

The Method of Risk Envelope in Estimation of Linear Functionals

G. K. Golubev

*Institute for Information Transmission Problems, RAS, Moscow;
Université de Provence, Marseille, France
golubev@gyptis.univ-mrs.fr*

Received April 23, 2003; in final form, June 17, 2003

Abstract—The problem of estimating a linear functional in a linear Gaussian model is considered. For the estimation, the class of projection estimators is used. The problem is to choose the optimal estimate from this class on the basis of observations. The solution of this problem is based on the principle of risk envelope minimization.

1. INTRODUCTION

One of the approaches to solution of the problem of choosing the best estimate from a given family of estimators is discussed by the example of estimating a linear functional. This problem plays one of the fundamental roles not only in modern statistics but also in learning theory, which is intensively developing (see, e.g., [1]). In general, the problem of our interest is the following. Assume that we want to estimate an unknown parameter $\theta \in \mathbb{R}^d$ from observations of a random variable X with probability distribution $\mathbf{P}_\theta(x)$, $x \in \mathbb{R}^n$. Assume also that for the estimation of the parameter θ we can use estimates from a given collection (class) of estimators $\hat{\Theta} = \{\hat{\theta}^1(X), \hat{\theta}^2(X), \dots\}$ only. The problem is to find, using the observation of X , an estimate from $\hat{\Theta}$ that estimates the unknown parameter θ in the best way. Here, the risk of the estimate $\hat{\theta}(X)$ is measured by the quantity

$$r(\hat{\theta}, \theta) = \mathbf{E}_\theta \ell(\hat{\theta}(X), \theta),$$

where $\ell(\cdot, \cdot): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ is a loss function and \mathbf{E}_θ is averaging over the measure \mathbf{P}_θ . Speaking more precisely, we want to find an integer-valued function $K(X)$ such that the risk $r(\hat{\theta}^K, \theta)$ is minimal for any $\theta \in \mathbb{R}^d$.

The answer to the question why precisely this approach is used in statistics can be explained by the fact that, in many problems of interest from a practical point of view, the dimension d is rather large; therefore, in this case, it would be somewhat naive to expect sufficiently reliable a priori information on the estimated parameter. On the other hand, using classes of estimators allows one to extract missing a priori information from observations. In addition, of course, the class $\hat{\Theta}$ should be neither too poor nor extremely rich. For example, if this set is close to the class of all estimates, then in such a situation we undoubtedly cannot find a good estimate. The point is that, in the case where all estimates could be used, we could take $\hat{\theta}(X) = 0$. This estimate is obviously very bad, but its risk is equal to 0 for $\theta = 0$. Therefore, the problem of finding the best estimate for all θ in the class of all estimators does not have a reasonable solution. However, using not too rich classes of estimators allows us to make this problem completely sensible. In general, it is difficult to exactly say how rich the class of estimators can be, and here we do not study this (very interesting) question. Some ideas and approaches to solution of this problem can be found, for example, in [1, 2].

One of the well-known approaches to solution of the problem of choosing the best estimate from a given class in classical nonparametric statistics is the *principle of unbiased risk estimation*. Roughly, its motivation is the following. Assume that there exists an unbiased estimate of the risk $r(\hat{\theta}^k, \theta)$, i.e., there exists a function $\bar{r}(k, X)$ such that $r(\hat{\theta}^k, \theta) = \mathbf{E}_\theta \bar{r}(k, X) + r_0(\theta)$ for all $\theta \in \mathbb{R}^d$. It is fundamentally important that the function $r_0(\theta)$ in this representation does not depend on the index k . According to the principle of unbiased risk estimation, the index of the best estimate is defined as

$$K(X) = \arg \min_k \bar{r}(k, X).$$

This idea is extensively used in statistics. For example, the Akaike criterion [3], Mallows criterion [4], and all cross-validation methods [5] are based on this principle. Mathematical results concerning the quality of these methods were, apparently, first obtained in [6]. However, one certainly cannot say that the principle of unbiased risk estimation is applicable to any practical situation. It can perform very well in certain problems but lead to a catastrophe in some others.

Consider a simple example. Let X be a Gaussian vector with components

$$X_i = \theta_i + \sigma_i \xi_i, \quad i = 1, 2, \dots, \quad (1)$$

where ξ_i is a white Gaussian noise (i.e., ξ_i are independent Gaussian random variables with $\mathbf{E} \xi_i = 0$ and $\mathbf{E} \xi_i^2 = 1$) and σ_i are known quantities. To estimate the vector $\theta = (\theta_1, \theta_2, \dots)^T$, we will use the class of projection estimators $\hat{\theta}^k(X)$, $k = 1, 2, \dots$, defined by the equalities

$$\hat{\theta}_i^k(X) = \mathbf{1}\{i \leq k\} X_i.$$

In addition, estimation quality for the vector θ is measured by the quadratic loss function

$$\ell(\hat{\theta}, \theta) = \sum_{i=1}^{\infty} (\hat{\theta}_i - \theta_i)^2.$$

It can easily be checked that the risk of the estimate $\hat{\theta}^k(X)$ is computed as

$$r(\hat{\theta}^k, \theta) = \sum_{i=k+1}^{\infty} \theta_i^2 + \sum_{i=1}^k \sigma_i^2 = \sum_{i=1}^k [\sigma_i^2 - \theta_i^2] + \sum_{i=1}^{\infty} \theta_i^2.$$

From this it is clear that, to construct an unbiased estimate of the risk, it is sufficient to construct such an estimate for θ_i^2 . We can do this very simply, taking $X_i^2 - \sigma_i^2$ as an estimate. Hence, the unbiased estimate of the risk has the form

$$\bar{r}(k, X) = \sum_{i=1}^k [2\sigma_i^2 - X_i^2],$$

and, according to the principle of unbiased risk estimation, we take the projection estimate with the index

$$K(X) = \arg \min_k \left\{ \sum_{i=1}^k [2\sigma_i^2 - X_i^2] \right\}. \quad (2)$$

In many cases this principle gives rather good results. In particular, this holds for the case where $\sigma_i = \sigma$. Denote by $r(\theta) = \min_k r(\hat{\theta}^k, \theta)$ the risk of the best projection estimate. Then, uniformly in all $\theta \in \ell_2(1, \infty)$ (see [7]), we have the inequality

$$r(\hat{\theta}^K, \theta) - r(\theta) \leq C \sqrt{\sigma^2 r(\theta)},$$

where C is a constant. In statistics, inequalities of such kind are called oracle inequalities since this inequality claims in fact that the difference between $r(\hat{\theta}^K, \theta)$ and the risk $r(\theta)$ (which can be interpreted as the risk of an oracle) is relatively small as compared with $r(\theta)$ in the cases where $r(\theta) \gg \sigma^2$. At the same time, this difference has order σ^2 if $r(\theta) \sim \sigma^2$.

To understand why choosing a projection estimate based on the unbiased risk estimation can be bad, consider the situation where

$$\sigma_i^2 = \sigma^2 a^i, \quad a > 1.$$

Consider for simplicity how the principle of unbiased risk estimation works in the case where all $\theta_i = 0$. In this situation, we choose the estimate with the index

$$K(X) = \arg \min_k \left\{ -\sum_{i=1}^k \sigma_i^2 \xi_i^2 + 2 \sum_{i=1}^k \sigma_i^2 \right\}. \tag{3}$$

Note that, for the case where σ_i^2 grows exponentially, we have the inequality

$$\sum_{i=1}^k \sigma_i^2 \leq \frac{\sigma_k^2}{a-1},$$

and therefore

$$-\sum_{i=1}^k \sigma_i^2 \xi_i^2 + 2 \sum_{i=1}^k \sigma_i^2 \leq \sigma_k^2 [-\xi_k^2 + 2/(a-1)]. \tag{4}$$

For any $T > 0$ we almost surely have

$$\min_{k>T} \sigma_k^2 \left\{ \frac{2}{a-1} - \xi_k^2 \right\} = -\infty$$

since the probability of the event $\xi_k^2 > 3/(a-1)$ is strongly positive for any k . From (4) and (3), we immediately obtain

$$\mathbf{P}\{K(X) \leq T\} \leq \mathbf{P}\left\{ -\sum_{i=1}^T \sigma_i^2 \xi_i^2 + 2 \sum_{i=1}^T \sigma_i^2 \leq \min_{k>T} \left[-\sigma_k^2 \xi_k^2 + \frac{2\sigma_k^2}{a-1} \right] \right\} = 0$$

for any $T > 0$. Therefore, it is obvious that the risk of the estimate $\hat{\theta}^K$ is infinitely large. In other words, the considered method of choosing an estimate is catastrophically bad in the case where σ_i grows exponentially.

Note that there are many other statistical problems in which the principle of unbiased risk estimation leads to very bad estimates or cannot be applied at all. In particular, one of such problems is that of estimating a linear functional. The simplest variant of this problem is to estimate

$$S(\theta) = \sum_{i=1}^{\infty} \theta_i$$

from observations (1). The present paper is precisely devoted to this problem. In contrast to the problem of estimating a vector as a whole, the problem of estimating a linear functional actually occurs rather rarely in practical applications, though, of course, problems are known in which this problem arises in a natural way, for example, the Wicksell problem [8]. Statistical literature on estimation of linear functionals is very extensive (see, e.g., [9, 10]) in spite of the relatively narrow scope of its applications. In this paper we use the problem of estimating a linear functional

as a model example allowing us to demonstrate the universality of the risk envelope method. Essentially, this method is a generalization of the principle of unbiased risk estimation.

The essence of the suggested method is very simple. It consists of two steps:

- construction of the risk envelope;
- minimization of the risk envelope on the basis of observations.

The risk envelope for a family of estimates $\{\theta^1(X), \theta^2(X), \dots\}$ is defined as a deterministic (independent of the observations X) function $L(k, \theta)$ for which the condition

$$\mathbf{E}_\theta \sup_k \left\{ \ell(\hat{\theta}^k(X), \theta) - L(k, \theta) \right\} \leq 0 \quad (5)$$

holds for all $\theta \in \mathbb{R}^d$.

It is obvious that the risk envelope allows one to easily control the risk of any method of choosing the estimate $K(X)$

$$r(\hat{\theta}^K, \theta) \leq \mathbf{E}_\theta L(K, \theta). \quad (6)$$

It is seen from this inequality that the smaller the envelope, the better we can control the risk of our method. On the other hand, we of course would like to find a method for choosing the estimate that minimizes the envelope $L(k, \theta)$ over all k . This functional depends on the parameter estimated, and we have to minimize it on the basis of observations X . In fact, this is the most difficult problem. The point is that the set of envelopes is vast and contains envelopes which, at first sight, look very good but which are absolutely impossible to stochastically minimize. The idea is to find a compromise: to choose, perhaps, not the optimal envelope but such that its minimization could be realized with acceptable quality. The main goal of this paper is to demonstrate how such a compromise can be realized in a particular problem.

2. ESTIMATION OF A LINEAR FUNCTIONAL

Our goal is to estimate a linear functional $S(\theta) = \sum_{i=1}^{\infty} \theta_i$ from observations (1). We assume that all σ_i are identical and equal σ . As a family of estimators, we use projection estimates

$$\hat{S}^k(X) = \sum_{i=1}^{w_k} X_i, \quad (7)$$

where

$$w_k = (1 + \alpha)^k, \quad \alpha > 0, \quad k = 0, 1, 2, \dots \quad (8)$$

The purport of introducing the sequence w_k with exponential growth is in the fact that only for the family of projection estimators generated by such kind of sequences we can construct relatively good methods of risk envelope minimization.

As a loss function, let us take

$$\ell(\hat{S}, S) = |\hat{S}(X) - S(\theta)|.$$

One can also consider (without essential problems) other loss functions but, to simplify technical details, we limit ourselves to this simplest case.

2.1. Risk Envelope

Computation of the risk envelope is rather simple. Define for brevity

$$S_\ell(\theta) = \sum_{i=w_\ell}^{w_{\ell+1}-1} \theta_i.$$

In what follows, we denote by C all constants whose exact values are unimportant.

Lemma 1. For any $\alpha \in (0, 1]$, the function

$$L(k, \theta) = \left| \sum_{j=k+1}^{\infty} S_j(\theta) \right| + \sigma \sqrt{w_k \log(w_k)} + \sigma \sqrt{\frac{w_k}{\alpha}} + C\sigma$$

is a risk envelope.

Proof. It is obvious that we have the inequality

$$|\widehat{S}^k(\theta) - S(\theta)| \leq \left| \sum_{j=k+1}^{\infty} S_j(\theta) \right| + \sigma \left| \sum_{i=1}^{w_k} \xi_i \right|. \tag{9}$$

Let us use two banal inequalities:

$$x - y \leq x \mathbf{1}\{x > y\}, \quad y \geq 0,$$

and the fact that for any $x_i \geq 0$ we have

$$\max\{x_1, \dots, x_n\} \leq \sum_{i=1}^n x_i.$$

Therefore,

$$\begin{aligned} \mathbf{E} \max_k \left\{ \left| \sum_{i=1}^{w_k} \xi_i \right| - \sqrt{w_k \log(w_k)} - \sqrt{\frac{w_k}{\alpha}} \right\} \\ \leq \sum_{k=0}^{\infty} \mathbf{E} \left| \sum_{i=1}^{w_k} \xi_i \right| \mathbf{1} \left\{ \left| \sum_{i=1}^{w_k} \xi_i \right| \geq \sqrt{w_k \log(w_k)} + \sqrt{\frac{w_k}{\alpha}} \right\} \\ = \sum_{k=0}^{\infty} \sqrt{w_k} \mathbf{E} |\xi_1| \mathbf{1} \left\{ |\xi_1| \geq \sqrt{\log(w_k)} + \sqrt{\frac{1}{\alpha}} \right\}. \end{aligned}$$

Further, integrating by parts, we obtain

$$\mathbf{E} |\xi_1| \mathbf{1} \left\{ |\xi_1| \geq \sqrt{\log(w_k)} + \sqrt{\frac{1}{\alpha}} \right\} \leq C \exp \left\{ -\frac{1}{2} \left[\sqrt{w_k} + \frac{1}{\sqrt{\alpha}} \right]^2 \right\}.$$

Thus,

$$\begin{aligned} \mathbf{E} \max_k \left\{ \left| \sum_{i=1}^{w_k} \xi_i \right| - \sqrt{w_k \log(w_k)} - \sqrt{\frac{w_k}{\alpha}} \right\} \\ \leq C \sum_{k=0}^{\infty} \sqrt{w_k} \exp \left\{ -\log(w_k)/2 - \sqrt{\log(w_k)/\alpha} - 1/(2\alpha) \right\} \\ \leq C \exp[-1/(2\alpha)] \sum_{k=0}^{\infty} \exp \left\{ -\sqrt{k \log(1 + \alpha)/\alpha} \right\} \leq \frac{C\alpha \exp[-1/(2\alpha)]}{\log(1 + \alpha)}. \end{aligned}$$

The statement of the lemma directly follows from this inequality, the definition of an envelope (5), and inequality (9).

2.2. Stochastic Minimization of the Risk Envelope

This is a simple but, in a sense, nontrivial problem. According to Lemma 1, we would like to find the index k that minimizes the envelope

$$L_0(k, \theta) = \left| \sum_{j=k+1}^{\infty} S_j(\theta) \right| + \sigma \sqrt{w_k \log(w_k)} + \sigma \sqrt{\frac{w_k}{\alpha}}.$$

The complexity of this problem consists only in that we do not know $S_j(\theta)$, and at first sight the situation looks hopeless since we arrive at the same problem as we started from. But, in fact, it is not too difficult to overcome this. Let us try to minimize a little larger quantity, namely,

$$\begin{aligned} L^*(k, \theta) &= \sum_{j=k+1}^{\infty} |S_j(\theta)| + \sigma \sqrt{w_k \log(w_k)} + \sigma \sqrt{\frac{w_k}{\alpha}} \\ &= \sum_{j=0}^{\infty} |S_j(\theta)| - \sum_{j=0}^k |S_j(\theta)| + \sigma \sqrt{w_k \log(w_k)} + \sigma \sqrt{\frac{w_k}{\alpha}}. \end{aligned} \quad (10)$$

Hence it is clear that it suffices to minimize over all k the following function:

$$- \sum_{j=0}^k |S_j(\theta)| + \sigma \sqrt{w_k \log(w_k)} + \sigma \sqrt{\frac{w_k}{\alpha}}.$$

But this is already a rather good problem, whose solution can easily be found. It suffices to estimate $|S_j(\theta)|$ from the observations

$$X_i = \theta_i + \sigma \xi_i, \quad i = w_j, \dots, w_{j+1} - 1,$$

and substitute the obtained estimates into the minimized functional. The problem can be simplified by summing up the observations X_i . In this case, we have to estimate $|S_j(\theta)|$ from the observation

$$Y_j = S_j(\theta) + \sigma_j \xi_j,$$

where $\sigma_j = \sigma \sqrt{w_{j+1} - w_j}$. As an estimate for $S_j(\theta)$, one can obviously take $|Y_j|$. Thus, we arrive at a method of estimating a linear functional described below.

2.3. Method of Estimating a Linear Functional

In this section, we describe a procedure of estimating a linear functional $S(\theta)$ from observations (1), which is based on the principle of risk envelope minimization. An estimate for $S(\theta)$ is constructed from the reduced data

$$Y_j = \sum_{i=w_j}^{w_{j+1}-1} X_i, \quad j = 0, 1, \dots$$

Define the quantity K_γ which minimizes the risk envelope $L^*(k, \theta)$ on the basis of observations:

$$K_\gamma = \arg \min_k \left\{ - \sum_{j=1}^k |Y_j| + \sigma \sqrt{(1 + \gamma) w_k \log(w_k)} + 5\sigma \sqrt{\frac{w_k}{\alpha}} \right\}. \quad (11)$$

Then the estimate is defined as follows:

$$\hat{S}(X) = \sum_{i=1}^{K_\gamma} Y_i. \quad (12)$$

The quality of this estimate is described by the following theorem.

Theorem 1. For the estimate $\widehat{S}(X)$ defined in (11) and (12), the inequality

$$\mathbf{E}_\theta |\widehat{S}(X) - S(\theta)| \leq \left(1 + \frac{C\sqrt{\alpha}}{\gamma}\right) \min_k \left\{ \sum_{j=k+1}^{\infty} |S_j(\theta)| + \sigma \sqrt{(1 + \gamma)w_k \log(w_k)} + 5\sigma \sqrt{\frac{w_k}{\alpha}} \right\} + C\sigma \quad (13)$$

is fulfilled uniformly in all $\theta \in \ell_2(1, \infty)$.

The proof of this result is based on simple probabilistic facts stated in Section 3.

The statistical meaning of Theorem 1 is, in essence, the following. Inequality (13) claims that on the basis of the observation X_i we can minimize the functional

$$L_*(\theta, k) = \sum_{j=k+1}^{\infty} |S_j(\theta)| + \sigma \sqrt{w_k \log(w_k)}$$

with small losses. To clarify the meaning of this claim, consider two situations: $\min_k L_*(\theta, k) \approx \sigma$ and $\min_k L_*(\theta, k) \gg \sigma$. In the first case, which corresponds to parametric estimation where only finitely many components of the vector θ are nonzero, we immediately obtain that the risk of our estimate has order σ . The second situation is typical for nonparametric estimation. In addition, the risk of our method can be made smaller than $(1 + \varepsilon) \min_k L_*(\theta, k)$, where $\varepsilon > 0$ is an arbitrary given number, by an appropriate choice of the parameters α and γ .

The question naturally arises whether it is possible to construct estimates that are better than $\widehat{S}(X)$ in both the parametric and nonparametric case. It turns out that the answer is negative. A discussion of the problem in what sense it is impossible to construct better estimates is beyond the framework of this paper. For details of some methods of constructing lower bounds in problems of such kind, we refer the reader to [11–13].

Some other possible applications of Theorem 1 are connected with adaptive minimax estimation. For example, if we assume that

$$\theta \in \Theta_{\mu,q}(L) = \left\{ \theta : \sum_{i=1}^{\infty} \theta_i^2 \exp(\mu i^q) \leq L \right\}, \quad \mu, q, L > 0,$$

then, using inequality (13), we can easily estimate the risk $\sup_{\theta \in \Theta_{\mu,q}(L)} \mathbf{E}_\theta |\widehat{S}(X) - S(\theta)|$ as $L/\sigma^2 \rightarrow \infty$.

Moreover, it is not difficult to verify that the estimate $\widehat{S}(X)$ is minimax for all classes $\Theta_{\mu,q}(L)$ with $\mu, q, L > 0$. Note that similar asymptotic results for estimation on the classes $\Theta_{\mu,q}(L)$ were obtained in [11] with the help of another method of estimation suggested in [14]. However, in contrast to the results of [11], which are of the asymptotically-minimax nature, Theorem 1 allows us to control the risk of the suggested method of estimation for all θ .

3. AUXILIARY RESULTS

Put

$$\eta_k = \sum_{i=0}^k \sigma_i (|\xi_i| - \mathbf{E} |\xi_i|),$$

where ξ_i are independent $\mathcal{N}(0, 1)$ random variables and $\sigma_i^2 = w_i - w_{i-1}$. First of all, we are interested in large deviations of η_k .

Lemma 2. For all $x \geq 0$, we have the inequality

$$\mathbf{P}\{\eta_k \geq x\} \leq \exp \left\{ -\frac{x^2}{2w_k} \right\}. \quad (14)$$

Proof. Note that

$$\begin{aligned} \mathbf{E} \exp\{\lambda \sigma_i |\xi_i|\} &= \frac{2}{\sqrt{2\pi}} \int_0^\infty \exp(\lambda \sigma_i x - x^2/2) dx \\ &= \exp\{\lambda^2 \sigma_i^2 / 2\} \left(1 + \frac{2}{\sqrt{2\pi}} \int_0^{\lambda \sigma_i} e^{-u^2/2} du \right) \\ &= \exp\left\{ \frac{\lambda^2 \sigma_i^2}{2} + \log \left[1 + \frac{2}{\sqrt{2\pi}} \int_0^{\lambda \sigma_i} e^{-u^2/2} du \right] \right\} \\ &\leq \exp\left\{ \frac{\lambda^2 \sigma_i^2}{2} + \lambda \sigma_i \mathbf{E} |\xi_1| \right\}. \end{aligned}$$

Therefore, using the exponential Chebyshev inequality, we obtain

$$\mathbf{P}\{\eta_k \geq x\} \leq \exp(-\lambda x) \mathbf{E} \exp\left\{ \lambda \sum_{i=0}^k \sigma_i [|\xi_i| - \mathbf{E} |\xi_i|] \right\} \leq \exp\left\{ -\lambda x + \frac{\lambda^2}{2} \sum_{i=0}^k \sigma_i^2 \right\}.$$

Choosing $\lambda = x/w_k$ in this inequality, we arrive at (14). \triangle

Lemma 3. For all $x \geq 1$ and $0 \leq q \leq 2$, we have

$$\mathbf{E} \eta_k^q \mathbf{1}\{\eta_k > x\sqrt{w_k}\} \leq C w_k^{q/2} x^q e^{-x^2/2}.$$

The proof directly follows from Lemma 2 and the formula of integration by parts. \triangle

Lemma 4. For any integer-valued random variable τ and any $p \in (1/2, 1]$, we have

$$\mathbf{E} \left(\eta_\tau - \sqrt{\frac{w_\tau}{\alpha}} \right) \leq \frac{C}{2p-1} (\mathbf{E} w_\tau^p)^{1/(2p)}, \quad (15)$$

where $\alpha \in (0, 1]$.

Proof. Let us first show that, for any $\mu > 0$, we have the following inequality:

$$\mathbf{E} \max_k \left[\eta_k - \sqrt{\frac{w_k}{\alpha}} - \mu w_k^p \right] \leq C \exp\left\{ \frac{2p}{1-2p} \log \frac{2p}{2p-1} \right\} \mu^{-1/(2p-1)}. \quad (16)$$

Take the integer $K_\mu = \min\{k : \sqrt{w_k} \geq \mu \sqrt{\alpha} w_k^p\}$. Note that we have the trivial inequality

$$\begin{aligned} \mathbf{E} \max_k \left[\eta_k - \sqrt{\frac{w_k}{\alpha}} - \mu w_k^p \right] &\leq \mathbf{E} \max_k \eta_k \mathbf{1}\left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} \\ &\leq \sum_{k=1}^{K_\mu} \mathbf{E} \eta_k \mathbf{1}\left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} + \sum_{k=K_\mu+1}^{\infty} \mathbf{E} \eta_k \mathbf{1}\left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\}. \end{aligned} \quad (17)$$

The first term on the right-hand side of this inequality can be estimated very simply. Using Lemma 3, we get

$$\begin{aligned} \sum_{k=1}^{K_\mu} \mathbf{E} \eta_k \mathbf{1}\left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} &\leq \sum_{k=1}^{K_\mu} w_k^{1/2} \left[\mu w_k^{p-1/2} + \frac{1}{\sqrt{\alpha}} \right] \exp\left\{ -\frac{1}{2} \left[\mu w_k^{p-1/2} + \frac{1}{\sqrt{\alpha}} \right]^2 \right\} \\ &\leq 2\alpha^{-1/2} \exp\left\{ -\frac{1}{2\alpha} \right\} \sum_{k=1}^{K_\mu} w_k^{1/2} \leq 6\alpha^{-1} \exp\left\{ -\frac{1}{2\alpha} \right\} w_{K_\mu}^{1/2} \\ &\leq C\alpha^{-1-1/(2p-1)} \exp\left\{ -\frac{1}{2\alpha} \right\} \mu^{-1/(2p-1)} \\ &= C \exp\left\{ -\frac{1}{2\alpha} + \left(1 + \frac{1}{2p-1} \right) \log \frac{1}{\alpha} \right\} \mu^{-1/(2p-1)}. \end{aligned} \quad (18)$$

It is not difficult to see that

$$\max_{\alpha} \left\{ -\frac{1}{2\alpha} + \left(1 + \frac{1}{2p-1}\right) \log \frac{1}{\alpha} \right\} = -\log \frac{2p}{2p-1} + \frac{2p}{2p-1} \log \frac{4p}{2p-1} \leq \frac{2p}{2p-1} \log \frac{2p}{2p-1}.$$

Therefore, substituting this inequality into the right-hand side of (18), we get

$$\sum_{k=1}^{K_{\mu}} \mathbf{E} \eta_k \mathbf{1} \left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} \leq C \exp \left\{ \frac{2p}{2p-1} \log \frac{2p}{2p-1} \right\} \mu^{-1/(2p-1)}. \quad (19)$$

To estimate the second term on the right-hand side of (17), we again use Lemma 3. Then we have

$$\begin{aligned} \sum_{k=K_{\mu}+1}^{\infty} \mathbf{E} \eta_k \mathbf{1} \left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} &\leq \mu \exp\{-1/(2\alpha)\} \sum_{k=K_{\mu}+1}^{\infty} w_k^p \exp\{-\mu^2 w_k^{2p-1}/2\} \\ &\leq \mu \exp\{-1/(2\alpha)\} \int_0^{\infty} (1+\alpha)^{xp} \exp\{-\mu^2(1+\alpha)^{x(2p-1)}/2\} dx. \end{aligned}$$

Making the change of variables $\mu^2(1+\alpha)^{x(2p-1)} = u$ in the integral, we can continue the above inequality as follows:

$$\sum_{k=K_{\mu}+1}^{\infty} \mathbf{E} \eta_k \mathbf{1} \left\{ \eta_k \geq \sqrt{\frac{w_k}{\alpha}} + \mu w_k^p \right\} \leq \frac{\mu^{-1/(2p-1)}}{\log(1+\alpha)(2p-1)} \int_0^{\infty} u^{(1-p)/(2p-1)} \exp(-u/2) du. \quad (20)$$

The integral on the right-hand side of this inequality is the gamma function, and it can be estimated by the Laplace method. In particular, we have the following inequality:

$$\int_0^{\infty} u^{(1-p)/(2p-1)} \exp(-u/2) du \leq C \exp \left\{ \frac{1}{2p-1} \log \frac{1}{2p-1} \right\}. \quad (21)$$

Inequality (16) automatically follows from (17) and (19)–(21).

After that, inequality (15) can be proved very easily. It suffices to note that, by (16), we have

$$\mathbf{E} \left[\eta_{\tau} - \sqrt{\frac{w_{\tau}}{\alpha}} \right] \leq \mu \mathbf{E} w_{\tau}^p + C \exp \left\{ \frac{2p}{2p-1} \log \frac{2p}{2p-1} \right\} \mu^{1/(2p-1)}.$$

Therefore, choosing

$$\mu = \frac{2p}{2p-1} [\mathbf{E} w_{\tau}^p]^{-(2p-1)/(2p)}$$

in this inequality, we get

$$\mathbf{E} \eta_{\tau} \leq \frac{C}{2p-1} [\mathbf{E} w_{\tau}^p]^{1/(2p)},$$

which proves the required inequality (15). \triangle

We need one more simple fact.

Lemma 5. *For any integer-valued random variable τ and any number $\gamma \in (0, 1]$, we have*

$$\mathbf{E} \eta_{\tau}^{1+\gamma} \mathbf{1} \left\{ \eta_{\tau} \geq \sqrt{(1+\gamma)w_{\tau} \log(w_{\tau})} + \sqrt{\frac{w_{\tau}}{\alpha}} \right\} \leq C. \quad (22)$$

Proof. Let us proceed in the same way as in the proof of Lemma 4. Using Lemma 3, we get

$$\begin{aligned}
& \mathbf{E} \eta_\tau^{1+\gamma} \mathbf{1} \left\{ \eta_\tau \geq \sqrt{(1+\gamma)w_\tau \log(w_\tau)} + \sqrt{\frac{w_\tau}{\alpha}} \right\} \\
& \leq \sum_{k=1}^{\infty} \mathbf{E} \eta_k^{1+\gamma} \mathbf{1} \left\{ \eta_k \geq \sqrt{(1+\gamma)w_k \log(w_k)} + \sqrt{\frac{w_k}{\alpha}} \right\} \\
& \leq C \sum_{k=1}^{\infty} \left[\sqrt{(1+\gamma) \log(w_k)} + \alpha^{-1/2} \right]^{1+\gamma} \exp \left\{ -\sqrt{(1+\gamma) \log(w_k)/\alpha} - \alpha^{-1/2} \right\} \\
& = C \exp[-1/(2\alpha)] \sum_{k=1}^{\infty} \left[\sqrt{(1+\gamma)k \log(1+\alpha)} + 1/(2\sqrt{\alpha}) \right]^{1+\gamma} \exp \left[-\sqrt{k \log(1+\alpha)/\alpha} \right] \\
& \leq C. \quad \triangle
\end{aligned}$$

4. PROOF OF THE THEOREM

Denote for brevity

$$\text{Pen}(k) = \sigma \sqrt{(1+\gamma)w_k \log(w_k)}.$$

Since $L^*(k, \theta)$ is a risk envelope, from (6), (10), and Lemma 1 we derive

$$\begin{aligned}
\mathbf{E}_\theta |\hat{S}(X) - S(\theta)| & \leq \mathbf{E}_\theta \left\{ \sum_{j=K_\gamma+1}^{\infty} |S_j(\theta)| + \text{Pen}(K_\gamma) + \sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} + C\sigma \\
& = \sum_{j=0}^{\infty} |S_j(\theta)| + \mathbf{E}_\theta \left\{ -\sum_{j=0}^{K_\gamma} |Y_j| + \text{Pen}(K_\gamma) + 5\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \\
& \quad + \mathbf{E}_\theta \left\{ \sum_{j=0}^{K_\gamma} [\mathbf{E}_\theta |Y_j| - |S_j(\theta)|] - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \\
& \quad + \mathbf{E}_\theta \left\{ \sum_{j=0}^{K_\gamma} [|Y_j| - \mathbf{E}_\theta |Y_j|] - \sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} + C\sigma. \tag{23}
\end{aligned}$$

Let us use the trivial inequality

$$-s \mathbf{E} |\xi| \leq |x| - \mathbf{E} |x + s\xi| \leq 0, \tag{24}$$

where $s \geq 0$ and $\xi \sim \mathcal{N}(0, 1)$. The validity of this relation can easily be checked if we note that the function

$$\varphi(x) = x - \mathbf{E} |x + s\xi|$$

does not decrease as $x \geq 0$ since its derivative $\varphi'(x) = 1 - \mathbf{E} \text{sign}(x + s\xi)$ is nonnegative.

Using inequality (24), we find

$$\begin{aligned}
\mathbf{E}_\theta \left\{ \sum_{j=0}^{K_\gamma} [\mathbf{E}_\theta |Y_j| - |S_j(\theta)|] - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} & \leq \mathbf{E}_\theta \left\{ \sigma \sum_{j=0}^{K_\gamma} \sqrt{w_{j+1} - w_j} - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \\
& = \mathbf{E}_\theta \left\{ \sigma \sum_{j=0}^{K_\gamma} \sqrt{\alpha(1+\alpha)^{j/2}} - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \\
& \leq \mathbf{E}_\theta \left\{ \sigma \frac{\sqrt{\alpha w_{K_\gamma}}}{\sqrt{1+\alpha} - 1} - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \leq 0. \tag{25}
\end{aligned}$$

Further, the definition of K_γ and inequality (24) directly imply the following inequality:

$$\begin{aligned} \sum_{j=0}^{\infty} |S_j(\theta)| + \mathbf{E}_\theta \left\{ - \sum_{j=0}^{K_\gamma} |Y_j| + \text{Pen}(K_\gamma) + 5\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \\ \leq \sum_{j=0}^{\infty} |S_j(\theta)| + \mathbf{E}_\theta \left\{ - \sum_{j=0}^{k_0} |Y_j| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} \right\} \\ \leq \sum_{j=k_0+1}^{\infty} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}}, \end{aligned} \quad (26)$$

where k_0 is an arbitrary integer.

Thus, it remains to estimate the next-to-last term on the right-hand side of (23). Applying Lemma 4, we immediately obtain

$$\mathbf{E}_\theta \left\{ \sum_{j=0}^{K_\gamma} [|Y_j| - \mathbf{E}_\theta |Y_j|] - \sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \right\} \leq C\sigma\alpha^{-\gamma}\gamma^{-1/2} \left[\mathbf{E}_\theta(w_{K_\gamma})^{(1+\gamma)/2} \right]^{1/(1+\gamma)}. \quad (27)$$

To upper bound the quantity $\mathbf{E}_\theta(w_{K_\gamma})^{(1+\gamma)/2}$, let us use the definition of K_γ . It is obvious that the inequality

$$- \sum_{j=0}^{K_\gamma} |Y_j| + \text{Pen}(K_\gamma) + 5\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \leq - \sum_{j=0}^{k_0} |Y_j| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} \quad (28)$$

holds for any integer k_0 . Then, putting for brevity

$$S_j(\xi) = \sum_{k=w_j}^{w_{j+1}-1} \xi_k,$$

we immediately obtain from (28) that, for $K_\gamma \geq k_0$, the inequalities

$$\begin{aligned} \sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} &\leq \sum_{j=k_0+1}^{K_\gamma} |Y_j| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} - \text{Pen}(K_\gamma) - 4\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \\ &\leq \sum_{j=k_0+1}^{K_\gamma} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} + \sum_{j=k_0+1}^{K_\gamma} |S_j(\xi)| - \text{Pen}(K_\gamma) - 4\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \\ &\leq \sum_{j=k_0+1}^{\infty} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} \\ &\quad + \sum_{j=0}^{K_\gamma} \{ |S_j(\xi)| - \mathbf{E}_\theta |S_j(\xi)| \} - \text{Pen}(K_\gamma) - \sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \end{aligned} \quad (29)$$

are valid. Here, we have used the following fact proved above (see (25)):

$$\sum_{j=0}^{K_\gamma} \mathbf{E}_\theta |S_j(\xi)| - 3\sigma \sqrt{\frac{w_{K_\gamma}}{\alpha}} \leq 0.$$

Further, using the Hölder inequality and Lemma 5, we derive from (29) that

$$\begin{aligned}
& \frac{\sigma}{\sqrt{\alpha}} \left[\mathbf{E}_{\theta}(w_{K_{\gamma}})^{(1+\gamma)/2} \mathbf{1}\{K_{\gamma} \geq k_0\} \right]^{1/(1+\gamma)} \leq \sum_{j=k_0+1}^{\infty} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} \\
& + \left\{ \mathbf{E}_{\theta} \left[\sum_{j=0}^{K_{\gamma}} [|S_j(\xi)| - \mathbf{E}_{\theta} |S_j(\xi)|] \right]^{1+\gamma} \mathbf{1} \left\{ \sum_{j=0}^{K_{\gamma}} [|S_j(\xi)| - \mathbf{E}_{\theta} |S_j(\xi)|] \geq \text{Pen}(K_{\gamma}) + \sigma \sqrt{\frac{w_{K_{\gamma}}}{\alpha}} \right\} \right\}^{1/(1+\gamma)} \\
& \leq \sum_{j=k_0+1}^{\infty} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} + C\sigma. \quad (30)
\end{aligned}$$

Note also that (28) implies, for $K_{\gamma} \leq k_0$, the trivial inequality

$$\sigma \sqrt{\frac{w_{K_{\gamma}}}{\alpha}} \leq \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}}.$$

Therefore, it is evident that the inequality

$$\frac{\sigma}{\sqrt{\alpha}} \left[\mathbf{E}_{\theta}(w_{K_{\gamma}})^{(1+\gamma)/2} \mathbf{1}\{K_{\gamma} \leq k_0\} \right]^{1/(1+\gamma)} \leq \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}}$$

is valid; hence, by (27) and (30), we obtain

$$\mathbf{E}_{\theta} \left\{ \sum_{j=0}^{K_{\gamma}} [|Y_j| - \mathbf{E}_{\theta} |Y_j|] - \sqrt{\frac{w_{K_{\gamma}}}{\alpha}} \right\} \leq C\gamma^{-1}\alpha^{1/2} \left[\sum_{j=k_0+1}^{\infty} |S_j(\theta)| + \text{Pen}(k_0) + 5\sigma \sqrt{\frac{w_{k_0}}{\alpha}} + C\sigma \right].$$

Finally, substituting this inequality and inequalities (25) and (26) into (23), we complete the proof of the theorem. \triangle

In conclusion, the author would like to express his sincere gratitude to the reviewer who carefully read this paper and made a number of important and useful remarks, which allowed the author to correct inaccuracies in proofs.

REFERENCES

1. Vapnik, V.N., *Vosstanovlenie zavisimosti po empiricheskim dannym*, Moscow: Nauka, 1979. Translated under the title *Estimation of Dependences Based on Empirical Data*, New York: Springer, 1982.
2. Devroye, L., Györfi, L., and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, New York: Springer, 1996.
3. Akaike, H., Information Theory and an Extension of the Maximum Likelihood Principle, *Proc. 2nd Int. Symp. on Information Theory, Tsahkadsor, Armenia, USSR, 1971*, Petrov, P.N. and Csaki, F., Eds., Budapest: Akad. Kiado, 1973, pp. 267–281.
4. Mallows, C.V., Some Comments on C_p , *Technometrics*, 1973, vol. 15, pp. 661–675.
5. Green, P.J. and Silverman, B.W., *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman & Hall, 1994.
6. Shibata, R., An Optimal Selection of Regression Variables, *Biometrika*, 1981, vol. 68, pp. 45–54.
7. Kneip, A., Ordered Linear Smoothers, *Ann. Statist.*, 1994, vol. 22, no. 3, pp. 835–866.
8. Golubev, G.K. and Levit, B.Ya., Asymptotically Efficient Estimation in the Wicksell Problem, *Ann. Statist.*, 1998, vol. 26, no. 6, pp. 2407–2419.
9. Ibragimov, I.A. and Khas'minskii, R.Z., On Nonparametric Estimation of the Value of a Linear Functional in White Gaussian Noise, *Teor. Veroyatn. Primen.*, 1984, vol. 29, no. 1, pp. 18–32.

10. Donoho, D.L. and Low, M.G., Renormalization Exponents and Optimal Pointwise Rates of Convergence, *Ann. Statist.*, 1992, vol. 20, no. 2, pp. 944–970.
11. Lepski, O.V. and Levit, B.Ya., Adaptive Minimax Estimation of Infinitely Differentiable Functions, *Math. Methods Statist.*, 1998, vol. 7, no. 2, pp. 123–156.
12. Tsybakov, A., Pointwise and sup-Norm Sharp Adaptive Estimation of Functions on the Sobolev Classes, *Ann. Statist.*, 1998, vol. 26, no. 6, pp. 2420–2469.
13. Cavalier, L., Golubev, G., Lepski, O. and Tsybakov, A., Block Thresholding and Sharp Adaptive Estimation in Severely Ill-Posed Inverse Problems, *Teor. Veroyatn. Primen.*, 2003, vol. 48, no. 3, pp. 534–556.
14. Lepski, O.V., Asymptotically Minimax Adaptive Estimation. I: Upper Bounds. Optimality of Adaptive Estimates, *Teor. Veroyatn. Primen.*, 1991, vol. 36, no. 4, pp. 645–659.