Aggregation of ordered smoothers in colored noise*

E. A. Krymova
krymova@phystech.edu
Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia;
Institute for Information Transmission Problems, 19 Bolshoy Karetny per., build 1, Moscow, Russia

The paper is devoted to the problem of recovery of one-dimensional functions given a set of noisy observations. Suppose that in addition, one is given a fixed set of a finite number of function estimates. Based on this set of estimates, it is necessary to construct a new estimator, the risk of which would be close to the risk of the "best" estimate (so-called oracle) in a given set. The "best" estimator is a minimizer of the risk over the given set of function estimators. New oracle inequalities for aggregation of regression function estimates in assumption of heteroscedasic Gaussian noise, namely, correlated Gaussian noise with different variances at each design point, have been proved.

Keywords: *aggregation; exponential weighting; ordered smoothers; unbiased risk estimation*

DOI: 10.21469/22233792.1.13.01

Агрегация упорядоченных оценок в цветном шуме*

Е. А. Крымова
Московский физико-технический институт (ГУ)
Институт проблем передачи информации им. А. А. Харкевича РАН

Рассматривается задача восстановления функции регрессии по конечному числу наблюдений функции в гауссовским шуме, заданных в конечном числе детерминированных точек. Предположим, что помимо наблюдений функции исследователю заранее известен фиксированный набор из конечного числа оценок функции. На основе этого набора оценок требуется построить новую оценку, качество которой было бы сравнимо с наилучшей (в смысле среднеквадратичного риска) оценкой из заданного множества (с так называемым «оракулом»). В работе получены новые оракульные неравенства для экспоненциальной агрегации упорядоченных оценок функции регрессии в предположении гетероскедастичного шума, а именно: шум предполагается коррелированным (ковариационная матрица известна) и дисперсия его различна в каждой точке наблюдения.

Ключевые слова: *агрегация оценок; экспоненциальное взвешивание; упорядоченные оценки; несмещенное оценивание риска*

DOI: 10.21469/22233792.1.13.01

1 Introduction

The paper is devoted to estimation of noisy vector (sequence space model) given a set of linear estimators. The sequence space model plays significant role in nonparametric statistics. Many problems can be transformed to the sequence space model formulation with white (i. e., with noncorrelated identically distributed zero-mean noise) Gaussian noise or with colored (i. e., noncorrelated nonidentically distributed zero-mean) Gaussian noise. For example, very often, linear inverse problems are easily transformed into diagonal form with the help of singular

*This work is partially supported by RFBR research project 15-07-09121.

value decomposition [1]. In this paper, the generalization of such models for the correlated colored Gaussian noise assumption is considered. Throughout the paper, it is assumed that one is given a special set of linear estimators, namely, ordered smoothers as various methods in statistics can be proved to have properties of ordered smoothers (for example, smoothing splines [2, 3], spectral regularization methods [1, 4], etc.). There exist various approaches to construct estimates given a set of estimators. One can use a model selection approach and select one estimator, for example, by a method of the unbiased risk estimation [5] which goes back to [6, 7].

Another approach is to use aggregation, namely, using a convex combination of given estimators. This approach was firstly developed by Nemirovsky [8] and independently by Catoni [9]. To tune the weights of the linear combination, authors performed the sample splitting. Later, this method was extended to several statistical models (see, e. g., [10–13]).

One can avoid sample splitting with the help of the exponential weighting. This method originates from the solution of functional aggregation problem by penalized empirical risk minimization [12]. It has been shown that for this method, one can yield rather good oracle inequalities for certain statistical models [14–16].

The goal is to prove new oracle inequalities for aggregation of ordered smoothers in assumption of heteroscedasic Gaussian noise, namely, correlated Gaussian noise with different variances at each design point.

2 Problem Statement

This paper deals with a sequence space model

$$Y_i = \theta_i + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

where $(Y_1, \dots, Y_n)^\intercal$ is the vector of observation; and $(\xi_1, \dots, \xi_n)^\intercal$ is the zero-mean Gaussian vector with known $n \times n$ covariance matrix Σ. The goal is to estimate an unknown vector $\theta \in \mathbb{R}^n$ based on the data $Y = (Y_1, \dots, Y_n)^\intercal$.

Denote the diagonal elements of Σ by σ_i^2, $i = 1, \dots, n$. Let one impose the following conditions on the covariance matrix Σ .

1. The spectral norm is bounded from above:

$$\sigma^2_{\max} = \sup_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \Sigma x < \infty\,.$$

2. The smallest eigenvalue is bounded from below:

$$\sigma^2_{\min} = \inf_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \Sigma x > 0\,.$$

3. Assume also that

$$\sup_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \left[\Sigma \circ \Sigma\right] x < C_\circ^2$$

where $\circ$ is the Hadamard product and $C_\circ$ is the constant.

Let one denote the risk of an estimator $\hat\theta(Y) = (\hat\theta_1(Y), \dots, \hat\theta_n(Y))^\intercal$ by

$$R(\hat\theta, \theta) = \mathsf{E}_\theta \|\hat\theta(Y) - \theta\|^2. \tag{2}$$

Here, E_θ stands for the expectation with respect to the measure P_θ generated by the observations (1) where $\|\cdot\|$ denotes the norm in $\mathbb{R}^n$: $\|x\|^2 = \sum_{i=1}^n x_i^2$.

Throughout this paper, θ will be recovered with the help of linear estimates

$$\hat{\theta}_i^h(Y) = h_i Y_i,\ h \in \mathcal{H} \tag{3}$$

where $\mathcal{H}$ is the finite set of so-called *ordered smoothers*, which has the following definition.

Definition 1. *A set $\mathcal{H}$ is a set of ordered multipliers if*

- *$h_i \in [0,1],\ i = 1,\dots,n$ for all $h \in \mathcal{H}$;*
- *$h_{i+1} \leqslant h_i,\ i = 1,\dots,n$, for all $h \in \mathcal{H}$; and*
- *if for some integer k and some $h, g \in \mathcal{H}$, $h_k < g_k$, then $h_i \leqslant g_i$ for all $i = 1,\dots,n$.*

The last condition means that vectors in $\mathcal{H}$ are naturally ordered, since for any $h, g \in \mathcal{H}$, there are only two possibilities: $h_i \leqslant g_i$ or $h_i \geqslant g_i$ for all $i = 1,\dots,n$.

Substituting the linear model (3) into the risk definition (2), one obtains

$$R(\hat{\theta}^h, \theta) = \|(1-h)\cdot\theta\|^2 + \|\sigma \cdot h\|^2,$$

where $x \cdot y$ denotes the coordinate-wise product of vectors $x, y \in \mathbb{R}^n$, i. e., $z = x \cdot y$ means that $z_i = x_i y_i$, $i = 1,\dots,n$, and $\sigma = (\sigma_1,\dots,\sigma_n)^\intercal$. Since $R(\hat{\theta}^h, \theta)$ depends on $h \in \mathcal{H}$, one can minimize it over $h \in \mathcal{H}$. The minimal risk

$$r^{\mathcal{H}}(\theta) = \min_{h\in\mathcal{H}} R(\hat{\theta}^h, \theta)$$

is often called in the literature as the oracle risk [8, 9].

Naturally, it is not possible to use the estimate

$$\theta^*(Y) = h^* \cdot Y, \quad h^* = \arg\min_{h\in\mathcal{H}} R(\hat{\theta}^h, \theta)$$

because it depends on the unknown vector θ. But if one knew θ, it would be possible to point out the estimate with the least risk. That is why, the goal is to construct an estimator $\tilde{\theta}^{\mathcal{H}}(Y)$ based on the family of linear estimators $\hat{\theta}^h(Y),\ h \in \mathcal{H}$, which is close to the oracle risk. Formally, this means that the estimator $\tilde{\theta}^{\mathcal{H}}(Y)$ should satisfy the so-called oracle inequality

$$R(\tilde{\theta}^{\mathcal{H}}, \theta) \leqslant r^{\mathcal{H}}(\theta) + \tilde{\Delta}^{\mathcal{H}}(\theta)$$

which holds uniformly in $\theta \in \mathbb{R}^n$.

This inequality implies that the term $\tilde{\Delta}^{\mathcal{H}}$ is small with respect to the oracle risk uniformly in $\theta \in \mathbb{R}^n$. It is well known that in general, it is not possible to construct such an estimator [17]. But as it was shown in [17] for the set $\mathcal{H}$ of *ordered smoothers*, one can find an estimator which provides the following properties of the remainder term:

- $\tilde{\Delta}^{\mathcal{H}}(\theta) \leqslant \tilde{C} r^{\mathcal{H}}(\theta)$ for all $\theta \in \mathbb{R}^n$ where $\tilde{C} > 1$ is the constant; and
- $\tilde{\Delta}^{\mathcal{H}}(\theta) \ll r^{\mathcal{H}}(\theta)$ for all $\theta :\ r^{\mathcal{H}}(\theta) \gg \sigma^2$.

That is why, throughout this paper, it will be assumed that the set $\mathcal{H}$ contains solely ordered multipliers. Below, an example of ordered smoothers is given. Note that ordered smoothers are very common in statistics, e. g., smoothing splines [2, 3], spectral regularization methods [1, 4].

3 A Motivating Example

Consider the regression estimation problem in the case of colored noise. It is necessary to recover a one-dimensional function $f(x),\ x \in [0,1]$, given the noisy observations

$$Z_i = f(x_i) + \bar{\xi}(x_i), \quad i = 1,\dots,n, \tag{4}$$

where $x_i \in (0,1)$ and $\bar{\xi}_i(x)$ is the centered Gaussian random process with variance $\bar{\sigma}^2(x)$. Denote by $\bar{\Sigma}$ the covariance matrix of the vector $(\bar{\xi}(x_1), \dots, \bar{\xi}(x_n))^\intercal$.

Let one make use of the smoothing spline estimate, which is defined as follows:

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \sum_{i=1}^n [Z_i - f(x_i)]^2 + \alpha \int_0^1 [f^{(m)}(x)]^2 \right\} \tag{5}$$

where $f^{(m)}(\cdot)$ denotes the derivative of order m and $\alpha > 0$ is the smoothing parameter which is usually chosen with the help of the Generalized Cross Validation (see, e. g., [18]).

To transform this model into the model (1), consider the Demmler–Reinsch basis [19] $\psi_k(x),\ x \in [0,1],\ k = 1, \dots, n$, which has double orthogonality property

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}\,;$$

$$\int_0^1 \psi_k^{(m)}(x) \psi_l^{(m)}(x)\, dx = \delta_{kl} \lambda_k,\ k, l = 1, \dots, n,$$

where here and below $\langle u, v \rangle_n$ stands for the inner product

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^n u(x_i) v(x_i)$$

and λ_i are the eigenvalues of the basis.

It is assumed for definiteness that the eigenvalues λ_k are sorted in ascending order:

$$\lambda_1 \leqslant \dots \leqslant \lambda_n.$$

With this basis, one can represent the underlying function as follows:

$$f(x) = \sum_{k=1}^n \psi_k(x) \theta_k \tag{6}$$

and one gets from (4)

$$Y_k = \langle Z, \psi_k \rangle_n = \theta_k + \xi_k$$

where

$$\xi_k = \sum_{j=1}^n \bar{\xi}(x_k) \psi_k(x_j)\,. \tag{7}$$

Next, substituting (6) in (5), one arrives at

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \sum_{k=1}^n (Y_k - \theta_k)^2 + \alpha \sum_{k=1}^n \lambda_k \theta_k^2 \right\}.$$

Therefore,

$$\hat{f}_\alpha(x, Y) = \sum_{k=1}^n \hat{\theta}_k \psi_k(x)$$

where

$$\hat{\theta}_k = \frac{Y_k}{1 + \alpha \lambda_k}.$$

Thus, one may conclude that the models (1)–(3) and (4)–(5) become equivalent with

$$h_k = h_k^\alpha = \frac{1}{1+\alpha\lambda_k}.$$

The vector $\xi = (\xi_1, \dots, \xi_n)^\intercal$ is a Gaussian zero-mean vector with covariance matrix

$$\Sigma = \frac{1}{n^2}\Psi^\intercal\bar{\Sigma}\Psi$$

where matrix Ψ consists of the columns $(\psi_i(x_1), \dots, \psi_i(x_n))^\intercal$, $i = 1, \dots, n$.

From the orthogonality property of Demmler–Reinsch basis, it is easily seen that eigenvalues of matrix Σ are equal to $\sigma_i^2 = \bar{\sigma}^2(x_i)/n$. Thus, for fixed n, the matrix Σ has finite eigenvalues and the problem is equivalent to (1).

The most interesting case is when $\bar{\Sigma}$ is a diagonal matrix with diagonal elements $\bar{\sigma}^2(x_1), \dots, \bar{\sigma}^2(x_n)$. It is known that in the case of equidistant design, Demmler–Reinsch basis has the following asymptotic as $n, k \to \infty$ [2]:

$$\psi_k(x) \approx \sqrt{\frac{2}{n}}\cos(\pi k x).$$

After a transformation of the regression estimation problem (4) with the help of Demmler–Reinsch basis, one obtains the following covariance of the noise (7):

$$\mathsf{E}\xi_k\xi_j \approx \frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i)\cos(\pi(k-j)p).$$

Thus, matrix Σ approximately equals to a correlation matrix of a stationary Gaussian sequence with variance $\sum_{i=1}^n \sigma^2(x_i)/n$ and the problem (4) becomes equivalent to the problem of estimation of an unknown vector in assumption of stationary noise.

In practice, one has to estimate the unknown covariance in (4). For the model with stationary noise, it is easy to estimate variance σ^2 given the data, for example, by

$$\bar{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n-1}[Z_i - Z_{i+1}]^2.$$

4 Exponential Weighing of Ordered Smoothers

In what follows, the exponential weighting estimate is used:

$$\bar{\theta}(Y) = \sum_{h\in\mathcal{H}} w^h(Y)\hat{\theta}^h(Y)$$

where

$$w^h(Y) = \pi^h \exp\left[-\frac{\bar{r}(Y,\hat{\theta}^h)}{2\beta\sigma^2_{\max}}\right] \Big/ \sum_{g\in\mathcal{H}} \pi^g \exp\left[-\frac{\bar{r}(Y,\hat{\theta}^g)}{2\beta\sigma^2_{\max}}\right].$$

Here, parameter $\beta > 0$ is fixed and $\bar{r}(Y,\hat{\theta}^h)$ is the unbiased risk estimate of $\hat{\theta}^h(Y)$ defined by

$$\bar{r}(Y,\hat{\theta}^h) \stackrel{\text{def}}{=} \|Y - \hat{\theta}^h(Y)\|^2 + 2\sum_{i=1}^{n} h_i\sigma_i^2 - \sum_{i=1}^{n}\sigma_i^2.$$

In order to cover $\mathcal{H}$ with small and large cardinalities, make use of the special prior weights defined as follows:

$$\pi^h \stackrel{\text{def}}{=} 1 - \exp\left\{-\frac{\sum_{i=1}^n \sigma_i^2(h_i^+ - h_i)}{\beta\sigma_{\max}^2}\right\}. \tag{8}$$

Here,

$$h^+ = \min\{g \in \mathcal{H} : g > h\}, \quad \pi^{h_{\max}} = 1$$

where $h^{\max}$ is the maximal multiplier in $\mathcal{H}$. Along with these weights, one needs also the following condition which can be proved to be true for smoothing splines and spectral regularization methods.

Condition 1. There exists a constant $K_\circ \in (0, \infty)$ such that

$$\|h\|^2 - \|g\|^2 \geqslant K_\circ\big(\|h\|_1 - \|g\|_1\big) \tag{9}$$

for all $h \geqslant g$ from $\mathcal{H}$, where $\|\cdot\|_1$ stands for the l_1-norm in $\mathbb{R}^n$, i. e.,

$$\|h\|_1 = \sum_{i=1}^n |h_i|.$$

Mention the following oracle inequality [16] for the exponential weighting of ordered smoothers in the case of white Gaussian noise with variance σ^2 that is diagonal Σ with $\sigma_{\min} = \sigma_{\max} = \sigma$.

Theorem 1. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\bar\theta - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}\right)\right].$$

This oracle inequality outperforms (in the form of the remainder term) Kniep's oracle inequality [17].

Theorem 2. *Uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\hat h \cdot Y - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + K\sigma^2\sqrt{1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}}$$

where a minimizer of the unbiased risk estimate $\hat h = \operatorname{argmin}_{h\in\mathcal{H}} \bar r(Y, \hat\theta^h)$ corresponds to the case $\beta \to 0$ in exponential weighting and K is the generic constant.

The main result of this paper is the following new oracle inequality with remainder term of the same form as in [16] for the exponential weighting in the case of colored noise problem.

Theorem 3. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\bar\theta - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma_{\max}^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma_{\min}^2}\right)\right].$$

Here and in what follows, $C = C(C_\circ, K_\circ, \beta, \varkappa)$ denotes strictly positive and bounded constant depending on $C_\circ, K_\circ, \beta$, and $\varkappa$, where $\varkappa = \sigma_{\max}/\sigma_{\min}$.

For the case of stationary noise ξ with variance σ^2, one has $\sigma^2_{\min} = \sigma^2_{\max} = \sigma^2$ and the following

Corollary 1. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta \|\bar{\theta} - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}\right)\right].$$

Here and in what follows, $C = C(C_\circ, K_\circ, \beta)$ denotes strictly positive and bounded constants depending on $C_\circ, K_\circ$, and β.

5 Simulations

To find out what value of β is good from a practical viewpoint and to compare the cases of white and coloured Gaussian noise, a numerical experiment has been carried out. The present author compares the exponential weighting methods applied to the set of cubic smoothing splines (as ordered smoothers) for $\beta = \{0, 1, 2, 4\}$ and for the equidistant design:

$$\mathcal{H} = \left\{h : h_k = \frac{1}{1 + [\alpha(k-1)]^4},\ \alpha > 0\right\}$$

where an asymptotic formula for the eigenvalues of Demmler–Reinsch basis was used in the case of equidistant design: $\lambda_k \asymp (\pi k)^4,\ k \to \infty$.

The scheme of the experiment is the following. For a given $A \in [0, 300]$, 100 000 replications of the observations

$$Y_k = \theta_k(A) + \xi_k,\ k = 1, \ldots, 400,$$

are generated. Here, $\theta(A) \in R^{400}$ is the Gaussian vector with independent components and

$$\mathsf{E}\theta_k(A) = 0, \quad \mathsf{E}\theta_k^2(A) = A \exp\left(-\frac{k^2}{2\Omega^2}\right)$$

where $\Omega = 50$.

Two types of the noise ξ were considered:

1) standard Gaussian white noise ($\sigma_i = 1$); and
2) Gaussian vector with covariance matrix Σ with eigenvalues $\sigma_i = i/400$, $i = 1, \ldots, 400$.

Next, the mean oracle risk

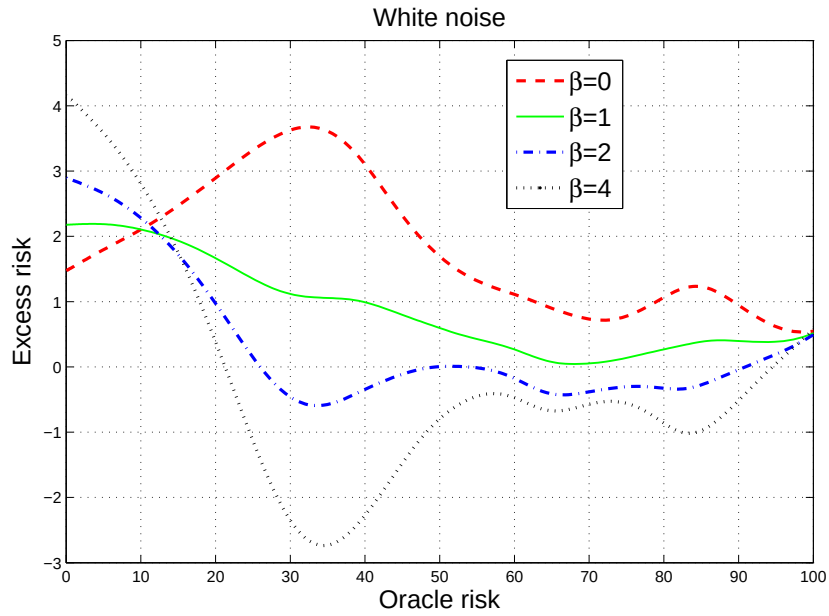
$$\bar{r}^{\mathcal{H}}(A) = \mathsf{E} \min_{h \in \mathcal{H}} \left\{ \|(1 - h) \cdot \theta(A)\|^2 + \|\sigma \cdot h\|^2 \right\}$$

and the mean excess risk

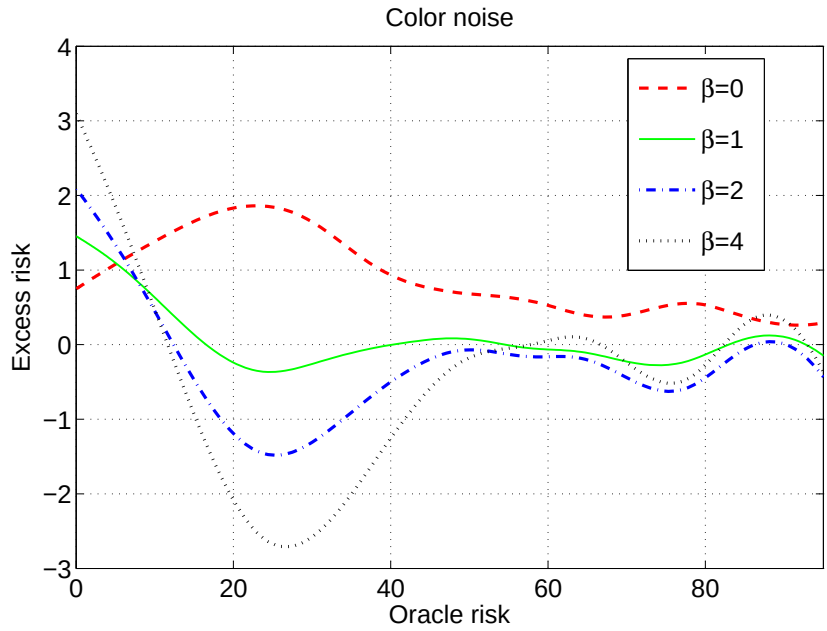
$$\bar{\Delta}_\beta(A) = \mathsf{E}\|\theta(A) - \bar{\theta}(Y)\|^2 - \bar{r}^{\mathcal{H}}(A)$$

were computed with the help of the Monte-Carlo method. Finally, the data $\{\bar{r}^{\mathcal{H}}(A),\ \bar{\Delta}_\beta(A),\ A \in [0, 300]\}$ are plotted in Fig. 1 to illustrate graphically the remainder term $\Delta_\beta(r^{\mathcal{H}}) = \mathsf{E}_\theta\|\bar{\theta} - \theta\|^2 - r^{\mathcal{H}}(\theta)$.

Looking at Fig. 1, one sees that there is no universal β minimizing the excess risk uniformly in θ. However, intuitively, it seems that a reasonable choice is $\beta \approx 1$ [15] but unfortunately, good oracle inequalities are not available for this case. Almost all methods demonstrate similar



(a)



(b)

Figure 1 Exponential weighting for the white (a) and colored (b) noise cases. The data $\{\bar{r}^{\mathcal{H}}(A), \bar{\Delta}_{\beta}(A),\ A \in [0, 300]\}$ that is the dependancy of excess risk on oracle risk is in the pictures

statistical performance (in Fig. 1, for the values of oracle risk bigger than 50). However, when $r^{\mathcal{H}}(\theta)/\sigma^2$ is not large, the exponential weighting works usually better (in Fig. 1, for the values of oracle risk from approximately 10 to 50).

6 Proofs

The main steps of the proof are based on a combination of methods for deriving oracle inequalities proposed in [16, 20]. Here, the main steps in the proof are sketched, all details are given below.

With the help of Stein's formula for the unbiased risk estimate, it can be shown that for $\beta \geqslant 4$,

$$\mathsf{E}_\theta \|\bar\theta - \theta\|^2 \leqslant \mathsf{E}_\theta \sum_{h\in\mathcal{H}} w^h(Y)\bar r(Y,\hat\theta^h) \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2_{\max}\mathsf{E}_\theta \sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} - 2\beta\sigma^2_{\max}\mathsf{E}_\theta \log\biggl\{\sum_{h\in\mathcal{H}}\pi^h \exp\biggl[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\biggr]\biggr\} \quad (10)$$

where $\hat h$ is the minimizer of the unbiased risk estimate $\hat h = \arg\min_{h\in\mathcal{H}} \bar r(Y,\hat\theta^h)$.

To control the right-hand side at this equation, make use of the ordering property of estimates $\hat\theta^h,\ h\in\mathcal{H}$. First, check that if π^h is defined by (8), then

$$\sum_{h\in\mathcal{H}}\pi^h \exp\biggl[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\biggr] \geqslant \sum_{h\geqslant\hat h}\pi^h \exp\biggl[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\biggr] \geqslant 1$$

and so, the last term in Eq. (10) is always negative.

The most difficult and delicate part of the proof is related to the average Kullback–Leibler divergence $\mathsf{E}_\theta\sum_{h\in\mathcal{H}} w^h(Y)\log(w^h(Y)/\pi^h)$. To compute a good lower bound for this value, follow the approach proposed in [20]. The main idea here is to make use of the following property of the unbiased risk estimate: for any sufficiently small $\varepsilon<1$, there exists $\hat h^\varepsilon$ depending on Y such that with probability 1, for all $h\geqslant\hat h^\varepsilon$,

$$\bar r(Y,\hat\theta^h) - \bar r(Y,\hat\theta^{\hat h}) \geqslant 2\beta\varepsilon\bigl[\|\sigma\cdot h\|^2 - \|\sigma\cdot\hat h\|^2\bigr] + 2\beta\sigma^2_{\min}\,.$$

This equation means that $w^h(Y)$ are exponentially decreasing for large h. With this property, one obtains the following entropy bound:

$$\sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} \leqslant \log\left[\sum_{h\leqslant\hat h^\varepsilon}\pi^h + \frac{C}{\varepsilon}\exp\left(\frac{C}{\varepsilon}\right)\right].$$

The rest of the proof consists in deriving the following bound from (9) and (8):

$$\sum_{h\leqslant\hat h^\varepsilon}\pi^h \leqslant 1 + \frac{\|\sigma\cdot\hat h^\varepsilon\|^2}{K_\circ\beta\sigma^2_{\max}}$$

and

$$\sqrt{\mathsf{E}_\theta\|\sigma\cdot\hat h^\varepsilon\|^2} \leqslant \sqrt{\frac{r^{\mathcal{H}}(\theta)}{1-2\beta\varepsilon}} + \frac{\sqrt{1+2\beta}}{1-2\beta\varepsilon}\frac{\sqrt{KC_\circ}}{\sigma^2_{\min}}.$$

Finally, combining the above equations, one arrives at (1).

7 Concluding Remarks

Based on the probabilistic properties of the unbiased risk estimate, the oracle inequality was proved for the method of aggregation of smoothing splines for the regression estimation problem in the case of colored noise. However, it seems that no good oracle inequalities are available for the reasonable choice of β parameter in the definition of aggregating weights. Numerical results demonstrate similar statistical performance for different choice of β parameter.

References

[1] Engl, H. W., M. Hanke, and A. Neubauer. 1996. *Regularization of inverse problems. Mathematics and its applications*. Dordrecht: Kluwer Academic Publishers Group. 375 p.

[2] Speckman, P. 1985. Spline smoothing and optimal rates of convergence in nonparametric regression. *Ann. Statist.* 13:970–983.

[3] Green, P. J., and B. W. Silverman. 1994. *Nonparametric regression and generalized linear models. A roughness penalty approach*. Chapman and Hall. 184 p.

[4] Tikhonov, A. N., and V. A. Arsenin. 1977. *Solution of ill-posed problems*. Scripta ser. in mathematics. Washington, DC – New York, NY: V. H. Winston & Sons – John Wiley & Sons. 258 p.

[5] Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* 9:1135–1151.

[6] Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *2nd Symposium (International) on Information Theory Proceedings*. 267–281.

[7] Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15:661–675.

[8] Nemirovski, A. 2000. *Topics in non-parametric statistics*. Lectures notes in mathematics ser. Berlin: Springer-Verlag. 197 p.

[9] Catoni, O. 2004. *Statistical learning theory and stochastic optimization*. Lectures notes in mathematics ser. Berlin: Springer-Verlag. 279 p.

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

[11] Lecué, G. 2007. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Stat.* 35:1698–1721.

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

[13] Rigollet, Ph., and A. Tsybakov. 2011. Sparse estimation by exponential weighting. arXiv:1108.5116v1 [math.ST].

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

[15] Dalayan, A., and J. Salmon. 2011. Sharp oracle inequalities for aggregation of affine estimators. arXiv:1104.3969v2 [math.ST].

[16] Chernousova, E., Yu. Golubev, and E. Krymova. 2013. Ordered smoothers with exponential weighting. *Electron. J. Statist.* 7.

[17] Kneip, A. 1994. Ordered linear smoothers. *Ann. Stat.* 22:835–866.

[18] Wahba, G. 1990. *Spline models for observational data*. Philadelphia, PA: SIAM. 161 p.

[19] Demmler, A., and C. Reinsch. 1975. Oscillation matrices with spline smoothing. *Numerische Mathematik* 24:375–382.

[20] Golubev, Yu. 2012. Exponential weighting and oracle inequalities for projection methods. *Problems Inform. Transmission* 3. arXiv:1206.4285.

Received June 15, 2015

Литература

[1] *Engl H. W., Hanke M., Neubauer A.* Regularization of inverse problems. Mathematics and its applications. — Dordrecht: Kluwer Academic Publishers Group, 1996. 375 p.

[2] *Speckman P.* Spline smoothing and optimal rates of convergence in nonparametric regression // Ann. Stat., 1985. No. 13. P. 970–983.

[3] *Green P. J., Silverman B. W.* Nonparametric regression and generalized linear models. A roughness penalty approach. — Chapman and Hall, 1994. 184 p.

[4] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979. 285 с.

[5] *Stein C.* Estimation of the mean of a multivariate normal distribution // Ann. Stat., 1981. No. 9. P. 1135–1151.

[6] *Akaike H.* Information theory and an extension of the maximum likelihood principle // 2nd Symposium (Intenational) on Information Theory Proceedings, 1973. P. 267–281.

[7] *Mallows C. L.* Some comments on C_p // Technometrics, 1973. No. 15. P. 661–675.

[8] *Nemirovski A.* Topics in non-parametric statistics. — Lectures notes in matematics ser. — Berlin: Springer-Verlag, 2000. 197 p.

[9] *Catoni O.* Statistical learning theory and stochastic optimization. — Lectures notes in mathematics ser. — Berlin: Springer-Verlag, 2004. 279 p.

[10] *Yang Y.* Aggregating regression procedures to improve performance // Bernoulli, 2004. No. 10. P. 25–47.

[11] *Lecué G.* Simultaneous adaptation to the margin and to complexity in classification // Ann. Stat., 2007. No. 35. P. 1698–1721.

[12] *Rigollet P., Tsybakov A. B.* Linear and convex aggregation of density estimators // Math. Methods Statist., 2007. No. 16. P. 260–280.

[13] *Rigollet Ph., Tsybakov A.* Sparse estimation by exponential weighting. 2011. arXiv:1108.5116v1 [math.ST].

[14] *Leung G., Barron A.* Information theory and mixing least-squares regressions // IEEE Trans. Inform. Theory, 2006. Vol. 52. No. 8. P. 3396–3410.

[15] *Dalayan A., Salmon J.* Sharp oracle inequalities for aggregation of affine estimators. 2011. arXiv:1104.3969v2 [math.ST].

[16] *Chernousova E., Golubev Yu., Krymova E.* Ordered smoothers with exponential weighting // Electron. J. Statist., 2013. No. 7.

[17] *Kneip A.* Ordered linear smoothers // Ann. Stat., 1994. No. 22. P. 835–866.

[18] *Wahba G.* Spline models for observational data. — Philadelphia, PA, USA: SIAM, 1990. 161 p.

[19] *Demmler A., Reinsch C.* Oscillation matrices with spline smoothing // Numerische Mathematik, 1975. No. 24. P. 375–382.

[20] *Голубев Г. К.* Экспоненциальное взвешивание и оракульные неравенства для проекционных оценок // Пробл. передачи информ., 2012. Т. 48. № 3. С. 83–95.

Поступила в редакцию 15.06.2015