Optimal aggregation of affine estimators

Joseph Salmon
LPMA
Université Paris Diderot Paris 7
name@math.jussieu.fr

Arnak Dalalyan
LIGM / IMAGINE
Université Paris Est / ENPC
name@imagine.enpc.fr

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 the least square regression, the kernel ridge regression, the shrinkage estimators, etc.—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, the most famous aggregation procedures such as *Boosting* (Freund, 1990), *Bagging* (Breiman, 1996) or *Random Forest* (Amit and Geman, 1997) were successfully used in practice for a large variety of applications. Moreover, the most recent Machine Learning competitions such as the Pascal VOC or the Netflix challenge were 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 (2000) was the first work concerned by the theoretical analysis of aggregation procedures. It was followed by a paper by Juditsky and Nemirovski (2000), as well as by a series of papers by Catoni (see Catoni (2004) for a comprehensive account) and Yang (2000, 2003, 2004). For the regression model, a significant progress was achieved by Tsybakov (2003) 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 was further developed by Lounici (2007), Rigollet and Tsybakov (2007), Lecué (2007), Bunea et al. (2007), 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 the 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 was reached by Leung and Barron (2006). Building on a very elegant but not very well known result of George (1986), 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 (2007, 2008), established the validity of Leung and Barron's result under more general (non Gaussian) noise distributions provided that 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.

Our interest in affine estimators is motivated by several reasons. First of all, affine estimators encompass many popular estimators such as the smoothing splines, the Pinsker estimator Pinsker (1980), Efromovich and Pinsker (1996), the local polynomial estimators, the non-local means Buades et al. (2005), Salmon and Le Pennec (2009), 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 (2009)—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 (cf. Juditsky and Nemirovski (2009) for precise definitions). It is worth noting that oracle inequalities for the penalized empirical risk minimizer were also established by Golubev (2010), and for the model selection by Arlot and Bach (2009), Baraud et al. (2010).

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. We assume that the design is deterministic and that the noise is Gaussian with a given covariance matrix. 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 deterministic estimators considered by Dalalyan and Tsybakov (2008) and of projection estimators treated by Leung and Barron (2006).

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:

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

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 diagonal (so it can be written $\Sigma = \operatorname{diag}(\sigma_1^2, \cdots, \sigma_n^2)$), with a known upper bound on the spectral norm $|||\Sigma|||$. In our case, $|||\Sigma||| = \max_{i=1,\cdots,n} \sigma_i^2$. 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.

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

$$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda, \tag{2}$$

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 $\boldsymbol{Y}$. It is well-known that the quadratic risk of the estimator (2) is given by

$$r_\lambda = \mathbb{E}\big(\|\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda\|_n^2\big) = \|(A_\lambda - I_{n\times n})f + \boldsymbol{b}_\lambda\|_n^2 + \frac{\operatorname{Tr}(A_\lambda \Sigma A_\lambda^\top)}{n} \tag{3}$$

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

$$\hat{r}_\lambda = \big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}(\Sigma A_\lambda) - \frac{1}{n}\sum_{i=1}^n \sigma_i^2 \tag{4}$$

is an unbiased estimator of r_λ (direct application of Stein's Lemma, cf. Appendix).

Let us describe now different families of linear and affine estimators successfully used in the statistical literature (cf., for instance, Arlot and Bach (2009)). Our results apply to all these families and lead 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 by Barron et al. (1999), 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 $A_\lambda = X_\lambda (X_\lambda^\top X_\lambda)^{-1} X_\lambda^\top$.

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 $A = \operatorname{diag}(a_1, \dots, a_n)$. 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 $w,\alpha > 0$. The set $\Lambda = (\mathbb{R}_+^*)^2$ indexes 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 (Shawe-Taylor and Cristianini, 2000, 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 proposed in Arlot and Bach (2009), 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

$$A_\lambda = \Big(\sum_{m=1}^M \lambda_m K_m\Big)\Big(\sum_{m=1}^M \lambda_m K_m + nI_{n\times n}\Big)^{-1}. \tag{5}$$

It is worth mentioning that the formulation in Eq.(5) can be linked to the group Lasso Yuan and Lin (2006) and to the multiple kernel Lanckriet et al. (2003/04) — see Bach (2008), Arlot and Bach (2009) for more details.

1.3 Organization of the paper

In Section 2, we introduce EWA and state a PAC-Bayes type bound assessing the optimality of EWA in combining affine estimators. As a consequence, we provide in Section 3 sharp oracle inequalities in various set-ups: ranging from finite to continuous families of constituent estimators and including the sparsity scenario. In Section 4, we apply our main results to prove that combining Pinsker's type filters with EWA leads to an asymptotically sharp adaptive procedure over the Sobolev ellipsoids. Section 5 is devoted to a numerical comparison of EWA with other classical filters (soft thresholding, blockwise shrinking, etc.), and illustrates the potential benefits of the aggregation. Some concluding remarks are presented in Section 6, while technical proofs are postponed to the Appendix.

2 Aggregation of estimators: main result

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 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 $\hat{\boldsymbol{f}}_{\text{aggr}}$, not necessarily in the family $\mathscr{F}_\Lambda$, that mimics the performance of the best constituent estimator. It is called *oracle* because of its dependence on the unknown function f. 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:

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\text{aggr}} - \boldsymbol{f}\|_n^2 \le C_n \inf_{\lambda\in\Lambda}\Big(\mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\Big) + R_n, \tag{6}$$

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. The residual term R_n depends on the complexity (size) of the family $\mathscr{F}_\Lambda$, as on the amount of noise, measured in term of variance in our context.

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:

$$\hat{\pi}(d\lambda) = \theta(\lambda)\pi(d\lambda) \qquad \text{with} \qquad \theta(\lambda) = \frac{\exp(-n\hat{r}_\lambda/\beta)}{\int_\Lambda \exp(-n\hat{r}_\omega/\beta)\pi(d\omega)}. \tag{7}$$

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 $\hat{\pi}$:

$$\hat{\boldsymbol{f}}_{\text{EWA}} = \int_\Lambda \hat{\boldsymbol{f}}_\lambda \, \hat{\pi}(d\lambda)\,. \tag{8}$$

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 (1986).

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, see for instance Zhang (2006), Audibert (2009).

2.2 Main result

To state our main result, 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 (PAC Bayesian Bound) *If either one of the following conditions is satisfied:*

$\mathbf{C}_1$: *The matrices A_λ are orthogonal projections (i.e., symmetric and idempotent) and the vectors $\boldsymbol{b}_\lambda$ satisfy $A_\lambda \boldsymbol{b}_\lambda = 0$, for all $\lambda \in \Lambda$.*

$\mathbf{C}_2$: *The matrices A_λ are all symmetric, positive semidefinite and satisfy $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$, $A_\lambda \Sigma = \Sigma A_\lambda$ for all $\lambda, \lambda' \in \Lambda$. All the vectors $\boldsymbol{b}_\lambda$ are zero.*

Then, the risk of the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ defined by Equations (7), (8) and (4) satisfies the inequality

$$r_{EWA} = \mathbb{E}(\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{p \in \mathscr{P}_\Lambda} \left(\int_\Lambda \mathbb{E}\left(\left\| \hat{\boldsymbol{f}}_\lambda - \boldsymbol{f} \right\|_n^2 \right) p(d\lambda) + \frac{\beta}{n} \mathscr{K}(p,\pi) \right) \tag{9}$$

provided that $\beta \ge \alpha |||\Sigma|||$, where $\alpha = 4$ if $\mathbf{C}_1$ holds true and $\alpha = 8$ if $\mathbf{C}_2$ holds true.

All the proofs of our results are given in the appendix, at the end of the paper.

Note also that the result of Theorem 1 applies to the estimator $\hat{\boldsymbol{f}}_{\text{EWA}}$ 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 risks $\hat{r}_\lambda$ that is used for defining $\hat{\boldsymbol{f}}_{\text{EWA}}$. The exponentially weighted aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ is easily extended to handle the more realistic situation where an unbiased estimate $\widehat{\Sigma}$, independent of $\boldsymbol{Y}$, of the covariance matrix Σ is available. Simply replace Σ by $\widehat{\Sigma}$ in the definition of the unbiased risk estimate (4). When the estimators $\hat{\boldsymbol{f}}_\lambda$ satisfy π-a.e. condition $\mathbf{C}_1$ or $\mathbf{C}_2$, choosing $\beta = \alpha |||\widehat{\Sigma}|||$, it can be checked that a claim similar to Theorem 1 remains valid.

Another observation is that using the extension of Stein's lemma presented in (Dalalyan and Tsybakov, 2008, Lemma 1), a result similar to Theorem 1 can be established for some specific non Gaussian noise distributions, provided that the components of the noise vector are independent.

3 Sharp oracle inequalities

In this section, we discuss consequences of the main result for specific choices of prior measures. Some of them are closely related to the oracle inequalities presented in Dalalyan and Tsybakov (2007, 2008), Alquier and Lounici (2010), Rigollet and Tsybakov (2011) especially when dealing with the sparsity scenario in the high dimensional framework.

3.1 Discrete oracle inequality

In order to demonstrate that Inequality (9) can be reformulated in terms of an OI as defined by (6), let us consider the simple 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 *If either one of the conditions* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *(cf. Theorem 1) is fulfilled and* π *is supported by* $\mathbb{N}$*, then the aggregate* $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ *defined by Equations (7), (8) and (4) satisfies the inequality*

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{j\in\mathbb{N}:\pi_j>0} \left(\mathbb{E}\|\hat{\boldsymbol{f}}_j - \boldsymbol{f}\|_n^2 + \frac{\beta \log(1/\pi_j)}{n} \right) \tag{10}$$

provided that $\beta \ge \alpha|||\Sigma|||$*, where* $\alpha = 4$ *if* $\mathbf{C}_1$ *holds true and* $\alpha = 8$ *if* $\mathbf{C}_2$ *holds true.*

Proof: It suffices to apply Theorem 1 and to bound the RHS from above by the minimum over all Dirac measures $p = \delta_j$ with j such that $\pi_j > 0$. ■

3.2 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 $d_2(\lambda, \Lambda)$ the largest real $\tau > 0$ such that the ball centered at λ with radius τ is included in Λ. In what follows, $\mathrm{Leb}(\cdot)$ 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* $\lambda \mapsto r_\lambda$ *is Lipschitz continuous, i.e.,* $|r_{\lambda'} - r_\lambda| \le L_r \|\lambda' - \lambda\|_2$*,* $\forall \lambda, \lambda' \in \Lambda$*. Under the conditions* $\mathbf{C}_1$ *or* $\mathbf{C}_2$ *aggregate* $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ *satisfies the inequality*

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2 \le \inf_{\lambda\in\Lambda} \left\{ \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 + \frac{\beta M}{n} \log\Big(\frac{\sqrt{M}}{2\min(n^{-1}, d_2(\lambda,\Lambda))} \Big) \right\} + \frac{L_r + \beta \log\big(\mathrm{Leb}(\Lambda)\big)}{n}. \tag{11}$$

for every $\beta \ge \alpha|||\Sigma|||$ *where* $\alpha = 4$ *if* $\mathbf{C}_1$ *holds true and* $\alpha = 8$ *if* $\mathbf{C}_2$ *holds true.*

3.3 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 $n/\max_i \sigma_i^2$). 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 $\lambda^* \in \Lambda$ such that the risk of $\hat{\boldsymbol{f}}_{\lambda^*}$ is small. Then, it is natural to choose a prior π that promotes the sparsity of λ. This can be done in the same vein as in Dalalyan and Tsybakov (2007, 2008), by means of the heavy tailed prior:

$$\pi(d\lambda) \propto \prod_{j=1}^M \frac{1}{(1 + |\lambda_j/\tau|^2)^2}\, \mathbb{1}_\Lambda(\lambda) d(\lambda), \tag{12}$$

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

Proposition 3 *Let* $\Lambda = \mathbb{R}^M$ *and let* π *be defined by (12). Assume that the mapping* $\lambda \mapsto r_\lambda$ *is continuously differentiable and, for some* $M \times M$ *matrix* $\mathscr{M}$*, satisfies:*

$$r_\lambda - r_{\lambda'} - \nabla r_{\lambda'}^\top (\lambda - \lambda') \le (\lambda - \lambda')^\top \mathscr{M} (\lambda - \lambda'), \qquad \forall \lambda, \lambda' \in \Lambda. \tag{13}$$

If either one of the conditions $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *(cf. Theorem 1) is fulfilled, then the aggregate* $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ *defined by Equations (7), (8) and (4) satisfies the inequality*

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{\lambda\in\mathbb{R}^M} \Big\{ \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 + \frac{4\beta}{n} \sum_{j=1}^M \log\Big(1 + \frac{|\lambda_j|}{\tau}\Big) \Big\} + \mathrm{Tr}(\mathscr{M})\tau^2 \tag{14}$$

provided that $\beta \ge \alpha|||\Sigma|||$*, where* $\alpha = 4$ *if* $\mathbf{C}_1$ *holds true and* $\alpha = 8$ *if* $\mathbf{C}_2$ *holds true.*

Let us discuss here some consequences of this sparsity oracle inequality. First of all, let us remark that in most cases $\mathrm{Tr}(\mathscr{M})$ is on the order of M and the choice $\tau = \sqrt{\beta/(nM)}$ ensures that the last term in the RHS of Eq. (14) 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, condition $\mathbf{C}_2$. 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, cf. Cohen (1966)). 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 $\lambda = (\lambda_1, \ldots, \lambda_M)$. If the collection $\{\hat{\boldsymbol{g}}_i\}$ satisfies condition $\mathbf{C}_2$, then it is also the case for the collection of their linear combinations $\{\hat{\boldsymbol{f}}_\lambda\}$. Moreover, the mapping $\lambda \mapsto r_\lambda$ is quadratic with the Hessian matrix $\nabla^2 r_\lambda$ 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 (13) holds with $\mathscr{M}$ being the Hessian divided by 2. Therefore, setting $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$, we get $\mathrm{Tr}(\mathscr{M}) \le |||\sum_{j=1}^M G_j^2|||(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2) \le M(\|f\|_n^2 + \|\boldsymbol{\sigma}\|_n^2)$, where the norm of a matrix is understood as its largest singular value. Applying Proposition 3 with $\tau = \sqrt{\beta/(nM)}$, we get for $\beta \ge 8|||\Sigma|||$,

$$\mathbb{E}\big(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big) \le \inf_{\alpha, j, j'} \mathbb{E}\|\alpha\hat{\boldsymbol{g}}_j + (1-\alpha)\hat{\boldsymbol{g}}_{j'} - \boldsymbol{f}\|_n^2 + \frac{8\beta}{n}\log\left(1 + \sqrt{\frac{Mn}{\beta}}\right) + \frac{\beta}{n}(\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2), \tag{15}$$

where the inf is taken over all $\alpha \in [0,1]$ and $j, j' \in \{1, \ldots, M\}$. This shows that, using EWA with a sufficiently large temperature, 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.

3.4 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 $\lambda = (a, b, k) \in [0,1]^2 \times \{1, \ldots, n\} := \Lambda$, $\hat{\boldsymbol{f}}_\lambda = A_\lambda\boldsymbol{Y}$ where $A_\lambda = \mathrm{diag}\big(a\mathbb{1}(i \le k) + b\mathbb{1}(i > k), i = 1, \cdots, n\big)$. 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 $A_\lambda\boldsymbol{Y}$. If Σ is a diagonal matrix, then for any $\lambda \in \Lambda$ and for any $\beta \ge 8|||\Sigma|||$,*

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le \mathbb{E}(\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2) + \frac{\beta}{n}\Big\{1 + \log\Big(\frac{n^2\|\boldsymbol{f}\|_n^2 + n\,\mathrm{Tr}(\Sigma)}{12\beta}\Big)\Big\}. \tag{16}$$

The proof of this result can be found in Dalalyan and Salmon (2011).

In the case $\Sigma = I_{n\times n}$, this result is comparable to (Leung, 2004, 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

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{Leung}} - \boldsymbol{f}\|_n^2) \le \mathbb{E}(\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2) + \frac{9}{n} + \frac{8}{n}\min_{K>0}\Big\{K \vee \Big(\log\frac{n-6}{K} - 1\Big)\Big\}. \tag{17}$$

4 Application to minimax adaptive estimation

In the celebrated paper Pinsker (1980)proved that in the model (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 sine transform of $\boldsymbol{f}$, hereafter denoted by $\mathscr{D}\boldsymbol{f}$. Pinsker's result—restricted to Sobolev ellipsoids $\mathscr{F}(\alpha, R) = \big\{\boldsymbol{f} \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha}\theta_k(\boldsymbol{f})^2 \le R\big\}$ and to the homoscedastic noise ($\Sigma = \sigma^2 I_{n\times n}$)—states that, as $n \to \infty$, the equivalences

$$\inf_{\hat{\boldsymbol{f}}} \sup_{\boldsymbol{f}\in\mathscr{F}(\alpha,R)} \mathbb{E}\big(\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|_n^2\big) \sim \inf_{A} \sup_{f\in\mathscr{F}(\alpha,R)} \mathbb{E}\big(\|A\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big) \tag{18}$$

$$\sim \inf_{w>0} \sup_{f\in\mathscr{F}(\alpha,R)} \mathbb{E}\big(\|A_{\alpha,w}\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big) \tag{19}$$

hold (Tsybakov, 2009, Theorem 3.2), where the first inf is taken over all possible estimators $\hat{f}$ and $A_{\alpha,w} = \mathscr{D}^\top\mathrm{diag}\big((1 - k^\alpha/w)_+; k = 1, \ldots, n\big)\mathscr{D}$ is the Pinsker filter in the discrete sine 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 (1984). The next result, that is a direct consequence of Theorem 1, shows that EWA with linear constituent estimators is also asymptotically sharp adaptive over Sobolev ellipsoids.

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

$$\pi(d\lambda) = \frac{2n_\sigma^{-\alpha/(2\alpha+1)}}{\left(1 + n_\sigma^{-\alpha/(2\alpha+1)} w\right)^3} e^{-\alpha} d\alpha dw, \tag{20}$$

where $n_\sigma = n/\sigma^2$. Then, in model (1) with homoscedastic errors, the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ 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[1] *on the family of classes $\{\mathscr{F}(\alpha, R) : \alpha > 0, R > 0\}$.*

It is worth noting that the exact minimax adaptivity property of our estimator $\hat{\boldsymbol{f}}_{\text{EWA}}$ 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 and Efromovich-Pinsker estimators or the step of discretization and the maximal value of bandwidth Cavalier et al. (2002). Another nice property of the estimator $\hat{\boldsymbol{f}}_{\text{EWA}}$ is that it does not require any pilot estimator based on the data splitting device Efromovich (1996), Yang (2004).

5 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:
`www.math.jussieu.fr/~salmon/code/index_codes.php`

We evaluate different estimation routines on several 1D signals, introduced by Donoho and Johnstone (1994, 1995) and considered as benchmark in literature on signal processing. The six signals we retained for our experiments because of their diversity are depicted in Figure 1. Since all these signals are non-smooth, we have also carried out experiments on their smoothed versions obtained by taking an antiderivative, see Figure 1. In what follows, the experiment on non-smooth 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 Sine Transform (DST) denoted by $\boldsymbol{\theta}(\boldsymbol{Y}) = \big(\theta_1(\boldsymbol{Y}), \ldots, \theta_n(\boldsymbol{Y})\big)^\top$.

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

Soft Thresholding (ST), Donoho and Johnstone (1994): For a given threshold parameter t, the soft thresholding estimator of the vector of DST coefficients $\theta_k(\boldsymbol{f})$ is defined by

$$\widehat{\theta}_k = \operatorname{sgn}\big(\theta_k(\boldsymbol{Y})\big)\big(|\theta_k(\boldsymbol{Y})| - \sigma t\big)_+. \tag{21}$$

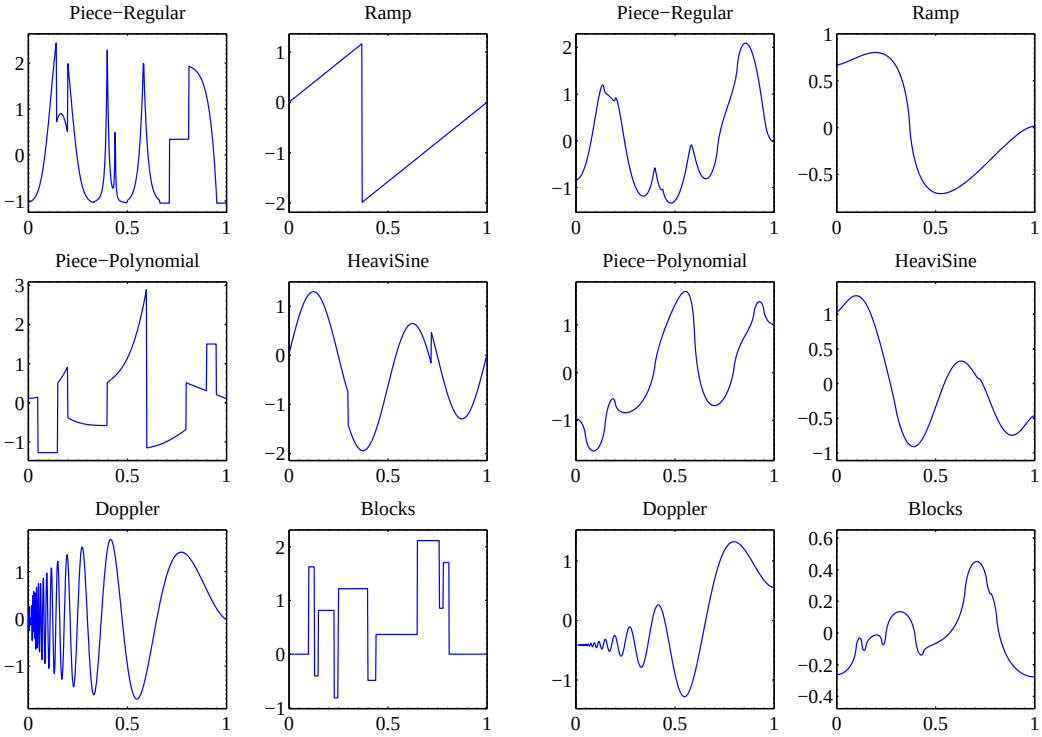
In our experiments, we use the threshold minimizing the estimated unbiased risk defined via Stein's lemma. This procedure is referred to as SURE-shrink in Donoho and Johnstone (1995).

Blockwise James-Stein (BJS) shrinkage, Cai (1999): 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 of true coefficients $\theta_{B_k}(\boldsymbol{f}) = \big(\theta_j(\boldsymbol{f})\big)_{j\in B_k}$ are estimated by shrinking the blocks of noisy coefficients $\theta_{B_k}(\boldsymbol{Y})$:

$$\widehat{\theta}_{B_k} = \left(1 - \frac{\lambda L\sigma^2}{S_k^2(\boldsymbol{Y})}\sigma\right)_+ \theta_{B_k}(\boldsymbol{Y}), \qquad k = 1, \cdots, N \tag{22}$$

where $S_k^2(\boldsymbol{Y}) = \|\theta_{B_k}(\boldsymbol{Y})\|_2^2$ and $\lambda = 4.50524$ as in Cai (1999).

[1] see (Tsybakov, 2009, Definition 3.8)



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

Figure 1: Test signals used in our experiment: Piece-Regular, Ramp, Piece-Polynomial, HeaviSine, Doppler and Blocks. (a) non-smooth (Experiment I) and (b) smooth (Experiment II).

Unbiased risk estimate (URE) minimization, Golubev (1992), Cavalier et al. (2002): it consists in using a Pinsker filter, as defined in Section 4, 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. The bi-dimensional grid used in all the experiments has 100×100 elements. We refer to Cavalier et al. (2002) for the closed-form formula of the unbiased risk estimator.

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 URE and 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 2) or discontinuities (cf. Figure 3). In most cases, the EWA method also outperforms 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 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 EWA as compared to ST and BJS procedures. The point is to stress the importance of having sharp adaptivity up to

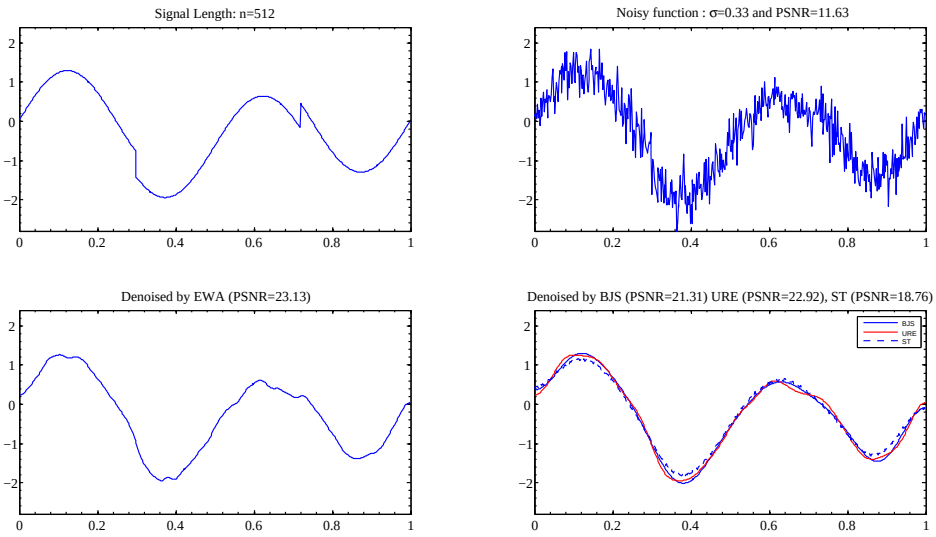


Figure 2: Heavisine. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation $\sigma = 0.33$ (right). The second row gives denoised version obtained by EWA (left), BJS, ST and URE (right). The PSNR is computed by the formula $\mathrm{PSNR} = 10\log_{10}\left(\max(f)^2/\mathrm{MSE}\right)$.

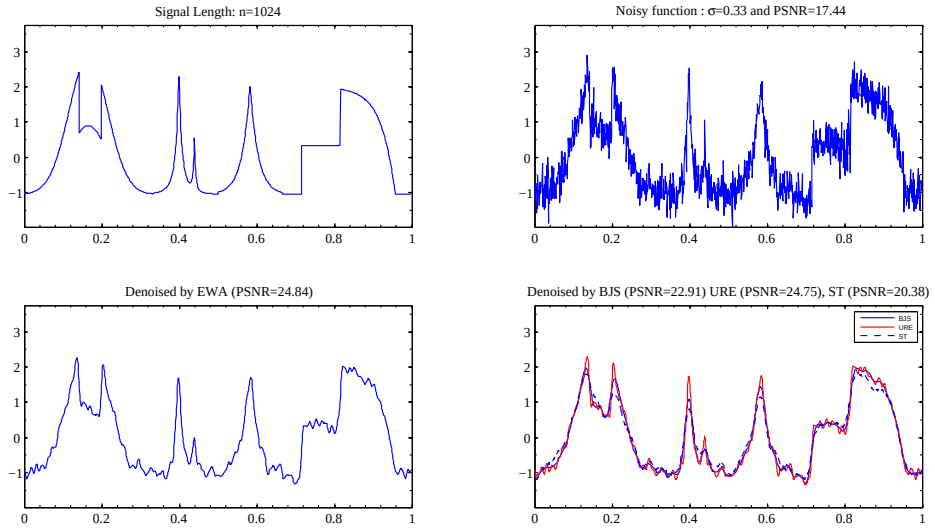


Figure 3: Piece-Regular. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation $\sigma = 0.33$ (right). The second row gives denoised version obtained by EWA (left) and by BJS, ST and URE (right). The PSNR is computed by the formula $\mathrm{PSNR} = 10\log_{10}\left(\max(f)^2/\mathrm{MSE}\right)$.

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 EWA and URE do attain.

6 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

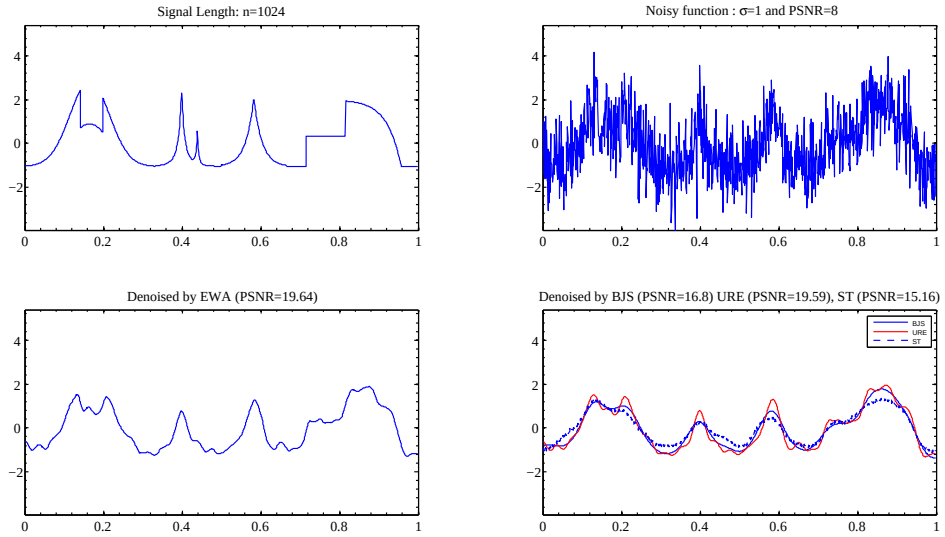


Figure 4: Piece-Regular. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation $\sigma = 1$ (right). The second row gives denoised version obtained by EWA (left) and by BJS, ST and URE (right). The PSNR is computed by the formula PSNR $= 10\log_{10}\left(\max(f)^2/\mathrm{MSE}\right)$.

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 EWA applied to the family of Pinsker's estimators produces an estimator, which is adaptive in the exact minimax sense. Next in our agenda is carrying out an experimental evaluation of the proposed aggregate using the approximation schemes described by Dalalyan and Tsybakov (2009), Rigollet and Tsybakov (2011) and Alquier and Lounici (2010). 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 (2008).

Acknowledgments

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

References

P. Alquier and K. Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Statist.*, 5:127–145, 2010.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9: 1545–1588, October 1997.

S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *NIPS*, pages 46–54, 2009.

J-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.

F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

Y. Baraud, Ch. Giraud, and S. Huet. Estimator selection in the gaussian setting. *submitted*, 2010.

A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

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: Comparison of several adaptive methods on the six (non-smooth) signals of interest. For each signal length n and each method, we give the average value of $n \times (\mathrm{MSE} - \mathrm{MSE}_{\mathrm{Oracle}})$ and the corresponding standard deviation below, for 1000 replications of the experiment. Negative values indicate that in some cases the EWA procedure has a smaller risk than that of the best linear estimator used for the aggregation, which is possible since the EWA itself is not a linear estimator.

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.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4): 1674–1697, 2007.

T. T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002.

A. Cohen. All admissible linear estimates of the mean vector. *The Annals of Mathematical Statistics*, 37 (2):458–463, 1966.

A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. Technical Report arXiv:1104.3969v2 [math.ST], April 2011.

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

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.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT*, 2009.

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: Comparison of several adaptive methods on the six smoothed signals of interest. For each signal length n and each method, we give the average value of $n(\mathrm{MSE}-\mathrm{MSE}_{\mathrm{Oracle}})$ and the corresponding standard deviation below, for 1000 replications of the experiment.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3): 425–455, 1994.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

S. Y. Efromovich. On nonparametric regression for IID observations in a general setting. *Ann. Statist.*, 24 (3):1125–1144, 1996.

S. Y. Efromovich and M. S. Pinsker. A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, 1(11):58–65, 1984.

S. Y. Efromovich and M. S. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996.

Y. Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the third annual workshop on Computational learning theory*, COLT, pages 202–216, 1990.

E. I. George. Minimax multiple shrinkage estimation. *Ann. Statist.*, 14(1):188–205, 1986.

Ch. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.

G. K. Golubev. Nonparametric estimation of smooth densities of a distribution in L_2. *Problemy Peredachi Informatsii*, 28(1):52–62, 1992.

Yuri Golubev. On universal oracle inequalities related to high-dimensional linear models. *Ann. Statist.*, 38(5):2751–2780, 2010.

A. B. Juditsky and A. S. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

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.

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.

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4): 1000–1022, 2007.

G. Leung. *Information Theory and Mixing Least Squares Regression*. PhD thesis, Yale University, 2004.

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8):3396–3410, 2006.

K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007.

A. S. Nemirovski. *Topics in non-parametric statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.

M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inf.*, 16 (2):52–68, 1980.

Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

Ph. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–471, 2011.

J. Salmon and E. Le Pennec. NL-Means and aggregation procedures. In *ICIP*, pages 2977–2980, 2009.

J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 2000.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

A. B. Tsybakov. Optimal rates of aggregation. In *COLT*, pages 303–313, 2003.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009.

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

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

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

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.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52(4):1307–1321, 2006.

Appendix

A Stein's Lemma with heteroscedastic noise

To define the EWA estimator, we first need to determine an unbiased risk estimate for any of the constituent estimators. We adapt a systematic method based on Stein's Lemma to the heteroscedastic framework. We recall this lemma given in Stein (1981), for our setting:

Stein's Lemma 1 *With the model* (1)*, if the estimator* $\hat{\boldsymbol{f}}$ *is almost everywhere differentiable in* Y *and if each* $\partial_{y_i}\hat{\boldsymbol{f}}_i$ *has finite first moment, then*

$$\hat{r} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}\|_n^2 + \frac{2}{n}\sum_{i=1}^n \sigma_i^2 \partial_{y_i}\hat{f}_i - \frac{1}{n}\sum_{i=1}^n \sigma_i^2, \tag{23}$$

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

Proof: For any $i = 1, \cdots, n$, one has

$$\mathbb{E}(Y_i - \hat{f}_i)^2 = \mathbb{E}(Y_i - f_i)^2 + \mathbb{E}(f_i - \hat{f}_i)^2 + 2\mathbb{E}\big[(Y_i - f_i)(f_i - \hat{f}_i)\big],$$

The following identity is the classical Stein Lemma (cf. Tsybakov (2009) p.157), based on integration by parts:

$$\mathbb{E}\big[(Y_i - f_i)\hat{f}_i\big] = \sigma_i^2 \mathbb{E}\big[\partial_{y_i} f_i\big]. \tag{24}$$

where the differentiation is according to Y_i. Using the last two displays, one has:

$$\mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{f}}\|_n^2 = \mathbb{E}\|\boldsymbol{Y} - \boldsymbol{f}\|_n^2 + \mathbb{E}\|f - \hat{\boldsymbol{f}}\|_n^2 - \frac{2}{n}\mathbb{E}\sum_{i=1}^n \sigma_i^2 \partial_{y_i} f_i\,, \tag{25}$$

leading to the announced unbiased risk estimate. ■

B Main Result

Now, we can apply Stein's Lemma for any estimator $\hat{\boldsymbol{f}}_\lambda$, so that we can build $\hat{r}_\lambda$ for any $\lambda \in \Lambda$. 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 $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$. Affine estimators can be defined by

$$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda, \tag{26}$$

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 design points $x_1, \cdots, x_n$ but not on the data vector $\boldsymbol{Y}$. It is easy to check that the the divergence term $\sum_{i=1}^n \sigma_i^2 \partial_{y_i}\hat{\boldsymbol{f}}_i$ in Stein's Lemma is simply $\mathrm{Tr}(\Sigma A_\lambda)$ for affine estimators. Then $\hat{r}_\lambda$, defined by

$$\hat{r}_\lambda = \big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\sum_{i=1}^n \sigma_i^2 \tag{27}$$

is an unbiased estimator of r_λ.

In order to state our main result, 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 $p, p' \in \mathscr{P}_\Lambda$.

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

Theorem 1 *If either one of the following conditions is satisfied:*

$\mathbf{C}_1$: *The matrices* A_λ *are orthogonal projections (i.e., symmetric and idempotent) and the vectors* $\boldsymbol{b}_\lambda$ *satisfy* $A_\lambda \boldsymbol{b}_\lambda = 0$*, for all* $\lambda \in \Lambda$*.*

$\mathbf{C}_2$: *The matrices* A_λ *are all symmetric, positive semidefinite and satisfy* $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$*,* $A_\lambda \Sigma = \Sigma A_\lambda$ *for all* $\lambda, \lambda' \in \Lambda$*. All the vectors* $\boldsymbol{b}_\lambda$ *are zero.*

Then, the aggregate $\hat{\boldsymbol{f}}_{\mathrm{EWA}}$ *defined by Equations (7), (8) and (4) satisfies the inequality*

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

provided that $\beta \ge \alpha |||\Sigma|||$*, where* $\alpha = 4$ *if* $\mathbf{C}_1$ *holds true and* $\alpha = 8$ *if* $\mathbf{C}_2$ *holds true.*

Proof:[when $\mathbf{C}_2$ is satisfied] According to Stein's lemma, the quantity

$$\hat{r}_{\mathrm{EWA}} = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 + \frac{2}{n}\sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{f}_{\mathrm{EWA},i} - \frac{1}{n}\sum_{i=1}^n \sigma_i^2 \tag{28}$$

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

$$\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 = \int_\Lambda \Big(\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\|_n^2 - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2\Big)\theta(\lambda)\pi(d\lambda). \tag{29}$$

By interchanging the integral and differential operators, we get the following expression for the derivatives of $\hat{f}_{\mathrm{EWA},i}$:

$$\partial_{y_i} \hat{f}_{\mathrm{EWA},i} = \int_\Lambda \big(\partial_{y_i} \hat{f}_{\lambda,i}\big)\theta(\lambda)\,\pi(d\lambda) + \int_\Lambda \hat{f}_{\lambda,i}\big(\partial_{y_i}\theta(\lambda)\big)\pi(d\lambda). \tag{30}$$

Let us defined $A_{\mathrm{EWA}} \triangleq \int_\Lambda A_\lambda \theta(\lambda)\pi(d\lambda)$. With this notation, the last equality, combined with Equations (4), (28), (29) and the fact that $\sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{f}_{\lambda,i} = \mathrm{Tr}(\Sigma A_\lambda)$, implies that

$$\hat{r}_{\mathrm{EWA}} = \int_\Lambda \big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2\big)\theta(\lambda)\,\pi(d\lambda) + \frac{2}{n}\sum_{i=1}^n \sigma_i^2 \int_\Lambda \hat{f}_{\lambda,i}\big(\partial_{y_i}\theta(\lambda)\big)\pi(d\lambda).$$

Taking into account that $\int_\Lambda \hat{f}_{\mathrm{EWA},i}\big(\partial_{y_i}\theta(\lambda)\big)\pi(d\lambda) = \hat{f}_{\mathrm{EWA},i}\partial_{y_i}\big(\int_\Lambda \theta(\lambda)\pi(d\lambda)\big) = 0$, we come up with the following expression for the unbiased risk estimate:

$$\hat{r}_{\mathrm{EWA}} = \int_\Lambda \Big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 + 2\big\langle \nabla_{\boldsymbol{Y}} \log\theta(\lambda) | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\big\rangle_n\Big)\theta(\lambda)\,\pi(d\lambda) \tag{31}$$

$$= \int_\Lambda \Big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 - 2n\beta^{-1}\big\langle \nabla_{\boldsymbol{Y}} \hat{r}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}})\big\rangle_n\Big)\theta(\lambda)\,\pi(d\lambda). \tag{32}$$

Note that, so far, the precise form of the constituent estimators has not been exploited. This form is important for computing $\nabla_{\boldsymbol{Y}} \hat{r}_\lambda$. In view of Equations (26) and (4), as well as the assumptions $A_\lambda^\top = A_\lambda$ and $\boldsymbol{b}_\lambda \equiv 0$ (holding thanks to $\mathbf{C}_2$), we get

$$\nabla_{\boldsymbol{Y}} \hat{r}_\lambda = \frac{2}{n}(I_{n\times n} - A_\lambda)^\top (I_{n\times n} - A_\lambda)\boldsymbol{Y} - \frac{2}{n}(I_{n\times n} - A_\lambda)^\top \boldsymbol{b}_\lambda = \frac{2}{n}(I_{n\times n} - A_\lambda)^2 \boldsymbol{Y}. \tag{33}$$

In what follows, we use the shorthand $I = I_{n\times n}$. Using this notation and Eq. (33), we get

$$\hat{r}_{\mathrm{EWA}} = \int_\Lambda \Big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 - \frac{4}{\beta}\big\langle (I - A_\lambda)^2 \boldsymbol{Y} | \Sigma(A_\lambda - A_{\mathrm{EWA}})\boldsymbol{Y}\big\rangle_n\Big)\theta(\lambda)\,\pi(d\lambda). \tag{34}$$

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 formula to $P = A_\lambda$ and $Q = A_{\mathrm{EWA}}$ we get the following expression: $\big\langle (I - A_\lambda)^2 \boldsymbol{Y} | \Sigma(A_\lambda - A_{\mathrm{EWA}})\boldsymbol{Y}\big\rangle_n = \big\langle (I - A_{\mathrm{EWA}})^2 \boldsymbol{Y} | \Sigma(A_\lambda - A_{\mathrm{EWA}})\boldsymbol{Y}\big\rangle_n - 2\big\langle \big(I - \frac{\hat{A}+A_\lambda}{2}\big)(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\big\rangle_n$. When one integrates over Λ with respect to the measure $\theta\cdot\pi$, the term of the first scalar product in the RHS of the last equation vanishes. On the other hand, positive semidefiniteness of matrices A_λ implies that of the matrix A_{EWA} and, therefore, $\big\langle \big(I - \frac{\hat{A}+A_\lambda}{2}\big)(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\big\rangle_n \le \langle (A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\big\rangle_n$. This inequality, in conjunction with (34) implies that

$$\begin{aligned}
\hat{r}_{\mathrm{EWA}} &\le \int_\Lambda \Big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 + \frac{8}{\beta}\langle (A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y} | \Sigma(A_{\mathrm{EWA}} - A_\lambda)\boldsymbol{Y}\big\rangle_n\Big)\theta(\lambda)\,\pi(d\lambda) \\
&= \int_\Lambda \Big(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2 + \frac{8}{\beta}\langle \hat{\boldsymbol{f}}_{\mathrm{EWA}} - \hat{\boldsymbol{f}}_\lambda | \Sigma(\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \hat{\boldsymbol{f}}_\lambda)\big\rangle_n\Big)\theta(\lambda)\,\pi(d\lambda) \\
&\le \int_\Lambda \Big(\hat{r}_\lambda - \Big(1 - \frac{8\max_i \sigma_i^2}{\beta}\Big)\|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\mathrm{EWA}}\|_n^2\Big)\theta(\lambda)\,\pi(d\lambda).
\end{aligned}$$

Taking into account the fact that $\beta \ge 8\max_i \sigma_i^2$, we get $\hat{r}_{\mathrm{EWA}} \le \int_\Lambda \hat{r}_\lambda \theta(\lambda)\,\pi(d\lambda) \le \int_\Lambda \hat{r}_\lambda \hat{\pi}(d\lambda) + \frac{\beta}{n}\mathscr{K}(\hat{\pi},\pi)$. 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 $p \in \mathscr{P}_\Lambda$ (see for instance Catoni (2004) p.160). Thus, for every $p \in \mathscr{P}_\Lambda$, it holds that

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

Taking the expectation of both sides, the desired result follows with Fatou's lemma. ■

Proof:[when $\mathbf{C}_1$ is satisfied] We can do the same calculation as when C_2 is satisfied until (32). In view of Equations (2) and (4), as well as the assumptions $A_\lambda^2 = A_\lambda^\top = A_\lambda$ and $A_\lambda^\top \boldsymbol{b}_\lambda \equiv 0$, we get

$$\nabla_{\boldsymbol{Y}} \hat{r}_\lambda = \frac{2}{n}(I_{n\times n} - A_\lambda)^\top (I_{n\times n} - A_\lambda)\boldsymbol{Y} - \frac{2}{n}(I_{n\times n} - A_\lambda)^\top \boldsymbol{b}_\lambda = \frac{2}{n}(I_{n\times n} - A_\lambda)\boldsymbol{Y} - \frac{2}{n}\boldsymbol{b}_\lambda. \tag{35}$$

Using the same shorthand $I = I_{n\times n}$ with Eq. (35) we come up with

$$\hat{r}_{\text{EWA}} = \int_\Lambda \left(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}}\|_n^2 - \frac{4}{\beta}\big\langle \boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}})\big\rangle_n \right) \theta(\lambda)\,\pi(d\lambda). \tag{36}$$

Now, since $\hat{\boldsymbol{f}}$ is the expectation of $\hat{\boldsymbol{f}}_\lambda$ with respect to the measure $\theta\cdot\pi$, we have

$$\begin{aligned}
\hat{r}_{\text{EWA}} &= \int_\Lambda \left(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}}\|_n^2 + \frac{4}{\beta}\langle \boldsymbol{Y} - \hat{\boldsymbol{f}}_{\text{EWA}} + \hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda | \Sigma(\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}})\rangle_n \right) \theta(\lambda)\,\pi(d\lambda) \\
&= \int_\Lambda \left(\hat{r}_\lambda - \|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}}\|_n^2 + \frac{4}{\beta}\langle \hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda | \Sigma(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)\rangle_n \right) \theta(\lambda)\,\pi(d\lambda) \\
&\le \int_\Lambda \left(\hat{r}_\lambda - \Big(1 - \frac{4\max_i \sigma_i^2}{\beta}\Big)\|\hat{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_{\text{EWA}}\|_n^2 \right) \theta(\lambda)\,\pi(d\lambda).
\end{aligned}$$

Taking into account the fact that $\beta \ge 4\max_i \sigma_i^2$, we get the same results as with condition $\mathbf{C}_2$: $\hat{r}_{\text{EWA}} \le \int_\Lambda \hat{r}_\lambda \theta(\lambda)\,\pi(d\lambda) \le \int_\Lambda \hat{r}_\lambda \hat{\pi}(d\lambda) + \frac{\beta}{n}\mathcal{K}(\hat{\pi},\pi)$. The end of the proof is unchanged and leads to the same general result as with condition $\mathbf{C}_2$, except for the choice of α. ■

C Continuous oracle inequality

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 $\lambda \mapsto r_\lambda$ is Lipschitz continuous, i.e., $|r_{\lambda'} - r_\lambda| \le L_r\|\lambda' - \lambda\|_2$, $\forall \lambda, \lambda' \in \Lambda$. Under the conditions $\mathbf{C}_1$ or $\mathbf{C}_1$ aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ satisfies the inequality*

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

for every $\beta \ge \alpha\|\Sigma\|$ where $\alpha = 4$ if $\mathbf{C}_1$ holds true and $\alpha = 8$ if $\mathbf{C}_2$ holds true.

Proof: It suffices to apply Theorem 1 and to bound from above the RHS of inequality (9)

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

Then, the RHS of the last inequality can be bounded from above by the minimum over all measures having as density $p_{\lambda_0,\tau_0}(\lambda) = \mathbb{1}_{B_{\lambda_0}(\tau_0)}(\lambda)/\text{Leb}(B_{\lambda_0}(\tau_0))$, with $\lambda_0 \in \Lambda$ and $\tau_0 = \min(1/n, d_2(\lambda_0,\Lambda))$ (hence $B_{\lambda_0}(\tau_0) \subset \Lambda$). Using the Lipschitz condition on r_λ, the bound on the risk becomes

$$\begin{aligned}
\mathbb{E}(\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2) &\le \int_\Lambda \big[|r_\lambda - r_{\lambda_0}| + r_{\lambda_0}\big]\, p_{\lambda_0,\tau_0}(d\lambda) + \frac{\beta}{n}\mathscr{K}(p_{\lambda_0,\tau_0},\pi) \\
\mathbb{E}(\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2) &\le r_{\lambda_0} + L_r\int_\Lambda \|\lambda - \lambda_0\|_2\, p_{\lambda_0,\tau_0}(d\lambda) + \frac{\beta}{n}\mathscr{K}(p_{\lambda_0,\tau_0},\pi) \\
\mathbb{E}(\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2) &\le r_{\lambda_0} + L_r\tau_0 + \frac{\beta}{n}\mathscr{K}(p_{\lambda_0,\tau_0},\pi)
\end{aligned} \tag{37}$$

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. π and the Kullback-Leibler divergence between these measures equals $\log\{\text{Leb}(\Lambda)/\text{Leb}\big(B_{\lambda_0}(\tau_0)\big)\}$. 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}: $\text{Leb}\big(B_{\lambda_0}(\tau_0)\big) \ge (2\tau_0/\sqrt{M})^M$. By combining this with inequality (37) the results of Proposition 2 is straightforward.

■

D Sparsity oracle inequality

Let us choose a prior π that promotes the sparsity of λ. This can be done in the same vein as in Dalalyan and Tsybakov Dalalyan and Tsybakov (2007, 2008), by means of the heavy tailed prior (Student $t(3)$ distribution):

$$\pi(d\lambda) \propto \prod_{j=1}^{M} \frac{1}{(1+|\lambda_j/\tau|^2)^2}\, \mathbb{1}_\Lambda(\lambda), \tag{38}$$

where $\tau > 0$ is a tuning parameter, that takes small values.

Proposition 3 *Let $\Lambda = \mathbb{R}^M$ and let π be defined by (12). Assume that the mapping $\lambda \mapsto r_\lambda$ is continuously differentiable and, for some $M \times M$ matrix $\mathscr{M}$, satisfies:*

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

If either one of the conditions $\mathbf{C}_1$ and $\mathbf{C}_2$ (cf. Theorem 1) is fulfilled, then the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ defined by Equations (7) and (4) satisfies the inequality

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

provided that $\beta \ge \alpha \max_{i=1,\dots,n} \sigma_i^2$, where $\alpha = 4$ if $\mathbf{C}_1$ holds true and $\alpha = 8$ if $\mathbf{C}_2$ holds true.

Proof: The proof is a simplified version of proofs given in Dalalyan and Tsybakov (2007, 2008), since Λ is the whole space, $\Lambda = \mathbb{R}^M$ instead of a bounded subset of $\mathbb{R}^M$.

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

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

By choosing $p_{\lambda^*}(\lambda) = \pi(\lambda - \lambda^*)$ for any $\lambda \in \mathbb{R}^M$, 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

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2) \le \inf_{\lambda^* \in \mathbb{R}^M} \left(r_{\lambda^*} + \mathrm{Tr}(\mathscr{M})\tau^2 + \frac{\beta}{n}\, \mathscr{K}(p_{\lambda^*}, \pi) \right) \tag{39}$$

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

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

We bound the quotient in the above equality by

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

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

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

and combine with Inequality (39), this ends the proof of the proposition. ■

E Application to minimax estimation

Let us denote by $\theta_k(f) = \langle f|\varphi_k\rangle_n$ the coefficients of the (orthogonal) discrete sine transform of f and the Sobolev ellipsoids $\mathscr{F}(\alpha,R) = \{f \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha}\theta_k(f)^2 \le R\}$. Assume in this section that the noise is homoscedastic ($\Sigma = \sigma^2 I_{n\times n}$) 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 sine basis.

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

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

where $n_\sigma = n/\sigma^2$. Then, in model (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 on the family of classes $\{\mathscr{F}(\alpha,R) : \alpha > 0, R > 0\}$.

Proof: 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 $\lambda_0 = (\alpha_0, w_0) \in \Lambda$ 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

$$\int_{\mathbb{R}^2} \alpha\, p_\pi(\alpha,w)\, d\alpha\, dw = 1, \qquad \int_{\mathbb{R}^2} w\, p_\pi(\alpha,w)\, d\alpha\, dw \le n_\sigma^{1/2}. \tag{40}$$

Let τ be a positive number such that $\tau \le \min(1, \alpha_0/(2\log w_0))$ and choose $p_{\lambda_0,\tau}$ as a translation/dilatation of π, concentrating on λ_0 when $\tau \to 0$:

$$p_{\lambda_0,\tau}(d\lambda) = p_\pi\Big(\frac{\lambda-\lambda_0}{\tau}\Big)\frac{d\lambda}{\tau^2}.$$

In view of Theorem 1,

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2) \le r_{\lambda_0} + \int_{\mathbb{R}^2} \big|r_{\alpha,w} - r_{\alpha_0,w_0}\big| p_{\lambda_0,\tau}(d\lambda) + \frac{\beta}{n}\mathscr{K}(p_{\lambda_0,\tau},\pi). \tag{41}$$

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 our choice of estimator, 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

$$n|r_{\alpha,w} - r_{\alpha,w_0}| \le 2\sum_{k=1}^n (f_k^2+\sigma^2)\big|(1-k^\alpha/w)_+ - (1-k^\alpha/w_0)_+\big|. \tag{42}$$

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

$$\Big|\Big(1-\frac{k^\alpha}{w}\Big)_+ - \Big(1-\frac{k^\alpha}{w_0}\Big)_+\Big| \le \Big|\frac{k^\alpha}{w} - \frac{k^\alpha}{w_0}\Big| \mathbb{1}_{\{k^\alpha \le w\}} \le \frac{w-w_0}{w_0}. \tag{43}$$

By virtue of Inequalities (42) and (43) we get

$$|r_{\alpha,w} - r_{\alpha,w_0}| \le 2n^{-1}\sum_{k=1}^n (f_k^2+\sigma^2)\frac{(w-w_0)}{w_0} \le 2(R_0+\sigma^2)\frac{w-w_0}{w_0}. \tag{44}$$

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

$$\begin{aligned} |r_{\alpha,w_0} - r_{\alpha_0,w_0}| &\le 2n^{-1}\sum_{k=1}^n (f_k^2+\sigma^2)\frac{k^\alpha - k^{\alpha_0}}{w_0}\mathbb{1}_{\{k^{\alpha_0}\le w_0\}} \\ &\le 2(R_0+\sigma^2)\big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}} - 1\big). \end{aligned} \tag{45}$$

Recall that we aim to bound the second term in the RHS of (41). To this end, we need an accurate upper bound on the integrals of the RHSs of (44) and (45) w.r.t. the probability measure $p_{\lambda_0,\tau}$. For the first one, we get

$$\begin{aligned}\int \big|r_{\alpha,w} - r_{\alpha,w_0}\big|\, p_{\lambda_0,\tau}(d\lambda) &\leq 2(R_0+\sigma^2)w_0^{-1}\int_{\mathbb{R}^2}(w-w_0)p_{\lambda_0,\tau}(d\lambda)\\ &\leq 4n_\sigma^{1/2}w_0^{-1}\tau(R_0+\sigma^2). \end{aligned} \tag{46}$$

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

$$\begin{aligned}\int_{\mathbb{R}^2}|r_{\alpha,w_0} - r_{\alpha_0,w_0}|p_{\lambda_0,\tau}(d\lambda) &\leq 2(R_0+\sigma^2)\int_{\mathbb{R}^2}\big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}}-1\big)p_{\lambda_0,\tau}(d\lambda)\\ &= \frac{2\tau(R_0+\sigma^2)\log w_0}{\alpha_0-\tau\log w_0}\\ &\leq 4\tau(R_0+\sigma^2)\alpha_0^{-1}\log w_0, \end{aligned} \tag{47}$$

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

The last term to bound in inequality (41) requires the evaluation of the Kullback-Leibler divergence between $p_{\lambda_0,\tau}$ and π. It can be done as follows:

$$\begin{aligned}\mathscr{K}(p_{\lambda_0,\tau},\pi) &= \int_{\mathbb{R}^2}\log\Big(\frac{e^{-\frac{\alpha-\alpha_0}{\tau}}p_w\big(\frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)}\tau}\big)}{e^{-\alpha}p_w\big(\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\big)}\frac{1}{\tau^2}\Big)p_{\lambda_0,\tau}(d\lambda)\\ &= \int_{\mathbb{R}^2}\Big\{\alpha-\frac{\alpha-\alpha_0}{\tau}+\log\frac{p_w\big(\frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)}\tau}\big)}{p_w\big(\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\big)}\Big\}p_{\lambda_0,\tau}(d\lambda)-2\log(\tau)\\ &\leq \alpha_0+(\tau-1)+\int_{\mathbb{R}_+^2}\log\Big(1+\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\Big)^3 p_{\lambda_0,\tau}(d\lambda)-2\log(\tau). \end{aligned}$$

where the third equality is derived thanks to Eq. (40) 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 $w_0+\tau n_\sigma^{\alpha/(2\alpha+1)}u \leq n_\sigma^{\alpha/(2\alpha+1)}(w_0+u)$, we get

$$\begin{aligned}\int_{\mathbb{R}_+^2}\log\Big(1+\frac{w}{n_\sigma^{\alpha/(2\alpha+1)}}\Big)^3 p_{\lambda_0,\tau}(d\lambda) &\leq 3\int_{\mathbb{R}_+}\log\big(1+w_0+u\big)p_w(u)du\\ &\leq 3\log\Big(1+w_0+\int_{\mathbb{R}_+}up_w(u)du\Big)\\ &= 3\log(2+w_0). \end{aligned}$$

Eventually, we can reformulate our bound on the risk of the EWA given in (41), leading to

$$\mathbb{E}(\|\hat{\boldsymbol{f}}_{\mathrm{EWA}}-\boldsymbol{f}\|_n^2) \leq r_{\lambda_0}+4\tau(R_0+\sigma^2)\Big(\frac{n_\sigma^{1/2}}{w_0}+\frac{\log w_0}{\alpha_0}\Big)+\frac{8\sigma^2(\alpha_0+3\log\big(\frac{2+w_0}{\tau}\big))}{n}. \tag{48}$$

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 (see, for instance, Tsybakov (2009), Theorem 3.2)

$$\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)} r_{\lambda_0} = (1+o_n(1))\min_{\hat{\boldsymbol{f}}}\max_{\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)}\mathbb{E}(\|\hat{\boldsymbol{f}}-\boldsymbol{f}\|_n^2).$$

In view of this result, taking the max over $\boldsymbol{f}\in\mathscr{F}(\alpha_0,R_0)$ in (48), we get

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

This leads to the desired result in view of the relation

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

which follows from (Tsybakov, 2009, Theorem 3.1). ■