Sharp Oracle Inequalities for Aggregation of Affine Estimators

Arnak S. Dalalyan, Joseph Salmon

SHARP ORACLE INEQUALITIES FOR AGGREGATION OF AFFINE ESTIMATORS

ARNAK S. DALALYAN AND JOSEPH SALMON

ABSTRACT. We consider the problem of combining a (possibly uncountably infinite) set of affine estimators in non-parametric regression model with heteroscedastic Gaussian noise. Focusing on the exponentially weighted aggregate, we prove a PAC-Bayesian type inequality that leads to sharp oracle inequalities in discrete but also in continuous settings. The framework is general enough to cover the combinations of various procedures such as least square regression, kernel ridge regression, shrinking estimators and many other estimators used in the literature on statistical inverse problems. As a consequence, we show that the proposed aggregate provides an adaptive estimator in the exact minimax sense without neither discretizing the range of tuning parameters nor splitting the set of observations. We also illustrate numerically the good performance achieved by the exponentially weighted aggregate.

1. INTRODUCTION

There is a growing empirical evidence of superiority of aggregated statistical procedures, also referred to as *blending*, *stacked generalization*, or *ensemble methods*, with respect to "pure" ones. Since their introduction in the 1990's, famous aggregation procedures such as *Boosting* [30], *Bagging* [9] or *Random Forest* [2] have been successfully used in practice for a large variety of applications. Moreover, most recent Machine Learning competitions such as Pascal VOC or Netflix challenge have been won by procedures combining different types of classifiers / predictors / estimators. It is therefore of central interest to understand from a theoretical point of view what kind of aggregation strategies should be used for getting the best possible combination of the available statistical procedures.

1.1. Historical remarks and motivation.

In the statistical literature, to the best of our knowledge, the lecture notes of Nemirovski [48] was the first work concerned by the theoretical analysis of aggregation procedures. It was followed by a paper by Juditsky and Nemirovski [38], as well as by a series of papers by Catoni (see [13] for a comprehensive account) and Yang [60, 61, 62]. For the regression model, a significant progress has been achieved by Tsybakov [58] with introducing the notion of optimal rates of aggregation and proposing aggregation-rate-optimal procedures for the tasks of linear, convex and model selection aggregation. This point has been further developed by [11, 43, 46, 51], especially in the context of high dimension with sparsity constraints.

From a practical point of view, an important limitation of the previously cited results on aggregation is that they are valid under the assumption that the aggregated procedures are deterministic (or random, but independent of the data used for the aggregation). In the Gaussian sequence model, a breakthrough has been reached by Leung and Barron [45]. Building on very elegant but not very well known results by George [31], they established sharp oracle inequalities for the exponentially weighted aggregate (EWA) under the condition that the aggregated estimators are obtained from the data vector by orthogonally projecting it on some linear subspaces. Dalalyan and Tsybakov [21, 22] have shown that the result of [45] remains valid under more general (non Gaussian) noise distributions and when the constituent estimators are independent of the data

used for the aggregation. A natural question arises whether a similar result can be proved for a larger family of constituent estimators containing projection estimators and deterministic ones as specific examples. The main aim of the present paper is to answer this question by considering families of affine estimators.

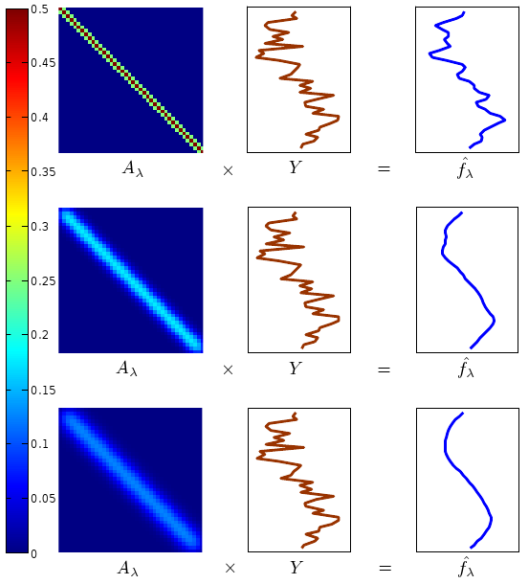


FIGURE 1. The effect of the smoothing matrix A_λ on the resulting estimator. In this example, the true signal is the sine function over $[-\pi,\pi]$ and the three matrices represented in the leftmost column are some powers of one convolution matrix. Large powers correspond to stronger smoothing. One clearly see that the third matrix leads to an almost perfect recovery of the original signal.

Our interest in affine estimators is motivated by several reasons. First of all, affine estimators encompass many popular estimators such as smoothing splines, the Pinsker estimator [28, 49], local polynomial estimators, non-local means [10, 53], etc. For instance, it is known that if the underlying (unobserved) signal belongs to a Sobolev ball, then the (linear) Pinsker estimator is asymptotically minimax up to the optimal constant, while the best projection estimator is only rate-minimax. A second motivation is that—as proved by Juditsky and Nemirovski [39]—the set of signals that are well estimated by linear estimators is very rich. It contains, for instance, sampled smooth functions, sampled modulated smooth functions and sampled harmonic functions. One can add to this set the family of piecewise constant functions as well, as demonstrated in [50] with natural application in magnetic resonance imaging. It is worth noting that oracle inequalities for penalized empirical risk minimizer has also been established by Golubev [37], and for model selection by Arlot and Bach [3], Baraud et al. [7].

In the present work, we establish sharp oracle inequalities in the statistical model of heteroscedastic regression, under various conditions on the constituent estimators assumed to be affine functions of the data. Our results provide theoretical guarantees of optimality, in terms of the expected loss, for the exponentially weighted aggregate. They have the advantage of covering in a unified fashion the particular cases of frozen estimators considered by Dalalyan and Tsybakov [22] and of projection estimators treated by Leung and Barron [45].

We will mainly focus on the theoretical guarantees expressed in terms of oracle inequalities for the expected squared loss. Interestingly, despite the fact that several recent papers [3, 7, 20, 35] discuss the paradigm of competing against the best linear procedure from a given family, none of them provide oracle inequalities with leading constant equal to one. Furthermore, most existing results involve some constants depending on different parameters of the setup. In contrast, the oracle inequality that we prove herein is with leading constant one and admits a very simple formulation. It is established for a (suitably symmetrized, if necessary) exponentially weighted aggregate [13, 23, 31] with an arbitrary prior and a temperature parameter which is not too small. The result is completely nonasymptotic and leads to asymptotically optimal residual term in the case where the sample size, as well as the cardinality of the family of competitors, tend to infinity. In its general form, the residual term is similar to those obtained in PAC-Bayes setting [42, 47, 54] in that it is proportional to the Kullback-Leibler divergence between two probability distributions.

Note also that the problem of competing against the best procedure in a given family has been extensively studied in the context of online learning and prediction with expert advice [18, 19, 40]. A remarkable connection between the results on online learning and the statistical oracle inequalities has been recently established by [33].

1.2. **Notation.** Throughout this work, we focus on the heteroscedastic regression model with Gaussian additive noise. More precisely, we assume that we are given a vector $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ obeying the model:

$$y_i = f_i + \xi_i, \quad \text{for } i = 1, \dots, n, \tag{1.1}$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$ is a centered Gaussian random vector, $f_i = \mathbf{f}(x_i)$ where $\mathbf{f}$ is an unknown function $\mathscr{X} \to \mathbb{R}$ and $x_1, \dots, x_n \in \mathscr{X}$ are deterministic points. Here, no assumption is made on the set $\mathscr{X}$. Our objective is to recover the vector $\boldsymbol{f} = (f_1, \dots, f_n)$, often referred to as *signal*, based on the data $y_1, \dots, y_n$. In our work, the noise covariance matrix $\Sigma = \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top]$ is assumed to be finite with a known upper bound on its spectral norm. We measure the performance of an estimator $\hat{\boldsymbol{f}}$ by its expected empirical quadratic loss: $r = \mathbb{E}(\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2)$ where $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f_i - \hat{f}_i)^2$. We also denote by $\langle \cdot | \cdot \rangle_n$ the corresponding empirical inner product.

Let us describe now different families of linear and affine estimators successfully used in the statistical literature. Our results apply to all these families leading to a procedure that behaves nearly as well as the best one of the family.

Ordinary least squares: Let $\{\mathscr{S}_\lambda : \lambda \in \Lambda\}$ be a set of linear subspaces of $\mathbb{R}^n$. A well known family of affine estimators, successfully used in the context of model selection [8], is the set of orthogonal projections onto $\mathscr{S}_\lambda$. In the case of a family of linear regression models with design matrices X_λ, one has $A_\lambda = X_\lambda (X_\lambda^\top X_\lambda)^+ X_\lambda^\top$, where $(X_\lambda^\top X_\lambda)^+$ stands for the Moore-Penrose pseudoinverse of $X_\lambda^\top X_\lambda$.

Diagonal filters: Another set of common estimators are the so called diagonal filters $\hat{\boldsymbol{f}} = A\boldsymbol{Y}$, where A is a diagonal matrix $A = \mathrm{diag}(a_1, \dots, a_n)$. Popular examples include:

- ✓ Ordered projections : $a_k = \mathbb{1}_{(k \le \lambda)}$ for some integer λ (where $\mathbb{1}_{(\cdot)}$ is the indicator function). Those weights are also called truncated SVD or spectral cut-off. In this case the natural parametrization is $\Lambda = \{1, \dots, n\}$, indexing the number of elements conserved.
- ✓ Block projections: $a_k = \mathbb{1}_{(k \le w_1)} + \sum_{j=1}^{m-1} \lambda_j \mathbb{1}_{(w_j \le k \le w_{j+1})}$, $k = 1, \dots, n$, where $\lambda_j \in \{0, 1\}$. Here the natural parametrization is $\Lambda = \{0,1\}^{m-1}$, indexing subsets of $\{1, m-1\}$.
- ✓ Tikhonov-Philipps filter: $a_k = \frac{1}{1+(k/w)^\alpha}$, where $w, \alpha > 0$. In this case, $\Lambda = (\mathbb{R}_+^*)^2$, indexing continuously the smoothing parameters.

✓ Pinsker filter: $a_k = \left(1 - \frac{k^\alpha}{w}\right)_+$, where $x_+ = \max(x, 0)$ and $w, \alpha > 0$. In this case also $\Lambda = (\mathbb{R}_+^*)^2$.

Kernel ridge regression: Assume that we have a positive definite kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and we aim at estimating the true function f in the associated reproducing kernel Hilbert space $(\mathscr{H}_k, \|\cdot\|_k)$. The kernel ridge estimator is obtained by minimizing the criterion $\|\boldsymbol{Y} - \boldsymbol{f}\|_n^2 + \lambda \|\boldsymbol{f}\|_k^2$ w.r.t. $f \in \mathscr{H}_k$ (see [55, page 118]). Denoting by K the $n \times n$ kernel-matrix with element $K_{i,j} = k(x_i, x_j)$, the unique solution $\hat{\boldsymbol{f}}$ is a linear estimate of the data, $\hat{\boldsymbol{f}} = A_\lambda \boldsymbol{Y}$, with $A_\lambda = K(K + n\lambda I_{n\times n})^{-1}$, where $I_{n\times n}$ is the identity matrix of size $n \times n$.

Multiple Kernel learning: As described in [3], it is also possible to handle the case of several kernels $k_1, \dots, k_M$, with associated positive definite matrices $K_1, \dots, K_M$. For a parameter $\lambda = (\lambda_1, \dots, \lambda_M) \in \Lambda = \mathbb{R}_+^M$ one can define the estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y}$ with

$$A_\lambda = \Big(\sum_{m=1}^M \lambda_m K_m \Big) \Big(\sum_{m=1}^M \lambda_m K_m + n I_{n\times n} \Big)^{-1}. \tag{1.2}$$

It is worth mentioning that the formulation in Eq.(1.2) can be linked to the group Lasso [63] and to the multiple kernel introduced in [41] — see [3, 6] for more details.

Moving averages: If we think of coordinates of $\boldsymbol{f}$ as some values assigned to the vertices of an undirected graph, satisfying the property that two nodes are connected if the corresponding values of $\boldsymbol{f}$ are close, then it is natural to estimate f_i by averaging out the values Y_j for indices j that are connected to i. The resulting estimator is a linear one with a matrice $A = (a_{ij})_{i,j=1}^n$ such that $a_{ij} = \mathbb{1}_{V_i}(j)/n_i$, where V_i is the set of neighbors of the node i in the graph and n_i is the cardinality of V_i.

Non-local means: In recent years, a signal denoising method—termed non-local means (NLM)—has become quite popular in image processing [10]. This method removes the noise by exploiting the signal self-similarities and has been shown to be tied in with the exponentially weighted aggregate [53]. We briefly define the NLM procedure in the case of one-dimensional signals.

Assume that a vector $\boldsymbol{Y} = (y_1, \dots, y_n)$ given by (1.1) is observed with $f_i = F(i/n)$, $i = 1, \dots, n$, for some function $F : [0, 1] \to \mathbb{R}$. For a fixed "patch-size" $k \in \{1, \dots, n\}$, let us define $\boldsymbol{f}_{[i]} = (f_i, f_{i+1}, \dots, f_{i+k-1})$ and $\boldsymbol{Y}_{[i]} = (y_i, y_{i+1}, \dots, y_{i+k-1})$ for every $i = 1, \dots, n-k+1$. The vectors $\boldsymbol{f}_{[i]}$ and $\boldsymbol{Y}_{[i]}$ are respectively called *true patch* and *noisy patch*. The NLM consists in regarding the noisy patches $\boldsymbol{Y}_{[i]}$ as constituent estimators for estimating the true patch $\boldsymbol{f}_{[i_0]}$ by applying the EWA. One easily checks that the constituent estimators $\boldsymbol{Y}_{[i]}$ are affine in $\boldsymbol{Y}_{[i_0]}$, that is $\boldsymbol{Y}_{[i]} = A_i \boldsymbol{Y}_{[i_0]} + \boldsymbol{b}_i$ with A_i and $\boldsymbol{b}_i$ independent of $\boldsymbol{Y}_{[i_0]}$. Indeed, if the distance between i and i_0 is larger than k, then $\boldsymbol{Y}_{[i]}$ is independent of $\boldsymbol{Y}_{[i_0]}$ and, therefore, $A_i = 0$ and $\boldsymbol{b}_i = \boldsymbol{Y}_{[i]}$. If $|i - i_0| < k$, then the matrix A_i is a suitably chosen shift matrix and $\boldsymbol{b}_i$ is the projection of $\boldsymbol{Y}_{[i]}$ onto the orthogonal complement of the image of A_i.

1.3. **Organization of the paper.** In Section 2, we introduce EWA and state a PAC-Bayes type bound in expectation assessing optimality properties of EWA in combining affine estimators. The extension of these results to the case of a grouped aggregation—in relation with the ill-posed inverse problems—is discussed in Section 3. As a consequence, we provide in Section 4 sharp oracle inequalities in various set-ups: ranging from finite to continuous families of constituent estimators and including the sparsity scenario. In Section 5, we apply our main results to prove that combining Pinsker's type filters with EWA leads to an asymptotically sharp adaptive procedure over Sobolev ellipsoids. Section 6 is devoted to numerical comparison of EWA with other classical filters (soft thresholding, blockwise shrinking, etc.), and illustrates the potential benefits

of aggregating. Conclusion is presented in Section 7, while technical proofs are postponed to the Appendix.

2. Aggregation of estimators: main results

In this section we describe the statistical framework for aggregating estimators and we also introduce the exponentially weighted aggregate. The task of aggregation consists in estimating $\boldsymbol{f}$ by a suitable combination of the elements of a family of *constituent estimators* $\mathscr{F}_\Lambda = (\hat{\boldsymbol{f}}_\lambda)_{\lambda\in\Lambda} \in \mathbb{R}^n$. The target objective of the aggregation is to build an aggregate $\hat{\boldsymbol{f}}_{\text{aggr}}$ that mimics the performance of the best constituent estimator, called *oracle* (because of its dependence on the unknown function $\mathbf{f}$). In what follows, we assume that Λ is a measurable subset of $\mathbb{R}^M$, for some $M \in \mathbb{N}$.

The theoretical tool commonly used for evaluating the quality of an aggregation procedure is the oracle inequality (OI), generally written in the following form:

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\text{aggr}} - \boldsymbol{f}\|_n^2 \le C_n \inf_{\lambda\in\Lambda}\left(\mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\right) + R_n, \tag{2.1}$$

with *residual* term R_n tending to zero, and *leading constant* C_n being bounded. The OIs with leading constant one are of central theoretical interest since they allow to bound the excess risk and to assess the aggregation-rate-optimality.

2.1. Exponentially Weighted Aggregate (EWA).

Let $r_\lambda = \mathbb{E}(\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2)$ denote the risk of the estimator $\hat{\boldsymbol{f}}_\lambda$, for any $\lambda \in \Lambda$, and let $\hat{r}_\lambda$ be an estimator of r_λ. The precise form of $\hat{r}_\lambda$ strongly depends on the nature of the constituent estimators. For any probability distribution π over the set Λ and for any $\beta > 0$, we define the probability measure of exponential weights, $\hat{\pi}$, by the following formula:

$$\hat{\pi}(d\lambda) = \theta(\lambda)\pi(d\lambda) \qquad \text{with} \qquad \theta(\lambda) = \frac{\exp(-n\hat{r}_\lambda/\beta)}{\int_\Lambda \exp(-n\hat{r}_\omega/\beta)\pi(d\omega)}. \tag{2.2}$$

The corresponding exponentially weighted aggregate, henceforth denoted by $\hat{\boldsymbol{f}}_{\text{EWA}}$, is the expectation of the $\hat{\boldsymbol{f}}_\lambda$ w.r.t. the probability measure $\hat{\pi}$:

$$\hat{\boldsymbol{f}}_{\text{EWA}} = \int_\Lambda \hat{\boldsymbol{f}}_\lambda \, \hat{\pi}(d\lambda)\,. \tag{2.3}$$

It is convenient and customary to use the terminology of Bayesian statistics: the measure π is called *prior*, the measure $\hat{\pi}$ is called *posterior* and the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ is then the *posterior mean*. The parameter β will be referred to as the *temperature parameter*. In the framework of aggregating statistical procedures, the use of such an aggregate can be traced back to George [31, 32].

The interpretation of the weights $\theta(\lambda)$ is simple: they up-weight estimators all the more that their performance, measured in terms of the risk estimate $\hat{r}_\lambda$, is good. The temperature parameter reflects the confidence we have in this criterion: if the temperature is small ($\beta \approx 0$) the distribution concentrates on the estimators achieving the smallest value for $\hat{r}_\lambda$, assigning almost zero weights to the other estimators. On the other hand, if $\beta \to +\infty$ then the probability distribution over Λ is simply the prior π, and the data do not modify our confidence in the estimators. It should also be noted that averaging w.r.t. the posterior $\hat{\pi}$ is not the only way of constructing an estimator of $\boldsymbol{f}$, some alternative estimators based on $\hat{\pi}$ have been studied, for instance, by Audibert in [4, 5].

2.2. **Main results.** In this paper, we only focus on *affine estimators* $\hat{\boldsymbol{f}}_\lambda$, *i.e.,* estimators that can be written as affine transforms of the data $\boldsymbol{Y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$. Using the convention that all vectors are one-column matrices, affine estimators can be defined by

$$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda, \tag{2.4}$$

where the $n \times n$ real matrix A_λ and the vector $\boldsymbol{b}_\lambda \in \mathbb{R}^n$ are deterministic. This means that the entries of A_λ and $\boldsymbol{b}_\lambda$ may depend on the points $x_1, \dots, x_n$ but not on the data vector $\boldsymbol{Y}$. Let $I_{n\times n}$ denote the identity matrix of size $n \times n$. It is well-known (see Section A for details) that the risk of the estimator (2.4) is given by

$$r_\lambda = \mathbb{E}[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2] = \|(A_\lambda - I_{n\times n})\boldsymbol{f} + \boldsymbol{b}_\lambda\|_n^2 + \frac{\mathrm{Tr}(A_\lambda \Sigma A_\lambda^\top)}{n} \tag{2.5}$$

and that $\hat{r}_\lambda^{\mathrm{unb}}$, defined by

$$\hat{r}_\lambda^{\mathrm{unb}} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 + \frac{2}{n}\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\mathrm{Tr}[\Sigma] \tag{2.6}$$

is an unbiased estimator of r_λ.

To state our main results, we denote by $\mathscr{P}_\Lambda$ the set of all probability measures on Λ and by $\mathscr{K}(p, p')$ the Kullback-Leibler divergence between two probability measures $p, p' \in \mathscr{P}_\Lambda$:

$$\mathscr{K}(p, p') = \begin{cases} \int_\Lambda \log\left(\frac{dp}{dp'}(\lambda)\right) p(d\lambda) & \text{if } p \ll p', \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 1. *Assume that the matrices A_λ are all symmetric and satisfy $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$, $A_\lambda \Sigma + \Sigma A_\lambda \succeq 0$ and $A_{\lambda'} \boldsymbol{b}_\lambda = 0$ for all $\lambda, \lambda' \in \Lambda$. Then, the aggregate $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ defined by Equations (2.2), (2.3) and the unbiased risk estimate $\hat{r}_\lambda = \hat{r}_\lambda^{\mathrm{unb}}$ (2.6) satisfies the inequality*

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{p\in\mathscr{P}_\Lambda} \left(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 \, p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p, \pi) \right) \tag{2.7}$$

provided that $\beta \ge 8|||\Sigma|||$, where $|||\Sigma|||$ stands for the spectral nor of Σ.

Remark 1. *The simplest setting in which all the conditions of Theorem 1 are fulfilled is when the matrices A_λ and Σ are all diagonal, or diagonalizable in a common base.*

This theorem, as we will see in Section 5, allows us to propose a new adaptive estimator, in the exact minimax sense, over the collection of all Sobolev ellipsoids. It also suggests a new method for efficiently combining varying-block-shrinkage shrinkage estimators, as described in Section 4.4.

The result of Theorem 1 applies to the estimator $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ that uses the full knowledge of the covariance matrix Σ. Indeed, even if for the choice of β only an upper bound on the spectral norm of Σ is required, the entire matrix Σ enters in the definition of the unbiased risk $\hat{r}_\lambda^{\mathrm{unb}}$ that is used for defining $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$. We will discuss in Section 7 some extensions of the proposed methodology to the case of unknown Σ.

Theorem 1 gives already satisfactory answers to a certain number of questions; however, it leaves open the issue of aggregating affine estimators defined via non-commuting matrices. For example, the previous results do not allow us to evaluate the MSE of the EWA when each A_λ is a convex or linear combination of a fixed family of projection matrices on non-orthogonal linear subspaces. In order to cover such kind of situations, we develop a theory that recommends to

use the EWA with an adjusted risk estimate:

$$\hat{r}_\lambda^{\mathrm{adj}} = \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 + \frac{2}{n}\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\mathrm{Tr}[\Sigma]}_{\hat{r}_\lambda^{\mathrm{unb}}} + \frac{1}{n}\boldsymbol{Y}^\top(A_\lambda - A_\lambda^2)\boldsymbol{Y}. \tag{2.8}$$

We still assume that A_λ is symmetric and positive semidefinite, as well as that $b_\lambda \in \mathrm{Ker}(A_\lambda)$ for every λ. One can notice that the adjusted risk estimate $\hat{r}_\lambda^{\mathrm{adj}}$ coincides with the unbiased risk estimate $\hat{r}_\lambda^{\mathrm{unb}}$ if and only if the matrix A_λ is an orthogonal projector.

Theorem 2. *If the matrices A_λ are all symmetric with $A_\lambda \preceq I_{n\times n}$ and $\boldsymbol{b}_\lambda \in Ker(A_\lambda)$ for every $\lambda \in \Lambda$, then the aggregate $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ defined by Equations (2.2), (2.3) and the adjusted risk estimate $\hat{r}_\lambda = \hat{r}_\lambda^{\mathrm{adj}}$ (2.8) satisfies the inequality*

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{p\in\mathscr{P}_\Lambda}\Big\{\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\, p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi) + \int_\Lambda\Big(\frac{1}{n}\boldsymbol{f}^\top(A_\lambda - A_\lambda^2)\boldsymbol{f} + \frac{1}{n}\mathrm{Tr}\big[\Sigma(A_\lambda - A_\lambda^2)\big]\Big)p(d\lambda)\Big\}$$

provided that $\beta \ge 4|||\Sigma|||$.

A first observation that one can make is that, in the particular case of a finite collection of projection estimators (*i.e.*, $A_\lambda = A_\lambda^2$ and $\boldsymbol{b}_\lambda = 0$ for every λ) this result reduces to [45, Corollary 6]. Furthermore, Theorem 2 handles the general noise covariances while [45] deals only with i.i.d. Gaussian noise.

An important situation that is covered by Theorem 2 but not by Theorem 1 concerns the case when the signals of interest $\boldsymbol{f}$ are smooth or sparse in a basis $\mathscr{B}_{\mathrm{sig}}$ which is different from the basis $\mathscr{B}_{\mathrm{noise}}$ orthogonalizing the covariance matrix Σ. In such a situation, one may be interested in considering matrices A_λ that are diagonalizable in the basis $\mathscr{B}_{\mathrm{sig}}$ which, in general, do not commute with Σ.

Remark 2. *We decided in this paper to focus on the case of Gaussian errors, in order to put the emphasis on the possibility of efficiently aggregating almost any family of affine estimators without spending time and space on other technical aspects. Most results stated in this section can be generalized to other noise distributions by following the approach developed in [22].*

Remark 3. *An equivalent and, perhaps, more convenient way of writing the risk bound of Theorem 2 is the following:*

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{p\in\mathscr{P}_\Lambda}\Big\{\frac{1}{n}\int_\Lambda\Big(\boldsymbol{f}^\top(I_{n\times n} - A_\lambda)\boldsymbol{f} + \mathrm{Tr}\big[\Sigma A_\lambda\big]\Big)p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi)\Big\}.$$

We opted for not stating Theorem 2 in this form, in order to stress the relation between the risk of the aggregate and those of constituent estimators.

All the results presented so far concern the situation when the matrices A_λ are symmetric. However, using the last theorem it is possible to propose an estimator that is almost as accurate as the best affine estimator $A_\lambda\boldsymbol{Y} + \boldsymbol{b}_\lambda$ even if the matrices A_λ are not symmetric. Interestingly, the estimator enjoying this property is not obtained by aggregating the original estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda\boldsymbol{Y} + \boldsymbol{b}_\lambda$ but the "symmetrized" estimators $\tilde{\boldsymbol{f}}_\lambda = \tilde{A}_\lambda\boldsymbol{Y} + \boldsymbol{b}_\lambda$ where $\tilde{A}_\lambda = A_\lambda + A_\lambda^\top - A_\lambda^\top A_\lambda$. Besides symmetry, an advantage of the matrices $\tilde{A}_\lambda$, as compared to the matrices A_λ, is that they automatically satisfy the contraction condition $\tilde{A}_\lambda \preceq I_{n\times n}$ required by Theorem 2.

Corollary 1. *Let $\{A_\lambda : \lambda \in \Lambda\}$ be any family of $n \times n$ matrices and $\{\boldsymbol{b}_\lambda : \lambda \in \Lambda\}$ be a set of vectors of $\mathbb{R}^n$ satisfying $A_\lambda \boldsymbol{b}_\lambda = A_\lambda^\top \boldsymbol{b}_\lambda = 0$ for every $\lambda \in \Lambda$. Assume in addition that Λ is equiped with a σ-algebra so that the mapping $\lambda \mapsto (A_\lambda, \boldsymbol{b}_\lambda)$ is measurable. Let $\tilde{\boldsymbol{f}}_{\mathrm{EWA}}$ denote the exponentially weighted aggregate of estimators $\tilde{\boldsymbol{f}}_\lambda = (A_\lambda + A_\lambda^\top - A_\lambda^\top A_\lambda)\boldsymbol{Y} + \boldsymbol{b}_\lambda$ with the weights (2.2) defined via the risk estimate $\hat{r}_\lambda^{\mathrm{unb}}$. Then, for every $\beta \geq 4|||\Sigma|||$, it holds that*

$$\mathbb{E}\big[\|\tilde{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big] \leq \inf_{p \in \mathscr{P}_\Lambda} \left\{ \int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big] p(d\lambda) + \frac{\beta}{n} \mathscr{K}(p,\pi) \right\} + \frac{\beta}{n} \log\Big[\int_\Lambda e^{\frac{2}{\beta}\mathrm{Tr}[\Sigma(A_\lambda - A_\lambda^\top A_\lambda)]} \pi(d\lambda) \Big].$$

The proof of this corollary is very simple: it consists in applying Theorem 2 to the affine estimators $\tilde{\boldsymbol{f}}_\lambda$ with the prior $\pi(d\lambda)$ replaced by $e^{\frac{2}{\beta}\mathrm{Tr}[\Sigma(A_\lambda - A_\lambda^\top A_\lambda)]}\pi(d\lambda) / \int_\Lambda e^{\frac{2}{\beta}\mathrm{Tr}[\Sigma(A_w - A_w^\top A_w)]}\pi(dw)$.

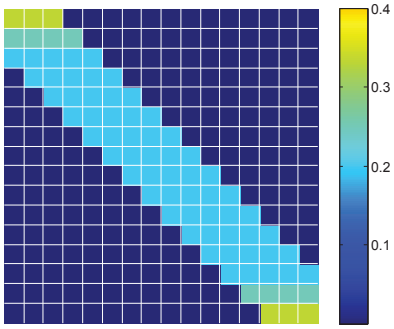
Remark 4. *It follows from Corollary 1 that*

$$\mathbb{E}\big[\|\tilde{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big] \leq \inf_{p \in \mathscr{P}_\Lambda} \left\{ \int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big] p(d\lambda) + \frac{\beta}{n} \mathscr{K}(p,\pi) \right\} \tag{2.9}$$

not only when the prior π is supported by the set of projection matrices, but under more general condition

(C) $$\pi\big\{\lambda \in \Lambda : \mathrm{Tr}(\Sigma A_\lambda) \leq \mathrm{Tr}(\Sigma A_\lambda^\top A_\lambda)\big\} = 1.$$

If the matrix Σ is diagonal, a notable example of linear estimators that satisfy this condition are Nadaraya-Watson estimators with rectangular kernel, also called moving averages or nearest neighbor filters (in the case of a regularly spaced design). Below is a visual illustration of a matrix defining a moving average estimator:



*Under a little bit more stringent assumption of homoscedasticity, i.e., when $\Sigma = \sigma^2 I_{n\times n}$, if the matrices A_λ are such that all the non-zero elements of each row are equal and sum up to one (or a quantity larger than one) then $\mathrm{Tr}(A_\lambda) \leq \mathrm{Tr}(A_\lambda^\top A_\lambda)$ and **(C)** is fulfilled.*
*Another class of matrices for which **(C)** holds true are those having only zeros on the main diagonal.*

3. Ill-posed inverse problems and group-weighting

As explained in [14, 17], the model of heteroscedastic regression is well suited for describing inverse problems. In fact, let T be a known linear operator on some Hilbert space $\mathscr{H}$, equipped with an inner product $\langle \cdot | \cdot \rangle_{\mathscr{H}}$. For some $h \in \mathscr{H}$, let Y be the random process indexed by $g \in \mathscr{H}$ such that

$$Y = Th + \varepsilon\xi \quad \Longleftrightarrow \quad \Big(Y(g) = \big\langle Th | g \big\rangle_{\mathscr{H}} + \varepsilon\xi(g), \quad \forall g \in \mathscr{H}\Big), \tag{3.1}$$

where $\varepsilon > 0$ is the noise magnitude and ξ is the white Gaussian noise on $\mathscr{H}$, *i.e.*, for any $g_1, \dots, g_k \in \mathscr{H}$ the vector $\big(Y(g_1), \dots, Y(g_k)\big)$ is Gaussian with zero mean and covariance matrix $\{\langle g_i | g_j \rangle_{\mathscr{H}}\}$.

Input:: data vector $\boldsymbol{Y} \in \mathbb{R}^n$, $n \times n$ noise covariance matrix Σ and a family of linear smoothers $\{\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y}; \lambda \in \Lambda\}$.
Output:: estimator $\tilde{\boldsymbol{f}}_{\mathrm{EWA}}$ of the true function $\boldsymbol{f}$.
Parameter:: prior probability distribution π on Λ, temperature parameter $\beta > 0$.
Strategy::

(1) For every λ, compute the risk estimate $\hat{r}_\lambda^{\mathrm{unb}} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 + \frac{2}{n}\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\mathrm{Tr}[\Sigma]$.
(2) Define the prob. distribution $\hat{\pi}(d\lambda) = \theta(\lambda)\pi(d\lambda)$ with $\theta(\lambda) \propto \exp(-n\hat{r}_\lambda^{\mathrm{unb}}/\beta)$.
(3) For every λ, build the symmetrized linear smoothers $\tilde{\boldsymbol{f}}_\lambda = (A_\lambda + A_\lambda^\top - A_\lambda^\top A_\lambda)\boldsymbol{Y}$.
(4) Average out the symmetrized smoothers w.r.t. posterior $\hat{\pi}$: $\tilde{\boldsymbol{f}}_{\mathrm{EWA}} = \int_\Lambda \tilde{\boldsymbol{f}}_\lambda \hat{\pi}(d\lambda)$.

FIGURE 2. The symmetrized exponentially weighted aggregation strategy for competing against the best linear smoother in a given family.

The statistical problem is then the following: estimate the element h assuming that the value of Y for any given g can be measured.

It is customary to use as "probe elements" g the eigenvectors of the adjoint of T, denoted by T^*. Suppose that the operator T^*T is compact, then one has the singular value decomposition

$$T\phi_k = b_k\psi_k, \quad T^*\psi_k = b_k\phi_k, \qquad k \in \mathbb{N}, \tag{3.2}$$

where b_k are the singular values, $\{\psi_k\}$ is an orthonormal basis in $\mathrm{Range}(T) \subset \mathscr{H}$ and $\{\phi_k\}$ is the corresponding orthonormal basis in $\mathscr{H}$. In view of (3.1), it holds that:

$$Y(\psi_k) = \langle h|\phi_k\rangle_{\mathscr{H}} b_k + \varepsilon\xi(\psi_k), \qquad k \in \mathbb{N}. \tag{3.3}$$

Since in practice only a finite number of measurements can be computed, it is natural to assume that the values $Y(\psi_k)$ are available only for k smaller than some integer n. Under the assumption that $b_k \neq 0$ the last equation is equivalent to (1.1) with the choice $f_i = \langle h|\phi_i\rangle_{\mathscr{H}}$ and $\Sigma = \mathrm{diag}(\sigma_i^2, i = 1, \cdots)$ where $\sigma_i = \varepsilon b_i^{-1}$. Important examples of inverse problems to which this statistical model has been successfully applied are derivative estimation, deconvolution with known kernel, computerized tomography—see [14] and the references therein for more applications.

For very mildly ill-posed inverse problems, *i.e.*, when the singular values b_k of the operator T decrease to zero not faster than any negative power of k, the approach presented in previous section will lead to satisfactory results. Indeed, choosing $\beta = 8|||\Sigma|||$ or $\beta = 4|||\Sigma|||$, the remainder term in (2.7) and (2.9) becomes—up to a logarithmic factor—proportional to $\max_{1\le k\le n} b_k^{-2}/n$, which is the optimal rate in the case of very mild ill-posedness.

However, even for mildly ill-posed inverse problems, the approach developped in previous section becomes obsolete since the remainder blows up when n increases to infinity. Furthermore, this is not an artefact of our theoretical results, but is a drawback of the aggregation strategy adopted in the previous section. Indeed, the posterior probability measure $\hat{\pi}$ defined by (2.2) can be seen as the solution of the entropy-penalized empirical risk minimization problem:

$$\hat{\pi}_n = \arg\inf_p \Big\{ \int_\Lambda \hat{r}_\lambda \, p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi) \Big\}, \tag{3.4}$$

where the inf is taken over the set of all probability distributions. This means that the same regularization parameter β is employed for estimating both the coefficients $f_i = \langle h|\phi_i\rangle_{\mathscr{H}}$ corrupted by noise of small magnitude and those corrupted by large noise. Since we place ourselves in the setting of known operator T and, therefore, known noise levels, such a uniform treatment of all

coefficients f_i is not reasonable. It is more natural to upweight the regularization term in the case of large noise downweighting the data fidelity term and, conversely, to downweight the regularization in the case of small noise upweighting the data fidelity term. This observation leads us to the grouped version of the exponentially weighted aggregate.

Let us consider a partition $B_1,\dots,B_m$ of the set $\{1,\dots,n\}$: $B_j=\{T_j+1,\dots,T_{j+1}\}$, for some integers $0=T_1<T_2<\ldots<T_{m+1}=n$. To each element B_j of this partition, we associate the data sub-vector $\boldsymbol{Y}^j=(Y_i : i\in B_j)$ and the sub-vector of true function $\boldsymbol{f}^j=(f_i : i\in B_j)$. As in previous section, we are concerned by the aggregation of affine estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y}+\boldsymbol{b}_\lambda$ but we will assume here that the matrices A_λ are block-diagonal:

$$A_\lambda=\begin{bmatrix} A_\lambda^1 & 0 & \dots & 0\\ 0 & A_\lambda^2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & A_\lambda^m\end{bmatrix}, \qquad \text{with} \quad A_\lambda^j\in\mathbb{R}^{(T_{j+1}-T_j)\times(T_{j+1}-T_j)}.$$

Similarly, we define $\hat{\boldsymbol{f}}_\lambda^j$ and $\boldsymbol{b}_\lambda^j$ as the subvectors of $\hat{\boldsymbol{f}}_\lambda$ and $\boldsymbol{b}_\lambda$, respectively, corresponding to the indices belonging to B_j. We will also assume that the noise covariance matrix Σ is block-diagonal with $(T_{j+1}-T_j)\times(T_{j+1}-T_j)$ blocks Σ^j. These notation imply in particular that $\hat{\boldsymbol{f}}_\lambda^j = A_\lambda^j\boldsymbol{Y}^j+\boldsymbol{b}_\lambda^j$ for every $j=1,\dots,m$. Moreover, the unbiased risk estimate $\hat r_\lambda^{\mathrm{unb}}$ of $\hat{\boldsymbol{f}}_\lambda$ can be decomposed into the sum of unbiased risk estimates $\hat r_\lambda^{j,\mathrm{unb}}$ of $\hat{\boldsymbol{f}}_\lambda^j$; namely $\hat r_\lambda^{\mathrm{unb}}=\sum_{j=1}^m \hat r_\lambda^{j,\mathrm{unb}}$, where

$$\hat r_\lambda^{j,\mathrm{unb}} = \|\boldsymbol{Y}^j-\hat{\boldsymbol{f}}_\lambda^j\|+\frac{2}{n}\operatorname{Tr}(\Sigma^j A_\lambda^j)-\frac{1}{n}\operatorname{Tr}[\Sigma^j], \qquad j=1,\dots,m.$$

To state the analogues of Theorem 1 and Remark 4 we introduce the following two settings.

Setting 1:: All the matrices A_λ^j are symmetric and satisfy $A_\lambda^j A_{\lambda'}^j = A_{\lambda'}^j A_\lambda^j$, $A_\lambda^j\Sigma^j+\Sigma^j A_\lambda^j\succeq 0$ and $A_{\lambda'}^j\boldsymbol{b}_\lambda^j=0$ for all $\lambda,\lambda'\in\Lambda$ and for all $j\in\{1,\dots,m\}$. For a vector of temperature parameters $\boldsymbol{\beta}=(\beta_1,\dots,\beta_m)^\top$ and for a prior π, we define the group exponentially weighted aggregate (GEWA) as $\hat{\boldsymbol{f}}^j_{\mathrm{GEWA}}=\int_\Lambda \hat{\boldsymbol{f}}_\lambda^j\hat\pi^j(d\lambda)$, where

$$\hat\pi^j(d\lambda)=\theta^j(\lambda)\pi(d\lambda) \qquad \text{with} \qquad \theta^j(\lambda)=\frac{\exp(-n\hat r_\lambda^{j,\mathrm{unb}}/\beta_j)}{\int_\Lambda \exp(-n\hat r_\omega^{j,\mathrm{unb}}/\beta_j)\pi(d\omega)}. \tag{3.5}$$

Setting 2:: For every $j=1,\dots,m$ and for every λ belonging to a set of π-measure one, the matrices A_λ satisfy the inequality $\operatorname{Tr}(\Sigma^j A_\lambda^j)\le\operatorname{Tr}(\Sigma^j(A_\lambda^j)^\top A_\lambda^j)$ while the vectors $\boldsymbol{b}_\lambda$ are such that $A_\lambda^j\boldsymbol{b}_\lambda^j=(A_\lambda^j)^\top\boldsymbol{b}_\lambda^j=0$. In this case, for a vector of temperature parameters $\boldsymbol{\beta}=(\beta_1,\dots,\beta_m)^\top$ and for a prior π, we define the group exponentially weighted aggregate (GEWA) as $\hat{\boldsymbol{f}}^j_{\mathrm{GEWA}}=\int_\Lambda \tilde{\boldsymbol{f}}_\lambda^j\hat\pi^j(d\lambda)$, where $\tilde{\boldsymbol{f}}_\lambda^j=(A_\lambda^j+(A_\lambda^j)^\top-(A_\lambda^j)^\top A_\lambda^j)\boldsymbol{Y}^j+\boldsymbol{b}_\lambda^j$ and $\hat\pi^j$ is defined by (3.5).

Theorem 3. *Under Setting 1, if $\beta_j\ge 8|||\Sigma^j|||$ for every $j=1,\dots,m$, then*

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}}-\boldsymbol{f}\|_n^2\big)\le\sum_{j=1}^m\inf_{p_j\in\mathscr{P}_\Lambda}\left(\int_\Lambda\mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j-\boldsymbol{f}^j\|_n^2\,p_j(d\lambda)+\frac{\beta_j}{n}\,\mathscr{K}(p_j,\pi)\right). \tag{3.6}$$

Furthermore, the same inequality holds true under Setting 2 provided that $\beta_j\ge 4|||\Sigma^j|||$ for every $j=1,\dots,m$.

As we shall see it in Section 5, this theorem allows us to propose an estimator of the unknown signal which is adaptive w.r.t. the smoothness properties of the underlying signal and achieves

the minimax rates and constants over the Sobolev ellipsoids provided the operator T is mildly ill-posed, *i.e.*, its singular values decrease at most polynomially.

4. Examples of sharp oracle inequalities

In this section, we discuss consequences of the main result for specific choices of prior measures. For coveying the main messages of this section it is enough to focus on the Settings 1 and 2 in the case of only one group ($m = 1$). In this situation, the estimators $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ and $\hat{\boldsymbol{f}}_{\mathrm{GEWA}}$ coincide.

4.1. **Discrete oracle inequality.**

In order to demonstrate that Inequality (3.6) can be reformulated in terms of an OI as defined by (2.1), let us consider the case when the prior π is discrete. That is, we assume that $\pi(\Lambda_0) = 1$ for a countable set $\Lambda_0 \subset \Lambda$. Without loss of generality, we assume that $\Lambda_0 = \mathbb{N}$. Then, the following result holds true.

Proposition 1. *Under Setting 1 with $m = 1$ and $\beta = \beta_1$ if π is supported by $\mathbb{N}$, then the aggregate $\hat{\boldsymbol{f}}_{\mathrm{GEWA}}$ satisfies the inequality*

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}} - \boldsymbol{f}\|_n^2) \le \inf_{\ell \in \mathbb{N}: \pi_\ell > 0} \left(\mathbb{E}\|\hat{\boldsymbol{f}}_\ell - \boldsymbol{f}\|_n^2 + \frac{\beta \log(1/\pi_\ell)}{n} \right) \tag{4.1}$$

provided that $\beta \ge 8|||\Sigma|||$. Furthermore, the same inequality holds true under Setting 2 provided that $\beta \ge 4|||\Sigma|||$.

Proof. It suffices to apply Theorem 1 and to bound the RHS from above by the minimum over all Dirac measures $p = \delta_\ell$ with ℓ such that $\pi_\ell > 0$. □

This inequality can be compared to Corollary 2 in [7, Section 4.3]. Our inequality has the advantage of having factor one both in front of the expectation of the LHS of (4.1) and in front of the inf of the RHS. It should be noted, however, that we consider the noise covariance matrix as known, whereas [7] estimates the noise covariance along with the regression function.

4.2. **Continuous oracle inequality.**

It may be useful in practice to combine a family of affine estimators indexed by an open subset of $\mathbb{R}^M$, for some integer $M > 0$, for instance when the aim is to build an estimator that is nearly as accurate as the best kernel estimator with fixed kernel and varying bandwidth. In order to state an oracle inequality in such a "continuous" setup, let us denote by $d_2(\boldsymbol{\lambda}, \Lambda)$ the largest real $\tau > 0$ such that the ball centered at $\boldsymbol{\lambda}$ with radius τ is included in Λ. In what follows, Leb$(\cdot)$ stands for the Lebesgue measure.

Proposition 2. *Let $\Lambda \subset \mathbb{R}^M$ be an open and bounded set and let π be the uniform probability on Λ. Assume that the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is Lipschitz continuous, i.e., $|r_{\boldsymbol{\lambda}'} - r_{\boldsymbol{\lambda}}| \le L_r \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\|_2$, $\forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$. Under Setting 1 with $m = 1$ and $\beta = \beta_1 \ge 8|||\Sigma|||$ the aggregate $\hat{\boldsymbol{f}}_{\mathrm{GEWA}}$ satisfies the inequality*

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2 \le \inf_{\boldsymbol{\lambda} \in \Lambda} \left\{ \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 + \frac{\beta M}{n} \log\left(\frac{\sqrt{M}}{2\min(n^{-1}, d_2(\boldsymbol{\lambda}, \Lambda))} \right) \right\} + \frac{L_r + \beta \log\big(\mathrm{Leb}(\Lambda)\big)}{n}. \tag{4.2}$$

Furthermore, the same inequality holds true under Setting 2 for every $\beta \ge 4|||\Sigma|||$.

Proof. Let us denote by $B_{\boldsymbol{\lambda}}(\tau)$ the Euclidean ball in $\mathbb{R}^M$ with radius $\tau > 0$ and centered at $\boldsymbol{\lambda} \in \mathbb{R}^M$. It suffices to apply Theorem 1 and to bound the RHS in Inequality (2.7) from above by the minimum over all measures having as density $p_{\boldsymbol{\lambda}_0, \tau_0}(\boldsymbol{\lambda}) = \mathbb{1}_{B_{\boldsymbol{\lambda}_0}(\tau_0)}(\boldsymbol{\lambda}) / \mathrm{Leb}(B_{\boldsymbol{\lambda}_0}(\tau_0))$. For a choice $\boldsymbol{\lambda}_0 = \min(n^{-1}, d_2(\boldsymbol{\lambda}, \Lambda))$ such that $B_{\boldsymbol{\lambda}_0}(\tau_0) \subset \Lambda$, the measure $p_{\boldsymbol{\lambda}_0, \tau_0}(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$ is absolutely continuous w.r.t. the uniform prior π and the Kullback-Leibler divergence between these two measures

equals $\log\{\mathrm{Leb}(\Lambda)/\mathrm{Leb}\big(B_{\boldsymbol{\lambda}_0}(\tau_0)\big)\}$. Using the obvious inequality $\mathrm{Leb}\big(B_{\boldsymbol{\lambda}_0}(\tau_0)\big) \geq (\frac{2\tau_0}{\sqrt{M}})^M$ and the Lipschitz condition, we get the desired inequality. □

Note that it is not very stringent to require that the risk function $r_{\boldsymbol{\lambda}}$ is Lipschitz continuous, especially that this condition needs not be satisfied uniformly in $\boldsymbol{f}$. As an example, let us consider the ridge regression: for a given design matrix $X \in \mathbb{R}^{n\times p}$, $A_\lambda = X(X^\top X + \gamma_n \lambda I_{n\times n})^{-1} X^\top$ with $\lambda \in [\lambda_*, \lambda^*]$, where γ_n is a given normalization factor typically set to n or $\sqrt{n}$, $\lambda_* > 0$ and $\lambda^* \in [\lambda_*, \infty]$. One easily checks that the Lipschitz continuity of the risk function is satisfied with $L_r = L_r(f) = 4\lambda_*^{-1}\|\boldsymbol{f}\|_n^2 + 2\,\mathrm{Tr}(\Sigma/n)$.

4.3. **Sparsity oracle inequality.** The continuous oracle inequality stated in previous subsection is well adapted to the case where the dimension M of Λ is small compared to the sample size n (or, more precisely, the signal to noise ratio $n/|||\Sigma|||$). If this is not the case, the choice of the prior should be done more carefully. For instance, consider the case of a set $\Lambda \subset \mathbb{R}^M$ with large M under the sparsity scenario: there is a sparse vector $\boldsymbol{\lambda}^* \in \Lambda$ such that the risk of $\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}^*}$ is small. Then, it is natural to choose a prior π that promotes the sparsity of $\boldsymbol{\lambda}$. This can be done in the same vein as in [21, 22], by means of the heavy tailed prior:

$$\pi(d\boldsymbol{\lambda}) \propto \prod_{j=1}^{M} \frac{1}{(1+|\lambda_j/\tau|^2)^2}\, \mathbb{1}_\Lambda(\boldsymbol{\lambda}), \tag{4.3}$$

where $\tau > 0$ is a tuning parameter.

Proposition 3. *Let $\Lambda = \mathbb{R}^M$ and let π be defined by (4.3). Assume that the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is continuously differentiable and, for some $M \times M$ matrix $\mathscr{M}$, satisfies:*

$$r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}'} - \nabla r_{\boldsymbol{\lambda}'}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}') \leq (\boldsymbol{\lambda} - \boldsymbol{\lambda}')^\top \mathscr{M} (\boldsymbol{\lambda} - \boldsymbol{\lambda}'), \qquad \forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda. \tag{4.4}$$

Under Setting 1 if $\beta \geq 8|||\Sigma|||$, then the aggregate $\hat{\boldsymbol{f}}_{\mathrm{EWA}} = \hat{\boldsymbol{f}}_{\mathrm{GEWA}}$ satisfies the inequality

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \leq \inf_{\boldsymbol{\lambda}\in\mathbb{R}^M} \Big\{\mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 + \frac{4\beta}{n}\sum_{j=1}^{M} \log\Big(1 + \frac{|\lambda_j|}{\tau}\Big)\Big\} + \mathrm{Tr}(\mathscr{M})\tau^2. \tag{4.5}$$

Moreover, the same inequality holds true under Setting 2 provided that $\beta \geq 4|||\Sigma|||$.

Let us discuss here some consequences of this sparsity oracle inequality. First of all, let us remark that in most cases—see, for instance, [22, 23] in the case of frozen estimators—$\mathrm{Tr}(\mathscr{M})$ is on the order of M and the choice $\tau = \sqrt{\beta/(nM)}$ ensures that the last term in the RHS of Eq. (4.5) decreases at the parametric rate $1/n$. This is the choice we recommend for practical applications.

Assume now that we are given a large number of linear estimators $\hat{\boldsymbol{g}}_1 = G_1\boldsymbol{Y}, \ldots, \hat{\boldsymbol{g}}_M = G_M\boldsymbol{Y}$ satisfying, for instance, conditions of Setting 1. We will focus on matrices G_j having a spectral norm bounded by one (it is well known that the failure of this condition makes the linear estimator inadmissible). Assume furthermore that our aim is to propose an estimator that mimics the behavior of the best possible convex combination of a pair of estimators chosen among $\hat{\boldsymbol{g}}_1, \ldots, \hat{\boldsymbol{g}}_M$. This task can be accomplished in the framework of the present paper by setting $\Lambda = \mathbb{R}^M$ and $\hat{\boldsymbol{f}}_\lambda = \lambda_1\hat{\boldsymbol{g}}_1 + \ldots \lambda_M\hat{\boldsymbol{g}}_M$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)$. If $\{\hat{\boldsymbol{g}}_i\}$ satisfy conditions of Setting 1, then it is also the case for $\{\hat{\boldsymbol{f}}_\lambda\}$. Moreover, the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is quadratic with the Hessian matrix $\nabla^2 r_{\boldsymbol{\lambda}}$ given by the entries $2\langle G_j\boldsymbol{f} | G_{j'}\boldsymbol{f}\rangle_n + \frac{2}{n}\mathrm{Tr}(G_{j'}\Sigma G_j)$, $j, j' = 1, \ldots, M$. This implies that Inequality (4.4) holds with $\mathscr{M}$ being the Hessian divided by 2. Therefore, denoting by σ_i^2 the ith diagonal entry of Σ and setting $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$, we get $\mathrm{Tr}(\mathscr{M}) \leq |||\sum_{j=1}^M G_j^2|||(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2) \leq M(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2)$, where

the norm of a matrix is understood as its largest singular value. Applying Proposition 3 with $\tau = \sqrt{\beta/(nM)}$, we get

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2 \le \inf_{\alpha, j, j'} \mathbb{E}\|\alpha\hat{\boldsymbol{g}}_j + (1-\alpha)\hat{\boldsymbol{g}}_{j'} - \boldsymbol{f}\|_n^2 + \frac{8\beta}{n}\log\Big(1 + \frac{Mn}{\beta}\Big) + \frac{\beta}{n}(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2), \tag{4.6}$$

where the inf is taken over all $\alpha \in [0,1]$ and $j, j' \in \{1, \dots, M\}$. We restrict our choice to $\boldsymbol{\lambda}$ having at most two non-zero coefficients, λ_{i_0} and λ_{j_0}, that are non-negative and sum to one: $\lambda_{i_0} + \lambda_{j_0} = 1$. Then, the summation in the RHS of inequality (4.5) has simply two terms controlled by the following inequality

$$\log\big(1 + \lambda_{i_0}\tau^{-1}\big) + \log\big(1 + \lambda_{j_0}\tau^{-1}\big) \le 2\log\big(1 + \tau^{-1}\big). \tag{4.7}$$

In practice, $\tau^2 = \beta/(Mn) < 1$ so $\log\big(1+\tau^{-1}\big) \le \log\big(1+\tau^{-2}\big)$ and Inequality (4.6) holds true.

This shows that, using the EWA, one can achieve the best possible risk over the convex combinations of a pair of linear estimators—selected from a large (but finite) family—at the price of a residual term that decreases at the parametric rate up to a log factor.

4.4. **Oracle inequalities for varying-block-shrinkage estimators.** Let us consider now the problem of aggregation of two-block shrinkage estimators. It means that the constituent estimators have the following form: for $\boldsymbol{\lambda} = (a, b, k) \in [0,1]^2 \times \{1, \dots, n\} := \Lambda$, $\hat{\boldsymbol{f}}_\lambda = A_{\boldsymbol{\lambda}}\boldsymbol{Y}$ where $A_{\boldsymbol{\lambda}} = \mathrm{diag}\big(a\mathbb{1}(i \le k) + b\mathbb{1}(i > k), i = 1, \cdots, n\big)$. Let us choose the prior π as the uniform probability distribution on the set Λ.

Proposition 4. *Let $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ be the exponentially weighted aggregate having as constituent estimators two-block shrinkage estimators $A_{\boldsymbol{\lambda}}\boldsymbol{Y}$. If Σ is a diagonal matrix, then for any $\boldsymbol{\lambda} \in \Lambda$ and for any $\beta \ge 8|||\Sigma|||$,*

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \mathbb{E}(\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}} - \boldsymbol{f}\|_n^2) + \frac{\beta}{n}\Big\{1 + \log\Big(\frac{n^2\|\boldsymbol{f}\|_n^2 + n\,\mathrm{Tr}(\Sigma)}{12\beta}\Big)\Big\}. \tag{4.8}$$

In the case $\Sigma = I_{n\times n}$, this result is comparable to [44, page 20, Theorem 2.49], which states that in the model of homoscedastic regression ($\Sigma = I_{n\times n}$), the EWA acting on two-block positive-part James-Stein shrinkage estimators satisfies, for any $k = 3, \cdots, n-3$, and for $\beta = 8$, the oracle inequality

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{Leung}} - \boldsymbol{f}\|_n^2) \le \mathbb{E}(\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}} - \boldsymbol{f}\|_n^2) + \frac{9}{n} + \frac{8}{n}\min_{K>0}\Big\{K \vee \Big(\log\frac{n-6}{K} - 1\Big)\Big\}. \tag{4.9}$$

5. Application to minimax adaptive estimation

In the celebrated paper [49], Pinsker proved that in the model (1.1) the minimax risk over ellipsoids can be asymptotically attained by a linear estimator. Let us denote by $\theta_k(\boldsymbol{f}) = \langle \boldsymbol{f} | \varphi_k \rangle_n$ the coefficients of the (orthogonal) discrete cosine[1] transform of $\boldsymbol{f}$, hereafter denoted by $\mathscr{D}\boldsymbol{f}$. Pinsker's result—restricted to Sobolev ellipsoids $\mathscr{F}_{\mathscr{D}}(\alpha, R) = \big\{\boldsymbol{f} \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha}\theta_k(\boldsymbol{f})^2 \le R\big\}$ —states

[1] The results of this section hold true not only for the discrete sine transform, but for any linear transform $\mathscr{D}$ such that $\mathscr{D}\mathscr{D}^\top = \mathscr{D}^\top\mathscr{D} = n^{-1}I_{n\times n}$.

that, as $n \to \infty$, the equivalences

$$\inf_{\hat{f}} \sup_{f \in \mathscr{F}_{\mathscr{D}}(\alpha,R)} \mathbb{E}\big(\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|_n^2\big) \sim \inf_{A} \sup_{f \in \mathscr{F}_{\mathscr{D}}(\alpha,R)} \mathbb{E}\big(\|A\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big) \tag{5.1}$$

$$\sim \inf_{w>0} \sup_{f \in \mathscr{F}_{\mathscr{D}}(\alpha,R)} \mathbb{E}\big(\|A_{\alpha,w}\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big) \tag{5.2}$$

hold [59, Theorem 3.2], where the first inf is taken over all possible estimators $\hat{f}$ and $A_{\alpha,w} = \mathscr{D}^\top \mathrm{diag}\big((1-k^\alpha/w)_+; k=1,\dots,n\big)\mathscr{D}$ is the Pinsker filter in the discrete cosine basis. In simple words, this implies that the (asymptotically) minimax estimator can be chosen from the quite narrow class of linear estimators with Pinsker's filter. However, it should be emphasized that the minimax linear estimator depends on the parameters α and R, that are generally unknown. An (adaptive) estimator, that does not depend on (α, R) and is asymptotically minimax over a large scale of Sobolev ellipsoids has been proposed by Efromovich and Pinsker [27]. The next result, that is a direct consequence of Theorem 1, shows that the EWA with linear constituent estimators is also asymptotically sharp adaptive over Sobolev ellipsoids.

Proposition 5. *Let $\boldsymbol{\lambda} = (\alpha, w) \in \Lambda = \mathbb{R}_+^2$ and consider the prior*

$$\pi(d\boldsymbol{\lambda}) = \frac{2n_\sigma^{-\alpha/(2\alpha+1)}}{\big(1 + n_\sigma^{-\alpha/(2\alpha+1)} w\big)^3} e^{-\alpha} d\alpha\, dw, \tag{5.3}$$

where $n_\sigma = n/\sigma^2$. Then, in model (1.1) with homoscedastic errors, the aggregate $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ based on the temperature $\beta = 8\sigma^2$ and the constituent estimators $\hat{\boldsymbol{f}}_{\alpha,w} = A_{\alpha,w}\boldsymbol{Y}$ (with $A_{\alpha,w}$ being the Pinsker filter) is adaptive in the exact minimax sense[2] on the family of classes $\{\mathscr{F}_{\mathscr{D}}(\alpha,R) : \alpha > 0, R > 0\}$.

It is worth noting that the exact minimax adaptivity property of our estimator $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ is achieved without any tuning parameter. All previously proposed methods that are provably adaptive in exact minimax sense depend on some parameters such as the lengths of blocks for blockwise Stein [16] and Efromovich-Pinsker [28] estimators or the step of discretization and the maximal value of bandwidth [17]. Another nice property of the estimator $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ is that it does not require any pilot estimator based on the data splitting device [29, 62].

We now turn to the setup of heteroscedastic regression, which corresponds to ill-posed inverse problems as described in Section 3. To achieve adaptivity in the exact minimax sense, we make use of $\hat{\boldsymbol{f}}_{\mathrm{GEWA}}$, the grouped version of the exponentially weighted aggregate. We assume hereafter that the matrix Σ is diagonal with diagonal entries $\sigma_1, \dots, \sigma_n$ satisfying the following property:

$$\exists \sigma_*, \gamma > 0 \qquad \text{such that} \qquad \sigma_k^2 = \sigma_*^2 k^{2\gamma}(1 + o_k(1)) \qquad \text{as} \qquad k \to \infty. \tag{5.4}$$

This kind of problems arise when T is a differential operator or the Radon transform [14, Section 1.3]. To handle such a situation, we define the groups in the same spirit as the weakly geometrically increasing blocks in [15]. Let $\nu = \nu_n$ be a positive integer that increases as $n \to \infty$. Set $\rho_n = \nu_n^{-1/3}$ and define

$$T_j = \begin{cases} (1+\nu_n)^{j-1} - 1, & j = 1, 2, \\ T_{j-1} + \lfloor \nu_n \rho_n (1+\rho_n)^{j-2} \rfloor, & j = 3, 4, \dots, \end{cases} \tag{5.5}$$

where $\lfloor x \rfloor$ stands for the largest integer strictly smaller than x. Let m be the smallest integer j such that $T_j \geq n$. We redefine $T_{m+1} = n$ and set $B_j = \{T_j + 1, \dots, T_{j+1}\}$ for all $j = 1, \dots, m$.

[2] see [59, Definition 3.8]

Proposition 6. *Let the groups $B_1,\dots,B_m$ be defined as above with ν_n satisfying $\log\nu_n/\log n\to\infty$ and $\nu_n\to\infty$ as $n\to\infty$. Let $\boldsymbol{\lambda}=(\alpha,w)\in\Lambda=\mathbb{R}_+^2$ and consider the prior*

$$\pi(d\boldsymbol{\lambda})=\frac{2n^{-\alpha/(2\alpha+2\gamma+1)}}{\left(1+n^{-\alpha/(2\alpha+2\gamma+1)}w\right)^3}e^{-\alpha}d\alpha\,dw. \tag{5.6}$$

Then, in model (1.1) with diagonal covariance matrix $\Sigma=diag(\sigma_k;1\le k\le n)$ satisfying condition (5.4), the aggregate $\hat{\boldsymbol{f}}_{\mathrm{GEWA}}$ based on the temperatures $\beta_j=8\max_{i\in B_j}\sigma_i^2$ and the constituent estimators $\hat{\boldsymbol{f}}_{\alpha,w}=A_{\alpha,w}\boldsymbol{Y}$ (with $A_{\alpha,w}$ being the Pinsker filter) is adaptive in the exact minimax sense on the family of classes $\{\mathscr{F}(\alpha,R):\alpha>0,R>0\}$.

Note that this result provides an estimator which atteins the optimal constant in the minimax sense when the unknown signal lies in an ellipsoid. This property is possible because of the fact that the minimax estimators over the ellipsoids are linear. For other type of subsets of $\mathbb{R}^n$, such as hyper-rectangles, Besov bodies and so on, this is not true anymore. However, as proved by Donoho et al. [26], for orthosymmetric quadratically convex sets the minimax linear estimators have a risk which is within 25% of the minimax risk among all estimates. Therefore, following the approach developed here for the set of ellipsoids, it is also possible to prove that the aggregate GEWA can lead to an adaptive estimator whose risk is within a factor 5/4 of the minimax risk, for example, for a broad class of hyperrectangles.

6. Experiments

In this section we present some numerical experiments on synthetic data, by focusing only on the case of homoscedastic Gaussian noise ($\Sigma=\sigma^2 I_{n\times n}$) with known variance. Following the philosophy of reproducible research, a toolbox is made available freely for download at the address `http://imagine.enpc.fr/~dalalyan/AffineAggr.html`.

We evaluate different estimation routines on several 1D signals, introduced by Donoho and Johnstone [24, 25] and considered as a benchmark in the literature on signal processing. The six signals we retained for our experiments because of their diversity are depicted in Figure 3. Since all these signals are nonsmooth, we have also carried out experiments on their smoothed versions obtained by taking the antiderivative, see Figure 3. In what follows, the experiment on nonsmooth signals will be referred to as Experiment I, whereas the experiment on their smoothed counterparts will be referred to as Experiment II. In both cases, prior to applying estimation routines, we normalize the (true) sampled signal to have an empirical norm equal to one and use the Discrete Cosine Transform (DCT) denoted by $\boldsymbol{\theta}(\boldsymbol{Y})=\big(\theta_1(\boldsymbol{Y}),\dots,\theta_n(\boldsymbol{Y})\big)^\top$.

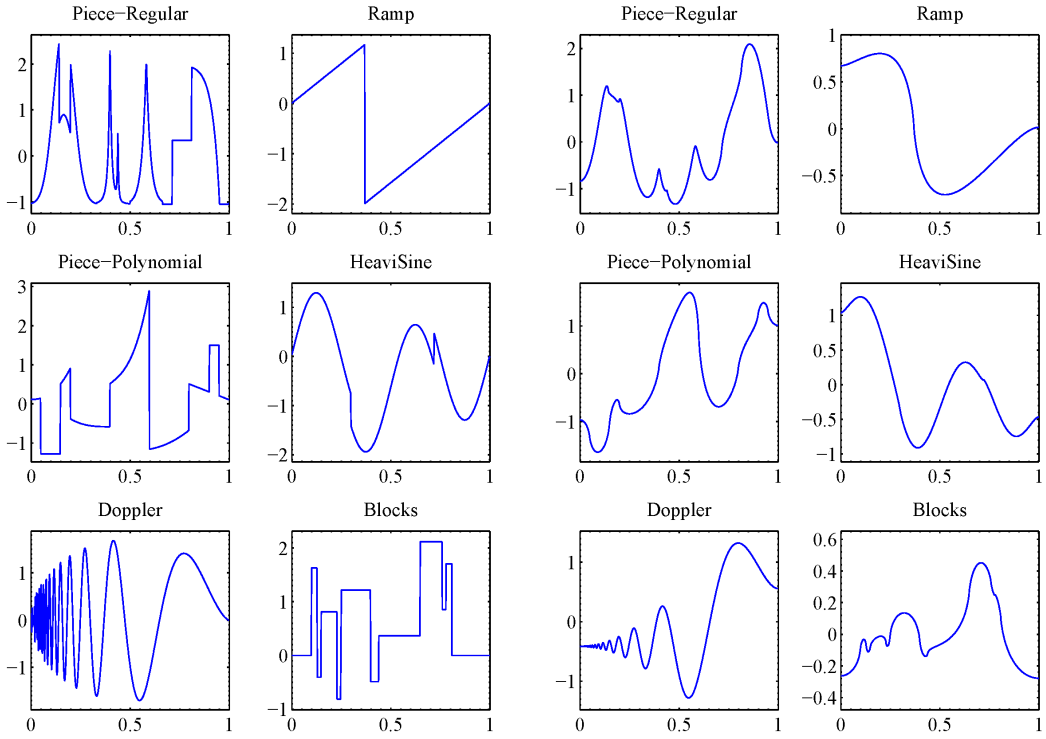
The four estimation routines—including the EWA—used in our experiments are detailed below:

Soft-Thresholding (ST) [24]:: For a given shrinkage parameter t, the Soft-Thresholding estimator of the vector of DCT coefficients $\theta_k(\boldsymbol{f})$ is defined by

$$\widehat{\theta}_k=\operatorname{sgn}\big(\theta_k(\boldsymbol{Y})\big)\big(|\theta_k(\boldsymbol{Y})|-\sigma t\big)_+. \tag{6.1}$$

In our experiments, we use the threshold minimizing the estimated unbiased risk defined via Stein's lemma. This procedure is referred to as SURE-shrink [25].

Blockwise James-Stein (BJS) shrinkage [12]:: The set of indices $\{1,\cdots,n\}$ is partitioned into $N=[n/\log(n)]$ non-overlapping blocks $B_1,B_2,\cdots B_N$ of equal size L. (If n is not a multiple of N, the last block may be of smaller size than all the others.) The corresponding blocks



(a) Test signals in Experiment I (b) Test signals in Experiment II

FIGURE 3. Test signals used in our experiments: Piece-Regular, Ramp, Piece-Polynomial, HeaviSine, Doppler and Blocks. (a) non-smooth (Experiment I) and (b) smooth (Experiment II).

of true coefficients $\theta_{B_k}(\boldsymbol{f}) = \big(\theta_j(\boldsymbol{f})\big)_{j\in B_k}$ are then estimated by:

$$\widehat{\theta}_{B_k} = \left(1 - \frac{\lambda L\sigma^2}{S_k^2(\boldsymbol{Y})}\right)_+ \theta_{B_k}(\boldsymbol{Y}), \qquad k = 1, \cdots, N \tag{6.2}$$

where $\theta_{B_k}(\boldsymbol{Y})$ are the blocks of noisy coefficients, $S_k^2 = \|\theta_{B_k}(\boldsymbol{Y})\|_2^2$ and $\lambda = 4.50524$ as suggested in [12].

Unbiased risk estimate (URE) minimization [17, 36] with Pinsker's filters:: This method consists in using a Pinsker filter, as defined in Section 5 above, with a data-driven choice of parameters α and w. This choice is done by minimizing an unbiased estimate of the risk over a suitably chosen grid for the values of α and w. Here, we use geometric grids ranging from 0.1 to 100 for α and from 1 to n for w. Thus, the bi-dimensional grid used in all the experiments has 100×100 elements. We refer to [17] for the closed-form formula of the unbiased risk estimator and further details.

EWA on Pinsker's filters:: We consider the same finite family of linear smoothers—defined by Pinsker's filters—as in the URE routine described above. According to Proposition 1, this leads to an estimator which is nearly as accurate as the best Pinsker's estimator in the given finite family.

To report the result of our experiments, we have also computed the best linear smoother based on a Pinsker filter chosen among the candidates that we used for defining the URE and the EWA

routines. By best smoother we mean the one minimizing the squared error, which can be computed since we know the ground truth. This pseudo-estimator will be referred to as oracle. The results summarized in Table 1 for Experiment I and Table 2 for Experiment II correspond to the average over 1000 trials of the mean squared error (MSE) from which we subtract the MSE of the oracle and multiply the resulting difference by the sample size. We report the results for $\sigma = 0.33$ and for $n \in \{2^8, 2^9, 2^{10}, 2^{11}\}$.

Simulations show that EWA and URE have very comparable performances and are significantly more accurate than Soft-Thresholding and Block James-Stein (see Table 1) for every size n of signals considered. The improvement is particularly important when the signal has large peaks (cf. Figure 4) or discontinuities (cf. Figure 5). In most cases, the EWA method also outperforms the URE, but this difference is much less pronounced. One can also observe that in the case of smooth signals, the difference of the MSEs between the EWA and the oracle, multiplied by n, remains nearly constant when n varies. This is in perfect agreement with our theoretical results in which the residual term decreases to zero inversely proportionally to the sample size.

Of course, Soft-Thresholding and blockwise James-Stein procedures have been designed for being applied to the wavelet transform of a Besov smooth function, rather than to the Fourier transform of a Sobolev-smooth function. However, the point here is not to demonstrate the superiority of the EWA as compared to ST and BJS procedures. The point is to stress the importance of having sharp adaptivity up to optimal constant and not simply adaptivity in the sense of rate of convergence. Indeed, the procedures ST and BJS are provably rate-adaptive when applied to Fourier transform of a Sobolev-smooth function, but they are not sharp adaptive—they do not attain the optimal constant—whereas the EWA and URE procedures do attain.

7. Summary and future work

In this paper, we have addressed the problem of aggregating a set of affine estimators in the context of regression with fixed design and heteroscedastic noise. Under some assumptions on the constituent estimators, we have proven that the EWA with a suitably chosen temperature parameter satisfies PAC-Bayesian type inequality, from which different types of oracle inequalities have been deduced. All these inequalities are with leading constant one and with rate-optimal residual term. As a by-product of our results, we have shown that the EWA applied to the family of Pinsker's estimators produces an estimator, which is adaptive in the exact minimax sense.

Although only the case of known covariance matrix is considered in the present work, the results are easy to extend for handling the more realistic situation where an unbiased estimate $\widehat{\Sigma}$, independent of $\boldsymbol{Y}$, of the covariance matrix Σ is available. One should merely replace Σ by $\widehat{\Sigma}$ in the definition of the unbiased risk estimate (2.6) and leave the remaining steps unchanged. For example, when the matrices A_λ satisfy condition **(C)**, the claim of Remark 4 remains valid and can be proved along the lines of Appendix A.

Next in our agenda is carrying out an experimental evaluation of the proposed aggregate using the approximation schemes described by Dalalyan and Tsybakov [23], Rigollet and Tsybakov [52] and Alquier and Lounici [1], with a special focus on the problems involving large scale data. It will also be interesting to extend the results of this work to the case of the unknown noise variance in the same vein as in Giraud [34].

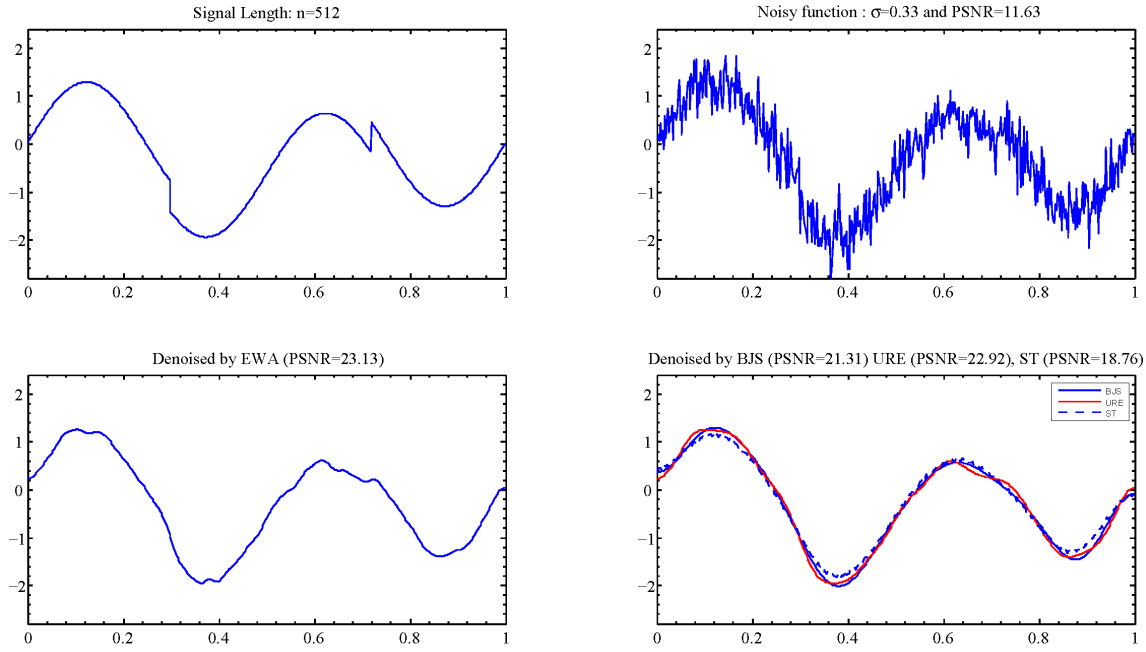


FIGURE 4. This figure presents the results of signal denoising for function Heavisine. The first row is the true signal (left) and a noisy version of it (right), where the noise is white Gaussian with standard deviation $\sigma = 0.33$. The second row presents the denoised signals obtained by EWA (left) and by BJS, ST and URE (right)

APPENDIX A. PROOFS OF MAIN THEOREMS

In this section we give the detailed proofs of the results stated in the manuscript.

A.1. **Stein's lemma.** The proof of our main results rely on the well-known Stein lemma [56, 57] providing an unbiased risk estimate for any estimator that depends sufficiently smoothly on the data vector $\boldsymbol{Y}$. For the convenience of the reader, we recall Stein's lemma in the case of heteroscedastic Gaussian regression.

Lemma 1. *Let $\boldsymbol{Y}$ be random vector drawn form the Gaussian distribution $\mathcal{N}_n(\boldsymbol{f}, \Sigma)$. If the estimator $\hat{\boldsymbol{f}}$ is a.e. differentiable in $\boldsymbol{Y}$ and the elements of the matrix $\nabla \cdot \hat{\boldsymbol{f}}^\top := (\partial_i \hat{\boldsymbol{f}}_j)$ have finite first moment, then*

$$\hat{r} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}\|_n^2 + \frac{2}{n}\operatorname{Tr}[\Sigma(\nabla \cdot \hat{\boldsymbol{f}}^\top)] - \frac{1}{n}\operatorname{Tr}[\Sigma],$$

is an unbiased estimate of r, i.e., $\mathbb{E}\hat{r} = r$.

The proof can be found in [59, p.157]. We apply Stein's lemma to affine estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$, where A_λ is an $n \times n$ deterministic real matrix and $\boldsymbol{b}_\lambda \in \mathbb{R}^n$ is a deterministic vector. We get that

$$\hat{r}_\lambda^{\mathrm{unb}} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 + \frac{2}{n}\operatorname{Tr}[\Sigma A_\lambda] - \frac{1}{n}\operatorname{Tr}[\Sigma]$$

is an unbiased estimator of the risk

$$r_\lambda = \mathbb{E}[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2] = \|(A_\lambda - I_{n\times n})\boldsymbol{f} + \boldsymbol{b}_\lambda\|_n^2 + \frac{1}{n}\operatorname{Tr}[A_\lambda \Sigma A_\lambda^\top].$$

n	EWA	URE	BJS	ST	EWA	URE	BJS	ST
		Blocks				**Doppler**		
256	0.051	0.245	9.617	4.846	0.062	0.212	13.233	6.036
	(0.42)	(0.39)	(1.78)	(1.29)	(0.35)	(0.31)	(2.11)	(1.23)
512	-0.052	0.302	13.807	9.256	-0.100	0.205	17.080	12.620
	(0.35)	(0.50)	(2.16)	(1.70)	(0.30)	(0.39)	(2.29)	(1.75)
1024	-0.050	0.299	19.984	17.569	-0.107	0.270	21.862	23.006
	(0.36)	(0.46)	(2.68)	(2.17)	(0.35)	(0.41)	(2.92)	(2.35)
2048	-0.007	0.362	28.948	30.447	-0.150	0.234	28.733	38.671
	(0.42)	(0.57)	(3.31)	(2.96)	(0.34)	(0.42)	(3.19)	(3.02)
		HeaviSine				**Piece-Regular**		
256	-0.060	0.247	1.155	3.966	-0.069	0.248	8.883	4.879
	(0.19)	(0.42)	(0.57)	(1.12)	(0.32)	(0.40)	(1.76)	(1.20)
512	-0.079	0.215	2.064	5.889	-0.105	0.237	12.147	9.793
	(0.19)	(0.39)	(0.86)	(1.36)	(0.30)	(0.37)	(2.28)	(1.64)
1024	-0.059	0.240	3.120	8.685	-0.092	0.291	15.207	16.798
	(0.23)	(0.36)	(1.20)	(1.64)	(0.34)	(0.46)	(2.18)	(2.13)
2048	-0.051	0.278	4.858	12.667	-0.059	0.283	21.543	27.387
	(0.25)	(0.48)	(1.42)	(2.03)	(0.34)	(0.54)	(2.47)	(2.77)
		Ramp				**Piece-Polynomial**		
256	0.038	0.294	6.933	5.644	0.017	0.203	12.201	3.988
	(0.37)	(0.47)	(1.54)	(1.20)	(0.37)	(0.37)	(1.81)	(1.19)
512	0.010	0.293	9.712	9.977	-0.078	0.312	17.765	9.031
	(0.36)	(0.51)	(1.76)	(1.67)	(0.35)	(0.49)	(2.72)	(1.62)
1024	-0.002	0.300	13.656	16.790	-0.026	0.321	23.321	17.565
	(0.30)	(0.45)	(2.25)	(2.06)	(0.38)	(0.48)	(2.96)	(2.28)
2048	0.007	0.312	19.113	27.315	-0.007	0.314	31.550	29.461
	(0.34)	(0.50)	(2.68)	(2.61)	(0.41)	(0.49)	(3.05)	(2.95)

TABLE 1. Comparing several adaptive methods on the six (non-smooth) signals of interest. For each sample size and each method, we report the average value of $\mathrm{MSE}-\mathrm{MSE}_{\mathrm{Oracle}}$ and the corresponding standard deviation (in parentheses), for 1000 replications of the experiment.

A.2. **An auxiliary result.** Prior to proceeding with the proof of main theorems, we prove an important auxiliary result which is the central ingredient of the proofs for our main results.

Lemma 2. *Let assumptions of Lemma 1 be satisfied. Let* $\{\hat{\boldsymbol{f}}_\lambda : \lambda \in \Lambda\}$ *be a family of estimators of* $\boldsymbol{f}$ *and* $\{\hat{r}_\lambda : \lambda \in \Lambda\}$ *a family of risk estimates such that the mapping* $\boldsymbol{Y} \mapsto (\hat{\boldsymbol{f}}_\lambda, \hat{r}_\lambda)$ *is a.e. differentiable for every* $\lambda \in \Lambda$. *Let* $\hat{r}_\lambda^{\mathrm{unb}}$ *be the unbiased risk estimate of* $\hat{\boldsymbol{f}}_\lambda$ *given by Stein's lemma.*

(1) *For every* $\pi \in \mathscr{P}_\Lambda$ *and for any* $\beta > 0$, *the estimator* $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ *defined as the average of* $\hat{\boldsymbol{f}}_\lambda$ *w.r.t. to the probability measure*

$$\hat{\pi}(\boldsymbol{Y}, d\lambda) = \theta(\boldsymbol{Y}, \lambda)\pi(d\lambda) \quad \textit{with} \quad \theta(\boldsymbol{Y}, \lambda) \propto \exp\big\{-n\hat{r}_\lambda(\boldsymbol{Y})/\beta\big\}$$

admits

$$\hat{r}_{\mathrm{EWA}} = \int_\Lambda \Big(\hat{r}_\lambda^{\mathrm{unb}} - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 - \frac{2n}{\beta}\langle \nabla_{\boldsymbol{Y}}\hat{r}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n\Big)\hat{\pi}(d\lambda)$$

n	EWA	URE	BJS	ST	EWA	URE	BJS	ST
		Blocks				**Doppler**		
256	0.387	0.216	0.216	2.278	0.214	0.237	1.608	2.777
	(0.43)	(0.40)	(0.24)	(0.98)	(0.23)	(0.40)	(0.73)	(1.04)
512	0.170	0.209	0.650	3.193	0.165	0.250	1.200	3.682
	(0.20)	(0.41)	(0.25)	(1.07)	(0.20)	(0.44)	(0.48)	(1.24)
1024	0.162	0.226	1.282	4.507	0.147	0.229	1.842	5.043
	(0.18)	(0.41)	(0.44)	(1.28)	(0.19)	(0.45)	(0.86)	(1.43)
2048	0.120	0.220	1.574	6.107	0.138	0.229	1.864	6.584
	(0.17)	(0.37)	(0.55)	(1.55)	(0.20)	(0.40)	(1.07)	(1.58)
		HeaviSine				**Piece-Regular**		
256	0.217	0.207	1.399	2.496	0.269	0.279	2.120	2.053
	(0.16)	(0.42)	(0.54)	(0.96)	(0.27)	(0.49)	(1.09)	(0.95)
512	0.206	0.221	0.024	3.045	0.216	0.248	2.045	2.883
	(0.18)	(0.43)	(0.26)	(1.10)	(0.20)	(0.45)	(1.17)	(1.13)
1024	0.179	0.200	0.113	3.905	0.183	0.228	1.251	3.780
	(0.18)	(0.50)	(0.27)	(1.27)	(0.20)	(0.41)	(0.70)	(1.37)
2048	0.162	0.189	0.421	5.019	0.145	0.223	1.650	4.992
	(0.15)	(0.37)	(0.27)	(1.53)	(0.19)	(0.42)	(1.12)	(1.42)
		Ramp				**Piece-Polynomial**		
256	0.162	0.200	0.339	2.770	0.215	0.257	1.486	2.649
	(0.16)	(0.38)	(0.24)	(1.00)	(0.25)	(0.48)	(0.68)	(1.01)
512	0.150	0.215	0.425	3.658	0.170	0.243	1.865	3.683
	(0.18)	(0.38)	(0.23)	(1.20)	(0.20)	(0.46)	(0.84)	(1.20)
1024	0.146	0.211	0.935	4.815	0.179	0.236	1.547	5.017
	(0.18)	(0.39)	(0.33)	(1.35)	(0.20)	(0.47)	(1.02)	(1.38)
2048	0.141	0.221	1.316	6.432	0.165	0.210	2.246	6.628
	(0.20)	(0.43)	(0.42)	(1.54)	(0.20)	(0.39)	(1.15)	(1.70)

TABLE 2. Comparing several adaptive methods on the six smoothed signals of interest. For each sample size and each method, we report the average value of $\mathrm{MSE}-\mathrm{MSE}_{\mathrm{Oracle}}$ and the corresponding standard deviation (in parentheses), for 1000 replications of the experiment.

as unbiased estimator of the risk.

(2) *If furthermore* $\hat{r}_\lambda \geq \hat{r}^{\mathrm{unb}}_\lambda$, $\forall \lambda \in \Lambda$ *and* $\int_\Lambda \langle \nabla \hat{r}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n \hat{\pi}(d\lambda) \geq -a \int_\Lambda \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 \hat{\pi}(d\lambda)$ *for some constant* $a > 0$*, then for every* $\beta \geq 2a$ *it holds that*

$$\mathbb{E}[\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2] \leq \inf_{p \in \mathscr{P}_\Lambda} \Big\{ \int_\Lambda \mathbb{E}[\hat{r}_\lambda]\, p(d\lambda) + \frac{\beta \mathscr{K}(p,\pi)}{n} \Big\}. \tag{A.1}$$

Proof. According to the Stein lemma, the quantity

$$\hat{r}_{\mathrm{EWA}} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 + \frac{2}{n} \mathrm{Tr}[\Sigma(\nabla \cdot \hat{\boldsymbol{f}}_{\mathrm{EWA}}(\boldsymbol{Y}))] - \frac{1}{n}\mathrm{Tr}[\Sigma] \tag{A.2}$$

is an unbiased estimate of the risk $r_n = \mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2)$. Using simple algebra, one checks that

$$\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 = \int_\Lambda \Big(\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 \Big) \hat{\pi}(d\lambda). \tag{A.3}$$

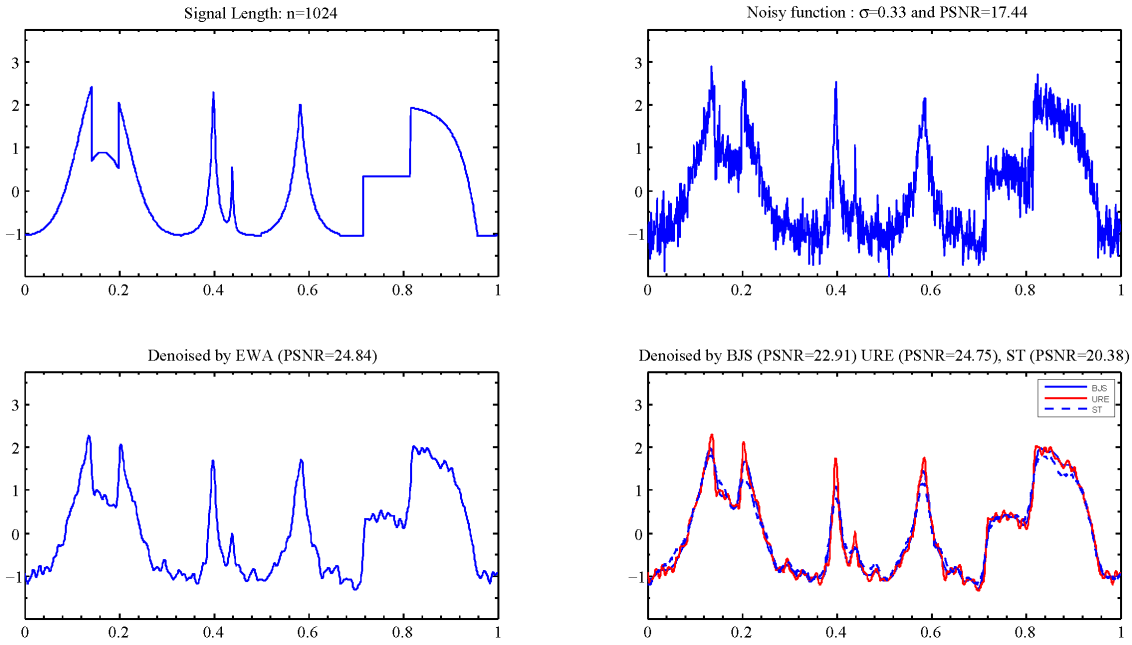


FIGURE 5. This figure presents the results of signal denoising for function Piece-Regular. The first row is the true signal (left) and a noisy version of it (right), where the noise is white Gaussian with standard deviation $\sigma = 0.33$. The second row presents the denoised signals obtained by EWA (left) and by BJS, ST and URE (right)

By interchanging the integral and differential operators, we get the following relation: $\partial_{y_i} \hat{\boldsymbol{f}}_{\mathrm{EWA},j} = \int_\Lambda \{(\partial_{y_j} \hat{f}_{\lambda,j}(\boldsymbol{Y}))\theta(\boldsymbol{Y},\lambda) + \hat{f}_{\lambda,j}(\boldsymbol{Y})(\partial_{y_i}\theta(\boldsymbol{Y},\lambda))\}\pi(d\lambda)$. This equality, combined with Equations (A.2) and (A.3) implies that

$$\hat{r}_{\mathrm{EWA}} = \int_\Lambda \big(\hat{r}_\lambda^{\mathrm{unb}} - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2\big)\hat{\pi}(d\lambda) + \frac{2}{n}\int_\Lambda \mathrm{Tr}[\Sigma \hat{\boldsymbol{f}}_\lambda \nabla_{\boldsymbol{Y}}\theta(\boldsymbol{Y},\lambda)^\top]\pi(d\lambda).$$

Taking into account that differentiation and integration can be interchanged, $\int_\Lambda \hat{\boldsymbol{f}}_{\mathrm{EWA}}\big(\nabla_{\boldsymbol{Y}}\theta(\boldsymbol{Y},\lambda)\big)^\top \pi(d\lambda) = \hat{\boldsymbol{f}}_{\mathrm{EWA}}\nabla_{\boldsymbol{Y}}\big(\int_\Lambda \theta(\boldsymbol{Y},\lambda)\pi(d\lambda)\big) = 0$, and we come up with the following expression for the unbiased risk estimate:

$$\begin{aligned}\hat{r}_{\mathrm{EWA}} &= \int_\Lambda \Big(\hat{r}_\lambda^{\mathrm{unb}} - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_n\|_n^2 + 2\langle \nabla_{\boldsymbol{Y}} \log\theta(\lambda) | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n\Big)\hat{\pi}(d\lambda)\\ &= \int_\Lambda \Big(\hat{r}_\lambda^{\mathrm{unb}} - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 - 2n\beta^{-1}\langle \nabla_{\boldsymbol{Y}}\hat{r}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n\Big)\hat{\pi}(d\lambda).\end{aligned}$$

This completes the proof of the first assertion of the lemma.

To prove the second assertion, let us observe that under the required condition and in view of the first assertion, for every $\beta \geq 2a$ it holds that $\hat{r}_{\mathrm{EWA}} \leq \int_\Lambda \hat{r}_\lambda^{\mathrm{unb}}\hat{\pi}(d\lambda) \leq \int_\Lambda \hat{r}_\lambda \hat{\pi}(d\lambda) \leq \int_\Lambda \hat{r}_\lambda \hat{\pi}(d\lambda) + \frac{\beta}{n}\mathscr{K}(\hat{\pi},\pi)$. To conclude, it suffices to remark that $\hat{\pi}$ is the probability measure minimizing the criterion $\int_\Lambda \hat{r}_\lambda p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi)$ among all $p \in \mathscr{P}_\Lambda$. Thus, for every $p \in \mathscr{P}_\Lambda$, it holds that

$$\hat{r}_{\mathrm{EWA}} \leq \int_\Lambda \hat{r}_\lambda p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi).$$

Taking the expectation of both sides, the desired result follows. □

A.3. **Proof of Theorem 1.** In what follows, we use the matrix shorthands $I = I_{n\times n}$ and $A_{\mathrm{EWA}} \triangleq \int_\Lambda A_\lambda \theta(\lambda)\pi(d\lambda)$. We apply Lemma 2 with $\hat r_\lambda = \hat r_\lambda^{\mathrm{unb}}$. To check the conditions of the second part, note that in view of Equations (2.4) and (2.6), as well as the assumptions $A_\lambda^\top = A_\lambda$ and $A_{\lambda'}\boldsymbol{b}_\lambda = 0$, we get

$$\nabla_{\boldsymbol{Y}} \hat r_\lambda^{\mathrm{unb}} = \frac{2}{n}(I-A_\lambda)^\top (I-A_\lambda)\boldsymbol{Y} - \frac{2}{n}(I-A_\lambda)^\top \boldsymbol{b}_\lambda = \frac{2}{n}(I-A_\lambda)^2\boldsymbol{Y} - \frac{2}{n}\boldsymbol{b}_\lambda.$$

Recall now that for any pair of commuting matrices P and Q the identity $(I-P)^2 = (I-Q)^2 + 2\big(I - \frac{P+Q}{2}\big)(Q-P)$ holds true. Applying this identity to $P = A_\lambda$ and $Q = A_{\mathrm{EWA}}$ we get the following relation: $\langle (I-A_\lambda)^2\boldsymbol{Y} | \Sigma(A_\lambda - A_{\mathrm{EWA}})\boldsymbol{Y}\rangle_n = \langle (I-A_{\mathrm{EWA}})^2\boldsymbol{Y} | \Sigma(A_\lambda - A_{\mathrm{EWA}})\boldsymbol{Y}\rangle_n - 2\langle \big(I - \frac{A_{\mathrm{EWA}}+A_\lambda}{2}\big)(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\rangle_n$. When one integrates over Λ with respect to the measure $\hat\pi$, the term of the first scalar product in the RHS of the last equation vanishes. On the other hand, positive semi-definiteness of matrices A_λ implies the one of the matrix A_{EWA} and, therefore, $\langle \big(I - \frac{A_{\mathrm{EWA}}+A_\lambda}{2}\big)(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\rangle_n \le \langle (A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\rangle_n$. This inequality implies that

$$\int_\Lambda \langle \nabla \hat r_\lambda^{\mathrm{unb}} | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n \hat\pi(d\lambda) \ge -4\int_\Lambda \|\Sigma^{1/2}(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\|_n^2 \hat\pi(d\lambda).$$

Therefore, the claim of Theorem 1 holds true for every $\beta \ge 8|||\Sigma|||$.

A.4. **Proof of Theorem 2.** Let now $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$ with symmetric $A_\lambda \preceq I_{n\times n}$ and $\boldsymbol{b}_\lambda \in \mathrm{Ker}(A_\lambda)$. According to the definition:

$$\hat r_\lambda^{\mathrm{adj}} = \hat r_\lambda^{\mathrm{unb}} + \frac{1}{n}\boldsymbol{Y}^\top (A_\lambda - A_\lambda^2)\boldsymbol{Y}.$$

One easily checks that $\hat r_\lambda^{\mathrm{adj}} \ge \hat r_\lambda^{\mathrm{unb}}$ for every λ and that

$$\begin{aligned}\int_\Lambda \langle \nabla \hat r_\lambda^{\mathrm{adj}} | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n \hat\pi(d\lambda) &= \int_\Lambda \langle 2(\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda) | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\rangle_n \hat\pi(d\lambda)\\ &= -2\int_\Lambda \|\Sigma^{1/2}(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\|_n^2 \hat\pi(d\lambda).\end{aligned}$$

Therefore, all the conditions required in the second part of Lemma 2 are fulfilled as soon as $\beta \ge 4|||\Sigma|||$. Applying this lemma, we get the desired result.

Appendix B. Proofs of propositions

B.1. **Proof of Proposition 2.** It suffices to apply Theorem 1 and to bound from above the RHS of inequality (2.7)

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{p\in\mathscr{P}_\Lambda}\left(\int_\Lambda \big[|r_\lambda - r_{\lambda_0}| + r_{\lambda_0}\big]\, p(d\lambda) + \frac{\beta}{n}\mathscr{K}(p,\pi)\right)$$

Then, the RHS of the last inequality can be bounded from above by the minimum over all measures having $p_{\lambda_0,\tau_0}(\lambda) = \mathbb{1}_{B_{\lambda_0}(\tau_0)}(\lambda)/\mathrm{Leb}(B_{\lambda_0}(\tau_0))$ as density. Assume moreover that λ_0 is such that

$B_{\lambda_0}(\tau_0) \subset \Lambda$, then using the Lipschitz condition on r_λ, the bound on the risk becomes

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{\substack{\lambda_0 \in \Lambda \\ B_{\lambda_0}(\tau_0) \subset \Lambda}} \left(\int_\Lambda \left[|r_\lambda - r_{\lambda_0}| + r_{\lambda_0} \right] p_{\lambda_0,\tau_0}(d\lambda) + \frac{\beta}{n} \mathscr{K}(p_{\lambda_0,\tau_0}, \pi) \right)$$

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{\substack{\lambda_0 \in \Lambda \\ B_{\lambda_0}(\tau_0) \subset \Lambda}} \left(r_{\lambda_0} + L_r \int_\Lambda \|\lambda - \lambda_0\|_2 \, p_{\lambda_0,\tau_0}(d\lambda) + \frac{\beta}{n} \mathscr{K}(p_{\lambda_0,\tau_0}, \pi) \right)$$

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{\substack{\lambda_0 \in \Lambda \\ B_{\lambda_0}(\tau_0) \subset \Lambda}} \left(r_{\lambda_0} + L_r \tau_0 + \frac{\beta}{n} \mathscr{K}(p_{\lambda_0,\tau_0}, \pi) \right) \tag{B.1}$$

Now, since λ_0 is such that $B_{\lambda_0}(\tau_0) \subset \Lambda$, the measure $p_{\lambda_0,\tau_0}(\lambda)\, d\lambda$ is absolutely continuous w.r.t. the uniform prior π over Λ and the Kullback-Leibler divergence between these measures equals $\log\{\mathrm{Leb}(\Lambda)/\mathrm{Leb}(B_{\lambda_0}(\tau_0))\}$. By the simple inequality $\|x\|_2^2 \le M\|x\|_\infty^2$ for any $x \in \mathbb{R}^M$, one can see that the Euclidean ball of radius τ_0 contains the hypercube of width $\frac{2\tau_0}{\sqrt{M}}$. So we have the following lower bound for the volume B_{λ_0}: $\mathrm{Leb}(B_{\lambda_0}(\tau_0)) \ge (\frac{2\tau_0}{\sqrt{M}})^M$. By combining this with inequality (B.1) the results of Proposition 2 is straightforward.

B.2. **Proof of Proposition 3.** We begin the proof as for the previous proposition, but pushing the development of the function $\lambda \to r_\lambda$ up to second order. So, for any $\lambda^* \in \mathbb{R}^M$, we have

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2 \le \inf_{\lambda^* \in \mathbb{R}^M} \Big(r_{\lambda^*} + \int_\Lambda \left(\nabla r_{\lambda^*}^\top (\lambda - \lambda^*) + (\lambda - \lambda^*)^\top \mathscr{M} (\lambda - \lambda^*) \right) p_{\lambda^*}(d\lambda) + \frac{\beta}{n} \mathscr{K}(p_{\lambda^*}, \pi) \Big)$$

By choosing $p_{\lambda^*}(\lambda) = \pi(\lambda - \lambda^*)$ for any $\lambda \in \mathbb{R}$, the second term in the last display vanishes since the distribution π is symmetric. The third term is computed thanks to the moment of order 2 of a scaled Student $t(3)$ distribution. Recall that if T is drawn from the scaled Student $t(3)$ distribution, its distribution function is $u \to 2/[\pi(1+u^2)^2]$, and that $\mathbb{E}T^2 = 1$. Thus, we have that $\int_\Lambda \lambda_1^2 \pi(\lambda) d\lambda = \tau^2$. We can then bound the risk of the EWA estimator by

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{\lambda^* \in \mathbb{R}^M} \left(r_{\lambda^*} + \mathrm{Tr}(\mathscr{M})\tau^2 + \frac{\beta}{n} \mathscr{K}(p_{\lambda^*}, \pi) \right) \tag{B.2}$$

So far, the particular choice of heavy tailed prior has not been used. This choice is important to control the Kullback-Leibler divergence between two translated versions of the same distribution

$$\mathscr{K}(p_{\lambda^*}, \pi) = \int_\Lambda \log \left[\prod_{j=1}^M \frac{(\tau^2 + \lambda_j^2)^2}{(\tau^2 + (\lambda_j - \lambda_j^*)^2)^2} \right] p_{\lambda^*}(d\lambda)$$

$$\mathscr{K}(p_{\lambda^*}, \pi) = 2 \sum_{j=1}^M \int_\Lambda \log \left[\frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \right] p_{\lambda^*}(d\lambda).$$

We bound the quotient in the above equality by

$$\frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} = 1 + \frac{2\tau(\lambda_j - \lambda_j^*)}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \frac{\lambda_j^*}{\tau} + \frac{\lambda_j^*}{\tau^2 + (\lambda_j - \lambda_j^*)^2}$$

$$\frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \le 1 + \left| \frac{\lambda_j^*}{\tau} \right| + \left(\frac{\lambda_j^*}{\tau} \right)^2 \le \left(1 + \left| \frac{\lambda_j^*}{\tau} \right| \right)^2.$$

Since the last inequality is independent of λ and p_{λ^*} is a probability measure, the integral disappears in the previous bound on the Kullback-Leibler divergence. So we eventually get

$$\mathscr{K}(p_{\lambda^*},\pi) \le 4\sum_{j=1}^{M}\log\left(1+\left|\frac{\lambda_j^*}{\tau}\right|\right).$$

Combining this with Inequality (B.2), leads the desired result.

B.3. **Proof of Proposition 4.** To simplify the notation, we set $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_n)$, the vector containing the standard deviations of the errors ξ_i. Let τ be a small positive number, the precise value of which will be given later. Let $\boldsymbol{\lambda}_0=(a_0,b_0,k_0)\in[\tau,1-\tau]^2\times\{1,\dots,n\}$ be some fixed element. Let us define $p_0(d\boldsymbol{\lambda})=\mathbb{1}_{[a_0-\tau,a_0+\tau]}(a)\mathbb{1}_{[b_0-\tau,b_0+\tau]}(b)\mathbb{1}(k=k_0)(2\tau)^{-2}dadb$.

Note that for any $\boldsymbol{\lambda}=(a,b,k)$, the risk of the estimator $A_{\boldsymbol{\lambda}}\boldsymbol{Y}$ is

$$r_{\boldsymbol{\lambda}}=\frac{1}{n}\sum_{i=1}^{k}\left((1-a)^2f_i^2+a^2\sigma_i^2\right)+\frac{1}{n}\sum_{i=k+1}^{n}\left((1-b)^2f_i^2+b^2\sigma_i^2\right).$$

In particular, the difference between the risks $r_{\boldsymbol{\lambda}}$ and $r_{\boldsymbol{\lambda}'}$—for two different parameters $\boldsymbol{\lambda}=(a,b,k)$ and $\boldsymbol{\lambda}_0=(a_0,b_0,k_0)$ such that $k=k_0$ is the same in the two cases—can be rewritten as follows:

$$\begin{aligned}r_{\boldsymbol{\lambda}}-r_{\boldsymbol{\lambda}_0}=&\frac{1}{n}\sum_{i=1}^{k}\left[2(a_0-a)\{(1-a_0)f_i^2-a_0\sigma_i^2\}+(a-a_0)^2\{f_i^2+\sigma_i^2\}\right]\\&+\frac{1}{n}\sum_{i=k+1}^{n}\left[2(b_0-b)\{(1-b_0)f_i^2-b_0\sigma_i^2\}+(b-b_0)^2\{f_i^2+\sigma_i^2\}\right]\end{aligned}$$

So, if we integrate w.r.t. the measure $p_0(d\boldsymbol{\lambda})$, the terms that are linear in $a-a_0$ and $b-b_0$ disappear and we get

$$\int_\Lambda(r_{\boldsymbol{\lambda}}-r_{\boldsymbol{\lambda}_0})\,p(d\boldsymbol{\lambda})=\frac{1}{n}\sum_{i=1}^{n}\{f_i^2+\sigma_i^2\}\int_{-\tau}^{\tau}u^2\frac{du}{2\tau}=\frac{\tau^2}{3}(\|\boldsymbol{f}\|_n^2+\|\boldsymbol{\sigma}\|_n^2). \tag{B.3}$$

Concerning the Kullback-Leibler divergence between p_0 and π, it can be computed as follows:

$$\begin{aligned}\mathscr{K}(p,\pi)&=\sum_{k=1}^{n}\int\int\log\left(\frac{p_0(da,db,k)}{\pi(da,db,k)}\right)p_0(da,db,k)\\&=\int_{a_0-\tau}^{a_0+\tau}\int_{b_0-\tau}^{b_0+\tau}\log\left(\frac{n}{4\tau^2}\mathbb{1}_{[a_0-\tau,a_0+\tau]}(a)\mathbb{1}_{[b_0-\tau,b_0+\tau]}(b)\right)\frac{da}{2\tau}\frac{db}{2\tau}\\&=\log\left(\frac{n}{4\tau^2}\right).\end{aligned} \tag{B.4}$$

Now we can use Equation (2.7) with our choice for p_0 and π. In view of the computations we have just done, we get

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}}-\boldsymbol{f}\|_n^2)&\le\inf_{p\in\mathscr{P}_\Lambda}\left(\int_\Lambda r_{\boldsymbol{\lambda}}\,p(d\boldsymbol{\lambda})+\frac{\beta}{n}\mathscr{K}(p,\pi)\right)\\&\le\int_\Lambda r_{\boldsymbol{\lambda}}\,p_0(d\boldsymbol{\lambda})+\frac{\beta}{n}\mathscr{K}(p_0,\pi)\\&=r_{\boldsymbol{\lambda}_0}+\int_\Lambda(r_{\boldsymbol{\lambda}}-r_{\boldsymbol{\lambda}_0})p_0(d\boldsymbol{\lambda})+\frac{\beta}{n}\log\left(\frac{n}{4\tau^2}\right)\\&=r_{\boldsymbol{\lambda}_0}+\frac{\tau^2(\|\boldsymbol{f}\|_n^2+\|\boldsymbol{\sigma}\|_n^2)}{3}+\frac{\beta}{n}\log\left(\frac{n}{4\tau^2}\right).\end{aligned} \tag{B.5}$$

The last expression, considered as a function of τ, admits as global minimum $\tau^2_{\min} = 3\beta/n(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2)$. Replacing this value in (B.5), we get the risk bound:

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \mathbb{E}(\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}_0} - \boldsymbol{f}\|_n^2) + \frac{\beta}{n}\Big\{1 + \log\Big(\frac{n^2(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2)}{12\beta}\Big)\Big\}. \tag{B.6}$$

Now, the desired result follows from the obvious equality $n\|\boldsymbol{\sigma}\|_n^2 = \mathrm{Tr}(\Sigma)$.

B.4. **Proof of Proposition 5.** We assume, without loss of generality, that the matrix $n^{1/2}\mathscr{D}$ coincides with the identity matrix. First, let us fix $\alpha_0 > 0$ and $R_0 > 0$, such that $n^{-1/2}\boldsymbol{f} \in \mathscr{F}(\alpha_0, R_0)$ and define $\boldsymbol{\lambda}_0 = (\alpha_0, w_0) \in \Lambda$ with w_0 chosen such that the Pinsker estimator $\hat{\boldsymbol{f}}_{\alpha_0,w_0}$ is minimax over the ellipsoid $\mathscr{F}(\alpha_0, R_0)$.

In what follows, we set $n_\sigma = n/\sigma^2$ and denote by p_π the density of π w.r.t. the Lebesgue measure on $\mathbb{R}^2_+$: $p_\pi(\alpha, w) = e^{-\alpha} n_\sigma^{-\alpha/(2\alpha+1)} p_w(w n_\sigma^{-\alpha/(2\alpha+1)})$, where p_w is a probability density function supported by $(0,\infty)$ such that $\int u p_w(u)\,du = 1$. One easily checks that

$$\int_{\mathbb{R}^2} \alpha\, p_\pi(\alpha, w)\, d\alpha dw = 1, \qquad \int_{\mathbb{R}^2} w\, p_\pi(\alpha, w)\, d\alpha dw \le n_\sigma^{1/2}. \tag{B.7}$$

Let τ be a positive number such that $\tau \le \min(1, \alpha_0/(2\log w_0))$ and choose $p_{\boldsymbol{\lambda}_0,\tau}$ as a translation/dilatation of π, concentrating on $\boldsymbol{\lambda}_0$ when $\tau \to 0$:

$$p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) = p_\pi\Big(\frac{\boldsymbol{\lambda} - \boldsymbol{\lambda}_0}{\tau}\Big)\frac{d\boldsymbol{\lambda}}{\tau^2}.$$

In view of Theorem 1,

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le r_{\boldsymbol{\lambda}_0} + \int_{\mathbb{R}^2} |r_{\alpha,w} - r_{\alpha_0,w_0}| p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) + \frac{\beta}{n}\mathscr{K}(p_{\boldsymbol{\lambda}_0,\tau}, \pi). \tag{B.8}$$

Let us decompose the term $r_{\alpha,w} - r_{\alpha_0,w_0}$ into two pieces: $r_{\alpha,w} - r_{\alpha_0,w_0} = \{r_{\alpha,w} - r_{\alpha,w_0}\} + \{r_{\alpha,w_0} - r_{\alpha_0,w_0}\}$ and find upper bounds for the resulting terms. With the choice of estimator we did, the difference between the risk functions at (α, w) and (α, w_0) is:

$$\begin{aligned} n(r_{\alpha,w} - r_{\alpha,w_0}) = &\sum_{k=1}^n \big[\big((1 - k^\alpha/w)_+ - 1\big)^2 - \big((1 - k^\alpha/w_0)_+ - 1\big)^2\big] f_k^2 \\ &+ \sum_{k=1}^n \big[\big((1 - k^\alpha/w)_+\big)^2 - \big((1 - k^\alpha/w_0)_+\big)^2\big]\sigma^2 \end{aligned}$$

Since the weights of the Pinsker estimators are in $[0,1]$, we have

$$n|r_{\alpha,w} - r_{\alpha,w_0}| \le 2\sum_{k=1}^n (f_k^2 + \sigma^2)\big|(1 - k^\alpha/w)_+ - (1 - k^\alpha/w_0)_+\big|. \tag{B.9}$$

For any $x, y \in \mathbb{R}$, the inequality $|x_+ - y_+| \le |x - y|$ is obvious. Combined with $\alpha_0 \le \alpha$ and $w_0 \le w$, we have that

$$\Big|\Big(1 - \frac{k^\alpha}{w}\Big)_+ - \Big(1 - \frac{k^\alpha}{w_0}\Big)_+\Big| \le \Big|\frac{k^\alpha}{w} - \frac{k^\alpha}{w_0}\Big| \mathbb{1}_{\{k^\alpha \le w\}} \le \frac{w - w_0}{w_0}. \tag{B.10}$$

By virtue of Inequalities (B.9) and (B.10) we get

$$|r_{\alpha,w} - r_{\alpha,w_0}| \le 2n^{-1}\sum_{k=1}^n (f_k^2 + \sigma^2)\frac{(w - w_0)}{w_0} \le 2(R_0 + \sigma^2)\frac{w - w_0}{w_0}. \tag{B.11}$$

Similar calculations lead to a bound for the other absolute difference between risk functions:

$$\begin{aligned} |r_{\alpha,w_0} - r_{\alpha_0,w_0}| &\le 2n^{-1}\sum_{k=1}^{n}(f_k^2+\sigma^2)\frac{k^\alpha - k^{\alpha_0}}{w_0}\mathbb{1}_{\{k^{\alpha_0}\le w_0\}} \\ &\le 2(R_0+\sigma^2)\big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}}-1\big). \end{aligned} \tag{B.12}$$

Recall that we aim to bound the second term in the RHS of (B.8). To this end, we need an accurate upper bound on the integrals of the RHSs of (B.11) and (B.12) w.r.t. the probability measure $p_{\boldsymbol{\lambda}_0,\tau}$. For the first one, we get

$$\begin{aligned} \int \big|r_{\alpha,w} - r_{\alpha,w_0}\big|\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) &\le 2(R_0+\sigma^2)\, w_0^{-1}\int_{\mathbb{R}^2}(w-w_0)\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) \\ &\le 4n_\sigma^{1/2} w_0^{-1}\tau(R_0+\sigma^2). \end{aligned} \tag{B.13}$$

Similar arguments apply to bound the integral of the second difference between risk functions:

$$\begin{aligned} \int_{\mathbb{R}^2}|r_{\alpha,w_0} - r_{\alpha_0,w_0}|\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) &\le 2(R_0+\sigma^2)\int_{\mathbb{R}^2}\big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}}-1\big)p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) \\ &= \frac{2\tau(R_0+\sigma^2)\log w_0}{\alpha_0 - \tau\log w_0} \\ &\le 4\tau(R_0+\sigma^2)\alpha_0^{-1}\log w_0, \end{aligned} \tag{B.14}$$

where we used the inequality $\tau \le \alpha_0/(2\log w_0)$.

The last term to bound in inequality (B.8) requires the evaluation of the Kullback-Leibler divergence between $p_{\boldsymbol{\lambda}_0,\tau}$ and π. It can be done as follows:

$$\begin{aligned} \mathscr{K}(p_{\boldsymbol{\lambda}_0,\tau},\pi) &= \int_{\mathbb{R}^2}\log\Big(\frac{e^{-\frac{\alpha-\alpha_0}{\tau}}\, p_w\big(\frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)}\tau}\big)}{e^{-\alpha}\, p_w\big(\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\big)}\frac{1}{\tau^2}\Big)p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) \\ &= \int_{\mathbb{R}^2}\Big\{\alpha - \frac{\alpha-\alpha_0}{\tau} + \log\frac{p_w\big(\frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)}\tau}\big)}{p_w\big(\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\big)}\Big\}p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) - 2\log(\tau) \\ &\le \alpha_0 + (\tau-1) + \int_{\mathbb{R}_+^2}\log\Big(1+\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\Big)^3 p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) - 2\log(\tau). \end{aligned}$$

where the third equality is derived thanks to Eq. (B.7) and the obvious relation $\|p_w\|_\infty = 2$. Now, making the change of variable $w = w_0 + \tau n_\sigma^{\alpha/(2\alpha+1)}u$ and using the fact that $w_0 + \tau n_\sigma^{\alpha/(2\alpha+1)}u \le n_\sigma^{\alpha/(2\alpha+1)}(w_0+u)$, we get

$$\begin{aligned} \int_{\mathbb{R}_+^2}\log\Big(1+\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\Big)^3 p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) &\le 3\int_{\mathbb{R}_+}\log\big(1+w_0+u\big)p_w(u)du \\ &\le 3\log\Big(1+w_0+\int_{\mathbb{R}_+}u\,p_w(u)du\Big) \\ &= 3\log(2+w_0). \end{aligned}$$

Eventually, we can reformulate our bound on the risk of the EWA given in (B.8), leading to

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le r_{\boldsymbol{\lambda}_0} + 4\tau(R_0+\sigma^2)\Big(\frac{n_\sigma^{1/2}}{w_0} + \frac{\log w_0}{\alpha_0}\Big) + \frac{8\sigma^2\big(\alpha_0 + 3\log\big(\frac{2+w_0}{\tau}\big)\big)}{n}. \tag{B.15}$$

To conclude the proof of the proposition, we set

$$\tau = \frac{\alpha_0}{n_\sigma^2 + \alpha_0 + 2\log w_0}, \qquad w_0 = \Big(\frac{R_0(\alpha_0+1)(2\alpha_0+1)}{\alpha_0}\Big)^{\frac{\alpha_0}{2\alpha_0+1}} n_\sigma^{\frac{\alpha_0}{2\alpha_0+1}}.$$

According to Pinsker's theorem,

$$\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)} r_{\boldsymbol{\lambda}_0} = (1+o_n(1))\min_{\hat{\boldsymbol{f}}}\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)}\mathbb{E}(\|\hat{\boldsymbol{f}}-\boldsymbol{f}\|_n^2).$$

In view of this result, taking the max over $\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)$ in (B.15), we get

$$\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R)}\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}}-\boldsymbol{f}\|_n^2) \le (1+o_n(1))\min_{\hat{\boldsymbol{f}}}\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R)}\mathbb{E}(\|\hat{\boldsymbol{f}}-\boldsymbol{f}\|_n^2) + O\Big(\frac{\log n}{n}\Big).$$

This leads to the desired result in view of the relation

$$\liminf_{n\to\infty}\min_{\hat{\boldsymbol{f}}}\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R)} n^{\frac{2\alpha_0}{2\alpha_0+1}}\mathbb{E}(\|\hat{\boldsymbol{f}}-\boldsymbol{f}\|_n^2) > 0.$$

B.5. **Proof of Proposition 6.** It is clear that all the conditions required in Setting 1 are fulfilled and we can apply Theorem 3 that yields:

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}}-\boldsymbol{f}\|_n^2\big) \le \sum_{j=1}^m \inf_{p_j\in\mathscr{P}_\Lambda}\left(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j-\boldsymbol{f}^j\|_n^2\, p_j(d\boldsymbol{\lambda}) + \frac{\beta_j}{n}\,\mathscr{K}(p_j,\pi)\right). \tag{B.16}$$

Let now $\boldsymbol{\lambda}_0=(\alpha_0,w_0)$ be a pair of real numbers such that the estimator $\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}_0}$ is minimax over the Sobolev ellipsoid $\mathscr{F}(\alpha_0,R_0)$. In what follows, we denote by p_π the density of π w.r.t. the Lebesgue measure on $\mathbb{R}_+^2$: $p_\pi(\alpha,w)=e^{-\alpha}n^{-\alpha/(2\alpha+2\gamma+1)}p_w(wn^{-\alpha/(2\alpha+2\gamma+1)})$, where p_w is a probability density function supported by $(0,\infty)$ such that $\int u p_w(u)\,du=1$. Let τ be a positive number such that $\tau\le\min(1,\alpha_0/(2\log w_0))$ and choose $p_{\boldsymbol{\lambda}_0,\tau}$ as a translation/dilatation of π, concentrating on $\boldsymbol{\lambda}_0$ when $\tau\to 0$:

$$p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) = p_\pi\Big(\frac{\boldsymbol{\lambda}-\boldsymbol{\lambda}_0}{\tau}\Big)\frac{d\boldsymbol{\lambda}}{\tau^2}.$$

Let m_0 be a positive integer smaller than m the precise value of which will be given later. As an immediate consequence of (B.16) we get:

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}}-\boldsymbol{f}\|_n^2\big) \le \sum_{j=1}^{m_0}\left(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j-\boldsymbol{f}^j\|_n^2\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) + \frac{\beta_j}{n}\,\mathscr{K}(p_{\boldsymbol{\lambda}_0,\tau},\pi)\right) + \sum_{j=m_0+1}^m\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j-\boldsymbol{f}^j\|_n^2\,\pi(d\boldsymbol{\lambda}).$$

Repeating the arguments of the proof of Proposition 5, one can check that

$$\sum_{j=1}^{m_0}\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j-\boldsymbol{f}^j\|_n^2\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) \;\le\; r_{\boldsymbol{\lambda}_0} + 4\tau\Big(R_0+n^{-1}\sum_{i=1}^{T_{m_0+1}}\sigma_i^2\Big)(n^{1/2}w_0^{-1}+\alpha_0^{-1}\log w_0), \tag{B.17}$$

$$\mathscr{K}(p_{\boldsymbol{\lambda}_0,\tau},\pi) \;\le\; \alpha_0 + 3\log\Big(\frac{2+w_0}{\tau}\Big). \tag{B.18}$$

It is obvious that

$$\sum_{i=1}^{T_{m_0+1}}\sigma_i^2 \le C\sigma_*^2\sum_{i=1}^{T_{m_0+1}} i^{2\gamma} \le C\sigma_*^2 T_{m_0+1}^{2\gamma+1} \le C\sigma_*^2 n^{2\gamma+1}.$$

Furthermore, using the definition of weakly geometrically increasing groups, we get that

$$\frac{1}{2}\nu_n(1+\rho_n)^j \le T_{j+1} \le \nu_n(1+\rho_n)^j.$$

This implies that

$$\sum_{i=1}^{m_0}\beta_j^2 \le C\sigma_*^2\sum_{i=1}^{m_0} T_{j+1}^{2\gamma} \le C\sigma_*^2 m_0 T_{m_0+1}^{2\gamma} \le C\sigma_*^2 m_0 T_{m_0}^{2\gamma}.$$

Let now m_0 be chosen in such a way that $T_{m_0}^{2\gamma} \le n^{(2\gamma+0.5)/(2\alpha_0+2\gamma+1)} < T_{m_0+1}^{2\gamma}$. The condition $\log n = o(\log \nu_n)$ implies that $m_0 T_{m_0}^{2\gamma} = o(T_{m_0}^{2\gamma(2\gamma+1)/(2\gamma+0.5)}) = o(n^{(2\gamma+1)/(2\alpha_0+\gamma+1)})$. Therefore, setting $\tau = \frac{\alpha_0}{n^2+\alpha_0+2\log w_0}$, it holds that

$$\sum_{j=1}^{m_0}\left(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j - \boldsymbol{f}^j\|_n^2\, p_{\boldsymbol{\lambda}_0,\tau}(d\boldsymbol{\lambda}) + \frac{\beta_j}{n}\,\mathscr{K}(p_{\boldsymbol{\lambda}_0,\tau},\pi)\right) = o(n^{-2\alpha_0/(2\alpha_0+2\gamma+1)}),$$

as $n\to\infty$. Since the minimax risk over $\mathscr{F}(\alpha_0,R_0)$, as well as $r_{\boldsymbol{\lambda}_0}$, is on the order of $n^{-2\alpha_0/(2\alpha_0+2\gamma+1)}$, we get

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}} - \boldsymbol{f}\|_n^2\big) \le r_{\boldsymbol{\lambda}_0}(1+o(1)) + \sum_{j=m_0+1}^{m}\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j - \boldsymbol{f}^j\|_n^2\,\pi(d\boldsymbol{\lambda}).$$

Using similar arguments, one checks that the last sum is also $o(n^{-2\alpha_0/(2\alpha_0+2\gamma+1)})$ and the desired result follows.

ACKNOWLEDGMENTS

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under the grant PARCIMONIE.

REFERENCES

[1] P. Alquier and K. Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *hal-00465801, submitted*, 2010.
[2] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9:1545–1588, October 1997.
[3] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *NIPS*, pages 46–54, 2009.
[4] J-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *NIPS*, pages 41–48, Vancouver, Canada, Dec 2007.
[5] J-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 2008.
[6] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9: 1179–1225, 2008.
[7] Y. Baraud, Ch. Giraud, and S. Huet. Estimator selection in the gaussian setting. *submitted*, 2010.
[8] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
[9] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.
[10] A. Buades, B. Coll, and J-M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4(2):490–530, 2005.
[11] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.
[12] T. T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999. ISSN 0090-5364.
[13] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.
[14] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19, 2008.

[15] L. Cavalier and A. B. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.*, 10(3):247–282, 2001.
[16] L. Cavalier and A. B. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.
[17] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002.
[18] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.
[19] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66:321–352, March 2007.
[20] P-A. Cornillon, N. Hengartner, and E. Matzner-Løber. Recursive bias estimation for multivariate regression smoothers. *submitted*, 2009.
[21] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. In *COLT*, pages 97–111, 2007.
[22] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.
[23] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT*, 2009.
[24] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
[25] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.
[26] David L. Donoho, Richard C. Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, 18(3):1416–1437, 1990.
[27] S. Y. Efromovich and M. S. Pinsker. A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, 1(11):58–65, 1984.
[28] S. Y. Efromovich and M. S. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996.
[29] Sam Efromovich. On nonparametric regression for IID observations in a general setting. *Ann. Statist.*, 24(3):1125–1144, 1996.
[30] Y. Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the third annual workshop on Computational learning theory*, COLT, pages 202–216, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.
[31] Edward I. George. Minimax multiple shrinkage estimation. *Ann. Statist.*, 14(1):188–205, 1986. ISSN 0090-5364.
[32] Edward I. George. Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.*, 81(394): 437–445, 1986. ISSN 0162-1459.
[33] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *submitted*, 2011.
[34] Ch. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14 (4):1089–1107, 2008.
[35] A. Goldenshluger and O. V. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008. ISSN 1350-7265.
[36] G. K. Golubev. Nonparametric estimation of smooth densities of a distribution in L_2. *Problemy Peredachi Informatsii*, 28(1):52–62, 1992. ISSN 0555-2923.
[37] Y. Golubev. On universal oracle inequalities related to high dimensional linear models. *Ann. Statist. (To appear)*, 2010.

[38] A. B. Juditsky and A. S. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.
[39] A. B. Juditsky and A. S. Nemirovski. Nonparametric denoising of signals with unknown local structure. I. Oracle inequalities. *Appl. Comput. Harmon. Anal.*, 27(2):157–179, 2009.
[40] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Computational learning theory (EuroCOLT)*, volume 1572 of *Lecture Notes in Comput. Sci.*, pages 153–167. Springer, Berlin, 1999.
[41] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72 (electronic), 2003/04.
[42] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *NIPS*, pages 423–430, 2002.
[43] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.
[44] G. Leung. *Information Theory and Mixing Least Squares Regression*. PhD thesis, Yale University, 2004.
[45] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8):3396–3410, 2006.
[46] K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007.
[47] D. A. McAllester. Some pac-bayesian theorems. In *COLT*, pages 230–234, New York, NY, USA, 1998. ACM.
[48] A. S. Nemirovski. *Topics in non-parametric statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.
[49] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inf.*, 16(2):52–68, 1980.
[50] J. Polzehl and V. G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(2):335–354, 2000.
[51] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.
[52] Ph. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *arXiv:1003.2654, submitted*, 2010.
[53] J. Salmon and E. Le Pennec. NL-Means and aggregation procedures. In *ICIP*, pages 2977–2980, 2009.
[54] M. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.*, 3:233–269, March 2003.
[55] J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 2000.
[56] C. M. Stein. Estimation of the mean of a multivariate distribution. In *Proc. Prague Symp. Asymptotic Statist.*, 1973.
[57] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6): 1135–1151, 1981.
[58] A. B. Tsybakov. Optimal rates of aggregation. In *COLT*, pages 303–313, 2003.
[59] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.
[60] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1): 135–161, 2000.
[61] Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statist. Sinica*, 13 (3):783–809, 2003.

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

[63] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

ÉCOLE DES PONTS PARISTECH 6,, AV BLAISE PASCAL - CITÉ DESCARTES, CHAMPS-SUR-MARNE, 77455 MARNE-LA-VALLÉ CEDEX 2 -, FRANCE

E-mail address: dalalyan@imagine.enpc.fr

2 PL. JUSSIEU, BP 7012, 75251 PARIS CEDEX 05-, FRANCE

E-mail address: salmon@math.jussieu.fr