Submitted to the Annals of Statistics
arXiv: math.ST/1104.3969

SHARP ORACLE INEQUALITIES FOR AGGREGATION OF AFFINE ESTIMATORS

BY ARNAK S. DALALYAN AND JOSEPH SALMON

ENSAE-CREST and Université Paris Diderot

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* [7] 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, theoretical foundations of aggregation procedures were first studied by Nemirovski (Nemirovski [48], Juditsky and Nemirovski

*The authors thank Pierre Alquier for fruitful discussions.

†Research partially supported by ANR Parcimonie.

AMS 2000 subject classifications: Primary 62G08, ; secondary 62C20,62G05,62G20

Keywords and phrases: aggregation, regression, oracle inequalities, model selection, minimax risk, exponentially weighted aggregation

[37]) and independently by a series of papers by Catoni (see [11] for an account) and Yang [63, 64, 65]. For the regression model, a significant progress was achieved by Tsybakov [60] 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 in [46, 52, 9], especially in the context of high dimension with sparsity constraints and in [51] for Kullback-Leibler aggregation. However, it should be noted that the procedures proposed in [60] that provably achieve the lower bounds in convex and linear aggregation require full knowledge of design distribution. This limitation was overcome in the recent work [62].

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 aggregation). The generality of those results—almost no restriction on the constituent estimators—compensates to this practical limitation.

In the Gaussian sequence model, a breakthrough was reached by Leung and Barron [45]. Building on very elegant but not very well known results by George [32] [1], they established sharp oracle inequalities for the exponentially weighted aggregate (EWA) for constituent estimators obtained from the data vector by orthogonally projecting it on some linear subspaces. Dalalyan and Tsybakov [21, 22] showed 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.

Our interest in affine estimators is motivated by several reasons. First, affine estimators encompass many popular estimators such as smoothing splines, the Pinsker estimator [49, 29], local polynomial estimators, nonlocal means [8, 56], 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 [38]—the set of signals that are well

[1] Corollary 2 in [32] coincides with Theorem 1 in [45] in the case of exponential weights with temperature $\beta = 2\sigma^2$, cf. Eq. (2.2) below for a precise definition of exponential weights. Furthermore, to the best of our knowledge, [32] is the first reference using the Stein lemma for evaluating the expected risk of the exponentially weighted aggregate.

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 were also proved by Golubev [36], and for model selection by Arlot and Bach [3], Baraud, Giraud and Huet [5].

In the present work, we establish sharp oracle inequalities in the 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 expected loss, for the exponentially weighted aggregate. They have the advantage of covering in a unified fashion the particular cases of frozen estimators considered in [22] and of projection estimators treated in [45].

We focus on the theoretical guarantees expressed in terms of oracle inequalities for the expected squared loss. Interestingly, although several recent papers [3, 5, 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 simple formulation. It is established for (suitably symmetrized, if necessary) exponentially weighted aggregates [32, 11, 21] with an arbitrary prior and a temperature parameter which is not too small. The result is non-asymptotic but leads to asymptotically optimal residual term when the sample size, as well as the cardinality of the family of constituent estimators, tends to infinity. In its general form, the residual term is similar to those obtained in PAC-Bayes setting [47, 42, 57] in that it is proportional to the Kullback-Leibler divergence between two probability distributions.

The problem of competing against the best procedure in a given family was extensively studied in the context of online learning and prediction with expert advice [39, 16]. A connection between the results on online learning and statistical oracle inequalities was established by Gerchinovitz [33].

1.2. *Notation and examples of linear estimators.* Throughout this work, we focus on the heteroscedastic regression model with Gaussian additive noise. We assume 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 $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is a centered Gaussian random vector, $f_i = \mathbf{f}(x_i)$ where $\mathbf{f} : \mathcal{X} \to \mathbb{R}$ is an unknown function and $x_1, \ldots, x_n \in \mathcal{X}$ are deterministic points. Here, no assumption is made on the set $\mathcal{X}$. Our objective is to recover the vector $\boldsymbol{f} = (f_1, \ldots, f_n)^\top$, often referred to as *signal*, based on the data $y_1, \ldots, y_n$. 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 $|||\Sigma|||$. We denote by $\langle\cdot|\cdot\rangle_n$ the empirical inner product in $\mathbb{R}^n$: $\langle \boldsymbol{u}|\boldsymbol{v}\rangle_n = (1/n)\sum_{i=1}^n u_i v_i$. We measure the performance of an estimator $\hat{\boldsymbol{f}}$ by its expected empirical quadratic loss: $r = \mathbb{E}\big[\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2\big]$ where $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f_i - \hat{f}_i)^2$.

We only focus on the task of aggregating *affine estimators* $\hat{\boldsymbol{f}}_\lambda$ indexed by some parameter $\lambda \in \Lambda$. These estimators can be written as affine transforms of the data $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Using the convention that all vectors are one-column matrices, we have $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$, where the $n \times n$ real matrix A_λ and the vector $\boldsymbol{b}_\lambda \in \mathbb{R}^n$ are deterministic. It means the entries of A_λ and $\boldsymbol{b}_\lambda$ may depend on the points $x_1, \ldots, x_n$ but not on the data $\boldsymbol{Y}$. 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 (unknown) one of the family.

Ordinary least squares. Let $\{\mathcal{S}_\lambda : \lambda \in \Lambda\}$ 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 [6], is the set of orthogonal projections onto $\mathcal{S}_\lambda$. 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 pseudo-inverse of $X_\lambda^\top X_\lambda$.

Diagonal filters. Other common estimators are the so called diagonal filters corresponding to diagonal matrices $A = \operatorname{diag}(a_1, \ldots, a_n)$. Examples include:

- Ordered projections: $a_k = \mathbb{1}_{(k \le \lambda)}$ for some integer λ ($\mathbb{1}_{(\cdot)}$ is the indicator function). Those weights are also called truncated SVD (Singular Value Decomposition) or spectral cut-off. In this case a natural parametrization is $\Lambda = \{1, \ldots, n\}$, indexing the number of elements conserved.
- Block projections: $a_k = \mathbb{1}_{(k \le w_1)} + \sum_{j=1}^{m-1} \lambda_j \mathbb{1}_{(w_j \le k \le w_{j+1})}$, $k = 1, \ldots, n$, where $\lambda_j \in \{0, 1\}$. Here the natural parametrization is $\Lambda = \{0, 1\}^{m-1}$, indexing subsets of $\{1, \ldots, m-1\}$.
- 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: $a_k = \big(1 - \frac{k^\alpha}{w}\big)_+$, where $x_+ = \max(x, 0)$ and $(w, \alpha) =$

$\lambda \in \Lambda = (\mathbb{R}_+^*)^2$.

Kernel ridge regression. Assume that we have a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and we aim at estimating the true function f in the associated reproducing kernel Hilbert space $(\mathcal{H}_k, \|\cdot\|_k)$. The kernel ridge estimator is obtained by minimizing the criterion $\|\boldsymbol{Y} - \boldsymbol{f}\|_n^2 + \lambda\|\boldsymbol{f}\|_k^2$ w.r.t. $f \in \mathcal{H}_k$ (see [58, p. 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 $n \times n$ identity matrix.

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

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

It is worth mentioning that the formulation in Eq.(1.2) can be linked to the group Lasso [66] and to the multiple kernel learning introduced in [41]—see [3] 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 matrix $A = (a_{ij})_{i,j=1}^n$ 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.

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 strengths and limitations of the results are discussed in Section 3. The extension of these results to the case of grouped aggregation—in relation with ill-posed inverse problems—is developed in Section 4. As a consequence, we provide in Section 5 sharp oracle inequalities in various set-ups: ranging from finite to continuous families of constituent estimators and including sparse scenarii. In Section 6, we apply our main results to prove that combining Pinsker's type filters with EWA leads to asymptotically sharp adaptive procedures over Sobolev ellipsoids. Section 7 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 given in Section 8, while the proofs of some technical results (Propositions 2-6) are provided in the supplementary material [19].

2. Aggregation of estimators: main results. In this section, we describe the statistical framework for aggregating estimators and we 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* $\mathcal{F}_\Lambda = (\hat{\boldsymbol{f}}_\lambda)_{\lambda\in\Lambda} \in \mathbb{R}^n$. 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:

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{aggr}} - \boldsymbol{f}\|_n^2\big] \leq C_n \inf_{\lambda\in\Lambda} \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big] + R_n, \tag{2.1}$$

with *residual* term R_n tending to zero as $n \to \infty$, 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. They are often referred to as sharp OI.

2.1. *Exponentially Weighted Aggregate (EWA).* Let $r_\lambda = \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]$ 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 Λ and for any $\beta > 0$, we define the probability measure of exponential weights, $\hat{\pi}$, by

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

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

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

We will frequently 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 [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 influence our confidence in the estimators.

2.2. *Main results.* In this paper, we only focus on *affine estimators*

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

where the $n \times n$ real matrix A_λ and the vector $\boldsymbol{b}_\lambda \in \mathbb{R}^n$ are deterministic. Furthermore, we will assume that an unbiased estimator $\widehat{\Sigma}$ of the noise covariance matrix Σ is available. It is well-known (*cf.*, Appendix A for details) that the risk of the estimator (2.4) is given by

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

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

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

is an unbiased estimator of r_λ. Along with $\hat{r}_\lambda^{\text{unb}}$ we will use another estimator of the risk that we call adjusted risk estimate and define by:

$$\hat{r}_\lambda^{\text{adj}} = \underbrace{\big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}(\widehat{\Sigma} A_\lambda) - \frac{1}{n}\operatorname{Tr}[\widehat{\Sigma}]}_{\hat{r}_\lambda^{\text{unb}}} + \frac{1}{n}\boldsymbol{Y}^\top (A_\lambda - A_\lambda^2)\boldsymbol{Y}. \tag{2.7}$$

One can notice that the adjusted risk estimate $\hat{r}_\lambda^{\text{adj}}$ coincides with the unbiased risk estimate $\hat{r}_\lambda^{\text{unb}}$ if and only if the matrix A_λ is an orthogonal projector.

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

$$\mathcal{K}(p,p') = \begin{cases} \int_\Lambda \log\big(\frac{dp}{dp'}(\lambda)\big) p(d\lambda) & \text{if } p \text{ is absolutely continuous w.r.t. } p', \\ +\infty & \text{otherwise.} \end{cases}$$

We write $S_1 \preceq S_2$ (resp. $S_1 \succeq S_2$) for two symmetric matrices S_1 and S_2, when $S_2 - S_1$ (resp. $S_1 - S_2$) is semi-definite positive.

THEOREM 1. *Let all the matrices A_λ be symmetric and $\widehat{\Sigma}$ be unbiased and independent of $\boldsymbol{Y}$.*

i) *Assume that for all $\lambda, \lambda' \in \Lambda$, it holds that $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$, $A_\lambda \Sigma + \Sigma A_\lambda \succeq 0$ and $\boldsymbol{b}_\lambda = 0$. If $\beta \geq 8|||\Sigma|||$ then, the aggregate $\hat{\boldsymbol{f}}_{\rm EWA}$ defined by Eq. (2.2), (2.3) and the unbiased risk estimate $\hat{r}_\lambda = \hat{r}_\lambda^{\rm unb}$ (2.6) satisfies*

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\rm EWA} - \boldsymbol{f}\|_n^2\big] \leq \inf_{p \in \mathcal{P}_\Lambda} \left\{ \int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]\, p(d\lambda) + \frac{\beta}{n}\, \mathcal{K}(p, \pi) \right\}. \tag{2.8}$$

ii) *Assume that, for all $\lambda \in \Lambda$, $A_\lambda \preceq I_{n\times n}$ and $A_\lambda \boldsymbol{b}_\lambda = 0$. If $\beta \geq 4|||\Sigma|||$ then, the aggregate $\hat{\boldsymbol{f}}_{\rm EWA}$ defined by Eq. (2.2), (2.3) and the adjusted risk estimate $\hat{r}_\lambda = \hat{r}_\lambda^{\rm adj}$ (2.7) satisfies*

$$\begin{aligned}\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\rm EWA} - \boldsymbol{f}\|_n^2\big] \leq \inf_{p \in \mathcal{P}_\Lambda} \bigg\{ &\int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]\, p(d\lambda) + \frac{\beta}{n}\, \mathcal{K}(p, \pi) \\ &+ \frac{1}{n}\int_\Lambda \Big(\boldsymbol{f}^\top (A_\lambda - A_\lambda^2)\boldsymbol{f} + \operatorname{Tr}\big[\Sigma(A_\lambda - A_\lambda^2)\big]\Big) p(d\lambda)\bigg\}.\end{aligned}$$

The simplest setting in which all the conditions of part i) of Theorem 1 are fulfilled is when the matrices A_λ and Σ are all diagonal, or diagonalizable in a common base. This result, as we will see in Section 6, leads to a new estimator which is adaptive, in the exact minimax sense, over the collection of all Sobolev ellipsoids. It also suggests a new method for efficiently combining varying-block-shrinkage estimators, as described in Section 5.4.

However, part i) of Theorem 1 leaves open the issue of aggregating affine estimators defined via non-commuting matrices. In particular, it does not allow us to evaluate the MSE of EWA when each A_λ is a convex or linear combination of a fixed family of projection matrices on non-orthogonal linear subspaces. Such kind of situations may be handled via the result of part ii) of Theorem 1. One can observe 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), the result of part ii) offers an extension of [45, Corollary 6] to the case of general noise covariances ([45] deals only with i.i.d. noise).

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

REMARK 1. *While the results in [45] yield a sharp oracle inequality in the case of projection matrices* A_λ, *they are of no help in the case when the matrices* A_λ *are nearly idempotent and not exactly. Assertion ii) of Theorem 1 fills this gap by showing that if* $\max_\lambda \operatorname{Tr}[A_\lambda - A_\lambda^2] \le \delta$ *then* $\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\mathrm{EWA}} - \boldsymbol{f}\|_n^2\big]$ *is bounded by*

$$\inf_{p\in\mathcal{P}_\Lambda}\bigg\{\int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]\,p(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p,\pi)\bigg\} + \delta\big(\|\boldsymbol{f}\|_n^2 + n^{-1}\|\!|\Sigma|\!\|\big).$$

REMARK 2. *We have focused only on Gaussian errors to emphasize that it is possible to efficiently aggregate almost any family of affine estimators. We believe that by a suitable adaptation of the approach developed in [22], claims of Theorem 1 can be generalized—at least when* ξ_i *are independent with known variances—to some other common noise distributions.*

The results presented so far concern the situation when the matrices A_λ are symmetric. However, using the last part of Theorem 1, it is possible to propose an estimator of $\boldsymbol{f}$ 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 A_λ's, is that they automatically satisfy the contraction condition $\tilde{A}_\lambda \preceq I_{n\times n}$ required by part ii) of Theorem 1. We will refer to this method as Symmetrized Exponentially Weighted Aggregates (or SEWA) [20].

THEOREM 2. *Assume that the matrices* A_λ *and the vectors* $\boldsymbol{b}_\lambda$ *satisfy* $A_\lambda\boldsymbol{b}_\lambda = A_\lambda^\top\boldsymbol{b}_\lambda = 0$ *for every* $\lambda\in\Lambda$. *Assume in addition that* $\widehat{\Sigma}$ *is an unbiased estimator of* Σ *and is independent of* $\boldsymbol{Y}$. *Let* $\tilde{\boldsymbol{f}}_{\mathrm{SEWA}}$ *denote the exponentially weighted aggregate of the (symmetrized) 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, under the conditions* $\beta \ge 4\|\!|\Sigma|\!\|$ *and*

$$\pi\Big\{\lambda\in\Lambda : \operatorname{Tr}(\widehat{\Sigma}A_\lambda) \le \operatorname{Tr}(\widehat{\Sigma}A_\lambda^\top A_\lambda)\Big\} = 1 \quad a.s. \qquad \textbf{(C)}$$

it holds that

$$\mathbb{E}\big[\|\tilde{\boldsymbol{f}}_{\mathrm{SEWA}} - \boldsymbol{f}\|_n^2\big] \le \inf_{p\in\mathcal{P}_\Lambda}\bigg\{\int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]p(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p,\pi)\bigg\}. \qquad (2.9)$$

To understand the scope of condition **(C)**, let us present several cases of widely used linear estimators for which this condition is satisfied.

- The simplest class of matrices A_λ for which condition **(C)** holds true are orthogonal projections. Indeed, if A_λ is a projection matrix, it satisfies $A_\lambda^\top A_\lambda = A_\lambda$ and, therefore, $\mathrm{Tr}(\widehat{\Sigma} A_\lambda) = \mathrm{Tr}(\widehat{\Sigma} A_\lambda^\top A_\lambda)$.
- When the matrix $\widehat{\Sigma}$ is diagonal, then a sufficient condition for **(C)** is $a_{ii} \leq \sum_{j=1}^n a_{ji}^2$. Consequently, **(C)** holds true for matrices having only zeros on the main diagonal. For instance, the kNN filter in which the weight of the observation Y_i is replaced by zero, *i.e.*, $a_{ij} = \mathbf{1}_{j\in\{j_{i,1},\ldots,j_{i,k}\}}/k$ satisfies this condition.
- Under a little bit more stringent assumption of homoscedasticity, *i.e.*, when $\widehat{\Sigma} = \widehat{\sigma}^2 I_{n\times n}$, 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 $\mathrm{Tr}(A_\lambda) = \mathrm{Tr}(A_\lambda^\top A_\lambda)$ and **(C)** is fulfilled. A notable example of linear estimators that satisfy this condition are Nadaraya-Watson estimators with rectangular kernel and nearest neighbor filters.

3. Discussion. Before elaborating on the main results stated in the previous section, by extending them to inverse problems and by deriving adaptive procedures, let us discuss some aspects of the presented OIs.

3.1. *Assumptions on* Σ. In some rare situations, the matrix Σ is known and it is natural to use $\widehat{\Sigma} = \Sigma$ as unbiased estimator. Besides this not very realistic situation, there are at least two contexts in which it is reasonable to assume that an unbiased estimator of Σ, independent of $\boldsymbol{Y}$, is available.

The first case corresponds to problems in which a signal can be recorded several times by the same device, or once but by several identical devices. For instance, this is the case when an object is photographed many times by the same digital camera during a short time period. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ be the available signals, which can be considered as i.i.d. copies of an n-dimensional Gaussian vector with mean $\boldsymbol{f}$ and covariance matrix Σ_Z. Then, defining $\boldsymbol{Y} = (\boldsymbol{Z}_1 + \ldots + \boldsymbol{Z}_N)/N$ and $\widehat{\Sigma}_Z = (N-1)^{-1}(\boldsymbol{Z}_1\boldsymbol{Z}_1^\top + \ldots + \boldsymbol{Z}_N\boldsymbol{Z}_N^\top - N\boldsymbol{Y}\boldsymbol{Y}^\top)$, we find ourselves within the framework covered by previous theorems. Indeed, $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{f}, \Sigma_Y)$ with $\Sigma_Y = \Sigma_Z/N$ and $\widehat{\Sigma}_Y = \widehat{\Sigma}_Z/N$ is an unbiased estimate of Σ_Y, independent of $\boldsymbol{Y}$. Note that our theory applies in this setting for every integer $N \geq 2$.

The second case is when the dominating part of the noise comes from the device which is used for recording the signal. In this case, the practitioner can use the device in order to record a known signal, $\boldsymbol{g}$. In digital image processing, $\boldsymbol{g}$ can be a black picture. This will provide a noisy signal $\boldsymbol{Z}$ drawn from Gaussian distribution $\mathcal{N}_n(\boldsymbol{g}, \Sigma)$, independent of $\boldsymbol{Y}$ which is the signal of interest. Setting $\widehat{\Sigma} = (\boldsymbol{Z} - \boldsymbol{g})(\boldsymbol{Z} - \boldsymbol{g})^\top$, one ends up with an unbiased estimator of Σ, which is independent of $\boldsymbol{Y}$.

3.2. *OI in expectation versus OI with high probability.* All the results stated in this work provide sharp non-asymptotic bounds on the expected risk of EWA. It would be insightful to complement this study by risk bounds that hold true with high probability. However, it was recently proved in [17] that EWA is deviation suboptimal: there exist a family of constituent estimators and a constant $C > 0$ such that the difference between the risk of EWA and that of the best constituent estimator is larger than $C/\sqrt{n}$ with probability at least 0.06. Nevertheless, several empirical studies (see, for instance, [18]) demonstrated that EWA has often a smaller risk than some of its competitors, such as the empirical star procedure [4], which are provably optimal in the sense of OIs with high probability. Furthermore, numerical experiments carried out in Section 7 show that the standard-deviation of the risk of EWA is of the order of $1/n$. This suggests that under some conditions on the constituent estimators it might be possible to establish OIs for EWA that are similar to (2.8) but hold true with high probability. A step in proving this kind of results was done in [43, Thm. C] for the model of regression with random design.

3.3. *Relation to previous work and limits of our results.* The OI of the previous section require various conditions on the constituent estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$. One may wonder how general these conditions are and is it possible to extend these OIs to more general $\hat{\boldsymbol{f}}_\lambda$'s. Although this work does not answer this question, we can sketch some elements of response.

First of all, we stress that the conditions of the present paper relax significantly those of previous results existing in statistical literature. For instance, Kneip [40] considered only linear estimators, *i.e.*, $\boldsymbol{b}_\lambda \equiv 0$ and, more importantly, only ordered set of *commuting* matrices A_λ. The ordering assumption is dropped in Leung and Barron [45], in the case of projection matrices. Note that neither of these assumptions is satisfied for the families of Pinsker and Tikhonov-Philipps estimators. The present work strengthens existing results in considering more general, affine estimators extending both projection matrices and ordered commuting matrices.

Despite the advances achieved in this work, there are still interesting cases that are not covered by our theory. We now introduce a family of estimators commonly used in image processing, that do not satisfy our assumptions. In recent years, non-local means (NLM) became quite popular in image processing [8]. This method of signal denoising, shown to be tied in with EWA [56], removes noise by exploiting signals self-similarities. We briefly define the NLM procedure in the case of one-dimensional signals.

Assume that a vector $\boldsymbol{Y} = (y_1, \dots, y_n)^\top$ given by (1.1) is observed with

$f_i = \mathbf{f}(i/n)$, $i = 1, \ldots, n$, for some function $\mathbf{f} : [0,1] \to \mathbb{R}$. For a fixed "patch-size" $k \in \{1, \ldots, n\}$, let us define $\boldsymbol{f}_{[i]} = (f_i, f_{i+1}, \ldots, f_{i+k-1})^\top$ and $\boldsymbol{Y}_{[i]} = (y_i, y_{i+1}, \ldots, y_{i+k-1})^\top$ for every $i = 1, \ldots, n-k+1$. 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 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. Unfortunately, these matrices $\{A_i\}$ and vectors $\{b_i\}$ do not fit our framework, *i.e.*, the assumption $A_i b_i = A_i^\top b_i = 0$ is not satisfied.

Finally, our proof technique is specific to affine estimators. Its extension to estimators defined as a more complex function of the data will certainly require additional tools and is a challenging problem for future research. Yet, it seems unlikely to get sharp OIs with optimal remainder term for a fairly general family of constituent estimators (without data-splitting), since this generality inherently increases the risk of overfitting.

4. Ill-posed inverse problems and group-weighting. As explained in [15, 12], 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 $\mathcal{H}$, with inner product $\langle \cdot | \cdot \rangle_{\mathcal{H}}$. For some $h \in \mathcal{H}$, let Y be the random process indexed by $g \in \mathcal{H}$ such that

$$Y = Th + \varepsilon\xi \quad \Longleftrightarrow \quad \Big(Y(g) = \langle Th | g \rangle_{\mathcal{H}} + \varepsilon\xi(g), \quad \forall g \in \mathcal{H} \Big), \tag{4.1}$$

where $\varepsilon > 0$ is the noise magnitude and ξ is a white Gaussian noise on $\mathcal{H}$, *i.e.,* for any $g_1, \ldots, g_k \in \mathcal{H}$ the vector $\big(Y(g_1), \ldots, Y(g_k)\big)$ is Gaussian with zero mean and covariance matrix $\{\langle g_i | g_j \rangle_{\mathcal{H}}\}$. The problem is then the following: estimate the element h assuming the value of Y can be measured for any given g. It is customary to use as g, the eigenvectors of the adjoint T^* of T. Under the condition that the operator T^*T is compact, the SVD yields $T\phi_k = b_k \psi_k$ and $T^*\psi_k = b_k \phi_k$, for $k \in \mathbb{N}$, where b_k are the singular values, $\{\psi_k\}$ is an orthonormal basis in $\mathrm{Range}(T) \subset \mathcal{H}$ and $\{\phi_k\}$ is the corresponding orthonormal basis in $\mathcal{H}$. In view of (4.1), it holds that:

$$Y(\psi_k) = \langle h | \phi_k \rangle_{\mathcal{H}}\, b_k + \varepsilon\xi(\psi_k), \qquad k \in \mathbb{N}. \tag{4.2}$$

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 $f_i = \langle h|\phi_i\rangle_{\mathcal{H}}$ and $\Sigma = \operatorname{diag}(\sigma_i^2, i = 1, 2, \dots)$ for $\sigma_i = \varepsilon b_i^{-1}$. Examples of inverse problems to which this statistical model has been successfully applied are derivative estimation, deconvolution with known kernel, computerized tomography—see [12] and the references therein for more applications.

For very mildly ill-posed inverse problems, *i.e.*, when the singular values b_k of T tend to zero not faster than any negative power of k, the approach presented in Section 2 will lead to satisfactory results. Indeed, by choosing $\beta = 8\|\!|\Sigma|\!\|$ or $\beta = 4\|\!|\Sigma|\!\|$, the remainder term in (2.8) 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 developed in previous section becomes obsolete since the remainder blows up when n increases to infinity. Furthermore, this is not an artifact of our theoretical results, but rather 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:

$$\hat{\pi}_n = \arg\inf_p \left\{ \int_\Lambda \hat{r}_\lambda \, p(d\lambda) + \frac{\beta}{n}\mathcal{K}(p, \pi) \right\}, \tag{4.3}$$

where the inf is taken over the set of all probability distributions. It means the same regularization parameter β is employed for estimating both the coefficients $f_i = \langle h|\phi_i\rangle_{\mathcal{H}}$ 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 is unreasonable. 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. This motivates our interest in the grouped EWA (or GEWA).

Let us consider a partition $B_1, \dots, B_J$ of the set $\{1, \dots, n\}$: $B_j = \{T_j + 1, \dots, T_{j+1}\}$, for some integers $0 = T_1 < T_2 < \dots < T_{J+1} = n$. 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 sections, we are concerned by the aggregation of affine estimators

$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$ but here we will assume the matrices A_λ are block-diagonal:

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

Similarly, we define $\hat{\boldsymbol{f}}_\lambda^j$ and $\boldsymbol{b}_\lambda^j$ as the sub-vectors 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 and its unbiased estimate $\widehat{\Sigma}$ are block-diagonal with $(T_{j+1}-T_j)\times(T_{j+1}-T_j)$ blocks Σ^j and $\widehat{\Sigma}^j$, respectively. These notations imply in particular that $\hat{\boldsymbol{f}}_\lambda^j = A_\lambda^j \boldsymbol{Y}^j + \boldsymbol{b}_\lambda^j$ for every $j = 1, \dots, J$. Moreover, the unbiased risk estimate $\hat{r}_\lambda^{\text{unb}}$ of $\hat{\boldsymbol{f}}_\lambda$ can be decomposed into the sum of unbiased risk estimates $\hat{r}_\lambda^{j,\text{unb}}$ of $\hat{\boldsymbol{f}}_\lambda^j$; namely $\hat{r}_\lambda^{\text{unb}} = \sum_{j=1}^J \hat{r}_\lambda^{j,\text{unb}}$, where

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

To state the analogues of Theorems 1 and 2 we introduce the following settings.

Setting 1: For all $\lambda, \lambda' \in \Lambda$ and $j \in \{1, \dots, J\}$, 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 $\boldsymbol{b}_\lambda^j = 0$. For a temperature vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)^\top$ and a prior π, we define GEWA as $\hat{\boldsymbol{f}}_{\text{GEWA}}^j = \int_\Lambda \hat{\boldsymbol{f}}_\lambda^j \hat{\pi}^j(d\lambda)$, where $\hat{\pi}^j(d\lambda) = \theta^j(\lambda)\pi(d\lambda)$ with

$$\theta^j(\lambda) = \frac{\exp(-n\hat{r}_\lambda^{j,\text{unb}}/\beta_j)}{\int_\Lambda \exp(-n\hat{r}_\omega^{j,\text{unb}}/\beta_j)\pi(d\omega)}. \tag{4.4}$$

Setting 2: For every $j = 1, \dots, J$ and for every λ belonging to a set of π-measure one, the matrices A_λ satisfy a.s. the inequality $\text{Tr}(\widehat{\Sigma}^j A_\lambda^j) \le \text{Tr}(\widehat{\Sigma}^j (A_\lambda^j)^\top A_\lambda^j)$ 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 temperature vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)^\top$ and a prior π, we define GEWA as $\hat{\boldsymbol{f}}_{\text{GEWA}}^j = \int_\Lambda \tilde{\boldsymbol{f}}_\lambda^j \hat{\pi}^j(d\lambda)$, 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 (4.4). Note that this setting is the grouped version of the SEWA.

THEOREM 3. *Assume that $\widehat{\Sigma}$ is unbiased and independent of $\boldsymbol{Y}$. Under Setting 1, if $\beta_j \ge 8 |\!|\!| \Sigma^j |\!|\!|$ for all $j = 1, \dots, J$, then*

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}} - \boldsymbol{f}\|_n^2\big] \le \sum_{j=1}^J \inf_{p_j} \bigg\{ \int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda^j - \boldsymbol{f}^j\|_n^2\, p_j(d\lambda) + \frac{\beta_j}{n}\,\mathcal{K}(p_j, \pi) \bigg\}. \tag{4.5}$$

Under Setting 2, this inequality holds true if $\beta_j \geq 4|||\Sigma^j|||$ *for every* $j = 1, \ldots, J$.

As we shall see in Section 6, 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 that the operator T is mildly ill-posed, *i.e.*, its singular values decrease at most polynomially.

5. Examples of sharp oracle inequalities. In this section, we discuss consequences of the main result for specific choices of prior measures. For conveying the main messages of this section it is enough to focus on the Settings 1 and 2 in the case of only one group ($J = 1$).

5.1. *Discrete oracle inequality.* In order to demonstrate that Inequality (4.5) 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, $\pi(\Lambda_0) = 1$ for a countable set $\Lambda_0 \subset \Lambda$, and w.l.o.g $\Lambda_0 = \mathbb{N}$. Then, the following result holds true.

PROPOSITION 1. *Let* $\widehat{\Sigma}$ *be unbiased, independent of* $\boldsymbol{Y}$ *and* π *be supported by* $\mathbb{N}$. *Under Setting 1 with* $J = 1$ *and* $\beta = \beta_1 \geq 8|||\Sigma|||$ *the aggregate* $\hat{\boldsymbol{f}}_{\text{GEWA}}$ *satisfies the inequality*

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}} - \boldsymbol{f}\|_n^2\big] \leq \inf_{\ell \in \mathbb{N}: \pi_\ell > 0} \left(\mathbb{E}\big[\|\hat{\boldsymbol{f}}_\ell - \boldsymbol{f}\|_n^2\big] + \frac{\beta \log(1/\pi_\ell)}{n} \right). \tag{5.1}$$

Furthermore, (5.1) *holds true under Setting 2 for* $\beta \geq 4|||\Sigma|||$.

PROOF. It suffices to apply Thm. 3 and to upper-bound the right-hand side by the minimum over all Dirac measures $p = \delta_\ell$ such that $\pi_\ell > 0$. $\square$

This inequality can be compared to Corollary 2 in [5, Section 4.3]. Our result has the advantage of having factor one in front of the expectation of the left-hand side, while in [5] a constant much larger than 1 appears. However, it should be noted that the assumptions on the (estimated) noise covariance matrix are much weaker in [5].

5.2. *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 $M \in \mathbb{N}$ (*e.g.,* to build an estimator nearly as accurate as the best kernel estimator with fixed kernel and varying bandwidth). To state an oracle inequality in such a "continuous" setup, let us denote by $d_2(\boldsymbol{\lambda}, \partial\Lambda)$ the largest real $\tau > 0$ such that the ball centered at $\boldsymbol{\lambda}$ of radius τ—hereafter denoted by $B_{\boldsymbol{\lambda}}(\tau)$—is included in Λ. Let $\mathrm{Leb}(\cdot)$ be the Lebesgue measure in $\mathbb{R}^M$.

PROPOSITION 2. *Let $\widehat{\Sigma}$ be unbiased, independent of $\boldsymbol{Y}$. Let $\Lambda \subset \mathbb{R}^M$ be an open and bounded set and let π be the uniform distribution on Λ. Assume that the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is Lipschitz continuous, i.e., $|r_{\boldsymbol{\lambda}'} - r_{\boldsymbol{\lambda}}| \leq L_r \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\|_2$, $\forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$. Under Setting 1 with $J = 1$ and $\beta = \beta_1 \geq 8 |||\Sigma|||$ the aggregate $\hat{\boldsymbol{f}}_{\text{GEWA}}$ satisfies the inequality*

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\text{GEWA}} - \boldsymbol{f}\|_n^2 \leq \inf_{\boldsymbol{\lambda}\in\Lambda} \left\{ \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big] + \frac{\beta M}{n} \log\Big(\frac{\sqrt{M}}{2\min(n^{-1}, d_2(\boldsymbol{\lambda}, \partial\Lambda))}\Big)\right\} + \frac{L_r + \beta \log\big(\mathrm{Leb}(\Lambda)\big)}{n}. \tag{5.2}$$

Furthermore, (5.2) holds true under Setting 2 for every $\beta \geq 4|||\Sigma|||$.

PROOF. It suffices to apply assertion i) of Theorem 1 and to upper-bound the right-hand side in Ineq. (2.8) by the minimum over all measures having as density $p_{\boldsymbol{\lambda}^*,\tau^*}(\boldsymbol{\lambda}) = \mathbb{1}_{B_{\boldsymbol{\lambda}^*}(\tau^*)}(\boldsymbol{\lambda})/\mathrm{Leb}(B_{\boldsymbol{\lambda}^*}(\tau^*))$. Choosing $\tau^* = \min(n^{-1}, d_2(\boldsymbol{\lambda}^*, \partial\Lambda))$ such that $B_{\boldsymbol{\lambda}^*}(\tau^*) \subset \Lambda$, the measure $p_{\boldsymbol{\lambda}^*,\tau^*}(\boldsymbol{\lambda})d\boldsymbol{\lambda}$ is absolutely continuous w.r.t. the uniform prior π and the Kullback-Leibler divergence between these two measures equals $\log\big\{\mathrm{Leb}(\Lambda)/\mathrm{Leb}\big(B_{\boldsymbol{\lambda}^*}(\tau^*)\big)\big\}$. Using $\mathrm{Leb}\big(B_{\boldsymbol{\lambda}^*}(\tau^*)\big) \geq (2\tau^*/\sqrt{M})^M$ and the Lipschitz condition, we get the desired inequality. □

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

5.3. *Sparsity oracle inequality.* The continuous oracle inequality stated in the previous subsection is well adapted to the problems in which the dimension M of Λ is small w.r.t. the sample size n (or, more precisely, the signal to noise ratio $n/|||\Sigma|||$). When this is not the case, the choice of the prior should be done more carefully. For instance, consider $\Lambda \subset \mathbb{R}^M$ 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 that favors sparse $\boldsymbol{\lambda}$'s. This can be done in the same vein as in [21, 22, 23, 24], by means of the heavy tailed prior:

$$\pi(d\boldsymbol{\lambda}) \propto \prod_{m=1}^{M} \frac{1}{(1+|\lambda_m/\tau|^2)^2}\, \mathbb{1}_{\Lambda}(\boldsymbol{\lambda}), \tag{5.3}$$

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

PROPOSITION 3. *Let $\widehat{\Sigma}$ be unbiased, independent of $\boldsymbol{Y}$. Let $\Lambda = \mathbb{R}^M$ and let π be defined by (5.3). Assume that the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is continuously differentiable and, for some $M \times M$ matrix $\mathcal{M}$, satisfies:*

$$r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}'} - \nabla r_{\boldsymbol{\lambda}'}^{\top}(\boldsymbol{\lambda}-\boldsymbol{\lambda}') \le (\boldsymbol{\lambda}-\boldsymbol{\lambda}')^{\top}\mathcal{M}(\boldsymbol{\lambda}-\boldsymbol{\lambda}'), \qquad \forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda. \tag{5.4}$$

Under Setting 1 if $\beta \ge 8|\!|\!|\Sigma|\!|\!|$, then the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}} = \hat{\boldsymbol{f}}_{\text{GEWA}}$ satisfies:

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}} - \boldsymbol{f}\|_n^2\big] \le \inf_{\boldsymbol{\lambda}\in\mathbb{R}^M} \left\{ \mathbb{E}\|\hat{\boldsymbol{f}}_{\lambda} - \boldsymbol{f}\|_n^2 + \frac{4\beta}{n} \sum_{m=1}^{M} \log\left(1 + \frac{|\lambda_m|}{\tau}\right)\right\} + \mathrm{Tr}(\mathcal{M})\tau^2. \tag{5.5}$$

Moreover, (5.5) holds true under Setting 2 if $\beta \ge 4|\!|\!|\Sigma|\!|\!|$.

Let us discuss here some consequences of this sparsity oracle inequality. First of all, consider the case of (linearly) combining frozen estimators, *i.e.*, when $\hat{\boldsymbol{f}}_\lambda = \sum_{j=1}^{M} \lambda_j \varphi_j$ with some known functions φ_j. Then, it is clear that $r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}'} - \nabla r_{\boldsymbol{\lambda}'}^{\top}(\boldsymbol{\lambda}-\boldsymbol{\lambda}') = 2(\boldsymbol{\lambda}-\boldsymbol{\lambda}')^{\top}\Phi(\boldsymbol{\lambda}-\boldsymbol{\lambda}')$, where Φ is the Gram matrix defined by $\Phi_{i,j} = \langle \varphi_i | \varphi_j \rangle_n$. So the condition in Proposition 3 consists in bounding the Gram matrix of the atoms φ_j. Let us remark that in this case—see, for instance, [22, 24]—$\mathrm{Tr}(\mathcal{M})$ is on the order of M and the choice $\tau = \sqrt{\beta/(nM)}$ ensures that the last term in the right-hand side of Eq. (5.5) decreases at the parametric rate $1/n$. This is the choice we recommend for practical applications.

As a second example, let us consider the case of a large number of linear estimators $\hat{\boldsymbol{g}}_1 = G_1\boldsymbol{Y}, \ldots, \hat{\boldsymbol{g}}_M = G_M\boldsymbol{Y}$ satisfying conditions of Setting 1 and such that $\max_{m=1,\ldots,M} |\!|\!|G_m|\!|\!| \le 1$. Assume we aim at proposing an estimator mimicking 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 our framework 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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)$. Remark that if $\{\hat{\boldsymbol{g}}_m\}$ satisfies conditions of Setting 1, so does $\{\hat{\boldsymbol{f}}_\lambda\}$. Moreover, the mapping $\boldsymbol{\lambda} \mapsto r_{\boldsymbol{\lambda}}$ is quadratic with Hessian matrix $\nabla^2 r_{\boldsymbol{\lambda}}$ given by the entries $2\langle G_m\boldsymbol{f} | G_{m'}\boldsymbol{f}\rangle_n + \frac{2}{n}\,\mathrm{Tr}(G_{m'}\Sigma G_m)$, $m, m' =$

$1,\dots,M$. It implies that Inequality (5.4) holds with $\mathcal{M}=\nabla^2 r_{\boldsymbol{\lambda}}/2$. Therefore, denoting by σ_i^2 the ith diagonal entry of Σ and setting $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_n)$, we get $\mathrm{Tr}(\mathcal{M})\le |||\sum_{m=1}^M G_m^2|||\big[\|\boldsymbol{f}\|_n^2+\|\boldsymbol{\sigma}\|_n^2\big]\le M\big[\|\boldsymbol{f}\|_n^2+\|\boldsymbol{\sigma}\|_n^2\big]$. Applying Proposition 3 with $\tau=\sqrt{\beta/(nM)}$, we get

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}}-\boldsymbol{f}\|_n^2\big]\le\inf_{\alpha,m,m'}\mathbb{E}\big[\|\alpha\hat{\boldsymbol{g}}_m+(1-\alpha)\hat{\boldsymbol{g}}_{m'}-\boldsymbol{f}\|_n^2\big]$$
$$+\frac{8\beta}{n}\log\Big(1+\Big[\frac{Mn}{\beta}\Big]^{1/2}\Big)+\frac{\beta}{n}\big[\|\boldsymbol{f}\|_n^2+\|\boldsymbol{\sigma}\|_n^2\big],\quad(5.6)$$

where the inf is taken over all $\alpha\in[0,1]$ and $m,m'\in\{1,\dots,M\}$. This inequality is derived from (5.5) by upper-bounding the $\inf_{\boldsymbol{\lambda}\in\mathbb{R}^M}$ by the infimum over $\boldsymbol{\lambda}$'s having at most two non-zero coefficients, λ_{m_0} and $\lambda_{m_0'}$, that are non-negative and sum to one: $\lambda_{m_0}+\lambda_{m_0'}=1$. To get (5.6), one simply notes that only two terms of the sum $\sum_m\log\big(1+|\lambda_m|\tau^{-1}\big)$ are non-zero and each of them is not larger than $\log\big(1+\tau^{-1}\big)$. Thus, one can achieve using EWA 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.

5.4. *Oracle inequalities for varying-block-shrinkage estimators.* Let us consider now the problem of aggregation of two-block shrinkage estimators. This 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 uniform on Λ.

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

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}}-\boldsymbol{f}\|_n^2\big]\le\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}}-\boldsymbol{f}\|_n^2\big]+\frac{\beta}{n}\bigg\{1+\log\bigg(\frac{n^2\|\boldsymbol{f}\|_n^2+n\,\mathrm{Tr}(\Sigma)}{12\beta}\bigg)\bigg\}.\quad(5.7)$$

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

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{Leung}}-\boldsymbol{f}\|_n^2\big]\le\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}}-\boldsymbol{f}\|_n^2\big]+\frac{9}{n}+\frac{8}{n}\min_{K>0}\bigg\{K\vee\bigg(\log\frac{n-6}{K}-1\bigg)\bigg\}.\quad(5.8)$$

6. Application to minimax adaptive estimation. Pinsker proved in his celebrated paper [49], 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[2] (DCT) transform of $\boldsymbol{f}$, hereafter denoted by $\mathcal{D}\boldsymbol{f}$. Pinsker's result—restricted to Sobolev ellipsoids $\mathcal{F}_{\mathcal{D}}(\alpha, R) = \big\{\boldsymbol{f} \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha}\theta_k(\boldsymbol{f})^2 \le R\big\}$ —states that, as $n \to \infty$, the equivalences

$$\inf_{\hat{\boldsymbol{f}}} \sup_{\boldsymbol{f} \in \mathcal{F}_{\mathcal{D}}(\alpha,R)} \mathbb{E}\big[\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|_n^2\big] \sim \inf_{A} \sup_{f \in \mathcal{F}_{\mathcal{D}}(\alpha,R)} \mathbb{E}\big[\|A\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big] \tag{6.1}$$

$$\sim \inf_{w>0} \sup_{f \in \mathcal{F}_{\mathcal{D}}(\alpha,R)} \mathbb{E}\big[\|A_{\alpha,w}\boldsymbol{Y} - \boldsymbol{f}\|_n^2\big] \tag{6.2}$$

hold [61, Theorem 3.2], where the first inf is taken over all possible estimators $\hat{\boldsymbol{f}}$ and $A_{\alpha,w} = \mathcal{D}^\top \mathrm{diag}\,\big((1 - k^\alpha/w)_+; k = 1, \dots, n\big)\mathcal{D}$ 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 [28]. 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 $\boldsymbol{\lambda} = (\alpha, w) \in \Lambda = \mathbb{R}_+^2$ and consider the prior*

$$\pi(d\boldsymbol{\lambda}) = \frac{2n_\sigma^{-\alpha/(2\alpha+1)}}{\big(1 + n_\sigma^{-\alpha/(2\alpha+1)}w\big)^3} e^{-\alpha} d\alpha dw, \tag{6.3}$$

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[3] on the family of classes $\{\mathcal{F}_{\mathcal{D}}(\alpha, R) : \alpha > 0, R > 0\}$.

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

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

[3] see [61, Definition 3.8]

on some parameters such as the lengths of blocks for blockwise Stein [14] and Efromovich-Pinsker [29] estimators or the step of discretization and the maximal value of bandwidth [15]. 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 [31].

We now turn to the setup of heteroscedastic regression, which corresponds to ill-posed inverse problems as described in Section 4. To achieve adaptivity in the exact minimax sense, we make use of $\hat{\boldsymbol{f}}_{\text{GEWA}}$, the grouped version of the exponentially weighted aggregate. We assume hereafter that the matrix Σ is diagonal with diagonal entries $\sigma_1^2, \dots, \sigma_n^2$ satisfying the following property:

$$\exists\, \sigma_*, \gamma > 0 \quad \text{such that} \quad \sigma_k^2 = \sigma_*^2 k^{2\gamma}(1 + o_k(1)) \quad \text{as} \quad k \to \infty. \tag{6.4}$$

This kind of problems arise when T is a differential operator or the Radon transform [12, Section 1.3]. To handle such situations, we define the groups in the same spirit as the weakly geometrically increasing blocks in [13]. Let $\nu = \nu_n$ be a positive integer that increases as $n \to \infty$. Set $\rho_n = \nu_n^{-1/3}$ and define

$$T_j = \begin{cases} (1+\nu_n)^{j-1} - 1, & j = 1, 2, \\ T_{j-1} + \lfloor \nu_n \rho_n (1+\rho_n)^{j-2} \rfloor, & j = 3, 4, \dots, \end{cases} \tag{6.5}$$

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

PROPOSITION 6. *Let the groups $B_1, \dots, B_J$ 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*

$$\pi(d\boldsymbol{\lambda}) = \frac{2n^{-\alpha/(2\alpha+2\gamma+1)}}{\big(1 + n^{-\alpha/(2\alpha+2\gamma+1)} w\big)^3} e^{-\alpha} d\alpha dw. \tag{6.6}$$

Then, in model (1.1) with diagonal covariance matrix $\Sigma = diag(\sigma_k^2; 1 \leq k \leq n)$ satisfying condition (6.4), the aggregate $\hat{\boldsymbol{f}}_{\text{GEWA}}$ (under Setting 1) 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 $\{\mathcal{F}(\alpha, R) : \alpha > 0, R > 0\}$.

Note that this result provides an estimator attaining the optimal constant in the minimax sense when the unknown signal lies in an ellipsoid. This property holds because minimax estimators over the ellipsoids are linear. For other subsets of $\mathbb{R}^n$, such as hyper-rectangles, Besov bodies and so on, this is not true anymore. However, as proved by Donoho, Liu and MacGibbon [27], 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, it is also possible to prove that GEWA can lead to an adaptive estimator whose risk is within within 25% of the minimax risk, for a broad class of hyperrectangles.

7. 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. A toolbox is made available freely for download at `http://josephsalmon.eu/code/index_codes.php`. Additional details and numerical experiments can be found in [20, 55].

We evaluate different estimation routines on several 1D signals considered as a benchmark in the literature on signal processing [25]. The six signals we retained for our experiments because of their diversity are depicted in Figure 1. Since these signals are non-smooth, we have also carried out experiments on their smoothed versions obtained by taking the antiderivative. Experiments on non-smooth (resp. smooth) signals are referred to as Experiment I (resp. 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 DCT denoted by $\boldsymbol{\theta}(\boldsymbol{Y}) = \big(\theta_1(\boldsymbol{Y}), \dots, \theta_n(\boldsymbol{Y})\big)^\top$.

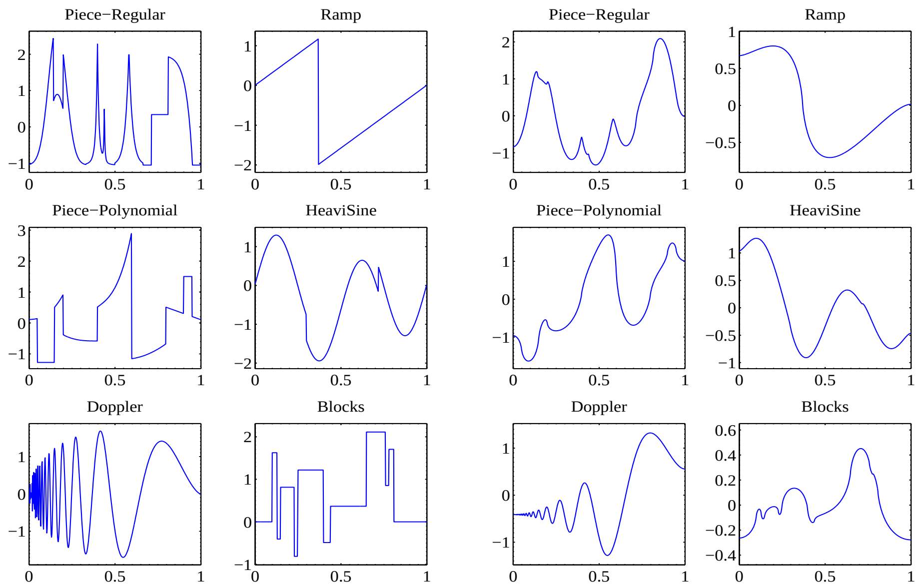
The four tested estimation routines—including EWA—are detailed below.

Soft-Thresholding (ST) [25]: For a given shrinkage parameter t, the soft-thresholding estimator is $\widehat{\theta}_k = \operatorname{sgn}\big(\theta_k(\boldsymbol{Y})\big)\big(|\theta_k(\boldsymbol{Y})| - \sigma t\big)_+$. We use the data-driven threshold minimizing the Stein unbiased risk estimate [26].

Blockwise James-Stein (BJS) shrinkage [10]: The set $\{1, \dots, n\}$ is partitioned into $N = [n/\log(n)]$ blocks $B_1, B_2, \dots, B_N$ of nearly equal size L. The corresponding blocks of true coefficients $\theta_{B_k}(\boldsymbol{f}) = \big(\theta_j(\boldsymbol{f})\big)_{j\in B_k}$ are then estimated by: $\widehat{\theta}_{B_k} = \Big(1 - \frac{\lambda L \sigma^2}{S_k^2(\boldsymbol{Y})}\Big)_+ \theta_{B_k}(\boldsymbol{Y})$, $k = 1, \dots, N$, with blocks of noisy coefficients $\theta_{B_k}(\boldsymbol{Y})$, $S_k^2 = \|\theta_{B_k}(\boldsymbol{Y})\|_2^2$ and $\lambda = 4.50524$.

Unbiased risk estimate (URE) minimization with Pinsker's filters [15]: Pinsker filter with a data-driven parameters α and w selected 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.

EWA on Pinsker's filters: We consider the same finite family of linear filters (defined by Pinsker's filters) as in the URE routine described above.



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

Fig 1 – 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).

According to Proposition 1, this leads to an estimator nearly as accurate as the best Pinsker's estimator in the given family.

To report the result of our experiments, we have also computed the best linear smoother, hereafter referred to as oracle, based on a Pinsker filter chosen among the candidates that we used for defining URE and EWA. By best smoother we mean the one minimizing the squared error (it can be computed since we know the ground truth). 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 (*cf.,* Table 1) for every size n of signals considered. Improvements are particularly important when signals have large peaks or discontinuities. In most cases, EWA also outperforms URE, but differences are less pronounced. One can also observe that for smooth signals, the difference of MSEs between EWA and the oracle, multiplied by n, remains nearly constant when n varies. This is in agreement with our theoretical results in which the residual term decreases to zero inversely proportionally to n.

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 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.

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

Evaluation of 4 adaptive methods on 6 (non-smooth) signals. For each sample size and each method, we report the average value of $n(\mathrm{MSE} - \mathrm{MSE}_{\mathrm{Oracle}})$ *and the corresponding standard deviation (in parentheses), for 1000 replications of the experiment.*

8. 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 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 rate-optimal residual term. As an application of our results, we have shown that EWA acting on Pinsker's estimators, produces an adaptive estimator 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 [24], Rigollet and Tsybakov [53, 54] and Alquier and Lounici [1], with a special focus on the problems involving large scale data.

Although we do not assume the covariance matrix Σ of the noise to be known, our approach relies on an unbiased estimator of Σ which is independent on the observed signal and on an upper bound on the largest singular value of Σ. In some applications, such an information may be hard to obtain and it can be helpful to relax the assumptions on $\widehat{\Sigma}$. This is another interesting avenue for future research for which, we believe, the approach developed by Giraud [34] can be of valuable guidance.

APPENDIX A: PROOFS OF MAIN THEOREMS

We develop now the detailed proofs of the results stated in the manuscript.

A.1. Stein's lemma. The proofs of our main results rely on Stein's lemma [59], recalled below, providing an unbiased risk estimate for any estimator that depends sufficiently smoothly on the data vector $\boldsymbol{Y}$.

LEMMA 1. *Let $\boldsymbol{Y}$ be a 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 $\boldsymbol{\nabla} \cdot \hat{\boldsymbol{f}}^\top := (\partial_i \hat{\boldsymbol{f}}_j)$ have finite first moment, then*

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

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

The proof can be found in [61, p.157]. We apply Stein's lemma to the affine estimators $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$, with A_λ an $n \times n$ deterministic real matrix and $\boldsymbol{b}_\lambda \in \mathbb{R}^n$ a deterministic vector. We get that if $\widehat{\Sigma}$ is an unbiased estimator of Σ, then $\hat{r}_\lambda^{\text{unb}} = \big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n} \operatorname{Tr}[\widehat{\Sigma} A_\lambda] - \frac{1}{n} \operatorname{Tr}[\widehat{\Sigma}]$ 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	0.387	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

Evaluation of 4 adaptive methods on 6 smoothed signals. For each sample size and each method, we report the average value of $n(\mathrm{MSE}-\mathrm{MSE}_{\mathrm{Oracle}})$ *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* $\hat r_\lambda^{\rm unb}$ *be the unbiased risk estimate of* $\hat{\boldsymbol f}_\lambda$ *given by Stein's lemma.*

1) *For every* $\pi\in\mathcal P_\Lambda$ *and for any* $\beta>0$, *the estimator* $\hat{\boldsymbol f}_{\rm EWA}$ *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

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

as unbiased estimator of the risk.

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

$$\mathbb E[\|\hat{\boldsymbol f}_{\rm EWA}-\boldsymbol f\|_n^2]\le \inf_{p\in\mathcal P_\Lambda}\Big\{\int_\Lambda \mathbb E[\hat r_\lambda]\,p(d\lambda)+\frac{\beta\mathcal K(p,\pi)}{n}\Big\}. \tag{A.1}$$

PROOF. According to the Stein lemma, the quantity

$$\hat r_{\rm EWA}=\|\boldsymbol Y-\hat{\boldsymbol f}_{\rm EWA}\|_n^2+\frac{2}{n}\,{\rm Tr}[\Sigma(\boldsymbol\nabla\cdot\hat{\boldsymbol f}_{\rm EWA}(\boldsymbol Y)]-\frac1n\,{\rm Tr}[\Sigma] \tag{A.2}$$

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

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

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

$$\hat r_{\rm EWA}=\int_\Lambda\big(\hat r_\lambda^{\rm unb}-\|\hat{\boldsymbol f}_\lambda-\hat{\boldsymbol f}_{\rm EWA}\|_n^2\big)\hat\pi(d\lambda)+\frac2n\int_\Lambda{\rm Tr}[\Sigma\hat{\boldsymbol f}_\lambda\nabla_{\boldsymbol Y}\theta(\boldsymbol Y,\lambda)^\top]\,\pi(d\lambda).$$

After having interchanged differentiation and integration, we obtain that $\int_\Lambda\hat{\boldsymbol f}_{\rm EWA}\big(\nabla_{\boldsymbol Y}\theta(\boldsymbol Y,\lambda)\big)^\top\pi(d\lambda)=\hat{\boldsymbol f}_{\rm EWA}\nabla_{\boldsymbol Y}\big(\int_\Lambda\theta(\boldsymbol Y,\lambda)\pi(d\lambda)\big)=0$ and, therefore, we come up with the following expression for $\hat r_{\rm EWA}$:

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

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 $\beta\ge 2a$ it holds that $\hat r_{\rm EWA}\le\int_\Lambda\hat r_\lambda^{\rm unb}\hat\pi(d\lambda)\le\int_\Lambda\hat r_\lambda\hat\pi(d\lambda)\le\int_\Lambda\hat r_\lambda\hat\pi(d\lambda)+\frac{\beta}{n}\mathcal K(\hat\pi,\pi)$. To conclude, it suffices to remark that $\hat\pi$ is the probability measure minimizing the criterion $\int_\Lambda\hat r_\lambda p(d\lambda)+\frac{\beta}{n}\mathcal K(p,\pi)$ among all $p\in\mathcal P_\Lambda$. Thus, for every $p\in\mathcal P_\Lambda$, we have

$$\hat r_{\rm EWA}\le\int_\Lambda\hat r_\lambda p(d\lambda)+\frac\beta n\,\mathcal K(p,\pi).$$

Taking the expectation of both sides, the desired result follows. $\square$

A.3. Proof of Theorem 1.

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

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

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 $Q = A_{\text{EWA}}$ (in view of the commuting property of the A_λ's) we get the following relation: $\big\langle (I-A_\lambda)^2\boldsymbol{Y} | \Sigma(A_\lambda - A_{\text{EWA}})\boldsymbol{Y}\big\rangle_n = \big\langle (I-A_{\text{EWA}})^2\boldsymbol{Y}|\Sigma(A_\lambda - A_{\text{EWA}})\boldsymbol{Y}\big\rangle_n - 2\big\langle \big(I - \frac{A_{\text{EWA}}+A_\lambda}{2}\big)(A_{\text{EWA}}-A_\lambda)\boldsymbol{Y}|\Sigma(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}\big\rangle_n$. When one integrates over Λ with respect to the measure $\hat\pi$, the term of the first scalar product in the right-hand side of the last equation vanishes. On the other hand,

$$\begin{aligned}
\langle A_\lambda(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}|\Sigma(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}\rangle_n
&= \langle A_\lambda(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)|\Sigma(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)\rangle_n\\
&= \langle (\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)|A_\lambda\Sigma(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)\rangle_n\\
&= \frac{1}{2n}(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)^\top (A_\lambda\Sigma + \Sigma A_\lambda)(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda) \ge 0.
\end{aligned}$$

Since positive semi-definiteness of matrices $\Sigma A_\lambda + A_\lambda\Sigma$ implies the one of the matrix $\Sigma A_{\text{EWA}} + \Sigma A_{\text{EWA}}$, we also have $\langle A_{\text{EWA}}(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}|\Sigma(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}\rangle_n \ge 0$. Therefore,

$$\begin{aligned}
\Big\langle \Big(I - \frac{A_{\text{EWA}} + A_\lambda}{2}\Big)(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}|\Sigma(A_{\text{EWA}} - A_\lambda)\boldsymbol{Y}\Big\rangle_n&\\
\le \langle (\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)|\Sigma(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)\rangle_n&\\
= \|\Sigma^{1/2}(\hat{\boldsymbol{f}}_{\text{EWA}} - \hat{\boldsymbol{f}}_\lambda)\|_n^2.&
\end{aligned}$$

This inequality implies that

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

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

Assertion ii). Let now $\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y} + \boldsymbol{b}_\lambda$ with symmetric $A_\lambda \preceq I_{n\times n}$ and $\boldsymbol{b}_\lambda \in \mathrm{Ker}(A_\lambda)$. Using 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 $\hat{r}_\lambda^{\mathrm{adj}} \geq \hat{r}_\lambda^{\mathrm{unb}}$ for every λ and that

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

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

A.4. Proof of Theorem 2. We apply the result of assertion ii) of Thm. 1 to the prior $\pi(d\lambda)$ replaced by the probability measure proportional to $e^{\frac{2}{\beta}\mathrm{Tr}[\widehat{\Sigma}(A_\lambda - A_\lambda^\top A_\lambda)]}\pi(d\lambda)$. This leads to

$$\begin{aligned}\mathbb{E}\big[\|\tilde{\boldsymbol{f}}_{\mathrm{SEWA}} - \boldsymbol{f}\|_n^2\big] \leq \inf_{p\in\mathcal{P}_\Lambda}&\bigg\{\int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big]p(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p,\pi)\bigg\}\\ &+\frac{\beta}{n}\mathbb{E}\bigg[\log\int_\Lambda e^{\frac{2}{\beta}\mathrm{Tr}[\widehat{\Sigma}(A_\lambda - A_\lambda^\top A_\lambda)]}\pi(d\lambda)\bigg].\end{aligned}$$

Condition **(C)** entails that the last term is always non-negative and the result follows.

A.5. Proof of Theorem 3. Let us place ourselves in Setting 1. It is clear that $\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\mathrm{GEWA}} - \boldsymbol{f}\|_n^2\big] = \sum_{j=1}^J \mathbb{E}\big[\|\hat{\boldsymbol{f}}^j_{\mathrm{GEWA}} - \boldsymbol{f}^j\|_n^2\big]$. For each $j \in \{1,\ldots,J\}$, since $\beta_j \geq 8|||\Sigma^j|||$, one can apply assertion i) of Thm. 1, which leads to the desired result. The case of Setting 2 is handled in the same manner.

SUPPLEMENTARY MATERIAL

Supplement A: Proofs of some propositions
(http://lib.stat.cmu.edu/aos/???/???). In this supplement, we present the detailed proofs of Propositions 2, 3, 4, 5 and 6.

REFERENCES

[1] Alquier, P. and Lounici, K. (2010). PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights. *Electron. J. Stat.* **5** 127-145.

[2] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput.* **9** 1545–1588.

[3] Arlot, S. and Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. In *NIPS* 46–54.
[4] Audibert, J.-Y. (2007). Progressive mixture rules are deviation suboptimal. In *NIPS* 41–48.
[5] Baraud, Y., Giraud, C. and Huet, S. (2010). Estimator selection in the Gaussian setting. *submitted*.
[6] Barron, A. R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.
[7] Breiman, L. (1996). Bagging Predictors. *Mach. Learn.* **24** 123–140.
[8] Buades, A., Coll, B. and Morel, J.-M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4** 490–530.
[9] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697.
[10] Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924.
[11] Catoni, O. (2004). *Statistical learning theory and stochastic optimization. Lecture Notes in Mathematics* **1851**. Springer-Verlag, Berlin.
[12] Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems* **24** 034004, 19.
[13] Cavalier, L. and Tsybakov, A. B. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.* **10** 247–282. Meeting on Mathematical Statistics (Marseille, 2000).
[14] Cavalier, L. and Tsybakov, A. B. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123** 323–354.
[15] Cavalier, L., Golubev, G. K., Picard, D. and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.* **30** 843–874.
[16] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge University Press.
[17] Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation Optimal Learning using Greedy Q-aggregation. *Ann. Statist., to appear.* arXiv:1203.2507.
[18] Dai, D. and Zhang, T. (2011). Greedy Model Averaging. In *NIPS* 1242–1250.
[19] Dalalyan, A. S. and Salmon, J. Supplement to "Sharp Oracle Inequalities for Aggregation of Affine Estimators".
[20] Dalalyan, A. S. and Salmon, J. (2011). Competing against the best nearest neighbor filter in regression. In *ALT. Lecture Notes in Computer Science* **6925** 129-143. Springer.
[21] Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by Exponential Weighting and Sharp Oracle Inequalities. In *COLT. Lecture Notes in Computer Science* **4539** 97-111. Springer.
[22] Dalalyan, A. S. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.
[23] Dalalyan, A. S. and Tsybakov, A. B. (2012a). Mirror Averaging with Sparsity Priors. *Bernoulli,* in press.
[24] Dalalyan, A. S. and Tsybakov, A. B. (2012b). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* **78** 1423-1443.
[25] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.
[26] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.
[27] Donoho, D. L., Liu, R. C. and MacGibbon, B. (1990). Minimax risk over hyper-

rectangles, and implications. *Ann. Statist.* **18** 1416–1437.
[28] Efromovich, S. Y. and Pinsker, M. S. (1984). A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.* **1** 58–65.
[29] Efromovich, S. Y. and Pinsker, M. S. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942.
[30] Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *COLT* 202–216.
[31] Gaïffas, S. and Lecué, G. (2011). Hyper-Sparse Optimal Aggregation. *J. Mach. Learn. Res.* **12** 1813–1833.
[32] George, E. I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14** 188–205.
[33] Gerchinovitz, S. (2011). Sparsity regret bounds for individual sequences in online linear regression. *J. Mach. Learn. Res. - Proceedings Track* **19** 377–396.
[34] Giraud, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089–1107.
[35] Goldenshluger, A. and Lepski, O. V. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190.
[36] Golubev, Y. (2010). On universal oracle inequalities related to high-dimensional linear models. *Ann. Statist.* **38** 2751–2780.
[37] Juditsky, A. B. and Nemirovski, A. S. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712.
[38] Juditsky, A. B. and Nemirovski, A. S. (2009). Nonparametric denoising of signals with unknown local structure. I. Oracle inequalities. *Appl. Comput. Harmon. Anal.* **27** 157–179.
[39] Kivinen, J. and Warmuth, M. K. (1999). Averaging expert predictions. In *Computational learning theory (EuroCOLT). Lecture Notes in Comput. Sci.* **1572** 153–167. Springer, Berlin.
[40] Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.
[41] Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L. and Jordan, M. (2003/04). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.* **5** 27–72 (electronic).
[42] Langford, J. and Shawe-Taylor, J. (2002). PAC-Bayes & Margins. In *NIPS* 423-430.
[43] Lecué, G. and Mendelson, S. (2010). On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli, to appear.*
[44] Leung, G. (2004). Information Theory and Mixing Least Squares Regression PhD thesis, Yale University.
[45] Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** 3396–3410.
[46] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.* **16** 246–259.
[47] McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *COLT* 230–234. ACM, New York, NY, USA.
[48] Nemirovski, A. S. (2000). *Topics in non-parametric statistics. Lecture Notes in Math.* **1738**. Springer, Berlin.
[49] Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inf.* **16** 52–68.
[50] Polzehl, J. and Spokoiny, V. G. (2000). Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 335–354.
[51] Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized lin-

ear models. *Ann. Statist.* **40** 639-665.

[52] Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280.

[53] Rigollet, P. and Tsybakov, A. B. (2011a). Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–471.

[54] Rigollet, P. and Tsybakov, A. B. (2011b). Sparse estimation by exponential weighting. *Submitted*.

[55] Salmon, J. and Dalalyan, A. S. (2011). Optimal aggregation of affine estimators. *Journal of Machine Learning Research - Proceedings Track* **19** 635-660.

[56] Salmon, J. and Le Pennec, E. (2009). NL-Means and aggregation procedures. In *ICIP* 2977–2980.

[57] Seeger, M. (2003). Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.* **3** 233–269.

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

[59] Stein, C. M. (1973). Estimation of the mean of a multivariate distribution. In *Proc. Prague Symp. Asymptotic Statist.*

[60] Tsybakov, A. B. (2003). Optimal Rates of Aggregation. In *COLT* 303–313.

[61] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation. Springer Series in Statistics*. Springer, New York.

[62] Wang, Z., Paterlini, S., Gao, F. and Yang, Y. (2012). Adaptive Minimax Estimation over Sparse l_q-Hulls. Technical Report, arXiv:1108.1961v4 [math.ST].

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

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

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

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

E-mail: name.surname@ensae.fr
URL: http://www.arnak-dalalyan.fr/

E-mail: name@math.jussieu.fr
URL: http://josephsalmon.eu/

Supplement to "Sharp Oracle Inequalities for Aggregation of Affine Estimators"

Arnak S. Dalalyan and Joseph Salmon

e-mail: name.surname@ensae.fr
url: http://www.arnak-dalalyan.fr/

e-mail: name@math.jussieu.fr
url: http://josephsalmon.eu

Abstract: This supplementary material contains the proofs of Propositions 2-6 presented in the paper Dalalyan and Salmon [1].

AMS 2000 subject classifications: Primary 62G08, ; secondary 62C20,62G05,62G20.
Keywords and phrases: aggregation, regression, oracle inequalities, model selection, minimax risk, exponentially weighted aggregation.

1. Proof of Proposition 2

Let us fix $\lambda^* \in \Lambda$. It suffices to apply [1, Thm. 3] and to upper-bound the right-hand side of inequality (2.8) from [1] as follows :

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

Then, the right-hand side of the last inequality can be bounded from above by the expression in parentheses evaluated at the probability measure p given by the density $p^*(\lambda) = \mathbb{1}_{B_{\lambda^*}(\tau^*)}(\lambda)/\text{Leb}(B_{\lambda^*}(\tau^*))$. Assume moreover that τ^* is such that $B_{\lambda^*}(\tau^*) \subset \Lambda$, then using the Lipschitz condition on r_λ, the bound on the risk becomes

$$\begin{aligned}\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] &\leq \int_\Lambda [|r_\lambda - r_{\lambda^*}| + r_{\lambda^*}]\, p^*(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p^*,\pi)\\ &\leq r_{\lambda^*} + L_r\int_\Lambda \|\lambda-\lambda^*\|_2\, p^*(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p^*,\pi)\\ &\leq r_{\lambda^*} + L_r\tau^* + \frac{\beta}{n}\,\mathcal{K}(p^*,\pi). \qquad (1.1)\end{aligned}$$

Now, since λ^* is such that $B_{\lambda^*}(\tau^*) \subset \Lambda$, the measure $p^*(\lambda)\,d\lambda$ is absolutely continuous w.r.t. the uniform prior π over Λ and the Kullback-Leibler divergence between these measures equals $\log\big\{\text{Leb}(\Lambda)/\text{Leb}\big(B_{\lambda^*}(\tau^*)\big)\big\}$. By the simple inequality $\|x\|_2^2 \leq M\|x\|_\infty^2$ for any $x \in \mathbb{R}^M$, one can see that the Euclidean ball of radius τ^* contains the hypercube of width $2\tau^*/\sqrt{M}$. So we have the following lower bound for the volume B_{λ^*}: $\text{Leb}\big(B_{\lambda^*}(\tau^*)\big) \geq (2\tau^*/\sqrt{M})^M$. By combining this with inequality (1.1) the results of Proposition 2 is straightforward.

2. Proof of Proposition 3

We start the proof as the one of the previous proposition, but pushing the development of the function $\lambda \mapsto r_\lambda$ up to second order. So, for any $\lambda^* \in \mathbb{R}^M$, the risk of EWA $\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big]$

is upper-bounded by

$$r_{\lambda^*} + \int_\Lambda \Big(\nabla r_{\lambda^*}^\top(\lambda-\lambda^*) + (\lambda-\lambda^*)^\top \mathcal{M}(\lambda-\lambda^*)\Big)\, p^*(d\lambda) + \frac{\beta}{n}\mathcal{K}(p^*,\pi).$$

By choosing $p^*(\lambda) = \pi(\lambda - \lambda^*)$ 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 ζ is drawn from the scaled Student $t(3)$ distribution, its density function is $u \to 2/[\pi(1+u^2)^2]$, and $\mathbb{E}[\zeta^2] = 1$. Thus, we have that $\int_\Lambda \lambda_1^2\pi(\lambda)d\lambda = \tau^2$. We can then bound the risk of EWA as follows:

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] \le \inf_{\lambda^*\in\mathbb{R}^M} \Big(r_{\lambda^*} + \operatorname{Tr}(\mathcal{M})\tau^2 + \frac{\beta}{n}\,\mathcal{K}(p^*,\pi)\Big) \tag{2.1}$$

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:

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

We bound the quotient in the above equality by

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

Since the last inequality is independent of λ and p^* is a probability measure, the integral disappears in the previous bound on the Kullback-Leibler divergence. So we get $\mathcal{K}(p^*,\pi) \le 4\sum_{m=1}^M \log\Big(1+\frac{|\lambda_m^*|}{\tau}\Big)$. This inequality combined with (2.1) leads to the desired result.

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 the probability density function $p_0(d\boldsymbol{\lambda}) = \mathbb{1}_{[a_0-\tau,a_0+\tau]}(a)\mathbb{1}_{[b_0-\tau,b_0+\tau]}(b)\mathbb{1}(k=k_0)(2\tau)^{-2}\,da\,db$.

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

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

In particular, the difference between the risks $r_{\boldsymbol{\lambda}}$ and $r_{\boldsymbol{\lambda}_0}$—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:

$$r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}_0} = \frac{1}{n}\sum_{i=1}^{k}\Big[2(a_0-a)\big\{(1-a_0)f_i^2 - a_0\sigma_i^2\big\} + (a-a_0)^2\{f_i^2+\sigma_i^2\}\Big]$$
$$+\frac{1}{n}\sum_{i=k+1}^{n}\Big[2(b_0-b)\big\{(1-b_0)f_i^2 - b_0\sigma_i^2\big\} + (b-b_0)^2\{f_i^2+\sigma_i^2\}\Big]$$

If we integrate w.r.t. the measure p_0, the terms linear in $a-a_0$ and $b-b_0$ disappear and we get

$$\int_\Lambda (r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}_0})\, p_0(d\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\{f_i^2+\sigma_i^2\}\int_{-\tau}^{\tau} u^2\frac{du}{2\tau} = \frac{\tau^2}{3}\big[\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2\big].$$

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

$$\begin{aligned}\mathcal{K}(p_0,\pi) &= \sum_{k=1}^{n}\int\int \log\left(\frac{p_0(da,db,k)}{\pi(da,db,k)}\right)p_0(da,db,k)\\ &= \int_{a_0-\tau}^{a_0+\tau}\int_{b_0-\tau}^{b_0+\tau}\log\left(\frac{n}{4\tau^2}\right)\frac{da}{2\tau}\frac{db}{2\tau} = \log\left(\frac{n}{4\tau^2}\right).\end{aligned} \tag{3.1}$$

We can use [1, Equation (2.8)] with our choice for p_0 and π, and in view of the last computations, we get

$$\begin{aligned}\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] &\le \int_\Lambda r_{\boldsymbol{\lambda}}\, p_0(d\boldsymbol{\lambda}) + \frac{\beta}{n}\mathcal{K}(p_0,\pi)\\ &= r_{\boldsymbol{\lambda}_0} + \int_\Lambda (r_{\boldsymbol{\lambda}} - r_{\boldsymbol{\lambda}_0})p_0(d\boldsymbol{\lambda}) + \frac{\beta}{n}\log\left(\frac{n}{4\tau^2}\right)\\ &= r_{\boldsymbol{\lambda}_0} + \frac{\tau^2\big[\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2\big]}{3} + \frac{\beta}{n}\log\left(\frac{n}{4\tau^2}\right).\end{aligned} \tag{3.2}$$

The last expression, considered as a function of τ, admits as global minimum $\tau^2_{\min} = 3(\beta/n)\big[\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2\big]$. Replacing this value in (3.2), we get the risk bound:

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] \le \mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}_0} - \boldsymbol{f}\|_n^2\big] + \frac{\beta}{n}\left\{1 + \log\left(\frac{n^2\big[\|\boldsymbol{f}\|_n^2 + \|\boldsymbol{\sigma}\|_n^2\big]}{12\beta}\right)\right\}. \tag{3.3}$$

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

4. Proof of Proposition 5

We assume, without loss of generality, that the matrix $n^{1/2}\mathcal{D}$ coincides with the identity matrix. First, let us fix $\alpha_0 > 0$ and $R_0 > 0$, such that $n^{-1/2}\boldsymbol{f} \in \mathcal{F}(\alpha_0, R_0)$ 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 $\mathcal{F}(\alpha_0, R_0)$.

In what follows, we set $n_\sigma = n/\sigma^2$, $n_{\sigma,\alpha} = n_\sigma^{-\alpha/(2\alpha+1)}$ and we denote by p_π the probability density function of π w.r.t. the Lebesgue measure on $\mathbb{R}_+^2$: $p_\pi(\alpha,w) = e^{-\alpha}n_{\sigma,\alpha}p_w(wn_{\sigma,\alpha})$, 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

$$\int_{\mathbb{R}^2} \alpha\, p_\pi(\alpha,w)\,d\alpha dw = 1, \qquad \int_{\mathbb{R}^2} w\, p_\pi(\alpha,w)\,d\alpha dw \le n_\sigma^{1/2}. \tag{4.1}$$

Let τ be a positive number such that $\tau \le \min(1,\alpha_0/(2\log w_0))$ and choose p_0 as a translation/dilatation of π, concentrating on $\boldsymbol{\lambda}_0$ when $\tau \to 0$:

$$p_0(d\boldsymbol{\lambda}) = p_\pi\Big(\frac{\boldsymbol{\lambda}-\boldsymbol{\lambda}_0}{\tau}\Big)\frac{d\boldsymbol{\lambda}}{\tau^2}\,.$$

In view of assertion i) of Theorem 1 from [1],

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] \le r_{\boldsymbol{\lambda}_0} + \int_{\mathbb{R}^2} \big|r_{\alpha,w} - r_{\alpha_0,w_0}\big| p_0(d\boldsymbol{\lambda}) + \frac{\beta}{n}\mathcal{K}(p_0,\pi)\,. \tag{4.2}$$

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

$$n|r_{\alpha,w} - r_{\alpha,w_0}| \le 2\sum_{k=1}^n (f_k^2+\sigma^2)\,|(1-k^\alpha/w)_+ - (1-k^\alpha/w_0)_+|\;. \tag{4.3}$$

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

$$\Big|\Big(1-\frac{k^\alpha}{w}\Big)_+ - \Big(1-\frac{k^\alpha}{w_0}\Big)_+\Big| \le \Big|\frac{k^\alpha}{w} - \frac{k^\alpha}{w_0}\Big|\mathbb{1}_{\{k^\alpha \le w\}} \le \frac{w-w_0}{w_0}. \tag{4.4}$$

By virtue of Inequalities (4.3) and (4.4) we get

$$|r_{\alpha,w} - r_{\alpha,w_0}| \le 2n^{-1}\sum_{k=1}^n (f_k^2+\sigma^2)\frac{(w-w_0)}{w_0} \le 2(R_0+\sigma^2)\frac{w-w_0}{w_0}. \tag{4.5}$$

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

$$|r_{\alpha,w_0} - r_{\alpha_0,w_0}| \le \frac{2}{n}\sum_{k^{\alpha_0}\le w_0} (f_k^2+\sigma^2)\frac{k^\alpha - k^{\alpha_0}}{w_0} \le 2(R_0+\sigma^2)\big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}} - 1\big). \tag{4.6}$$

Recall that we aim to bound the second term in the right-hand side of (4.2). To this end, we need an accurate upper bound on the integrals of the right-hand sides of (4.5) and (4.6) w.r.t. the probability measure p_0. For the first one, we get

$$\begin{aligned} \int |r_{\alpha,w} - r_{\alpha,w_0}|\, p_0(d\boldsymbol{\lambda}) &\le 2(R_0+\sigma^2)w_0^{-1}\int_{\mathbb{R}^2}(w-w_0)p_0(d\boldsymbol{\lambda}) \\ &\le \frac{4n_\sigma^{1/2}\tau}{w_0}(R_0+\sigma^2)\,. \end{aligned} \tag{4.7}$$

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

$$
\begin{aligned}
\int_{\mathbb{R}^2} |r_{\alpha,w_0} - r_{\alpha_0,w_0}| p_0(d\boldsymbol{\lambda}) &\leq 2(R_0+\sigma^2)\int_{\mathbb{R}^2} \big(w_0^{\frac{\alpha-\alpha_0}{\alpha_0}} - 1\big) p_0(d\boldsymbol{\lambda}) \\
&= \frac{2\tau(R_0+\sigma^2)\log w_0}{\alpha_0 - \tau\log w_0} \\
&\leq 4\tau(R_0+\sigma^2)\alpha_0^{-1}\log w_0, \qquad (4.8)
\end{aligned}
$$

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

The last term to bound in inequality (4.2) requires the evaluation of the Kullback-Leibler divergence between p_0 and π. It can be done as follows:

$$
\begin{aligned}
\mathcal{K}(p_0,\pi) &= \int_{\mathbb{R}^2} \log\bigg(\frac{e^{-(\alpha-\alpha_0)/\tau} p_w\big((w-w_0)n_{\sigma,\alpha}/\tau\big)}{e^{-\alpha} p_w\big(w n_{\sigma,\alpha}\big)}\frac{1}{\tau^2}\bigg) p_0(d\boldsymbol{\lambda}) \\
&= \int_{\mathbb{R}^2} \bigg\{\alpha - \frac{\alpha-\alpha_0}{\tau} + \log\frac{p_w\big((w-w_0)n_{\sigma,\alpha}/\tau\big)}{p_w\big(w n_{\sigma,\alpha}\big)}\bigg\} p_0(d\boldsymbol{\lambda}) - 2\log(\tau) \\
&\leq \alpha_0 + (\tau - 1) + \int_{\mathbb{R}^2_+} \log\big(1 + w n_{\sigma,\alpha}\big)^3 p_0(d\boldsymbol{\lambda}) - 2\log(\tau).
\end{aligned}
$$

where the third equality is derived thanks to Eq. (4.1) and the relation $\|p_w\|_\infty = 2$. Making the change of variable $w = w_0 + \tau u n_{\sigma,\alpha}^{-1}$ and using that $w_0 + \tau u n_{\sigma,\alpha}^{-1} \leq (w_0+u)n_{\sigma,\alpha}^{-1}$, we get

$$
\begin{aligned}
\int_{\mathbb{R}^2_+} \log\big(1 + w n_{\sigma,\alpha}\big)^3 p_0(d\boldsymbol{\lambda}) &\leq 3\int_{\mathbb{R}_+} \log\big(1 + w_0 + u\big) p_w(u)du \\
&\leq 3\log\bigg(1 + w_0 + \int_{\mathbb{R}_+} u p_w(u)du\bigg) \\
&= 3\log(2 + w_0).
\end{aligned}
$$

Eventually, we can reformulate our bound on the risk of EWA given in (4.2), leading to

$$
\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2\big] \leq r_{\boldsymbol{\lambda}_0} + 4\tau(R_0+\sigma^2)\Big(\frac{n_\sigma^{1/2}}{w_0} + \frac{\log w_0}{\alpha_0}\Big) + \frac{8\sigma^2\big(\alpha_0 + 3\log\big(\frac{2+w_0}{\tau}\big)\big)}{n}. \qquad (4.9)
$$

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,

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

In view of this result, taking the max over $\boldsymbol{f} \in \mathcal{F}(\alpha_0, R_0)$ in (4.9), we get

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

This leads to the desired result in view of the relation

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

5. Proof of Proposition 6

It is clear that all the conditions required in Setting 1 are fulfilled and we can apply [1, Thm. 3] that yields:

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}}-\boldsymbol{f}\|_n^2\big]\le\sum_{j=1}^{J}\inf_{p_j}\Big\{\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,p_j(d\boldsymbol{\lambda})+\frac{\beta_j}{n}\,\mathcal{K}(p_j,\pi)\Big\}. \tag{5.1}$$

We start by setting $n_{\alpha,\gamma}=n^{-\alpha/(2\alpha+2\gamma+1)}$. Let $\boldsymbol{\lambda}_0=(\alpha_0,w_0)$ be a pair of real numbers s.t. $\hat{\boldsymbol{f}}_{\boldsymbol{\lambda}_0}$ is minimax over the Sobolev ellipsoid $\mathcal{F}(\alpha_0,R_0)$ and let p_π be the density of π: $p_\pi(\alpha,w)=e^{-\alpha}n_{\alpha,\gamma}p_w(wn_{\alpha,\gamma})$, where p_w is a probability density function supported by $(0,\infty)$ such that $\int up_w(u)\,du=1$. Let τ be a non-negative number such that $\tau\le\min(1,\alpha_0/(2\log w_0))$ and choose p_0 as a translation and dilatation of π, concentrating on $\boldsymbol{\lambda}_0$ when $\tau\to 0$: $p_0(d\boldsymbol{\lambda})=p_\pi\Big(\frac{\boldsymbol{\lambda}-\boldsymbol{\lambda}_0}{\tau}\Big)\frac{d\boldsymbol{\lambda}}{\tau^2}$. Let J_0 be a positive integer smaller than J the precise value of which will be given later. As an immediate consequence of (5.1) we get:

$$\begin{aligned}\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}}-\boldsymbol{f}\|_n^2\big]\le&\sum_{j=1}^{J_0}\Big(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,p_0(d\boldsymbol{\lambda})+\frac{\beta_j}{n}\,\mathcal{K}(p_0,\pi)\Big)\\&+\sum_{j=J_0+1}^{J}\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,\pi(d\boldsymbol{\lambda}).\end{aligned}$$

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

$$\sum_{j=1}^{J_0}\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,p_0(d\boldsymbol{\lambda})\le r_{\boldsymbol{\lambda}_0}+4\tau\Big(R_0+\frac{1}{n}\sum_{i=1}^{T_{J_0+1}}\sigma_i^2\Big)\Big(\frac{n^{1/2}}{w_0}+\frac{\log w_0}{\alpha_0}\Big),$$
$$\mathcal{K}(p_0,\pi)\le\alpha_0+3\log\big(\frac{2+w_0}{\tau}\big).$$

One readily checks $\sum_{i=1}^{T_{J_0+1}}\sigma_i^2\le C\sigma_*^2\sum_{i=1}^{T_{J_0+1}}i^{2\gamma}\le C\sigma_*^2T_{J_0+1}^{2\gamma+1}\le C\sigma_*^2n^{2\gamma+1}$. Furthermore, using the definition of weakly geometrically increasing groups, we get $\frac{1}{2}\nu_n(1+\rho_n)^j\le T_{j+1}\le\nu_n(1+\rho_n)^j$. This implies that

$$\sum_{i=1}^{J_0}\beta_j^2\le C\sigma_*^2\sum_{i=1}^{J_0}T_{j+1}^{2\gamma}\le C\sigma_*^2J_0T_{J_0+1}^{2\gamma}\le C\sigma_*^2J_0T_{J_0}^{2\gamma}.$$

Let J_0 be chosen such that $T_{J_0}^{2\gamma}\le n^{(2\gamma+1/2)/(2\alpha_0+2\gamma+1)}<T_{J_0+1}^{2\gamma}$. The condition $\log n=o(\log\nu_n)$ implies that $J_0T_{J_0}^{2\gamma}=o(n^{(2\gamma+1)/(2\alpha_0+2\gamma+1)})$. Therefore, setting $\tau=\frac{\alpha_0}{n^2+\alpha_0+2\log w_0}$, we come up with

$$\sum_{j=1}^{J_0}\Big(\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,p_0(d\boldsymbol{\lambda})+\frac{\beta_j}{n}\,\mathcal{K}(p_0,\pi)\Big)=o(n^{-2\alpha_0/(2\alpha_0+2\gamma+1)}),$$

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

$$\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{GEWA}}-\boldsymbol{f}\|_n^2\big]\le r_{\boldsymbol{\lambda}_0}(1+o(1))+\sum_{j=J_0+1}^{J}\int_\Lambda \mathbb{E}\|\hat{\boldsymbol{f}}^j_\lambda-\boldsymbol{f}^j\|_n^2\,\pi(d\boldsymbol{\lambda}).$$

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

References

[1] Dalalyan, A. S. and Salmon, J. (2012). Sharp Oracle Inequalities for Aggregation of Affine Estimators. *Ann. Statist. (under revision).*