Competing against the best nearest neighbor filter in regression

Arnak S. Dalalyan and Joseph Salmon

[1] Université Paris Est, Ecole des Ponts ParisTech,
77455 Marne-la-Vallée Cedex 2, France,
dalalyan@imagine.enpc.fr
[2] Duke University, ECE Departement,
7PO Box 90291 Durham, NC 27708, USA,
joseph.salmon@duke.edu

Abstract. Designing statistical procedures that are provably almost as accurate as the best one in a given family is one of central topics in statistics and learning theory. Oracle inequalities offer then a convenient theoretical framework for evaluating different strategies, which can be roughly classified into two classes: selection and aggregation strategies. The ultimate goal is to design strategies satisfying oracle inequalities with leading constant one and rate-optimal residual term. In many recent papers, this problem is addressed in the case where the aim is to beat the best procedure from a given family of linear smoothers. However, the theory developed so far either does not cover the important case of nearest-neighbor smoothers or provides a sub-optimal oracle inequality with a leading constant considerably larger than one. In this paper, we prove a new oracle inequality with leading constant one that is valid under a general assumption on linear smoothers allowing, for instance, to compete against the best nearest-neighbor filters.

Keywords: adaptive smoothing, nonparametric regression, supervised learning

1 Introduction

Linear procedures are omnipresent in machine learning. Sliding windows estimators, nearest neighbor smoothers, support vector machines with L^2 loss, etc., are popular examples of learning procedures obtained from the data vector by a linear transform. However, the performance of these procedures is, in general, severely affected by the choice of various tuning parameters. A typical example is presented in Figure 1: among the three linear estimators of a regression function, the two up-most estimators perform very poorly while the third one leads to an almost perfect recovery. The goal of the present paper is to propose a strategy which is able to estimate a regression function almost as well as the best linear procedure in a given family. Such a family may be obtained by considering, for instance, different values for the number of neighbors in nearest neighbor smoothing. It is also possible to make vary the metric in which the proximity is measured.

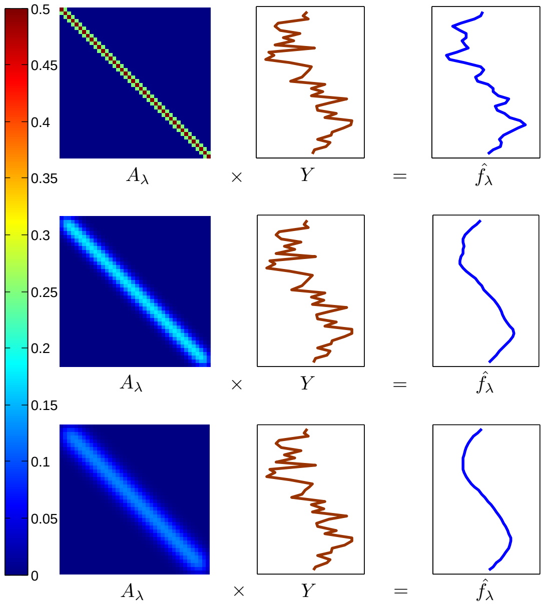


Fig. 1. The effect of the smoothing matrix A_λ on the resulting estimator. In this example, the true signal is the sine function over $[-\pi, \pi]$ and the three matrices represented in the leftmost column are some powers of one convolution matrix. Large powers correspond to stronger smoothing. It is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

We will mainly focus on the theoretical guarantees expressed in terms of oracle inequalities for the expected squared loss. Interestingly, despite the fact that several recent papers [1,4,18,11] discuss the paradigm of competing against the best linear procedure from a given family, none of them provide oracle inequalities with leading constant equal to one. Furthermore, most existing results involve some constants depending on different parameters of the setup. In contrast, the oracle inequality that we prove herein is with leading constant one and admits a very simple formulation. It is established for a suitably symmetrized version of the exponentially weighted aggregate [16,8,14] with arbitrary prior (see Figure 2) and a temperature parameter which is not too small. The result is completely non-asymptotic and leads to asymptotically optimal residual term in the case where the sample size, as well as the cardinality of the family of competitors, tends to infinity.

More precisely, let $\mathbf{f}$ be an unknown function defined on some set $\mathcal{X}$ (called feature space) and taking values in $\mathbb{R}$. To fix the ideas, assume that $\mathcal{X}$ is equipped with a metric d. We consider the setting where only noisy evaluations of the function $\mathbf{f}$ at n points $x_1, \ldots, x_n$ of $\mathcal{X}$ are available. The observations are then $\mathcal{D} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$. We are interested here in recovering the true values $\mathbf{f}(x_i)$ for $i = 1, \ldots, n$ based on the data $\mathcal{D}$. In this context, if we assume that $\mathbf{f}$ is smooth in some sense, a common estimator of $\mathbf{f}(x_i)$ is given by the k-nearest neighbor (kNN) filter $\hat{\mathbf{f}}_{k,d}(x_i)$. In order to define it, let us denote by $j_{i,0}, j_{i,1}, \ldots, j_{i,(n-1)}$ a permutation of $\{1,\ldots,n\}$ that leads to a rearrangement of the design points x_j from the closest to x_i to the farthest, *i.e.*, $0 = d(x_i, x_{j_{i,0}}) \leq d(x_i, x_{j_{i,1}}) \leq \ldots \leq d(x_i, x_{j_{i,(n-1)}})$. The kNN smoothing filter is then defined by

$$\hat{\mathbf{f}}_{k,d}(x_i) = \frac{1}{k}\sum_{m=0}^{k-1} Y_{j_{i,m}}. \tag{1}$$

In most applications, one can define different metrics d on the feature space and obtain different estimators of $\mathbf{f}(x_i)$ with very different statistical properties. The choice of the parameter k and the metric d that leads to the smallest possible estimation error is an important problem both from practical and theoretical viewpoints. A natural question arises: assume that we are given several metrics $d_1,\ldots,d_L$ on the feature space, is it possible to design a statistical procedure that estimates each of $\mathbf{f}(x_i)$ nearly as well as the best kNN-filter from the family $\{\hat{\mathbf{f}}_{k,d_\ell} : k = 1,\ldots,n; \ell = 1,\ldots,L\}$? We show that the answer to this question is affirmative, but there is a price to pay for not knowing the optimal metric and the optimal value of k. In the present work, we address this issues by aggregating the estimators $\hat{\mathbf{f}}_{k,d}$ over the set of all possible values of k and the metric d. Our results imply that the price to pay for not knowing the best values of these parameters is of the order $\log(L(n-1))/n$.

Note that the estimator (1) can be written as $\hat{\mathbf{f}}_{k,d}(x_i) = \sum_{j=1}^{n} a_{ij} Y_j$, with a_{ij} being equal to $1/k$ if $j \in \{j_{i,0},\ldots,j_{i,(k-1)}\}$ and 0 otherwise. Thus, the weights a_{ij} depend exclusively on the the features $x_1,\ldots,x_n$ and not on the Y_is. Therefore, the kNN filter is a particular instance of linear estimators defined by

$$\hat{\boldsymbol{f}} = \begin{bmatrix} \hat{\mathbf{f}}(x_1) \\ \vdots \\ \hat{\mathbf{f}}(x_n) \end{bmatrix} = A\boldsymbol{Y},$$

where A is a $n \times n$ weight matrix and $\boldsymbol{Y} = [Y_1,\ldots,Y_n]^\top$ is the vector of observed responses. The main results of this paper hold for this general class of estimators under some condition on the weight matrix A. This condition is satisfied for a kNN estimator whatever the metric d and the parameter k are.

From the perspective of learning theory, oracle inequalities constitute a valuable theoretical tool for assessing the performance of procedures elaborated in the context of agnostic learning introduced by [20], see also [19] for a recent contribution to the subject. Note also that the problem of competing against the best procedure in a given family has been extensively studied in the context of online learning and prediction with expert advice [21,9,10,5]. A remarkable connection between the results on online learning and the statistical oracle inequalities has been recently established by [17]. The case of linear estimators has been studied by [24,26,12] for projection matrices A and by [26,12] for diagonal matrices A. However, these result do not cover several important classes of linear estimators including the kNN filter.

We should mention that the numerical experiments we have carried out on a number of synthetic datasets have shown that the symmetrized exponentially

```
Input: data vector $\boldsymbol{Y} \in \mathbb{R}^n$, $n \times n$ noise covariance matrix $\boldsymbol{\Sigma}$ and a family of linear
    smoothers $\{\hat{\boldsymbol{f}}_\lambda = \mathsf{A}_\lambda \boldsymbol{Y}; \lambda \in \Lambda\}$.
Output: estimator $\hat{\boldsymbol{f}}_{\text{SEWA}}$ of the true function $\boldsymbol{f}$.
Parameter: prior probability distribution $\pi$ on $\Lambda$, temperature parameter $\beta > 0$.
Strategy:
  1. For every $\lambda$, compute the risk estimate $\hat{r}_\lambda^{\text{unb}} = \big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}(\Sigma\mathsf{A}_\lambda) - \frac{1}{n}\operatorname{Tr}[\Sigma]$.
  2. Define the probability distribution $\hat{\pi}(d\lambda) = \theta(\lambda)\pi(d\lambda)$ with $\theta(\lambda) \propto \exp(-n\hat{r}_\lambda^{\text{unb}}/\beta)$.
  3. For every $\lambda$, build the symmetrized linear smoothers $\tilde{\boldsymbol{f}}_\lambda = (\mathsf{A}_\lambda + \mathsf{A}_\lambda^\top - \mathsf{A}_\lambda^\top \mathsf{A}_\lambda)\boldsymbol{Y}$.
  4. Average out the symmetrized smoothers w.r.t. posterior $\hat{\pi}$: $\hat{\boldsymbol{f}}_{\text{SEWA}} = \int_\Lambda \tilde{\boldsymbol{f}}_\lambda \hat{\pi}(d\lambda)$.
```

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

weighted aggregate performs as predicted by our theoretical result. Interestingly, these experiments show also that the standard (non-symmetrized) exponentially weighted aggregate is not worse than the symmetrized one. However, we are not able so far to provide theoretical guarantees for the non-symmetrized strategy.

Outline. The rest of the paper is organized as follows. We introduce the main notation along with a short background on oracle inequalities and on linear filtering in Section 2. The main contribution of the paper, a sharp oracle inequality for the symmetrized exponentially weighted aggregate, is stated in Section 3, while Section 4 contains some numerical results. Section 5 summarizes the content of the paper and provides some perspectives. The proofs are postponed to the Appendix.

2 Notation and background

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

$$y_i = f_i + \xi_i, \quad \text{for} \;\; i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is a centered Gaussian random vector, $f_i = \mathbf{f}(x_i)$ where $\mathbf{f}$ is an unknown function $\mathcal{X} \to \mathbb{R}$ 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 $\boldsymbol{f} = (f_1, \ldots, f_n)$, often referred to as *signal*, based on the data $y_1, \ldots, y_n$. In our work the noise covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top]$ is assumed to be finite and known, with a possible extension to the case of estimated covariance matrix discussed in Remark 5. We measure the performance of an estimator $\hat{\boldsymbol{f}}$ by its expected empirical quadratic loss: $r = \mathbb{E}(\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2)$, where $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f_i - \hat{f}_i)^2$. We also denote by $\langle \cdot | \cdot \rangle_n$ the corresponding empirical inner product. For any matrix B, $|||\mathsf{B}|||$ stands for the spectral norm of B, *i.e.*, its largest singular value.

2.1 Oracle inequalities

In this subsection we describe the paradigm of selection/aggregation of estimators in a data-driven manner from a given family of estimators. The task of aggregation consists in estimating $\boldsymbol{f}$ by a suitable combination of the elements of a family of *constituent estimators* $\mathcal{F}_\Lambda = (\hat{\boldsymbol{f}}_\lambda)_{\lambda\in\Lambda} \in \mathbb{R}^n$, while the task of selection is just to choose a data-dependent value $\hat{\lambda}$ of λ for which the estimator $\hat{\boldsymbol{f}}_{\hat{\lambda}}$ is close to $\boldsymbol{f}$. The target objective of the selection/aggregation is to build an estimator $\hat{\boldsymbol{f}}_{\text{select}}/\hat{\boldsymbol{f}}_{\text{aggr}}$ that mimics the performance of the best constituent estimator, called *oracle* (because of its dependence on the unknown function f). In what follows, we assume that Λ is a measurable subset of $\mathbb{R}^M$, for some $M \in \mathbb{N}$.

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

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\text{aggr}} - \boldsymbol{f}\|_n^2 \leq C_n \inf_{\lambda\in\Lambda} \Big(\mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\Big) + R_n, \tag{3}$$

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

2.2 Nearest neighbor filtering

When the unknown function $\mathbf{f}$ is smooth or can be well approximated by a smooth function, it is reasonable to estimate it by computing the moving averages or k-Nearest Neighbor (kNN) filters, see *e.g.* [15]. More precisely, let us fix an index i and consider the problem of estimating the value f_i of $\mathbf{f}$ at x_i. Let $x_{j_1}, \ldots, x_{j_k}$ be the set of k points from $\{x_1, \ldots, x_n\}$ which are at smallest distance (in some metric) from x_i. The idea of kNN filtering is to estimate the unknown value f_i by taking the average of k values Y_{j_ℓ}, $\ell = 1, \ldots, k$. This approach is particularly popular, for instance, in stereo-vision for reconstructing 3D scenes from 3D point clouds.

A crucial point when estimating a function by kNN-filtering is the choice of the tuning parameter k. This parameter allows the user to control the trade-off between the bias and the variance of estimation. If the value of k is too small, the resulting estimator is very oscillating, whereas large values of k lead to over-smoothed estimators. Many strategies for selecting k in a data driven manner have been proposed in the literature [25,23,22,18,1]. However, to the best of our knowledge, none of these procedures is proved to satisfy a sharp oracle inequality in the sense made precise in the previous section. In the present work, we propose a strategy—for which a sharp oracle inequality is established—based on data-driven aggregation of kNN filters rather than on (data-driven) selection of the parameter k.

2.3 General linear smoothing

More generally, we will focus on *linear estimators* $\hat{\boldsymbol{f}}_\lambda$, *i.e.,* estimators that are linear 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, linear estimators can be defined by

$$\hat{\boldsymbol{f}}_\lambda = \mathsf{A}_\lambda \boldsymbol{Y}, \tag{4}$$

where the $n \times n$ real matrix A_λ is deterministic. This means that the entries of A_λ may depend on the points $x_1, \ldots, x_n$ but not on the data vector $\boldsymbol{Y}$. Let I_n denote the identity matrix of size $n \times n$. It is well-known that the risk of the estimator (4) is given by

$$\mathbb{E}[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2] = \|(\mathsf{A}_\lambda - \mathsf{I}_n)\boldsymbol{f}\|_n^2 + \frac{\operatorname{Tr}(\mathsf{A}_\lambda \Sigma \mathsf{A}_\lambda^\top)}{n} \tag{5}$$

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

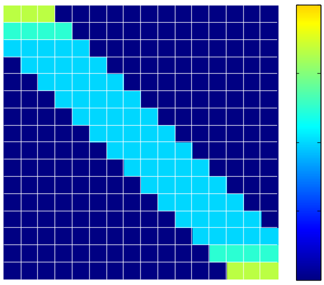
$$\hat{r}_\lambda^{\text{unb}} = \big\|\boldsymbol{Y} - \hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}(\Sigma \mathsf{A}_\lambda) - \frac{1}{n}\operatorname{Tr}[\Sigma] \tag{6}$$

is an unbiased estimator of $r_\lambda = \mathbb{E}[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2]$. In order to get a sharp oracle inequality with a simple residual term, we will need the following assumption.

[**C($\boldsymbol{\lambda}$)**] The matrix A_λ satisfies $\operatorname{Tr}(\Sigma \mathsf{A}_\lambda) \leq \operatorname{Tr}(\Sigma \mathsf{A}_\lambda^\top \mathsf{A}_\lambda)$.

Let us present now several classes of widely used linear estimators for which this condition is satisfied.

1. The simplest class of matrices A_λ for which condition **C**(λ) holds true are orthogonal projections. Indeed, if A_λ is a projection matrix, it satisfies $\mathsf{A}_\lambda^\top \mathsf{A}_\lambda = \mathsf{A}_\lambda$ and, therefore, $\operatorname{Tr}(\Sigma \mathsf{A}_\lambda) = \operatorname{Tr}(\Sigma \mathsf{A}_\lambda^\top \mathsf{A}_\lambda)$.
2. If the matrix Σ is diagonal, then a sufficient condition for **C**(λ) is $a_{ii} \leq \sum_{j=1}^n a_{ji}^2$. Consequently, for the matrices having only zeros on the main diagonal **C**(λ) holds true. 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.
3. Under a little bit more stringent assumption of homoscedasticity, *i.e.*, when $\Sigma = \sigma^2 \mathsf{I}_n$, 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 $\operatorname{Tr}(\mathsf{A}_\lambda) = \operatorname{Tr}(\mathsf{A}_\lambda^\top \mathsf{A}_\lambda)$ and **C**(λ) is fulfilled. A notable example of linear estimators that satisfy this condition are Nadaraya-Watson estimators with rectangular kernel and nearest neighbor filters. Below is a visual illustration of a matrix defining a Nadaraya-Watson estimator:



3 Main result

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

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

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

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

It is convenient and customary to use the terminology of Bayesian statistics: the measure π is called *prior*, the measure $\hat{\pi}$ is called *posterior* and the aggregate $\hat{\boldsymbol{f}}_{\text{EWA}}$ is then the *posterior mean*. The parameter β will be referred to as the *temperature parameter*. In the framework of aggregating statistical procedures, the use of such an aggregate can be traced back to [16].

The density $\theta(\cdot)$ assigns weights to the estimators according to their performance, measured in terms of the risk estimate $\hat{r}_\lambda$. The temperature parameter reflects the confidence we have in this criterion: if $\beta \approx 0$ the posterior concentrates on the estimators achieving the smallest value for $\hat{r}_\lambda$, whereas if $\beta \to +\infty$ then the posterior approaches to the prior π, and the data do not modify our confidence in the estimators. It should also be noted that averaging w.r.t. the posterior $\hat{\pi}$ is not the only way of constructing an estimator of $\boldsymbol{f}$ based on $\hat{\pi}$; some alternative randomized estimators have been studied, for instance, in [2].

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 $p, p' \in \mathcal{P}_\Lambda$:

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

Theorem 1. *Let $\{\mathsf{A}_\lambda : \lambda \in \Lambda\}$ be any family of $n \times n$ matrices satisfying condition* $\mathbf{C}(\lambda)$ *on a set of π-measure one. Let $\hat{\boldsymbol{f}}_{\text{SEWA}}$ denote the symmetrized exponentially weighted aggregate, i.e. the exponentially weighted aggregate acting on symmetrized estimators $\tilde{\boldsymbol{f}}_\lambda = (\mathsf{A}_\lambda + \mathsf{A}_\lambda^\top - \mathsf{A}_\lambda^\top \mathsf{A}_\lambda)\boldsymbol{Y}$ with the weights (7) defined via the risk estimate $\hat{r}_\lambda^{\text{unb}}$. Then, for every $\beta \geq 4|||\Sigma|||$, it holds that*

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

A first observation that one can make is that, in the particular case of a finite collection of projection estimators (*i.e.*, $\mathsf{A}_\lambda = \mathsf{A}_\lambda^\top = \mathsf{A}_\lambda^2$ for every λ) this result reduces to Corollary 6 in [24]. Furthermore, Theorem 1 handles the general noise covariances while [24] deals only with i.i.d. Gaussian noise.

Note also that the result of Theorem 1 applies to the estimator $\hat{\boldsymbol{f}}_{\text{EWA}}$ that uses the full knowledge of the covariance matrix Σ. Indeed, even if for the choice of β only an upper bound on the spectral norm of Σ is required, the entire matrix Σ enters in the definition of the unbiased risk $\hat{r}_\lambda^{\text{unb}}$ that is used for defining $\hat{\boldsymbol{f}}_{\text{SEWA}}$. We will discuss in Remark 5 some extensions of the proposed methodology to the case of unknown Σ.

Remark 1. We decided in this paper to focus on the case of Gaussian errors, in order to put the emphasis on the possibility of efficiently aggregating broad families of *linear estimators* without spending time and space on other technical aspects. The result stated in this section can be generalized to some other noise distributions by following the approach developed in [13].

Remark 2. We prove a result that is stronger than the one stated in Theorem 1. In particular, it holds for any matrices A_λ and boils down to the elegant inequality stated in Theorem 1 when condition $\mathbf{C}(\lambda)$ is π-a.e. satisfied. The precise form of this more general result is the following. Let $\hat{\boldsymbol{f}}_{\text{SEWA}}$ denote the aggregate defined in Figure 2. Then, for every $\beta \geq 4|||\Sigma|||$, the risk $\mathbb{E}\big[\|\hat{\boldsymbol{f}}_{\text{SEWA}} - \boldsymbol{f}\|_n^2\big]$ of $\hat{\boldsymbol{f}}_{\text{SEWA}}$ is bounded from above by

$$\inf_{p \in \mathcal{P}_\Lambda} \left\{ \int_\Lambda \mathbb{E}\big[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2\big] p(d\lambda) + \frac{\beta}{n}\, \mathcal{K}(p, \pi) \right\} + R_n \tag{9}$$

with the residual term

$$R_n = \frac{\beta}{n} \log\left[\int_\Lambda e^{\frac{2}{\beta}\operatorname{Tr}[\Sigma(\mathsf{A}_\lambda - \mathsf{A}_\lambda^\top \mathsf{A}_\lambda)]} \pi(d\lambda) \right].$$

Remark 3. Using the previous remark, one can also get the risk bound (9), when condition $\mathbf{C}(\lambda)$ is only approximately satisfied. More precisely, if condition $\mathbf{C}(\lambda)$ is replaced by :

$[\mathbf{C}(\boldsymbol{\lambda}, \boldsymbol{\varepsilon})]$ The matrix A_λ satisfies $\operatorname{Tr}(\Sigma \mathsf{A}_\lambda) \leq \operatorname{Tr}(\Sigma \mathsf{A}_\lambda^\top \mathsf{A}_\lambda) + \varepsilon$,

then the residual term R_n in Inequality (9) simply becomes $\frac{2\varepsilon}{n}$.

In order to demonstrate that Theorem 1 can be reformulated in terms of an OI as defined by (3), let us consider the case when the prior π is discrete. That is, we assume that $\pi(\Lambda_0) = 1$ for a countable set $\Lambda_0 \subset \Lambda$. Without loss of generality, we assume that $\Lambda_0 = \mathbb{N}$. Then, the following result holds true.

Proposition 1. *If π is supported by $\mathbb{N}$ and condition $\mathbf{C}(\lambda)$ is satisfied for every $\lambda \in \mathbb{N}$, then the aggregate $\hat{\boldsymbol{f}}_{\text{SEWA}}$ satisfies the inequality*

$$\mathbb{E}[\|\hat{\boldsymbol{f}}_{\text{SEWA}} - \boldsymbol{f}\|_n^2] \leq \inf_{\lambda : \pi_\lambda > 0} \left\{ \mathbb{E}\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2 + \frac{\beta \log(1/\pi_\lambda)}{n} \right\} \tag{10}$$

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

Proof. It suffices to apply Theorem 1 and to bound the right hand side from above by the minimum over all Dirac measures $p = \delta_\lambda$ with λ such that $\pi_\lambda > 0$.

This inequality can be compared to Corollary 2 in Section 4.3 of [4]. Our inequality has the advantage of being sharp, *i.e.*, having factor one both in front of the expectation of the LHS of (10) and in front of the inf of the RHS. To the best of our knowledge, there is no other result in the literature providing such a sharp OI for linear estimators which are not of projection type. In particular, in [4] the risk in the LHS of the OI is multiplied by a constant which is smaller than one and depends on different parameters of the problem. It should be noted, however, that we consider the noise covariance matrix as known, whereas [4] estimates the noise covariance along with the regression function.

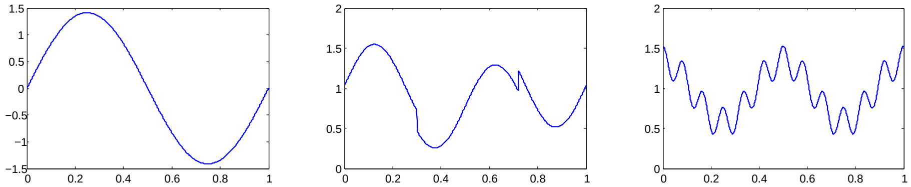


Fig. 3. Three test signals used in experimental evaluation. From left to right : the sine function, HeaviSine and Wave functions [7].

Remark 4. A particular case of Proposition 1 is the situation where π is the uniform probability over a finite set of cardinality M. In such a situation, the remainder term in (10) becomes of the form $(\beta \log M)/n$. The rate $(\log M)/n$ of the remainder term in the OI has been proven [28] unavoidable in the context of aggregating data-independent estimators. By similar arguments, it is possible to prove that this rate is optimal in the case of aggregating linear smoothers as well.

Remark 5. The symmetrized exponentially weighted aggregate $\hat{\boldsymbol{f}}_{\rm SEWA}$ is easily extended to handle the more realistic situation where an unbiased estimate $\widehat{\Sigma}$, independent of $\boldsymbol{Y}$, of the covariance matrix Σ is available. Simply replace Σ by $\widehat{\Sigma}$ in the definition of the unbiased risk estimate (6). When the matrices A_λ satisfy π-a.e. condition $\mathbf{C}(\lambda)$, it is easy to see that the claim of Theorem 1 remains valid. Of course, the condition $\beta \geq 4|||\Sigma|||$ should be replaced by $\beta \geq 4|||\widehat{\Sigma}|||$ and β should be replaced by $\mathbb{E}[\beta]$ in the right hand side of the oracle inequality.

4 Numerical experiments

We have implemented the symmetrized exponentially weighted aggregate (SEWA) in Matlab in the case of combining kNN filters with varying values of k. Along with SEWA we have also implemented the classical exponentially weighted aggregate (EWA) as defined for instance in [24,14] and the empirical risk minimization (ERM) algorithm, the latter consisting in choosing the value of k minimizing the unbiased estimator of the risk (6). Following the philosophy of reproducible research, a toolbox containing the code we used for getting the results reported in this section will be made available by the date of the conference at the authors' home pages.

In our experiments, we compared the aforementioned three strategies, ERM, EWA and SEWA, on three common 1D signals depicted in Figure 3. Each signal has been beforehand normalized to have an L^2 norm equal to one. We have chosen several sample sizes $n \in \{30, 50, 100\}$ and noise levels $\sigma^2 \in \{0.2, 0.5, 1, 1.5, 2\}$ and randomly generated the data vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ by the formula $Y_i = \mathbf{f}(i/n) + \epsilon_i$, where $(\epsilon_1, \ldots, \epsilon_n)$ is a Gaussian random vector $\mathcal{N}(0, \sigma^2 \mathrm{I}_n)$. We then computed the three estimators ERM, EWA and SEWA and repeated the experiment 10^4 times. As preliminary estimators we used the kNN filters with $k \in \{1, \ldots, [n/2]\}$. The prior was chosen to be uniform and the temperature parameter is the one suggested by the theory: $\beta = 4\sigma^2$. The medians and the inter-quartile ranges of the errors[3] $\|\hat{\boldsymbol{f}}_\bullet - \boldsymbol{f}\|_2$ are summarized in Tables 1, 2 and 3 below.

	$n = 30$			$n = 50$			$n = 100$		
	ERM	EWA	SEWA	ERM	EWA	SEWA	ERM	EWA	SEWA
$\sigma^2 = 0.2$	1.4397	1.3906	1.3469	1.5290	1.4685	1.4663	1.6768	1.5984	1.6471
	(0.40)	(0.40)	(0.35)	(0.40)	(0.39)	(0.38)	(0.40)	(0.38)	(0.38)
$\sigma^2 = 0.5$	2.0301	1.8806	1.7861	2.1395	2.0800	2.0086	2.3634	2.2661	2.2786
	(0.57)	(0.48)	(0.53)	(0.65)	(0.59)	(0.56)	(0.63)	(0.62)	(0.59)
$\sigma^2 = 1$	2.4966	2.2161	2.1933	2.8026	2.5501	2.4487	3.0561	2.9287	2.8590
	(0.69)	(0.61)	(0.71)	(0.81)	(0.67)	(0.74)	(0.93)	(0.83)	(0.81)
$\sigma^2 = 1.5$	2.7930	2.4966	2.5046	3.1521	2.8125	2.7660	3.5679	3.3088	3.2167
	(0.94)	(0.83)	(0.96)	(0.94)	(0.84)	(0.95)	(1.09)	(0.92)	(0.96)
$\sigma^2 = 2$	3.0113	2.7180	2.7793	3.3930	3.0757	3.0413	3.9748	3.5854	3.4970
	(1.08)	(1.02)	(1.17)	(1.10)	(0.93)	(1.06)	(1.19)	(1.00)	(1.09)

Table 1. Sine function: the values of the median error and the inter-quartile range (in parentheses) over 10^4 trials are reported.

A first observation is that the aggregation strategies, EWA and SEWA, are always better than the selection strategy ERM. This is essentially explained by a relative lack of stability of selection strategies thoroughly discussed in [6]. A second observation is that there is no clear winner among the aggregation strategies EWA and SEWA. Both of them are quite accurate with very little difference in the error of estimation. This raises the following question: is it possible to prove a sharp oracle inequality for the standard EWA without applying the symmetrization trick? To date, we are unable to answer this question.

It is important to stress that the medians reported in Tables 1-3 are those of estimation errors without normalization by the sample size n. Therefore, it is quite natural that these errors increase with n (more and more parameters are estimated). It is however clear from the reported results that the non-normalized

[3] In this expression the norm is the classical Euclidean one and $\hat{\boldsymbol{f}}_\bullet$ is either one of the estimators ERM, EWA or SEWA.

	$n = 30$			$n = 50$			$n = 100$		
	ERM	EWA	SEWA	ERM	EWA	SEWA	ERM	EWA	SEWA
$\sigma^2 = 0.2$	1.6552	1.5906	1.5708	1.8157	1.7274	1.7306	2.0170	1.9359	1.9921
	(0.37)	(0.36)	(0.35)	(0.37)	(0.39)	(0.38)	(0.39)	(0.37)	(0.39)
$\sigma^2 = 0.5$	2.2783	2.1604	2.0845	2.4834	2.3370	2.2589	2.7984	2.6620	2.6611
	(0.55)	(0.57)	(0.58)	(0.59)	(0.54)	(0.57)	(0.62)	(0.59)	(0.59)
$\sigma^2 = 1$	2.9039	2.7275	2.6416	3.1558	2.9446	2.8783	3.5533	3.3284	3.2715
	(0.82)	(0.81)	(0.85)	(0.85)	(0.83)	(0.84)	(0.86)	(0.80)	(0.82)
$\sigma^2 = 1.5$	3.3554	3.1526	3.0878	3.5758	3.3576	3.2583	4.0708	3.7886	3.7106
	(1.08)	(0.99)	(0.97)	(1.02)	(0.95)	(1.00)	(1.05)	(0.97)	(1.00)
$\sigma^2 = 2$	3.7266	3.4729	3.4443	4.0147	3.7368	3.6694	4.4888	4.1560	4.0723
	(1.34)	(1.19)	(1.22)	(1.30)	(1.23)	(1.24)	(1.24)	(1.13)	(1.16)

Table 2. HeaviSine function [7]: the values of the median error and the inter-quartile range (in parentheses) over 10^4 trials are reported.

	$n = 30$			$n = 50$			$n = 100$		
	ERM	EWA	SEWA	ERM	EWA	SEWA	ERM	EWA	SEWA
$\sigma^2 = 0.2$	1.4340	1.3814	1.3724	1.5887	1.5725	1.5580	1.9720	1.8696	1.8612
	(0.37)	(0.29)	(0.30)	(0.41)	(0.33)	(0.33)	(0.34)	(0.30)	(0.33)
$\sigma^2 = 0.5$	1.8300	1.6868	1.7159	2.1004	1.9571	1.9608	2.4045	2.3730	2.3462
	(0.45)	(0.41)	(0.47)	(0.53)	(0.41)	(0.47)	(0.67)	(0.49)	(0.52)
$\sigma^2 = 1$	2.1727	2.0073	2.0976	2.4719	2.2784	2.3351	2.9898	2.7755	2.7716
	(0.74)	(0.65)	(0.73)	(0.69)	(0.60)	(0.68)	(0.77)	(0.58)	(0.66)
$\sigma^2 = 1.5$	2.4395	2.2637	2.4013	2.7554	2.5266	2.6331	3.2993	3.0282	3.0761
	(1.00)	(0.84)	(0.94)	(0.93)	(0.77)	(0.89)	(0.88)	(0.72)	(0.83)
$\sigma^2 = 2$	2.6845	2.5068	2.6809	2.9950	2.7495	2.8961	3.5428	3.2290	3.3133
	(1.23)	(1.01)	(1.12)	(1.15)	(0.94)	(1.06)	(1.05)	(0.86)	(0.99)

Table 3. Wave function: the values of the median error and the inter-quartile range (in parentheses) over 10^4 trials are reported.

accuracy increases very slowly when n increases. This is in agreement with our theoretical result stating that the error increases at most logarithmically.

5 Conclusion and outlook

We have suggested a new strategy for aggregating linear smoothers in order to denoise a signal corrupted by an additive Gaussian noise. We proved a sharp oracle inequality for the proposed strategy, termed SEWA for symmetrized exponentially weighted aggregation. A few experimental results are also reported that allow to illustrate our theoretical result and to quantify the advantage of aggregation as compared to selection.

The SEWA results may have profitable application to classification and pattern recognition. As proved in [3], fast rates in classification can be obtained by plugging-in efficient regression estimators. We are experimenting with the use of a procedure analogous to SEWA to perform binary classification. The results, to date, have been as encouraging as in the regression case.

Another possible application of SEWA is for more accurate estimation of autoregressive models in time series. Picking the order of the autoregressive scheme is similar to estimating the parameter k in kNN filtering. Exponentially weighted aggregate should carry over to this area and may provide increased prediction accuracy.

Acknowledgments

This work has been partially supported by ANR grant Parcimonie.

References

1. S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *NIPS*, pages 46–54, 2009.
2. J-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.
3. J-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.
4. Y. Baraud, Ch. Giraud, and S. Huet. Estimator selection in the gaussian setting. *submitted*, 2010.
5. Sh. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.
6. L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.
7. J.B. Buckheit and D.L. Donoho. Wavelab and reproducible research. In *Wavelets and Statistics*, number 103 in Lect. Notes Statist., pages 55–81. Springer-Verlag, New York, 1995.
8. O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.
9. N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.
10. N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66:321–352, March 2007.
11. P-A Cornillon, N. Hengartner, and E. Matzner-Løber. Recursive bias estimation for multivariate regression smoothers. *submitted*, 2009.
12. A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *technical report*, arXiv:1104.3969v2 [math.ST], 2011.
13. A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.
14. A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT*, 2009.
15. L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.
16. E. I. George. Minimax multiple shrinkage estimation. *Ann. Statist.*, 14(1):188–205, 1986.

17. S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *submitted*, 2011.
18. A. Goldenshluger and O. V. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.
19. A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008.
20. M. J. Kearns, R. E. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.
21. J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Computational learning theory (EuroCOLT)*, volume 1572 of *Lecture Notes in Comput. Sci.*, pages 153–167. Springer, Berlin, 1999.
22. J. Lafferty and L. Wasserman. Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63, 2008.
23. O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.
24. G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8):3396–3410, 2006.
25. K-C. Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4):1352–1377, 1985.
26. J. Salmon and A. S. Dalalyan. Optimal aggregation of affine estimators. In *COLT*, 2011.
27. C. M. Stein. Estimation of the mean of a multivariate distribution. In *Proc. Prague Symp. Asymptotic Statist.*, 1973.
28. A. B. Tsybakov. Optimal rates of aggregation. In *COLT*, pages 303–313, 2003.

A Proof of Theorem 1

Stein's lemma

The proof of our main result relies on the well-known Stein lemma [27] providing an unbiased risk estimate for any estimator that depends sufficiently smoothly on the data vector $\boldsymbol{Y}$. For the convenience of the reader, we recall Stein's lemma in the case of heteroscedastic Gaussian regression.

Lemma 1. *Let $\boldsymbol{Y}$ be random vector drawn form the Gaussian distribution $\mathcal{N}_n(\boldsymbol{f},\Sigma)$. If the estimator $\hat{\boldsymbol{f}}$ is a.e. differentiable in $\boldsymbol{Y}$ and the elements of the matrix $\nabla\cdot\hat{\boldsymbol{f}}^\top := (\partial_i\hat{\boldsymbol{f}}_j)$ have finite first moment, then $\hat{r}_\Sigma = \|\boldsymbol{Y}-\hat{\boldsymbol{f}}\|_n^2 + \frac{2}{n}\operatorname{Tr}[\Sigma(\nabla\cdot\hat{\boldsymbol{f}}^\top)] - \frac{1}{n}\operatorname{Tr}[\Sigma]$, is an unbiased estimate of r, i.e., $\mathbb{E}\hat{r}_\Sigma = r$. Moreover, if $\widehat{\Sigma}$ is an unbiased estimator of Σ such that $\boldsymbol{Y}$ and $\widehat{\Sigma}$ are independent, then*

$$\hat{r} = \|\boldsymbol{Y}-\hat{\boldsymbol{f}}\|_n^2 + \frac{2}{n}\operatorname{Tr}[\widehat{\Sigma}(\nabla\cdot\hat{\boldsymbol{f}}^\top)] - \frac{1}{n}\operatorname{Tr}[\widehat{\Sigma}], \tag{11}$$

is an unbiased estimator of the risk r as well.

We apply Stein's lemma to the estimator $\hat{\boldsymbol{f}}_\lambda = \mathsf{A}_\lambda\boldsymbol{Y}$, where A_λ is an $n\times n$ matrix. We get that $\hat{r}^{\mathrm{unb}}_{\lambda,\Sigma} = \big\|\boldsymbol{Y}-\hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}[\Sigma\mathsf{A}_\lambda] - \frac{1}{n}\operatorname{Tr}[\Sigma]$ is an unbiased estimator of the risk $r_\lambda = \mathbb{E}[\|\hat{\boldsymbol{f}}_\lambda - \boldsymbol{f}\|_n^2] = \|(\mathsf{A}_\lambda - \mathsf{I}_n)\boldsymbol{f}\|_n^2 + \frac{1}{n}\operatorname{Tr}[\mathsf{A}_\lambda\Sigma\mathsf{A}_\lambda^\top]$.
Furthermore, if $\widehat{\Sigma}$ is an unbiased estimator of Σ then $\hat{r}^{\mathrm{unb}}_\lambda = \big\|\boldsymbol{Y}-\hat{\boldsymbol{f}}_\lambda\big\|_n^2 + \frac{2}{n}\operatorname{Tr}[\widehat{\Sigma}\mathsf{A}_\lambda] - \frac{1}{n}\operatorname{Tr}[\widehat{\Sigma}]$ is also an unbiased estimator of r_λ.

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 proof for our main result.

Lemma 2. *Let assumptions of Lemma 1 be satisfied. Let* $\{\tilde{\boldsymbol{f}}_\lambda : \lambda \in \Lambda\}$ *be a family of estimators of* $\boldsymbol{f}$ *and* $\{\widetilde{r}_\lambda : \lambda \in \Lambda\}$ *a family of risk estimates such that the mapping* $\boldsymbol{Y} \mapsto (\tilde{\boldsymbol{f}}_\lambda, \widetilde{r}_\lambda)$ *is a.e. differentiable* $\forall \lambda \in \Lambda$*. Let* $\tilde{r}_\lambda^{\text{unb}}$ *be the unbiased risk estimate of* $\tilde{\boldsymbol{f}}_\lambda$ *given by Stein's lemma (cf. Eq. (11)).*

1. *For every* $\mu \in \mathcal{P}_\Lambda$ *and for any* $\beta > 0$*, the estimator* $\tilde{\boldsymbol{f}}_{\text{EWA}}$ *defined as the average of* $\tilde{\boldsymbol{f}}_\lambda$ *w.r.t. the probability measure* $\hat{\mu}(\boldsymbol{Y}, d\lambda) = \theta(\boldsymbol{Y}, \lambda)\mu(d\lambda)$ *with* $\theta(\boldsymbol{Y}, \lambda) \propto \exp\big\{-n\widetilde{r}_\lambda(\boldsymbol{Y})/\beta\big\}$ *admits*
$$\hat{r}_{\text{EWA}} = \int_\Lambda \Big(\tilde{r}_\lambda^{\text{unb}} - \|\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2 - \frac{2n}{\beta}\big\langle \nabla_{\boldsymbol{Y}} \widetilde{r}_\lambda | \widehat{\Sigma}(\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}})\big\rangle_n\Big)\hat{\mu}(d\lambda)$$
as unbiased estimator of the risk.
2. *If furthermore* $\widetilde{r}_\lambda \geq \tilde{r}_\lambda^{\text{unb}}$*,* $\forall \lambda \in \Lambda$ *and* $\int_\Lambda \big\langle \nabla \widetilde{r}_\lambda | \widehat{\Sigma}(\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}})\big\rangle_n \hat{\mu}(d\lambda) \geq -a \int_\Lambda \|\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2 \hat{\mu}(d\lambda)$ *for some random* $a > 0$ *independent of* $\boldsymbol{Y}$*, then for every* $\beta \geq 2na$ *it holds that*
$$\mathbb{E}[\|\tilde{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2] \leq \inf_{p \in \mathcal{P}_\Lambda} \Big\{ \int_\Lambda \mathbb{E}[\widetilde{r}_\lambda] p(d\lambda) + \frac{\mathbb{E}[\beta]\mathcal{K}(p, \mu)}{n} \Big\}.$$

Proof. According to the Lemma 1, the quantity
$$\hat{r}_{\text{EWA}} = \|\boldsymbol{Y} - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2 + \frac{2}{n}\operatorname{Tr}[\widehat{\Sigma}(\boldsymbol{\nabla} \cdot \tilde{\boldsymbol{f}}_{\text{EWA}}(\boldsymbol{Y})] - \frac{1}{n}\operatorname{Tr}[\widehat{\Sigma}] \tag{12}$$
is an unbiased estimate of the risk $r_n = \mathbb{E}(\|\tilde{\boldsymbol{f}}_{\text{EWA}} - \boldsymbol{f}\|_n^2)$. Using simple algebra, one checks that
$$\|\boldsymbol{Y} - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2 = \int_\Lambda \Big(\|\boldsymbol{Y} - \tilde{\boldsymbol{f}}_\lambda\|_n^2 - \|\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2\Big)\,\hat{\mu}(d\lambda). \tag{13}$$
By interchanging the integral and differential operators, we get the following relation: $\partial_{y_i}\hat{\boldsymbol{f}}_{\text{EWA},j} = \int_\Lambda \big\{\big(\partial_{y_j}\tilde{\boldsymbol{f}}_\lambda^j(\boldsymbol{Y})\big)\,\theta(\boldsymbol{Y}, \lambda) + \tilde{\boldsymbol{f}}_\lambda^j(\boldsymbol{Y})\,\big(\partial_{y_i}\theta(\boldsymbol{Y}, \lambda)\big)\big\}\,\mu(d\lambda)$. This equality, combined with Equations (12) and (13) implies that
$$\hat{r}_{\text{EWA}} = \int_\Lambda \big(\tilde{r}_\lambda^{\text{unb}} - \|\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2\big)\hat{\mu}(d\lambda) + \frac{2}{n}\int_\Lambda \operatorname{Tr}[\widehat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{f}}_\lambda \nabla_{\boldsymbol{Y}}\theta(\boldsymbol{Y}, \lambda)^\top]\,\mu(d\lambda).$$
Taking into account the fact that the differentiation and the integration can be interchanged, $\int_\Lambda \tilde{\boldsymbol{f}}_{\text{EWA}}\big(\nabla_{\boldsymbol{Y}}\theta(\boldsymbol{Y}, \lambda)\big)^\top\mu(d\lambda) = \tilde{\boldsymbol{f}}_{\text{EWA}}\nabla_{\boldsymbol{Y}}\big(\int_\Lambda \theta(\boldsymbol{Y}, \lambda)\mu(d\lambda)\big) = 0$, and we come up with the following expression for the unbiased risk estimate:
$$\begin{aligned}\hat{r}_{\text{EWA}} &= \int_\Lambda \Big(\tilde{r}_\lambda^{\text{unb}} - \|\tilde{\boldsymbol{f}}_\lambda - \hat{\boldsymbol{f}}_n\|_n^2 + 2\big\langle \nabla_{\boldsymbol{Y}} \log\theta(\lambda) | \widehat{\Sigma}(\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}})\big\rangle_n\Big)\hat{\mu}(d\lambda)\\ &= \int_\Lambda \Big(\tilde{r}_\lambda^{\text{unb}} - \|\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}}\|_n^2 - 2n\beta^{-1}\big\langle \nabla_{\boldsymbol{Y}}\widetilde{r}_\lambda | \widehat{\Sigma}(\tilde{\boldsymbol{f}}_\lambda - \tilde{\boldsymbol{f}}_{\text{EWA}})\big\rangle_n\Big)\hat{\mu}(d\lambda).\end{aligned}$$

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 $\beta \ge 2na$ it holds that $\hat r_{\text{EWA}} \le \int_\Lambda \tilde r_\lambda^{\text{unb}} \hat\mu(d\lambda) \le \int_\Lambda \widetilde r_\lambda \hat\mu(d\lambda) \le \int_\Lambda \widetilde r_\lambda \hat\mu(d\lambda) + \frac{\beta}{n}\mathcal{K}(\hat\mu,\mu)$. To conclude, it suffices to remark that $\hat\mu$ is the probability measure minimizing the criterion $\int_\Lambda \widetilde r_\lambda p(d\lambda) + \frac{\beta}{n}\mathcal{K}(p,\mu)$ among all $p \in \mathcal{P}_\Lambda$. Thus, for every $p \in \mathcal{P}_\Lambda$, it holds that $\hat r_{\text{EWA}} \le \int_\Lambda \widetilde r_\lambda p(d\lambda) + \frac{\beta}{n}\,\mathcal{K}(p,\mu)$. Taking the expectation of both sides, the desired result follows.

Proof of Remark 2 and Theorem 1

Let now $\tilde{\boldsymbol f}_\lambda = \widetilde{\mathsf A}_\lambda \boldsymbol Y$ with a symmetric $\widetilde{\mathsf A}_\lambda = \mathsf A_\lambda + \mathsf A_\lambda^\top - \mathsf A_\lambda^\top \mathsf A_\lambda$. We apply Lemma 2 with the prior $\mu(d\lambda) \propto \exp\{2\,\mathrm{Tr}[\Sigma(\mathsf A_\lambda^\top - \mathsf A_\lambda^\top \mathsf A_\lambda)]/\beta\}\pi(d\lambda)$, with $\hat{\boldsymbol f}_\lambda = \mathsf A_\lambda \boldsymbol Y$ and with the risk estimate

$$\widetilde r_\lambda = \big\|\boldsymbol Y - \hat{\boldsymbol f}_\lambda\big\|_n^2 + \frac{2}{n}\,\mathrm{Tr}[\Sigma\widetilde{\mathsf A}_\lambda] - \frac{1}{n}\,\mathrm{Tr}[\Sigma] = \hat r_\lambda^{\text{unb}} + \frac{2}{n}\,\mathrm{Tr}[\Sigma \mathsf A_\lambda^\top \mathsf A_\lambda - \Sigma \mathsf A_\lambda]. \tag{14}$$

One easily checks that this choice leads to the posterior $\hat\mu$ that is equal to $\hat\pi$ defined in Figure 2. Therefore, the aggregate $\tilde{\boldsymbol f}_{\text{EWA}}$ based on the prior μ coincides with $\hat{\boldsymbol f}_{\text{SEWA}}$ based on the prior π. Thus we obtain the following inequality:

$$\mathbb{E}[\|\hat{\boldsymbol f}_{\text{SEWA}} - \boldsymbol f\|_n^2] \le \inf_{p\in\mathcal{P}_\Lambda} \Big\{ \int_\Lambda \mathbb{E}[\widetilde r_\lambda]\,p(d\lambda) + \frac{\beta\mathcal{K}(p,\mu)}{n} \Big\}. \tag{15}$$

Furthermore, easy algebra yields that all the conditions required in the second part of Lemma 2 are fulfilled with $a = \frac{2|\!|\!|\Sigma|\!|\!|}{n}$ as soon as $\beta \ge 4|\!|\!|\Sigma|\!|\!|$. Indeed, one can notice that $\nabla_{\boldsymbol Y} \widetilde r_\lambda = \frac{2}{n}(\boldsymbol Y - \tilde{\boldsymbol f}_\lambda)$. This leads to

$$\begin{aligned} \int_\Lambda \big\langle \nabla_{\boldsymbol Y}\widetilde r_\lambda | \Sigma(\tilde{\boldsymbol f}_\lambda - \tilde{\boldsymbol f}_{\text{EWA}})\big\rangle_n \hat\mu(d\lambda) &= \frac{2}{n}\int_\Lambda \big\langle \tilde{\boldsymbol f}_{\text{EWA}} - \tilde{\boldsymbol f}_\lambda | \Sigma(\tilde{\boldsymbol f}_\lambda - \tilde{\boldsymbol f}_{\text{EWA}})\big\rangle_n \hat\mu(d\lambda) \\ &\le \frac{2|\!|\!|\Sigma|\!|\!|}{n}\int_\Lambda \|\tilde{\boldsymbol f}_\lambda - \tilde{\boldsymbol f}_{\text{EWA}}\|_n^2 \hat\mu(d\lambda). \end{aligned} \tag{16}$$

Hence the conclusion of the second part of Lemma 2 holds true. To prove the claim of Remark 2, one can notice that:

$$\begin{aligned} \mathcal{K}(p,\mu) &= -\int_\Lambda \log\big(\frac{d\mu}{dp}(\lambda)\big)p(d\lambda) \\ &= \int_\Lambda \frac{2}{\beta}\,\mathrm{Tr}[\Sigma(\mathsf A_\lambda^\top \mathsf A_\lambda - \mathsf A_\lambda)]p(d\lambda) + \log\Big[\int_\Lambda e^{\frac{2}{\beta}\,\mathrm{Tr}[\Sigma(\mathsf A_\lambda - \mathsf A_\lambda^\top \mathsf A_\lambda)]}\pi(d\lambda)\Big] + \mathcal{K}(p,\pi). \end{aligned} \tag{17}$$

Then, by taking the expectation and combining together relations (14), (15) and (17), one gets $\mathbb{E}[\|\hat{\boldsymbol f}_{\text{SEWA}} - \boldsymbol f\|_n^2] \le \inf_{p\in\mathcal{P}_\Lambda}\Big\{\int_\Lambda r_\lambda p(d\lambda) + \frac{\beta\mathcal{K}(p,\pi)}{n}\Big\} + \mathrm{R}_n$, and the claim of Remark 2 follows.

Finally, if condition $\mathbf{C}(\lambda)$ is satisfied for π-almost all values of λ, then R_n is non-positive, and we get the sharp oracle inequality stated in Theorem 1.