PAC-Bayesian Aggregation of Affine Estimators

L. Montuelle and E. Le Pennec

Abstract Aggregating estimators using exponential weights depending on their risk appears optimal in expectation but not in probability. We use here a slight overpenalization to obtain oracle inequality in probability for such an explicit aggregation procedure. We focus on the fixed design regression framework and the aggregation of linear estimators and obtain results for a large family of linear estimators under a non-necessarily independent sub-Gaussian noise assumptions.

1 Introduction

We consider here a classical fixed design regression model

$$\forall i \in \{1, \dots, n\},\ Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function, x_i the fixed design points, and $W = (W_i)_{i \leq n}$ a centered sub-Gaussian noise. We assume that we have at hand a family of linear estimate $\{\hat{f}_t(Y) = A_t Y | A_t \in \mathscr{S}_n^+(\mathbb{R}), b_t \in \mathbb{R}^n, t \in \mathscr{T}\}$, for instance a family of projection estimator, of linear ordered smoother in a basis or in a family of basis. The most classical way to use such a family is to select one of the estimates according to the observations, for instance using a penalized empirical risk principle. A better way is to combine linearly those estimates with weights depending on the observation. A simple strategy is the Exponential Weighting Average in which all those estimates are averaged with a weight proportional to $\exp\left(-\frac{\widetilde{r}_t}{\beta}\right)\pi(t)$ where $\widetilde{r}_t$ is a (penalized) estimate of the risk of $\hat{f}_t$. This strategy is not new nor optimal

L. Montuelle
RTE, La Défense, France
e-mail: lucie.montuelle@rte-france.com

E. Le Pennec (✉)
CMAP/XPOP, École Polytechnique, Palaiseau, France
e-mail: erwan.le-pennec@polytechnique.edu

P. Bertail et al. (eds.), *Nonparametric Statistics*, Springer Proceedings
in Mathematics & Statistics 250, https://doi.org/10.1007/978-3-319-96941-1_9

as explained below but is widely used in practice. In this chapter, we analyze the performance of this simple EWA estimator by providing oracle inequalities in probability under mild sub-Gaussian assumption on the noise.

Our aim is to obtain the best possible estimate of the function f_0 at the grid points. This setting is probably one of the most common in statistics and many regression estimators are available in the literature. For non-parametric estimation, Nadaraya-Watson estimator [39, 52] and its fixed design counterpart [26] are widely used, just like projection estimators using trigonometric, wavelet [24] or spline [51] basis, for example. In the parametric framework, least squares or maximum likelihood estimators are commonly employed, sometimes with minimization constraints, leading to LASSO [47], ridge [34], elastic net [60], AIC [1], or BIC [45] estimates.

Facing this variety, the statistician may wonder which procedure provides the best estimation. Unfortunately, the answer depends on the data. For instance, a rectangular function is well approximated by wavelets but not by trigonometric functions. Since the best estimator is not known in advance, our aim is to mimic its performances in terms of risk. This is theoretically guaranteed by an oracle inequality:

$$R(f_0, \tilde{f}) \leq C_n \inf_{t \in \mathscr{T}} R(f_0, \hat{f}_t) + \epsilon_n$$

comparing the risk of the constructed estimator $\tilde{f}$ to the risk of the best available procedure in the collection $\{\hat{f}_t, t \in \mathscr{T}\}$. Our strategy is based on convex combination of these preliminary estimators and relies on PAC-Bayesian aggregation to obtain a single adaptive estimator. We focus on a wide family, commonly used in practice : affine estimators $\{\hat{f}_t(Y) = A_t(Y-b)+b+b_t | A_t \in \mathscr{S}_n^+(\mathbb{R}), b_t \in \mathbb{R}^n, t \in \mathscr{T}\}$ with $b \in \mathbb{R}^n$ a common recentering.

Aggregation procedures have been introduced by Vovk [50], Littlestone and Warmuth [37], Cesa-Bianchi et al. [13], Cesa-Bianchi and Lugosi [14]. They are a central ingredient of bagging [9], boosting [25, 44], or random forest ([3] or [10]; or more recently [6–8, 27]).

The general aggregation framework is detailed in [40] and studied in [11, 12] through a PAC-Bayesian framework as well as in [53–59]. See, for instance, [49] for a survey. Optimal rates of aggregation in regression and density estimation are studied by Tsybakov [48], Lounici [38], Rigollet and Tsybakov [42], Rigollet [41] and Lecué [35].

A way to translate the confidence of each preliminary estimate is to aggregate according to a measure exponentially decreasing when the estimate's risk rises. This widely used strategy is called exponentially weighted aggregation. More precisely, as explained before, the weight of each element $\hat{f}_t$ in the collection is proportional to $\exp\left(-\frac{\widetilde{r}_t}{\beta}\right)\pi(t)$ where $\widetilde{r}_t$ is a (penalized) estimate of the risk of $\hat{f}_t$, β is a positive parameter, called the temperature, that has to be calibrated and π is a prior measure over $\mathscr{T}$. The key property of exponential weights is that they explicitly minimize the aggregated risk penalized by the Kullback-Leibler divergence to the prior measure π[12]. Our aim is to give sufficient conditions on the risk estimate $\widetilde{r}_t$ and the

temperature β to obtain an oracle inequality for the risk of the aggregate. Note that when the family $\mathscr{T}$ is countable, the exponentially weighted aggregate is a weighted sum of the preliminary estimates.

This procedure has shown its efficiency, offering lower risk than model selection because we bet on several estimators. Aggregation of projections has already been addressed by Leung and Barron [36]. They have proved, by the mean of an oracle inequality, that the aggregate performs almost as well, in expectation, as the best projection in the collection. Those results have been extended to several settings and noise conditions [5, 18, 19, 21–23, 29, 30, 43, 46] under a *frozen* estimator assumption: they should not depend on the observed sample. This restriction, not present in the work by Leung and Barron [36], has been removed by Dalalyan and Salmon [20] within the context of affine estimator and exponentially weighted aggregation. Nevertheless, they make additional assumptions on the matrices A_t and the Gaussian noise to obtain an optimal oracle inequality in expectation for affine estimates. Very sharp results have been obtained in [15, 31] and [32]. Those papers, except the last one, study a risk in expectation.

Indeed, the Exponential Weighting Aggregation is not optimal anymore in probability. Dai et al. [17] have indeed proved the sub-optimality in deviation of exponential weighting, not allowing to obtain a sharp oracle inequality in probability. Under strong assumptions and independent noise, [4] provides a sharp oracle inequality with optimal rate for another aggregation procedure called Q-aggregation. It is similar to exponential weights but the criterion to minimize is modified and the weights no longer are explicit. Results for the original EWA scheme exist nevertheless but with a constant strictly larger than 1 in the oracle inequality. Dai [16] obtain, for instance, a result under a Gaussian white noise assumption by penalizing the risk in the weights and taking a temperature at least 20 times greater than the noise variance. Golubev and Ostobski [32] does not use an overpenalization but assumes some ordered structure on the estimate to obtain a result valid even for low temperature. An unpublished work, by Gerchinovitz [28], provides also weak oracle inequality with high probability for projection estimates on non-linear models. Alquier and Lounici [2] consider *frozen* and bounded preliminary estimators and obtain a sharp oracle inequality in deviation for the excess risk under a sparsity assumption, if the regression function is bounded, with again a modified version of exponential weights.

In this work, we will play on both the temperature and the penalization. We will be able to obtain oracle inequalities for the Exponential Weighting Aggregation under a general sub-Gaussian noise assumption that does not require a coordinate independent setting. We conduct an analysis of the relationship between the choice of the penalty and the minimal temperature. In particular, we show that there is a continuum between the usual noise based penalty and a sup norm type one allowing a *sharp* oracle inequality.

2 Framework and Estimate

Recall that we observe

$$\forall i \in \{1, \dots, n\},\ Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function and x_i the fixed grid points. Our only assumption will be on the noise. We do not assume any independence between the coordinates W_i but only that $W = (W_i)_{i \leq n} \in \mathbb{R}^n$ is a centered sub-Gaussian variable. More precisely, we assume that $\mathbb{E}(W) = 0$ and there exists $\sigma^2 \in \mathbb{R}^+$ such that

$$\forall \alpha \in \mathbb{R}^n,\ \mathbb{E}\left[\exp\left(\alpha^\top W\right)\right] \leq \exp\left(\frac{\sigma^2}{2}\|\alpha\|_2^2\right),$$

where $\|.\|_2$ is the usual euclidean norm in $\mathbb{R}^n$. If W is a centered Gaussian vector with covariance matrix Σ, then σ^2 is nothing but the largest eigenvalue of Σ.

The quality of our estimate will be measured through its error at the design points. More precisely, we will consider the classical euclidean loss, related to the squared norm

$$\|g\|_2^2 = \sum_{i=1}^{n} g(x_i)^2.$$

Thus, our unknown is the vector $(f_0(x_i))_{i=1}^n$ rather than the function f_0.

As announced, we will consider affine estimators $\hat{f}_t(Y) = A_t(Y - b) + b + b_t$ corresponding to affine smoothed projection.

We will assume that

$$\hat{f}_t(Y) = A_t(Y - b) + b + b_t = \sum_{i=1}^{n} \rho_{t,i}\langle Y - b, g_{t,i}\rangle g_{t,i} + b + b_t$$

where $(g_{t,i})_{i=1}^n$ is an orthonormal basis, $(\rho_{t,i})_{i=1}^n$ a sequence of non-negative real numbers, and $b_t \in \mathbb{R}^n$. By construction, A_t is thus a symmetric positive semi-definite real matrix. We assume furthermore that the matrix collection $\{A_t\}_{t \in \mathscr{T}}$ is such that $\sup_{t \in \mathscr{T}} \|A_t\|_2 \leq 1$. For the sake of simplicity, we only use the notation $\hat{f}_t(Y) = A_t(Y - b) + b + b_t$ in the following.

To define our estimate from the collection $\{\hat{f}_t(Y) = A_t Y + b_t | A_t \in \mathscr{S}_n^+(\mathbb{R}), b_t \in \mathbb{R}^n, t \in \mathscr{T}\}$, we specify the estimate $\widetilde{r}_t$ of the (penalized) risk of the estimator $\hat{f}_t(Y)$, choose a prior probability measure π over $\mathscr{T}$ and a temperature $\beta > 0$. We define the exponentially weighted measure ρ_{EWA}, a probability measure over $\mathscr{T}$, by

$$d\rho_{EWA}(t) = \frac{\exp\left(-\frac{1}{\beta}\widetilde{r}_t\right)}{\int \exp\left(-\frac{1}{\beta}\widetilde{r}_{t'}\right) d\pi(t')} d\pi(t)$$

and the exponentially weighted aggregate f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho_{EWA}(t)$. If $\mathscr{T}$ is countable, then

$$f_{EWA} = \sum_{t \in \mathscr{T}} \frac{e^{-\widetilde{r}_t/\beta}\pi_t}{\sum_{t' \in \mathscr{T}} e^{-\widetilde{r}_{t'}/\beta}\pi_{t'}} \hat{f}_t.$$

This construction naturally favors low risk estimates. When the temperature goes to zero, this estimator becomes very similar to the one minimizing the risk estimate while it becomes an indiscriminate average when β grows to infinity. The choice of the temperature appears thus to be crucial and a low temperature seems to be desirable.

Our choice for the risk estimate $\widetilde{r}_t$ is to use the classical Stein unbiased estimate, which is sufficient to obtain optimal oracle inequalities in expectation,

$$r_t = \|Y - \hat{f}_t(Y)\|_2^2 + 2\sigma^2 \mathrm{Tr}(A_t) - n\sigma^2$$

and add a penalty pen(t). We will consider simultaneously the case of a penalty independent of f_0 and the one where the penalty may depend on an upper bound of (kind of) sup norm.

More precisely, we allow the use, at least in the analysis, of an upper bound $\widetilde{\|f_0 - b\|_\infty}$ which can be thought as the supremum of the sup norm of the coefficients of f_0 in any basis appearing in $\mathscr{T}$. Indeed, we define $\widetilde{\|f_0 - b\|_\infty}$ as the smallest non-negative real number C such that for any $t \in \mathscr{T}$,

$$\|A_t(f_0 - b)\|_2^2 \leq C^2 \mathrm{Tr}(A_t^2).$$

By construction, $\widetilde{\|f_0 - b\|_\infty}$ is smaller than the sup norm of any coefficients of $f_0 - b$ in any basis appearing in the collection of estimators. Note that $\widetilde{\|f_0 - b\|_\infty}$ can also be upper bounded by $\|f_0 - b\|_1$, $\|f_0 - b\|_2$ or $\sqrt{n}\|f_0 - b\|_\infty$ where the ℓ_1 and sup norm can be taken in any basis.

Our aim is to obtain sufficient conditions on the penalty pen(t) and the temperature β so that an oracle inequality of type

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} (1+\epsilon) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + (1+\epsilon') \left(\int \mathrm{price}(t) d\mu(t) + 2\beta \mathrm{KL}(\mu, \pi) + \beta \ln \frac{1}{\eta} \right)$$

holds either in probability or in expectation. Here, ϵ and ϵ' are some small non-negative numbers possibly equal to 0 and price(t) a loss depending on the choice of pen(t) and β. When $\mathscr{T}$ is countable, such an oracle proves that the risk of

our aggregate estimate is of the same order as the one of the best estimate in the collection as it implies

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{t\in\mathscr{T}} \left\{ (1+\epsilon)\|f_0 - \hat{f}_t\|_2^2 + (1+\epsilon') \left(\text{price}(t) + \beta \ln \frac{1}{\pi(t)^2 \eta} \right) \right\}.$$

Before stating our more general result, which is in Sect. 4, we provide a comparison with some similar results in the literature on the countable $\mathscr{T}$ setting.

3 Penalization Strategies and Preliminary Results

The most similar result in the literature is the one from [16] which holds under a Gaussian white noise assumption and uses a penalty proportional to the known variance σ^2:

Proposition 3.1 ([16]) *If* $\text{pen}(t) = 2\sigma^2 \text{Tr}(A_t)$, *and* $\beta \geq 4\sigma^2 16$, *then for all* $\eta > 0$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \min_t \left\{ \left(1 + \frac{128\sigma^2}{3\beta}\right) \|f_0 - \hat{f}_t\|^2 + 8\sigma^2 \text{Tr}(A_t) + 3\beta \ln \frac{1}{\pi_t} + 3\beta \ln \frac{1}{\eta} \right\}.$$

Our result generalizes this result to the non-necessarily independent sub-Gaussian noise. We obtain

Proposition 3.2 *If* $\beta \geq 20\sigma^2$, *there exists* $\gamma \in [0, 1/2)$, *such that if* $\text{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2} \text{Tr}(A_t^2)\sigma^2$, *for any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \left(1 + \frac{4\gamma}{1-2\gamma}\right) \|f_0 - \hat{f}_t\|^2 + \left(1 + \frac{2\gamma}{1-2\gamma}\right) \left(\text{pen}(t) + 2\sigma^2 \text{Tr}(A_t) + 2\beta \ln \frac{1}{\pi_t} + \beta \ln \frac{1}{\eta} \right) \right\}.$$

The parameter γ is explicit and satisfies $\epsilon = O(\frac{\sigma^2}{\beta})$. We recover thus a similar weak oracle inequality under a weaker assumption on the noise. It should be noted that [4] obtains a sharp oracle inequality for a slightly different aggregation procedure but only under the very strong assumption that $\text{Tr}(A_t) \leq \ln \frac{1}{\pi(t)}$.

Following [33], a lower bound on the penalty that involves the sup norm of f_0, can be given. In that case, the oracle inequality is sharp as $\epsilon = \epsilon' = 0$. Furthermore, the parameter γ is not necessary and the minimum temperature is lower.

Proposition 3.3 *If $\beta > 4\sigma^2$, and*

$$\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2}\left(\sigma^2 \mathrm{Tr}(A_t^2) + 2\left[\|\widetilde{f_0 - b}\|_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right),$$

then for any $\eta > 0$, with probability at least $1 - \eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \|f_0 - \hat{f}_t\|^2 + 2\sigma^2 \mathrm{Tr}(A_t) + \frac{8\sigma^2}{\beta - 4\sigma^2}\left[\|\widetilde{f_0 - b}\|_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right] + \mathrm{pen}(t) + 2\beta \ln\frac{1}{\pi_t} + \beta \ln\frac{1}{\eta}\right\}.$$

We are now ready to state the central result of this contribution, which gives an explicit expression for γ and introduce an optimization parameter $\nu > 0$, from which this theorem can be deduced.

4 A General Oracle Inequality

We consider now the general case for which $\mathscr{T}$ is not necessarily countable. Recall that we have defined the exponentially weighted measure ρ_{EWA}, a probability measure over $\mathscr{T}$, by

$$d\rho_{EWA}(t) = \frac{\exp\left(-\frac{1}{\beta}\widetilde{r}_t\right)}{\int \exp\left(-\frac{1}{\beta}\widetilde{r}_{t'}\right) d\pi(t')} d\pi(t)$$

and the exponentially weighted aggregate f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho_{EWA}(t)$. Propositions 3.2 and 3.3 will be obtained as straightforward corollaries.

Our main contribution is the following two similar theorems:

Theorem 4.1 *For any $\beta \geq 20\sigma^2$, let*

$$\gamma = \frac{\beta - 12\sigma^2 - \sqrt{\beta - 4\sigma^2}\sqrt{\beta - 20\sigma^2}}{16\sigma^2}.$$

If for any $t \in \mathscr{T}$,

$$\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2}\sigma^2 \mathrm{Tr}(A_t^2),$$

then

- *for any* $\eta \in (0,1]$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} \left(1 + \frac{4\gamma}{1-2\gamma}\right) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \left(1 + \frac{2\gamma}{1-2\gamma}\right) \int \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) d\mu(t) + \beta \left(1 + \frac{2\gamma}{1-2\gamma}\right) \left(2\mathrm{KL}(\mu,\pi) + \ln \frac{1}{\eta}\right).$$

- *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} \left(1 + \frac{4\gamma}{1-2\gamma}\right) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \left(1 + \frac{2\gamma}{1-2\gamma}\right) \int \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) d\mu(t) + 2\beta \left(1 + \frac{2\gamma}{1-2\gamma}\right) \mathrm{KL}(\mu,\pi).$$

and

Theorem 4.2 *For any* $\delta \in [0,1]$, *if* $\beta > 4\sigma^2$, *If for any* $t \in \mathscr{T}$,

$$\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2} \left(\sigma^2 \mathrm{Tr}(A_t^2) + 2\left[\|\widetilde{f_0 - b}\|_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right),$$

then

- *for any* $\eta \in (0,1]$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) + \frac{8\sigma^2}{\beta - 4\sigma^2} \left[\|\widetilde{f_0 - b}\|_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right] d\mu(t) + \beta \left(2\mathrm{KL}(\mu,\pi) + \ln \frac{1}{\eta}\right).$$

- *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} \left(1 + \frac{4\gamma}{1-2\gamma}\right) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) + \frac{8\sigma^2}{\beta - 4\sigma^2} \left[\|\widetilde{f_0 - b}\|_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right] d\mu(t) + 2\beta \mathrm{KL}(\mu,\pi).$$

When $\mathscr{T}$ is discrete, one can replace the minimization over all the probability measures in $\mathscr{M}_+^1(\mathscr{T})$ by the minimization overall all Dirac measures δ_{f_t} with $t \in \mathscr{T}$. Propositions 3.2 and 3.3 are then straightforward corollaries. Note that the result in expectation is obtained with the same penalty, which is known not to be necessary, at least in the Gaussian case, as shown by Dalalyan and Salmon [20].

If we assume the penalty is given

$$\mathrm{pen}(t) = \kappa \mathrm{Tr}(A_t^2)\sigma^2,$$

one can rewrite the assumption in terms of κ. The weak oracle inequality holds for any temperature greater than $20\sigma^2$ as soon as $\kappa \geq \frac{4\sigma^2}{\beta - 4\sigma^2}$. While an exact oracle inequality holds for any vector f_0 and any temperature β greater than $4\sigma^2$ as soon as

$$\frac{\beta - 4\sigma^2}{4\sigma^2}\kappa - 1 \geq \frac{\|\widetilde{f_0 - b}\|_\infty^2 + \|b_t\|^2/\mathrm{Tr}(A_t^2)}{\sigma^2}.$$

For fixed κ and β, this corresponds to a low peak signal to noise ratio $\frac{\|\widetilde{f_0 - b}\|_\infty^2}{\sigma^2}$ up to the $\|b_t\|^2$ term which vanishes when $b_t = 0$. Note that similar results hold for a penalization scheme but with much larger constants and some logarithmic factor in n.

Finally, the minimal temperature of $20\sigma^2$ can be replaced by some smaller value if one further restricts the smoothed projections used. As it appears in the proof, the temperature can be replaced by $8\sigma^2$ or even $6\sigma^2$ when the smoothed projections are, respectively, classical projections and projections on the same basis. The question of the minimality of such temperature is still open. Note that in this proof, there is no loss due to the sub-Gaussianity assumption, since the same upper bound on the exponential moment of the deviation as in the Gaussian case is found, providing the same penalty and bound on temperature.

The two results can be combined in a single one producing weak oracle inequalities for a wider range of temperatures than Theorem 4.1. Our proof is available in an extended version of this contribution in which, we prove that a continuum between those two cases exists: a weak oracle inequality, with smaller leading constant than the one of Theorem 4.1, holds as soon as there exists $\delta \in [0, 1)$ such that $\beta \geq 4\sigma^2(1 + 4\delta)$ and

$$\frac{\beta - 4\sigma^2}{4\sigma^2}\kappa - 1 \geq (1 - \delta)(1 + 2\gamma)^2 \frac{\|\widetilde{f_0 - b}\|_\infty^2 + \|b_t\|^2/\mathrm{Tr}(A_t^2)}{\sigma^2},$$

where the signal to noise ratio guides the transition. The temperature required remains nevertheless always above $4\sigma^2$. The convex combination parameter δ measures the account for signal to noise ratio in the penalty.

Note that in practice, the temperature can often be chosen smaller. It is an open question whether the $4\sigma^2$ limit is an artifact of the proof or a real lower bound. In the Gaussian case, [32] have been able to show that this is mainly technical. Extending this result to our setting is still an open challenge.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* (pp. 267–281). Budapest: Akadémiai Kiadó.
2. Alquier, P., & Lounici, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics, 5*, 127–145.
3. Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*(7), 1545–1588.
4. Bellec, P. C. (2018). Optimal bounds for aggregation of affine estimators. *The Annals of Statistics, 46*(1), 30–59.
5. Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika, 98*(4), 791–806.
6. Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research, 13*, 1063–1095.
7. Biau, G., & Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis, 101*(10), 2499–2518.
8. Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research, 9*, 2015–2033.
9. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
10. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
11. Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Vol. 1851. Lecture notes in mathematics*. Berlin: Springer. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, 8–25 July 2001.
12. Catoni, O. (2007). *Pac-Bayesian supervised classification: The thermodynamics of statistical learning: Vol. 56. Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Beachwood, OH: Institute of Mathematical Statistics.
13. Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., & Warmuth M. K. (1997). How to use expert advice. *Journal of the ACM, 44*(3), 427–485.
14. Cesa-Bianchi, N., & Lugosi, G. (1999). On prediction of individual sequences. *The Annals of Statistics, 27*(6), 1865–1895.
15. Chernousova, E., Golubev, Y., & Krymova, E. (2013). Ordered smoothers with exponential weighting. *Electronic Journal of Statistics, 7*, 2395–2419.
16. Dai, D., Rigollet, P., Xia, L., & Zhang, T. (2014). Aggregation of affine estimators. *Electronic Journal of Statistics, 8*, 302–327.
17. Dai, D., Rigollet, P., & Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics, 40*(3), 1878–1905.
18. Dalalyan, A. S. (2012). SOCP based variance free Dantzig selector with application to robust estimation. *Comptes Rendus Mathematique Academie des Sciences, Paris, 350*(15–16), 785–788.
19. Dalalyan, A. S., Hebiri,M., Meziani, K., & Salmon, J. (2013). Learning heteroscedastic models by convex programming under group sparsity. *Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, 28*(3), 379–387.

20. Dalalyan, A. S., & Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics, 40*(4), 2327–2355.
21. Dalalyan, A. S., & Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In N. H. Bshouty & C. Gentile (Eds.), *Learning theory: Vol. 4539. Lecture notes in computer science* (pp. 97–111). Berlin: Springer.
22. Dalalyan, A. S., & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning, 72*(1–2), 39–61.
23. Dalalyan, A. S., & Tsybakov, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences, 78*(5), 1423–1443.
24. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., & Picard D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society Series B, 57*(2), 301–369.
25. Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation, 121*(2), 256–285.
26. Gasser, T., & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics, 11*(3), 171–185.
27. Genuer, R. (2011). *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. PhD thesis, Université Paris-Sud.
28. Gerchinovitz, S. (2011). *Prediction of Individual Sequences and Prediction in the Statistical Framework : Some Links Around Sparse Regression and Aggregation Techniques*. Thesis, Université Paris Sud.
29. Giraud, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli, 14*(4), 1089–1107.
30. Giraud, C., Huet, S., & Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statistical Science, 27*(4), 500–518.
31. Golubev, Y. (2012). Exponential weighting and oracle inequalities for projection estimates. *Problems of Information Transmission, 48*, 269–280.
32. Golubev, Y., & Ostobski, D. (2014). Concentration inequalities for the exponential weighting method. *Mathematical Methods of Statistics, 23*(1), 20–37.
33. Guedj, B., & Alquier, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics, 7*, 264–291.
34. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction. Springer series in statistics* (2nd ed.). New York: Springer.
35. Lecué, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli, 13*(4), 1000–1022.
36. Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory, 52*(8), 3396–3410.
37. Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation, 108*(2), 212–261.
38. Lounici, K. (2007). Generalized mirror averaging and *D*-convex aggregation. *Mathematical Methods of Statistics, 16*(3), 246–259.
39. Nadaraya, É. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications, 10*(1), 186–190.
40. Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998): Vol. 1738. Lecture notes in mathematics* (pp. 85–277). Berlin: Springer.
41. Rigollet, P. (2006). *Inégalités d'oracle, agrégration et adaptation*. PhD thesis, Université Pierre et Marie Curie- Paris VI.
42. Rigollet, P., & Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics, 16*(3), 260–280.
43. Rigollet, P., & Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science, 27*(4), 558–575.
44. Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227.
45. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

46. Sun, T., & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika, 99*(4), 879–898.
47. Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.
48. Tsybakov, A. B. (2003). Optimal rates of aggregation. In B. Schölkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines: Vol. 2777. Lecture notes in computer science* (pp. 303–313). Berlin/Heidelberg: Springer.
49. Tsybakov, A. B. (2008). Agrégation d'estimateurs et optimisation stochastique. *Journal de la Société Franēaise de Statistique & Review of Statistics and Its Application, 149*(1), 3–26.
50. Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT '90* (pp. 371–386). San Francisco, CA: Morgan Kaufmann Publishers Inc.
51. Wahba, G. (1990). *Spline models for observational data: Vol. 59. CBMS-NSF regional conference series in applied mathematics*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
52. Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A, 26*, 359–372.
53. Yang, Y. (2000). Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica, 10*(4), 1069–1089.
54. Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis, 74*(1), 135–161.
55. Yang, Y. (2000). Mixing strategies for density estimation. *The Annals of Statistics, 28*(1), 75–87.
56. Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association, 96*(454), 574–588.
57. Yang, Y. (2003). Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica, 13*(3), 783–809.
58. Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli, 10*(1), 25–47.
59. Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory, 20*(1), 176–222.
60. Zou, H., & Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*(2), 301–320.

Light- and Heavy-Tailed Density Estimation by Gamma-Weibull Kernel

L. Markovich

Abstract In our previous papers we focus on the gamma kernel estimators of density and its derivatives on positive semi-axis by dependent data by univariate and multivariate samples. We introduce the gamma product kernel estimators for the multivariate joint probability density function (pdf) with the nonnegative support and its partial derivatives by the multivariate dependent data with a strong mixing. The asymptotic behavior of the estimates and the optimal bandwidths in the sense of minimal mean integrated squared error (MISE) are obtained. However, it is impossible to fit accurately the tail of the heavy-tailed density by pure gamma kernel. Therefore, we construct the new kernel estimator as a combination of the asymmetric gamma and Weibull kernels, i.e. Gamma-Weibull kernel. The gamma kernel is nonnegative and it changes the shape depending on the position on the semi-axis and possesses good boundary properties for a wide class of densities. Thus, we use it to estimate the pdf near the zero boundary. The Weibull kernel is based on the Weibull distribution which can be heavy-tailed and hence, we use it to estimate the tail of the unknown pdf. The theoretical asymptotic properties of the proposed density estimator like the bias and the variance are derived. We obtain the optimal bandwidth selection for the estimate as a minimum of the MISE. The optimal rate of convergence of the MISE for the density is found.

L. Markovich (✉)
Moscow Institute of Physics and Technology , Dolgoprudny, Moscow Region, Russia

Institute for Information Transmission Problems, Moscow, Russia

V. A. Trapeznikov Institute of Control Sciences, Moscow, Russia
e-mail: kimo1@mail.ru

P. Bertail et al. (eds.), *Nonparametric Statistics*, Springer Proceedings
in Mathematics & Statistics 250, https://doi.org/10.1007/978-3-319-96941-1_10