tuning parameter of the method. As established in the aforementioned references, in different settings one can prove that EWA satisfies the inequality

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \min_{j=1,\dots,m} \Big(\|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*\|^2 + \beta \log(1/\pi_0(j)) \Big). \tag{1.2}$$

In particular, if π_0 is the uniform distribution over $\{1, \dots, m\}$, one obtains the rate-optimal remainder term $\beta \log m$ for the difference in (1.1).

As pointed out in some papers (8; 9; 6), it is helpful to extend the above-described framework to the case of aggregating a family of estimators which is potentially infinite. This is equivalent to considering a subset $S_0 \subset \mathbb{R}^n$ and aiming at finding an "optimal" way of combining all its elements in order to estimate $\boldsymbol{\theta}^*$. These types of considerations have led to the following extension of the estimator (1.2):

$$\widehat{\boldsymbol{\theta}} = \int_{\mathbb{R}^n} \boldsymbol{\theta}\, \widehat{\pi}(d\boldsymbol{\theta}), \qquad \text{with} \qquad \frac{d\widehat{\pi}}{d\pi_0}(\boldsymbol{\theta}) = \frac{\exp\{-\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2/\beta\}}{\int_{\mathbb{R}^n} \exp\{-\|\boldsymbol{Y} - \boldsymbol{u}\|^2/\beta\}\pi_0(d\boldsymbol{u})}. \tag{1.3}$$

Notice that this estimator is the Bayesian posterior mean in the case where $\boldsymbol{\xi}$ is drawn from the Gaussian distribution with zero mean and covariance matrix $(\beta/2)\mathbf{I}_n$. The goal of this note is to provide an alternative and simple proof of the fact that EWA $\widehat{\boldsymbol{\theta}}$ satisfies (1.2) and its extension to aggregating an infinite set, provided that the distribution of the noise ξ satisfies some suitable conditions. We also slightly extend the existing results by including noise distributions that are not symmetric with respect to the origin. This is particularly suitable for estimating the parameters of Bernoulli or binomial distributions.

Notation. We use boldface letters for vectors, which are always seen as one-column matrices. For any vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|_\infty$ are respectively the Euclidean norm and the sup-norm. By convention, throughout this work, $0 \cdot \infty = 0$. For a probability distribution π on $\mathbb{R}^n$, we denote by $\mathrm{Var}_\pi(\boldsymbol{\theta})$ the variance with respect to π defined by $\int_{\mathbb{R}^n} \|\boldsymbol{\theta}\|^2 \, \pi(d\boldsymbol{\theta}) - \| \int_{\mathbb{R}^n} \boldsymbol{\theta}\, \pi(d\boldsymbol{\theta})\|^2$. For two probability distributions μ and ν defined on the same probability space and such that μ is absolutely continuous with respect to ν, the Kullback-Leibler divergence is defined by $D_{\mathrm{KL}}(\mu||\nu) = \int \frac{d\mu}{d\nu}(x) \log \frac{d\mu}{d\nu}(x)\, \nu(dx)$.

2. Main result

This section is devoted to stating and briefly discussing the main result, the proof being postponed to Section 4 below. Prior to stating the result, we recall the Bernstein condition. For some $v > 0$ and $b \geqslant 0$, we say that a random variable η satisfies the (v, b)-Bernstein condition, if

$$\mathbf{E}[e^{t\eta}] \leqslant \exp\Big\{\frac{v^2t^2}{2(1-b|t|)}\Big\}, \qquad \forall t \in (-1/b, 1/b).$$

This condition is clearly on the distribution of the random variable. One can check that if η satisfies the (v, b)-Bernstein condition, then it is sub-exponential with zero mean, and the variance of η is at least equal to v. Many common distributions satisfy this assumption. For instance, any sub-Gaussian distribution with variance proxy τ satisfies the $(\tau, 0)$-Bernstein condition. Any random variable supported by $[-A, A]$ satisfies the Bernstein condition with $(v, b) = (A^2, 0)$ but also with $(v, b) = (\mathrm{Var}(\eta), A/3)$ (19). We will see that the latter is more useful for our purposes than the former.

Similarly, if $\mathcal{F}$ is a sigma-algebra and v and b are two $\mathcal{F}$- measurable random variables, we say that η is (v, b)-Bernstein conditionally to $\mathcal{F}$, if almost surely, the inequality $\mathbf{E}[e^{t\eta}|\mathcal{F}] \leqslant \exp\{v^2t^2/(1-b|t|)\}$ is satisfied for every $t \in \mathbb{R}$ such that $|t|b < 1$.

Theorem 1. *Let π_0 be a probability distribution supported by $S_0 \subset \mathbb{R}^n$ with a diameter measured in sup-norm bounded by $\mathcal{D}_0$. Assume that the distribution of $\boldsymbol{\xi}$ satisfies the following assumption: for some sigma algebra $\mathcal{F}$ and for some $b : [0, 1] \to [0, \infty)$ and continuously differentiable function $v : [0, 1] \to [0, \infty)$ vanishing at the origin, for every $\alpha \in (0, 1]$, there exists an n-dimensional random vector $\boldsymbol{\zeta}$ such that*

$$\mathbf{E}[\boldsymbol{\zeta}|\mathcal{F}] = 0, \qquad \boldsymbol{\xi} + \boldsymbol{\zeta} \overset{\mathscr{D}}{=} (1+\alpha)\boldsymbol{\xi}.$$

and, conditionally to $\mathcal{F}$, the entries ζ_i are independent and satisfy the $(v(\alpha), b(\alpha))$-Bernstein condition. Then, for every $\beta \geqslant 2b(0)\mathcal{D}_0$, we have

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_{\pi}\bigg\{\int_{\mathbb{R}^n}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2\,\pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0)\bigg\} + \bigg(\frac{2v'(0)}{\beta - 2b(0)\mathcal{D}_0} - 1\bigg)\mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})],$$

where the inf *is over all the probability distributions. As a consequence, for* $\beta \geqslant 2v'(0) + 2b(0)\mathcal{D}_0$*, we get*

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_{\pi} \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\}. \tag{2.1}$$

Let us briefly comment on this result. First, the link between (2.1) and (1.2) might not be easy to see. It is obtained by considering a prior distribution π_0 supported by the finite set $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m\}$ and by upper bounding the infimum in (2.1) by the minimum over all the Dirac measures $\delta_{\boldsymbol{\theta}_j}$. One easily checks that $D_{\mathrm{KL}}(\delta_{\boldsymbol{\theta}_j}||\pi_0) = \log(1/\pi_0(j))$, which allows to infer (1.2) from (2.1).

Second, one may wonder where the form of the upper bound in (2.1) comes from. The presence of the KL-divergence in this bound may seem surprising. The reason is that there is a deep connection between the KL-divergence and the exponential weights. Indeed, according to the Varadhan-Donsker variational formula, the "posterior" distribution $\widehat{\pi}$ defined in (1.3) is solution to following problem:

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi} \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{Y}\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\},$$

where the min is over all the probability distributions. This result will be the starting point of the proof.

Finally, one can wonder how restrictive the assumptions of this theorem are. We will show below that they are satisfied for a broad class of noise distributions.

3. Instantiation to some well-known noise distributions

The main theorem stated in the previous section requires a general and a rather abstract condition to be satisfied by the noise distribution. This section shows that many distributions encountered in applications satisfy this assumption with some parameters $v'(0)$ and $b(0)$ which are easy to determine.

3.1. **Centered Bernoulli noise.** Assume that each ξ_i is a centered Bernoulli random variable: it takes the value $1 - \rho_i$ with probability ρ_i and the value $-\rho_i$ with probability $1 - \rho_i$. Here, $\rho_i \in (0, 1)$. Then, one can set

$$\mathbf{P}\big(\zeta_i = \alpha\xi_i \,|\, \xi_i\big) = \frac{1 + \alpha - \alpha|\xi_i|}{\alpha + 1}, \quad \mathbf{P}\big(\zeta_i = -\operatorname{sgn}(\xi_i)(1 + \alpha - \alpha|\xi_i|) \,|\, \xi_i\big) = \frac{\alpha|\xi_i|}{\alpha + 1}.$$

We see that conditionally to ξ_i, the random variable ζ_i is zero mean and takes its values in an interval of length $\alpha(1 - \rho_i) + \alpha\rho_i + 1 = \alpha\rho_i + 1 + \alpha - \alpha\rho_i = 1 + \alpha$. This implies that ζ_i satisfies the $((1 + \alpha)^2/4, 0)$-Bernstein condition, conditionally to ξ_i. In other terms, ζ_i is sub-Gaussian with variance proxy $(1 + \alpha)^2/4$. However, this

does not help in applying Theorem 1, since the function $v(\alpha) = (1+\alpha)^2/4$ does not vanish at the origin. On the positive side, since the conditional variance of ζ_i given ξ_i is smaller than $\alpha(1+\alpha)$ and the support is included in $[-(1+\alpha),(1+\alpha)]$, the conditional distribution of ζ_i given ξ_i satisfies the Bernstein condition with $v(\alpha) = \alpha(1+\alpha)$ and $b(\alpha) = (1+\alpha)/3$, see (19, Exercise 2.8.5). This yields the following result.

Corollary 1. Let π_0 be a probability distribution supported by $S_0 \subset \mathbb{R}^n$ such that $\mathcal{D}_0 = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in S_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty < \infty$. Assume that $\boldsymbol{\xi}$ has independent entries ξ_i satisfying $\mathbf{P}(\xi_i = 1-\rho_i) = 1 - \mathbf{P}(\xi_i = -\rho_i) = \rho_i$ for some $\rho_i \in (0,1)$. Then, for every $\beta \geqslant (2/3)\mathcal{D}_0$, we have

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_\pi \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\} + \Big(\frac{6}{3\beta - 2\mathcal{D}_0} - 1 \Big) \mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})]. \tag{3.1}$$

In particular, if $\beta \geqslant 2 + (2/3)\mathcal{D}_0$, the last term in (3.1) is nonpositive and, therefore, can be neglected.

This corollary can be used in cases where the observations Y_i are independent Bernoulli random variables with mean θ_i^*. In such a situation, it is natural to choose a prior distribution π_0 that is concentrated on the unit hypercube $[0,1]^n$, the diameter of which in sup-norm is equal to 1. The corollary implies that in such a situation the inequality stated in (2.1) is true provided that $\beta \geqslant 8/3$. We refer the reader to (10) for an application of this result to graphon estimation.

3.2. **Gaussian noise.** In the case of the Gaussian noise $\boldsymbol{\xi}$ with independent entries having 0 mean and variance equal to σ_i^2, one can check that the conditions of Theorem 1 are satisfied with the random vector $\boldsymbol{\zeta}$ which is independent of $\boldsymbol{\xi}$ and has independent entries drawn from the Gaussian distribution $\mathcal{N}(0, (2\alpha + \alpha^2)\sigma_i^2)$. This means that in the Bernstein condition one can choose $\mathcal{F} = \sigma(\boldsymbol{\xi})$, $b = 0$ and $v(\alpha) = (2\alpha + \alpha^2) \max_{1 \leqslant i \leqslant n} \sigma_i^2$, which leads to the following result.

Corollary 2. Let π_0 be a probability distribution on $\mathbb{R}^n$. Assume that $\boldsymbol{\xi}$ has independent entries $\xi_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, \dots, n$. Then, for every $\beta > 0$, we have

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_\pi \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\} + \big(4\sigma^2\beta^{-1} - 1\big) \mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})], \tag{3.2}$$

where $\sigma = \max_{1 \leqslant i \leqslant n} \sigma_i$. In particular, if $\beta \geqslant 4\sigma^2$, the last term in (3.2) is nonpositive and, therefore, can be neglected.

Some preliminary versions of this result can be traced back to (12; 11). In the form (2.1), and with an extension to aggregation of projection estimators, the result appeared in (15). Further generalisations to various families of linear estimators have been explored in (7). The proof of the oracle inequality in all these papers is very specific to the Gaussian distribution since it is based on Stein's lemma (integration by parts for the Gaussian measure). The alternative proof presented in this work relies on techniques developed in (8; 5; 6).

3.3. **Bounded noise.** For every $a, b > 0$, let $\mathcal{B}(a, b)$ be the distribution of a random variable that takes the values a and $-b$ with probabilities $b/(a+b)$ and $a/(a+b)$, respectively. If the distribution of ξ_i can be written as a mixture of the distributions $\mathcal{B}(a, b)$ with a mixing distribution with bounded support, then our main theorem can be applied. More precisely, assume that the distribution of ξ_i is given by

$$p_{\xi_i}(dx) = \int_0^A \int_0^B \frac{b\delta_a(dx) + a\delta_{-b}(dx)}{a+b}\, \nu_i(da, db),$$

where ν_i is a probability distribution on $[0, A] \times [0, B]$. This means that $\xi_i = \eta_i^{\alpha_i, \beta_i}$ with random variables (α_i, β_i) drawn from ν_i and $\eta_i^{a,b}$ drawn from the binary distribution $\frac{b\delta_a(dx)+a\delta_{-b}(dx)}{a+b}$. Akin to the first subsection of this section, one can choose $\zeta_i^{a,b}$ so that $(1+\alpha)\eta_i^{a,b}$ has the same distribution as $\eta_i^{a,b} + \zeta_i^{a,b}$, for every pair (a, b). Then, clearly, $(1+\alpha)\xi_i$ has the same distribution as $\xi_i + \zeta_i^{\alpha,\beta}$. Let $\mathcal{F}$ be the sigma algebra generated by the random variables $\alpha, \beta, \{\eta_j^{a,b} : (a, b) \in [0, A] \times [0, B], i \in [n]\}$. Conditionally to $\mathcal{F}$, $\zeta_i^{a,b}$ is a binary random variable with zero mean and takes its values in the interval $[-B, A]$, it satisfies the Bernstein condition with $b(\alpha) = (A+B)(1+\alpha)/3$ and $v(\alpha) = (A+B)^2\alpha(1+\alpha)$. Therefore, we get the following result.

Corollary 3. Let π_0 be a probability distribution supported by $S_0 \subset \mathbb{R}^n$ such that $\mathcal{D}_0 = \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in S_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty < \infty$. Assume that $\boldsymbol{\xi}$ has independent entries ξ_i, $i = 1, \dots, n$, taking values in an interval I_i of length at most L. Then, for every $\beta \geqslant (2/3)L\mathcal{D}_0$, we have

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_\pi \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\} + \Big(\frac{6L^2}{3\beta - 2L\mathcal{D}_0} - 1 \Big) \mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})]. \tag{3.3}$$

In particular, if $\beta \geqslant 2L^2 + (2/3)L\mathcal{D}_0$, the last term in (3.3) is nonpositive and, therefore, can be neglected.

This result is well suited for the setting where the components Y_i of the observation $\boldsymbol{Y}$ are bounded. For instance, if we know that $\mathbf{P}(Y_i \in [0, L]) = 1$ for every $i \in \{1, \dots, n\}$, then it is also natural to choose a prior distribution satisfying $\mathcal{D}_0 = L$. Inequality (2.1) is then satisfied for every $\beta \geqslant (8/3)L^2$. Note that, to the best of our knowledge, this is the first time that such a precise bound is obtained for asymmetric noise distributions. The similar result established in (6, Theorem 2) deals with symmetric distributions only.

3.4. **Centered binomial noise.** Consider the case where ξ_i's are independent and drawn from a centered and scaled binomial distribution $a\mathcal{B}(k, \rho_i) - ak\rho_i$, where $a > 0$ is the scaling factor. This distribution is a particular case of distributions supported by a finite interval considered in the previous subsection. One can therefore apply the last corollary with $L = ak$. However, this leads to a bound which is too crude. Indeed, one can use the fact that ξ_i is equal in distribution to $a(\eta_1 + \ldots + \eta_k)$ where η_j's are iid centered Bernoulli variables. Defining $\bar{\zeta}_1, \dots, \bar{\zeta}_k$ as independent random variables satisfying

$$\mathbf{P}\big(\bar{\zeta}_j = \alpha\eta_j \,|\, \eta_j\big) = \frac{1 + \alpha - \alpha|\eta_j|}{\alpha + 1}, \quad \mathbf{P}\big(\bar{\zeta}_j = -\operatorname{sgn}(\eta_j)(1 + \alpha - \alpha|\eta_j|) \,|\, \eta_j\big) = \frac{\alpha|\eta_j|}{\alpha + 1},$$

one easily checks that $\eta_j + \bar{\zeta}_j$ has the same distribution as $(1 + \alpha)\eta_j$. Therefore, $\xi_i + \zeta_i$, for $\zeta_i = a(\bar{\zeta}_1 + \ldots + \bar{\zeta}_k)$, has the same distribution as $(1+\alpha)\xi_i$. Furthermore, conditionally to the sigma-algebra generated by $\{\eta_1, \dots, \eta_k\}$, ζ_i has zero mean and satisfies the Bernstein condition with $b(\alpha) = a(1 + \alpha)/3$ and $v(\alpha) = a^2 k\alpha(1 + \alpha)$.

Corollary 4. Let π_0 be a probability distribution supported by $S_0 \subset \mathbb{R}^n$ such that $\mathcal{D}_0 = \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in S_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty < \infty$. Assume that $\boldsymbol{\xi}$ has independent entries ξ_i, $i = 1, \dots, n$, drawn from the scaled and centered binomial distribution $a(\mathcal{B}(k, \rho_i) - k\rho_i))$. Then, for every $\beta \geqslant (2/3)a\mathcal{D}_0$, we have

$$\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_{\pi} \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi || \pi_0) \right\} + \left(\frac{6a^2 k}{3\beta - 2a\mathcal{D}_0} - 1 \right) \mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})]. \tag{3.4}$$

In particular, if $\beta \geqslant 2a^2 k + (2/3)a\mathcal{D}_0$, the last term in (3.4) is nonpositive and, therefore, can be neglected.

A typical application of this result concerns the case of observing the average of k Bernoulli variables, that is $Y_i \sim (1/k)\mathcal{B}(k, \theta_i^*)$. In this case, all the θ_i^* belong to $[0, 1]$ and, therefore, it is reasonable to choose a prior distribution π_0 supported by $[0, 1]^n$. This ensures that $\mathcal{D}_0 \leqslant 1$, and, therefore, inequality (2.1) follows from the last corollary provided that $\beta \geqslant 8/(3k)$ (this is obtained by choosing $a = 1/k$).

3.5. **Double exponential noise.** All the previous examples considered in this section are distributions with sub-exponential tails. Let us check that Theorem 1 can also be applied to some distributions that have heavier, say sub-exponential, tails. Let ξ_i be independent drawn from the Laplace distribution[2] with parameters $\mu_i > 0$, $i = 1, \dots, n$. Then, one can choose $\mathcal{F} = \mu(\boldsymbol{\xi})$ and $\zeta_1, \dots, \zeta_n$ to be independent, independent of $\boldsymbol{\xi}$, and drawn from the distribution $\frac{1}{(1+\alpha)^2}\delta_0 + \frac{2\alpha+\alpha^2}{(1+\alpha)^2}\mathsf{Lap}((1+\alpha)\mu_i)$. The fact that $\xi_i + \zeta_i$ has the same distribution as $(1+\alpha)\xi_i$ can be checked by computing the characteristic functions of these variables and by verifying that they are equal. As for the Bernstein condition, for every t such that $(1+\alpha)\mu_i|t| \leqslant 1$ we have

$$\begin{aligned}
\mathbf{E}[e^{t\zeta_i}] &= \frac{1}{(1+\alpha)^2} + \frac{2\alpha+\alpha^2}{(1+\alpha)^2} \times \frac{1}{1-(1+\alpha)^2 t^2 \mu_i^2} \\
&\qquad \big(p := 1-(1+\alpha)^{-2}, z := (1+\alpha)t\mu_i\big) \\
&= 1 - p + \frac{p}{1-z^2} = 1 + \frac{pz^2}{1-z^2} \leqslant 1 + \frac{pz^2}{1-|z|} \\
&\leqslant \exp\left\{\frac{pz^2}{1-|z|}\right\} = \exp\left\{\frac{\alpha(2+\alpha)\mu_i^2 t^2}{1-(1+\alpha)\mu_i|t|}\right\}
\end{aligned}$$

This means that the (conditional) Bernstein condition is satisfied with $v(\alpha) = \alpha(2+\alpha)\mu^2$ and $b(\alpha) = (1+\alpha)\mu$, where μ is the largest value among μ_i.

Corollary 5. Let π_0 be a probability distribution supported by $S_0 \subset \mathbb{R}^n$ such that $\mathcal{D}_0 = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in S_0} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty < \infty$. Assume that $\boldsymbol{\xi}$ has independent entries ξ_i, $i = 1, \dots, n$, drawn from the Laplace distribution $\mathsf{Lap}(\mu_i)$. Set $\mu = \max_{1 \leqslant i \leqslant n} \mu_i$. Then, for every $\beta \geqslant 2\mu\mathcal{D}_0$, we have

$$\begin{aligned}
\mathbf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] \leqslant \inf_{\pi} \left\{ \int_{\mathbb{R}^n} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \right\} \\
+ \Big(\frac{4\mu^2}{\beta - 2\mu\mathcal{D}_0} - 1\Big)\mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})]. \qquad (3.5)
\end{aligned}$$

In particular, if $\beta \geqslant 4\mu^2 + 2\mu\mathcal{D}_0$, the last term in (3.5) is nonpositive and, therefore, can be neglected.

The last claim improves on (9, Prop. 1), since the latter requires the condition $\beta \geqslant (16\mu^2) \vee (\sqrt{8}\,\mu\mathcal{D}_0)$.

Remark 1. *Let us finally remark that the construction of ζ_i's used in this section can be extended to the case where ξ_i's are scale-mixtures of Laplace distributions with a mixing density supported by a compact set. The only modification in the statement of the final result should be the definition of μ, which should correspond*

[2]This means that the density of ξ_i is equal to $(2\mu_i)^{-1}\exp(-|x|/\mu_i)$.

to the smallest real number such that the mixing density has no mass in (μ, ∞). *Similar extension can be carried out in the case of scale-mixtures of Gaussians.*

4. Proof of Theorem 1

Since $\widehat{\pi}$ minimizes the criterion $\pi \mapsto \int_{\mathbb{R}^n} \|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0)$, we have

$$\int_{\mathbb{R}^n} \|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 \, \widehat{\pi}(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\widehat{\pi}||\pi_0) \leqslant \int_{\mathbb{R}^n} \|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0)$$

for all densities π over $\mathbb{R}^n$. The KL-divergence being always nonnegative, we infer from the last display that

$$\begin{aligned} \|\boldsymbol{Y}-\widehat{\boldsymbol{\theta}}\|^2 &= \int_{\mathbb{R}^n} \|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 \, \widehat{\pi}(d\boldsymbol{\theta}) - \int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 \, \widehat{\pi}(d\boldsymbol{\theta}) \\ &\leqslant \int_{\mathbb{R}^n} \|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) - \int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 \, \widehat{\pi}(d\boldsymbol{\theta}). \end{aligned} \tag{4.1}$$

Using the decompositions $\|\boldsymbol{Y}-\widehat{\boldsymbol{\theta}}\|^2 = \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2 + 2(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top \boldsymbol{\xi} + \|\boldsymbol{\xi}\|^2$ and $\|\boldsymbol{Y}-\boldsymbol{\theta}\|^2 = \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 + 2(\boldsymbol{\theta}^*-\boldsymbol{\theta})^\top \boldsymbol{\xi} + \|\boldsymbol{\xi}\|^2$ and taking the expectation of the two sides of (4.1), we get

$$\begin{aligned} \mathbf{E}[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2] + 2\mathbf{E}[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top \boldsymbol{\xi}] &\leqslant \int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) \\ &\quad + \beta D_{\mathrm{KL}}(\pi||\pi_0) \mathbf{E}\bigg[\int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 \, \widehat{\pi}(d\boldsymbol{\theta}) \bigg] \end{aligned}$$

which can be equivalently written as

$$\begin{aligned} \mathbf{E}[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2] &\leqslant \int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \, \pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) \\ &\quad + 2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] - \int_{\mathbb{R}^n} \mathbf{E}[\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 \, \widehat{\pi}(\boldsymbol{\theta})] \, d\boldsymbol{\theta}. \end{aligned} \tag{4.2}$$

In addition, we have

$$2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] = \frac{\beta}{\alpha} \mathbf{E}\bigg[\int_{\mathbb{R}^n} \log e^{2(\alpha/\beta)\boldsymbol{\theta}^\top \boldsymbol{\xi}} \widehat{\pi}(d\boldsymbol{\theta}) \bigg],$$

where $\alpha > 0$ is an arbitrary number. Since the logarithm is concave, the Jensen inequality yields

$$
\begin{aligned}
2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] &\leqslant \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{2(\alpha/\beta)\boldsymbol{\theta}^\top\boldsymbol{\xi}}\widehat{\pi}(d\boldsymbol{\theta})\bigg)\bigg] \\
&= \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{2(\alpha/\beta)\boldsymbol{\theta}^\top\boldsymbol{\xi}-\|\boldsymbol{\theta}^*+\boldsymbol{\xi}-\boldsymbol{\theta}\|^2/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg) \\
&\quad - \log\bigg(\int_{\mathbb{R}^n} e^{-\|\boldsymbol{\theta}^*+\boldsymbol{\xi}-\boldsymbol{\theta}\|^2/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg)\bigg] \\
&= \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{(2(1+\alpha)\boldsymbol{\theta}^\top\boldsymbol{\xi}-\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|^2)/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg) \\
&\quad - \log\bigg(\int_{\mathbb{R}^n} e^{(2\boldsymbol{\theta}^\top\boldsymbol{\xi}-\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|^2)/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg)\bigg] \qquad (4.3)
\end{aligned}
$$

Let $\boldsymbol{\zeta} = \boldsymbol{\zeta}_\alpha$ be the n dimensional random vector the existence of which is required in the statement of the theorem. Recall that it satisfies

$$
\mathbf{E}[\boldsymbol{\zeta}|\mathcal{F}] = 0, \qquad \boldsymbol{\xi} + \boldsymbol{\zeta} \overset{\mathscr{D}}{=} (1+\alpha)\boldsymbol{\xi},
$$

These conditions imply that in the first expectation in (4.3), one can replace $(1+\alpha)\boldsymbol{\xi}$ by $\boldsymbol{\xi} + \boldsymbol{\zeta}$, which yields

$$
\begin{aligned}
2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] &\leqslant \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{(2\boldsymbol{\theta}^\top\boldsymbol{\xi}+2\boldsymbol{\theta}^\top\boldsymbol{\zeta}-\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|^2)/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg)\bigg] \\
&\quad - \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{(2\boldsymbol{\theta}^\top\boldsymbol{\xi}-\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|^2)/\beta}\,\pi_0(d\boldsymbol{\theta})\bigg)\bigg] \\
&= \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{2\boldsymbol{\theta}^\top\boldsymbol{\zeta}/\beta}\,\widehat{\pi}(d\boldsymbol{\theta})\bigg)\bigg] = \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} e^{2(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\zeta}/\beta}\,\widehat{\pi}(d\boldsymbol{\theta})\bigg)\bigg]. \qquad (4.4)
\end{aligned}
$$

Since conditionally to $\mathcal{F}$, ζ_i's are independent and each ζ_i satisfies the $(v(\alpha), b(\alpha))$-Bernstein condition, one can use the Jensen inequality to upper bound the expectation in (4.4) as follows

$$
\begin{aligned}
2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] &\leqslant \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} \mathbf{E}[e^{2(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\zeta}/\beta}|\mathcal{F}]\,\widehat{\pi}(d\boldsymbol{\theta})\bigg)\bigg] \\
&\leqslant \frac{\beta}{\alpha}\mathbf{E}\bigg[\log\bigg(\int_{\mathbb{R}^n} \exp\Big\{\frac{2\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 v(\alpha)}{\beta(\beta - 2b(\alpha)\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_\infty)}\Big\}\,\widehat{\pi}(d\boldsymbol{\theta})\bigg)\bigg] \qquad (4.5)
\end{aligned}
$$

for every β satisfying $\beta \geqslant 2b(\alpha)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty$ for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in S_0 := \mathrm{supp}(\pi_0)$. Note that for every $\boldsymbol{\theta} \in S_0$, we have $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_\infty \leqslant \mathcal{D}_0$. The inequality in (4.5) being true

for any $\alpha > 0$, one can check that

$$
\begin{aligned}
2\mathbf{E}[\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\xi}] &\leqslant \liminf_{\alpha\to 0} \frac{\beta}{\alpha}\mathbf{E}\biggl[\log\biggl(\int_{\mathbb{R}^n} \exp\Bigl\{\frac{2\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 v(\alpha)}{\beta(\beta-2b(\alpha)\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_\infty)}\Bigr\}\widehat{\pi}(d\boldsymbol{\theta})\biggr)\biggr] \\
&= \mathbf{E}\biggl[\int_{\mathbb{R}^n} \frac{2\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2 v'(0)}{\beta-2b(0)\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_\infty}\,\widehat{\pi}(d\boldsymbol{\theta})\biggr] \\
&\leqslant \frac{2v'(0)}{\beta-2b(0)\mathcal{D}_\infty(S_0)}\mathbf{E}\biggl[\int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|^2\,\widehat{\pi}(d\boldsymbol{\theta})\biggr]. \qquad (4.6)
\end{aligned}
$$

Combining (4.2) and (4.6), we see that

$$
\begin{aligned}
\mathbf{E}[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2] \leqslant \int_{\mathbb{R}^n} \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2\,\pi(d\boldsymbol{\theta}) + \beta D_{\mathrm{KL}}(\pi||\pi_0) + \\
+\biggl(\frac{2v'(0)}{\beta-2b(0)\mathcal{D}_\infty(S_0)}-1\biggr)\mathbf{E}[\mathrm{Var}_{\widehat{\pi}}(\boldsymbol{\vartheta})].
\end{aligned}
$$

This completes the proof.

Список литературы

[1] P. Alquier and Karim Lounici, "PAC-Bayesian bounds for sparse regression estimation with exponential weights", Electron. J. Stat., **5**, 127 -– 145 (2011).

[2] P. C. Bellec, "Optimal bounds for aggregation of affine estimators", Ann. Statist., **46** (1), 30 -– 59 (2018).

[3] F. Bunea, A. B. Tsybakov and M. H. Wegkamp, "Aggregation for gaussian regression", Ann. Statist., **35** (4), 1674 -– 1697 (2007).

[4] E. Chernousova, Yu. Golubev and E. Krymova, "Ordered smoothers with exponential weighting", Electron. J. Stat., **7**, 2395 -– 2419 (2013).

[5] A. S. Dalalyan and A. B. Tsybakov, "Sparse regression learning by aggregation and Langevin Monte-Carlo", COLT 2009 - The 22nd Conference on Learning Theory, Montreal, June 18-21, 1 -– 10 (2009).

[6] A. S. Dalalyan, "Exponential weights in multivariate regression and a low-rankness favoring prior", Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, **56** (2), 1465 -– 1483 (2020).

[7] A. S. Dalalyan and J. Salmon, "Sharp oracle inequalities for aggregation of affine estimators", The Annals of Statistics, **40** (4), 2327 -– 2355 (2012).

[8] A. S. Dalalyan and A. B. Tsybakov, "Aggregation by exponential weighting and sharp oracle inequalities", Learning theory, **4539** of Lecture Notes in Comput. Sci., 97 -– 111, Springer, Berlin (2007).

[9] A. S. Dalalyan and A. B. Tsybakov, "Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity", Machine Learning, **72** (1-2), 39 -– 61 (2008).

[10] E. Donier-Meroz, A. S. Dalalyan, F. Kramarz, Ph. Choné, and X. D'Haultfoeuille, "Graphon estimation in bipartite graphs with observable edge labels and unobservable node labels", Technical report (2023).

[11] E. I. George, "Combining minimax shrinkage estimators", J. Amer. Statist. Assoc., **81** (394), 437 -– 445 (1986).

[12] E. I. George, "Minimax multiple shrinkage estimation", The Annals of Statistics, **14**(1), 188 -– 205 (1986).

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

[14] G. Lecué and Sh. Mendelson, "On the optimality of the aggregate with exponential weights for low temperatures", Bernoulli, **19**(2), 646 -– 675 (2013).

[15] G. Leung and A.R. Barron, "Information theory and mixing least-squares regressions", IEEE Transactions on Information Theory, **52** (8), 3396 -– 3410 (2006).

[16] Ph. Rigollet and A. Tsybakov, "Exponential screening and optimal rates of sparse estimation", Ann. Statist., **39** (2), 731 -– 771 (2011).

[17] A. B. Tsybakov, "Optimal rates of aggregation", Bernhard Schölkopf and Manfred K. Warmuth, Learning Theory and Kernel Machines, 303 — 313, Berlin, Heidelberg (2003).

[18] A. B. Tsybakov, "Aggregation and minimax optimality in high-dimensional estimation", Proceedings of the International Congress of Mathematicians (Seoul, August 2014), **3**, 225 -– 246 (2014).

[19] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (2018).

[20] Y. Yang, "Combining different procedures for adaptive regression", J. Multivariate Anal., **74**(1), 135 -– 161 (2000).

[21] Y. Yang, "Regression with multiple candidate models: selecting or mixing?", Statist. Sinica, **13** (3), 783 -– 809 2003.

[22] Y. Yang, "Aggregating regression procedures to improve performance", Bernoulli, **10** (1), 25 -– 47 (2004).

Поступила 07 декабря 2022

После доработки 22 марта 2023

Принята к публикации 24 марта 2023