

# ON ORACLE INEQUALITIES RELATED TO HIGH DIMENSIONAL LINEAR MODELS

YURI GOLUBEV\*

**Abstract.** We consider the problem of estimating an unknown vector  $\theta$  from the noisy data  $Y = A\theta + \epsilon$ , where  $A$  is a known  $m \times n$  matrix and  $\epsilon$  is a white Gaussian noise. It is assumed that  $n$  is large and  $A$  is ill-posed. Therefore in order to estimate  $\theta$ , a spectral regularization method is used and our goal is to choose a spectral regularization parameter with the help of the data  $Y$ . We study data-driven regularization methods based on the empirical risk minimization principle and provide some new oracle inequalities related to this approach.

**Key words.** Spectral regularization, excess risk, ordered smoothers, empirical risk minimization principle, oracle inequality.

**AMS(MOS) subject classifications.** Primary 62G05, 62G20.

**1. Introduction and main results.** In this paper, we deal with a classical problem of recovering an unknown vector  $\theta = (\theta(1), \dots, \theta(n))^T \in \mathbb{R}^n$  from the noisy data

$$Y = A\theta + \epsilon, \tag{1.1}$$

where  $A$  is a known  $m \times n$  - matrix and  $\epsilon \in \mathbb{R}^m$  is a white Gaussian noise with a known variance  $\sigma^2 = \mathbf{E}\epsilon^2(k)$ ,  $k = 1, \dots, m$ .

The standard way to estimate  $\theta$  is based on the maximum likelihood estimator

$$\hat{\theta}_0 = \arg \min_{\theta \in \mathbb{R}^n} \|Y - A\theta\|^2,$$

where  $\|y\|^2 = \sum_{k=1}^m x^2(k)$ . It is easy to see that  $\hat{\theta}_0 = (A^T A)^{-1} A^T Y$  and the mean square risk of this estimator is computed as follows

$$\begin{aligned} \mathbf{E}\|\hat{\theta}_0 - \theta\|^2 &= \mathbf{E}\|(A^T A)^{-1} A^T \epsilon\|^2 = \sigma^2 \text{trace}[(A^T A)^{-1}] \\ &= \sigma^2 \sum_{k=1}^n \lambda_k, \end{aligned} \tag{1.2}$$

where  $\lambda_k$  and  $\phi_k \in \mathbb{R}^n$  are the eigenvalues and the eigenfunctions of  $(A^T A)^{-1}$ :

$$\lambda_k A^T A \phi_k = \phi_k.$$

In what follows, it is assumed that  $A$  is ill-posed i.e.,  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ . The equation (1.2) reveals the principal difficulty in  $\hat{\theta}_0$ : *its risk may be*

---

\*CNRS and Institute for Problems of Information Transmission CMI, 39 rue F. Joliot-Curie, 13453 Marseille, France.

very large when  $n$  is large or when  $A$  has a large condition number. In this paper, we suppose that  $n$  is large (it may be infinity), so the risk of  $\hat{\theta}_0$  is also large.

The basic method to improve  $\hat{\theta}_0$  is to make the variance  $\sigma^2 \sum_{k=1}^n \lambda_k$  smaller by suppressing large  $\lambda_k$ . The simplest way to implement this idea is to smooth  $\hat{\theta}_0$  with the help of a properly chosen  $n \times n$  - matrix  $H$  i.e., using a new estimator  $H\hat{\theta}_0$ . In this paper, we focus on with the following family of linear estimators

$$\hat{\theta}_\alpha = H_\alpha \hat{\theta}_0 = H_\alpha [(A^\top A)^{-1}] (A^\top A)^{-1} A^\top Y,$$

where  $H_\alpha(z)$  is an analytic function  $H_\alpha(z) = \sum_{k=0}^{\infty} h_\alpha(k) z^k$  such that  $\lim_{\alpha \rightarrow 0} H_\alpha(z) = 1$ ,  $\lim_{z \rightarrow \infty} H_\alpha(z) = 0$ . This method is called *spectral regularization* (see [3]). The regularization parameter  $\alpha$  controls the quality of  $\hat{\theta}_\alpha$ . Indeed, with an elementary algebra we get the standard bias-variance decomposition

$$\mathbb{E} \|\hat{\theta}_\alpha - \theta\|^2 = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k), \quad (1.3)$$

where  $\psi_k = A\phi_n / \|A\phi_k\|$  and  $\langle \theta, \psi_k \rangle = \sum_{l=1}^n \theta(l) \psi_k(l)$ . It is clear that the spectral regularization may substantially improve  $\hat{\theta}_0$  when  $\langle \theta, \psi_k \rangle^2$  is small for large  $k$ .

In practice, a good choice of  $H_\alpha(\cdot)$  is a delicate problem related to the numerical complexity  $\hat{\theta}_\alpha$ . For instance, to make use of the spectral cut-off regularization with  $H_\alpha(z) = \mathbf{1}\{\alpha z \leq 1\}$ , one has to compute the singular value decomposition (SVD) of  $A$ . For large  $n$  this numerical problem may be difficult or even infeasible.

The very popular Tikhonov's [7] regularization is defined by

$$\hat{\theta}_\alpha = \arg \min_{\theta} \left\{ \|Y - A\theta\|^2 + \alpha \|\theta\|^2 \right\}.$$

For this method we have  $H_\alpha(z) = 1/(1 + \alpha z)$ . Notice here that this regularization technique is good if  $A$  is really ill-posed. Indeed in view of (1.3),  $\hat{\theta}_\alpha$  may improve the banal estimator  $\hat{\theta}_0$  if

$$\sum_{k=1}^n \frac{\lambda_k}{[1 + \alpha \lambda_k]^2} \ll \sum_{k=1}^n \lambda_k.$$

This means for instance that for inverse problems with  $\lambda_k \approx 1$  Tikhonov's regularization makes no sense.

Another widespread regularization technique is due to Landweber. This method is based on a very simple idea: to find recursively a root of equation

$$A^\top Y = A^\top A\theta.$$

Notice that for positive  $a$  we can write  $A^\top Y = [A^\top A - aI]\theta + a\theta$ , or equivalently  $\theta = [I - a^{-1}A^\top A]\theta + a^{-1}A^\top Y$ . This formula motivates Landweber's iterations defined by

$$\theta_i = [I - a^{-1}A^\top A]\theta_{i-1} + a^{-1}A^\top Y.$$

It is easy to see that these iterations converge if  $a\lambda_1 < 1$ . It is also easy to check that

$$H_i(z) = 1 - \left(1 - \frac{1}{az}\right)^{i+1} \quad (1.4)$$

In spite of its iterative character, the numerical complexity of Landweber's iterations may be very high. Indeed, when the noise is small,  $H_i(z)$  should be 1, and (1.4) results in

$$i > \text{cond}(A) \stackrel{\text{def}}{=} \frac{\lambda_n}{\lambda_1}$$

So, if  $A$  is severely ill-posed, the number of iterations may be very large, thus making the method infeasible. A substantial improvement of Landweber's iterations is provided with the  $\nu$ -method (see e.g., [3]).

Whatever an inversion method is used, the principal question is how to choose its regularization parameter. Intuitively, (see (1.3)), this parameter should minimize in some sense the risk

$$L[\alpha, \theta] = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k).$$

In statistics, there are two main ideas to formalize this optimization problem:

- to assume that  $\theta$  belongs to a known set  $\Theta \in \mathbb{R}^n$  and to take

$$\alpha^* = \arg \min_{\alpha} \sup_{\theta \in \Theta} L[\alpha, \theta]$$

- to construct based on the data an "estimate"  $\hat{L}[\alpha, Y]$  of  $L[\alpha, \theta]$  and to compute

$$\hat{\alpha} = \arg \min_{\alpha} \hat{L}[\alpha, Y]$$

Statistical literature related to these approaches is so vast that it would be impractical to cite it here. We refer interested reader to [6] and [2] as typical representatives of its. Notice that the first approach is related to the theory of minimax estimation [4].

This paper focuses on the second approach, namely, it deals with data-driven regularization parameters computed with help of the empirical risk

minimization principle. This method says that the regularization parameter should be computed as follows

$$\hat{\alpha} = \arg \min_{\alpha} R_{Pen}[Y, \alpha],$$

where

$$R_{Pen}[Y, \alpha] = \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 + Pen(\alpha)\sigma^2 - \sigma^2 \sum_{k=1}^n \lambda_k,$$

and  $Pen(\cdot)$  is a given function  $\mathbb{R}^+ \rightarrow \mathbb{R}^+$ . A heuristic motivation behind this method is rather transparent. Indeed, the best regularization parameter is obviously given by

$$\alpha^* = \arg \min_{\alpha} \|\theta - \hat{\theta}_\alpha\|^2. \quad (1.5)$$

Evidently,  $\alpha^*$  cannot be used since it depends on  $\theta$  which is unknown. So, the first idea is replace  $\theta$  in (1.5) by  $\hat{\theta}_0$ . It is clear, that directly this idea doesn't work because  $\min_{\alpha} \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 = 0$ . Therefore we need to correct  $\|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2$  by an additional term, thus arriving at  $R_{Pen}[Y, \hat{\alpha}]$ . Intuitively, this idea assumes that the best  $Pen(\alpha)$  should be a minimal function such that uniformly in  $\theta \in \mathbb{R}^n$

$$\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \lesssim \mathbf{E}R_{Pen}[Y, \hat{\alpha}]. \quad (1.6)$$

Unfortunately, the mathematical formalization of “the best penalty” and (1.6) is a very delicate problem. We refer interested readers to [1], which provides a reasonable approach to this formalization.

In this paper, we assume that the penalty is given and our goal is to bound from above  $\mathbf{E}_\theta \|\hat{\theta}_{\hat{\alpha}} - \theta\|^2$ . The simplest way to analyze this risk is to use SVD. Let  $\lambda_k$  and  $\phi_k$  be the eigenfunctions and eigenvectors of  $(A^\top A)^{-1}$ . Denoting  $\psi_k = A\phi_k / \|A\phi_k\|$ , one checks easily

$$y(k) \stackrel{\text{def}}{=} \langle Y, \psi_k \rangle \sqrt{\lambda_k} = \langle \theta, \psi_k \rangle + \sigma \xi(k) \sqrt{\lambda_k}, \quad (1.7)$$

where  $\xi(k)$  are i.i.d.  $\mathcal{N}(0,1)$ . Notice that  $\hat{\theta}_\alpha$  admits the following representation

$$\langle \hat{\theta}_\alpha, \psi_k \rangle = H_\alpha(\lambda_k) y(k)$$

and

$$\|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 y^2(k) \quad (1.8)$$

$$\|\theta - \hat{\theta}_\alpha\|^2 = \sum_{k=1}^n [\theta(k) - H_\alpha(\lambda_k) y(k)]^2. \quad (1.9)$$

We have already mentioned that the main idea in the empirical risk minimization is to control  $\mathbf{E}_\theta \|\hat{\theta}_{\hat{\alpha}} - \theta\|^2$  with the help of  $\mathbf{E}_\theta R_{Pen}[Y, \hat{\alpha}]$ . Mathematically this idea can be expressed as follows. For any  $\mu > 0$  and a given penalty  $Pen(\cdot)$  define the excess risk by

$$\Delta_{Pen}(\mu) = \sup_{\theta \in \mathbb{R}^n} \left\{ \mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 - (1 + \mu) \mathbf{E}_\theta R_{Pen}[Y, \hat{\alpha}] \right\} \quad (1.10)$$

and we will say that the penalty is admissible if  $\Delta_{Pen}(\mu) < \infty$  for any  $\mu > 0$ .

If  $Pen(\alpha)$  is admissible, then we obtain immediately

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq (1 + \mu) \inf_{\alpha} \bar{R}_{Pen}[\theta, \alpha] + \Delta_{Pen}(\mu), \quad (1.11)$$

where

$$\begin{aligned} \bar{R}_{Pen}[\theta, \alpha] &= \mathbf{E}_\theta R_{Pen}[Y, \alpha] = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2 \\ &\quad + \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) + \sigma^2 Pen(\alpha) - 2\sigma^2 \sum_{k=1}^n \lambda_k H_\alpha(\lambda_k). \end{aligned}$$

This equation can be interpreted as a bias-variance decomposition of  $\hat{\theta}_{\hat{\alpha}}$  related to the empirical risk. The empirical bias term is given by  $\sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2$  and it coincides with the standard one. However the variance term differs from  $\sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k)$  and it is computed as follows

$$\Sigma_{Pen}(\alpha) = \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) + \sigma^2 Pen(\alpha) - 2\sigma^2 \sum_{k=1}^n \lambda_k H_\alpha(\lambda_k).$$

The inequality (1.11) can be rewritten in the form of an oracle inequality

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq r_{Pen}[\theta] + \inf_{\mu} \left\{ \mu r_{Pen}[\theta] + \Delta_{Pen}(\mu) \right\},$$

where  $r_{Pen}[\theta] = \inf_{\alpha} \bar{R}_{Pen}[\theta, \alpha]$  is the oracle risk. Thus, to control the risk of our data-driven method we need to compute the excess risk. Notice that when  $n$  is large the exact computation of the excess risk is infeasible: indeed, for given  $\theta$  with the Monte-Carlo method we can compute

$$D(\theta, \mu) = \mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 - (1 + \mu) \mathbf{E}_\theta R_{Pen}[Y, \hat{\alpha}]$$

but we cannot maximize this function numerically over  $\mathbb{R}^n$  for large  $n$ .

In order to overcome this difficulty, let us introduce

$$\begin{aligned} \Delta_{Pen}^C(\mu) &\stackrel{\text{def}}{=} \mathbf{E}_0 \sup_{\alpha} \left\{ 2(1 + \mu) \sum_{k=1}^n \xi^2(k) \lambda_k H_\alpha(\lambda_k) \right. \\ &\quad \left. - \mu \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \xi^2(k) + \frac{C \max_k \lambda_k H_\alpha^2(\lambda_k)}{\mu} - (1 + \mu) Pen(\alpha) \right\}. \end{aligned} \quad (1.12)$$

Notice that in contrast to the excess risk,  $\Delta_{Pen}^C(\mu)$  can be computed by the Monte-Carlo method, and we will show that for a sufficiently large class of spectral regularization methods  $\Delta_{Pen}(\mu) \leq \sigma^2 \Delta_{Pen}^C(\mu)$ . These regularization methods (smoothers) are called *ordered smoothers*. They were firstly introduced in [5].

DEFINITION 1.1. *The family of smoothers  $\{H_\alpha(\cdot), \alpha \geq 0\}$  is called ordered if:*

1. for all  $\alpha \geq 0$  and  $\lambda \geq 0$ ,  $0 \leq H_\alpha(\lambda) \leq 1$
2.  $H_{\alpha_1}(\lambda) \geq H_{\alpha_2}(\lambda)$ , for all  $\alpha_1 \leq \alpha_2$  and all  $\lambda > 0$ .

Typical examples of ordered smoothers are provided the Tikhonov regularization, the spectral cut-off method, the Landweber iterations.

The main result of this paper is given by the following theorem

THEOREM 1.1. *Let  $\{H_\alpha(\cdot), \alpha \geq 0\}$  be a family of ordered smoothers. Then for some  $C > 0$  and for all  $\mu > 0$*

$$\Delta_{Pen}(\mu) \leq \sigma^2 \Delta_{Pen}^C(\mu).$$

Let us illustrate how this theorem works. Suppose

$$Pen(\alpha) = \underline{Pen}(\alpha) = 2 \sum_{k=1}^n H_\alpha(\lambda_k) \lambda_k.$$

It is well known that this penalty is related to the unbiased risk estimation, since for given  $\alpha$

$$\mathbf{E}_\theta \|\theta - \hat{\theta}_\alpha\|^2 = \bar{R}_{\underline{Pen}}[\theta, \alpha] = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k).$$

Assume also that  $A$  is not severely ill-posed. More precisely, suppose there exists  $\kappa < 1$  such that for all  $\alpha \geq 0$

$$\max_k \lambda_k H_\alpha^2(\lambda_k) \leq \lambda_1 \left[ \frac{1}{\lambda_1} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right]^\kappa, \quad (1.13)$$

$$\sum_{k=1}^n \lambda_k^2 H_\alpha^2(\lambda_k) \leq \frac{\lambda_1^2}{1 - \kappa} \left[ \frac{1}{\lambda_1} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right]^{1+\kappa}. \quad (1.14)$$

It is easy to see that these conditions allow only a polynomial growth of  $\lambda_k$ . Indeed, if

$$\frac{\lambda_k}{\lambda_1} = k^m \quad \text{for some } m \in [0, \infty),$$

then for the spectral cut-off with  $H_\alpha(\lambda) = \mathbf{1}\{\alpha\lambda \leq 1\}$  it is easy to check that  $\kappa = m/(m+1)$ .

THEOREM 1.2. *Let  $\{H_\alpha(\cdot), \alpha \geq 0\}$  be a family of ordered smoothers and (1.13-1.14) hold true, then for some  $C > 1$  and any  $\mu \in (0, 1)$*

$$\Delta_{\underline{Pen}}(\mu) \leq \frac{C^{1/(1-\kappa)} \lambda_1 \sigma^2}{(1-\kappa)^{1/(1-\kappa)} \mu^{(1+\kappa)/(1-\kappa)}}.$$

This theorem improves oracle inequalities from [2] over the classes of ordered smoothers. It says that the upper bound for the excess risk doesn't depend on  $n$ . Moreover, it allows the minimization of the empirical risk over all possible regularization parameters, thus showing that a priori restrictions on  $\alpha$  are not essential. On the other hand, it reveals some serious difficulties related to the penalty  $\underline{Pen}(\alpha)$ : the upper bound for excess risk explodes as  $\kappa \rightarrow 1$ . It means that this penalty is not good for severely ill-posed inverse problems. More details about this effect along with an improved penalty can be found in [1].

## 2. Proofs.

**2.1. Ordered processes and their properties.** Our method of deriving oracle inequalities is related to a special class of random processes. Let  $\xi(t)$ ,  $t \geq 0$  be a random process with  $\mathbf{E}\xi(t) = 0$  and a finite variance  $\mathbf{E}\xi^2(t) = \sigma^2(t)$  which is assumed to be monotone

$$\sigma^2(t_2) \geq \sigma^2(t_1), \quad t_2 \geq t_1.$$

The process  $\xi(t)$ ,  $t \geq 0$ , is called *ordered* if it is separable and for all  $t_2 \geq t_1$

$$\mathbf{E}[\xi(t_2) - \xi(t_1)]^2 \leq \sigma^2(t_2) - \sigma^2(t_1). \quad (2.1)$$

Obviously, this condition can be rewritten as

$$\mathbf{E}\xi(t_2)\xi(t_1) \geq \min\{\mathbf{E}\xi^2(t_2), \mathbf{E}\xi^2(t_1)\}.$$

So, an ordered process can be viewed as a natural generalization of the Wiener process  $W(t)$  for which  $\mathbf{E}W(t_1)W(t_2) = \min\{\mathbf{E}W^2(t_1), \mathbf{E}W^2(t_2)\}$ .

Denote for brevity

$$\Delta_\xi(t_1, t_2) = \frac{\xi(t_1) - \xi(t_2)}{\sqrt{\mathbf{E}[\xi(t_1) - \xi(t_2)]^2}}.$$

The main property of ordered processes is given by the following lemma.

LEMMA 2.1. *Suppose there exists  $\lambda > 0$  such that*

$$\varphi(\lambda) \stackrel{\text{def}}{=} \sup_{t_1, t_2} \mathbf{E} \cosh[\lambda \Delta_\xi(t_1, t_2)] < \infty. \quad (2.2)$$

*Then there exists a constant  $C$  depending on  $\lambda$  such that for all  $T > 0$  and all  $p \geq 1$*

$$\left[ \mathbf{E} \sup_{t, s \in [0, T]} |\xi(t) - \xi(s)|^p \right]^{1/p} \leq Cp\sigma(T), \quad (2.3)$$

where  $C$  is a generic constant.

*Proof.* We provide the proof of (2.3) only for reader's convenience. In fact, its proof is standard and it is based on the classical chaining arguments [8]. For simplicity, we assume that  $\sigma^2(t)$  is a continuous function. Then for a given integer  $s \geq 0$  we can find points  $t_k^s$  on  $[0, T]$  such that

$$\sigma^2(t_k^s) = 2^{-s} k \sigma^2(T), \quad k = 0, \dots, 2^s - 1$$

and denote by  $\mathcal{T}^s$  the set of these points. Let  $u$  be an arbitrary point in  $\mathcal{T}^s$ . Then we can find a chain, i.e. the points  $\tau_j(u) \in \mathcal{T}^j$ ,  $j = 0, \dots, s$  such that

1.  $u = \tau_s(u)$ ,  $0 = \tau_0(u)$
2.  $|\sigma^2(\tau_j(u)) - \sigma^2(\tau_{j-1}(u))| \leq 2^{-j+1} \sigma^2(T)$ .

To verify that such points exist, one can imagine the standard binary tree with the nodes at the level  $j$  associated with the points  $t_l^j$ ,  $l = 0, \dots, 2^j - 1$ . It is clear that there exists a unique way connecting  $u \in \mathcal{T}^s$  and 0 (top of the tree). This way passes via nodes, which are denoted by  $\tau_k(u)$ . So, we can write

$$u = \sum_{k=0}^{s-1} [\tau_{k+1}(u) - \tau_k(u)],$$

and for arbitrary points  $u, v$  in  $\mathcal{T}^s$  we get

$$u - v = \sum_{k=0}^{s-1} [\tau_{k+1}(u) - \tau_k(u)] - \sum_{k=0}^{s-1} [\tau_{k+1}(v) - \tau_k(v)].$$

Therefore in view of (2.2), we have

$$\begin{aligned} & \mathbf{E}^{1/p} \sup_{u, v \in \mathcal{T}^s} |\xi(u) - \xi(v)|^p \\ & \leq 2 \sum_{k=0}^{s-1} \left[ \mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\xi(\tau_{k+1}(u)) - \xi(\tau_k(u))|^p \right]^{1/p} \\ & = 2 \sum_{k=0}^{s-1} \left[ \mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \right. \\ & \quad \left. \times [\mathbf{E}[\xi(\tau_{k+1}(u)) - \xi(\tau_k(u))]^2]^{p/2} \right]^{1/p} \\ & \leq 2\sigma(T) \sum_{k=0}^{s-1} 2^{-k} \left[ \mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \right]^{1/p}. \end{aligned} \tag{2.4}$$

Let  $L(x) = \log^p(x + e^{p-1})$ . Notice that for any  $p \geq 1$  this function is convex on  $(0, \infty)$ , since

$$L''(x) = \frac{p \log^{p-2}(x + e^{p-1})}{(x + e^{p-1})^2} [p - 1 - \log(x + e^{p-1})] \leq 0.$$

Therefore, using convexity of  $L(x)$  and (2.2), we obtain

$$\begin{aligned} & \mathbf{E}^{1/p} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \\ & \leq \frac{1}{\lambda} L \left[ \sum_{u \in \mathcal{T}^{k+1}} \mathbf{E} e^{\lambda |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|} \right] \\ & \leq \frac{1}{\lambda} \log [2^{k+2} \varphi(\lambda) + e^{p-1}] = \frac{(k+2) \log(2) + p - 1}{\lambda} + \frac{\log[\varphi(\lambda)]}{\lambda}. \end{aligned}$$

Substituting this in (2.4), we arrive at the inequality

$$\left[ \mathbf{E} \sup_{u, v \in \mathcal{T}^s} |\xi(u) - \xi(v)|^p \right]^{1/p} \leq C \sigma(T)$$

which proves the lemma, since by separability of  $\xi(t)$

$$\left[ \mathbf{E} \sup_{u, v \in [0, T]} |\xi(u) - \xi(v)|^p \right]^{1/p} = \limsup_{s \rightarrow \infty} \left[ \mathbf{E} \sup_{u, v \in \mathcal{T}^s} |\xi(u) - \xi(v)|^p \right]^{1/p}. \quad \square$$

Lemma 2.1 almost immediately results in the following fact:

LEMMA 2.2. *Let  $\xi(t)$  be an ordered process satisfying (2.1, 2.2) and such that  $\xi(0) = 0$ . Then there exists a constant  $C$  depending on  $\lambda$  such that for all  $\gamma > 0$*

$$\mathbf{E} \sup_{t \geq 0} [\xi(t) - \gamma \sigma^q(t)]_+^p \leq \frac{C [2q(p+2) - 4]^{q(p+2)-2}}{\gamma^{p/(q-1)}}, \quad (2.5)$$

where  $[x]_+ = \max(0, x)$ .

*Proof.* We will use the following form of the Markov inequality

$$\mathbf{E} \eta^p \mathbf{1}\{\eta > x\} \leq \frac{\mathbf{E} |\eta|^{p+d}}{x^d} \quad (2.6)$$

which immediately results from the banal inequality

$$\eta^p \mathbf{1}\{\eta > x\} \leq |\eta|^p |\eta/x|^d.$$

Without loss of generality, we may assume that  $\sigma^2(t)$  is continuous and such that  $\lim_{t \rightarrow \infty} \sigma^2(t) = \infty$ . Then for any integer  $k \geq 0$  we can find  $t_k(\gamma)$  such that

$$\sigma^{q-1}(t_k(\gamma)) = \frac{k}{\gamma}.$$

Using that  $f(x) = x^p \mathbf{1}\{x > x_0\}$  is monotone in  $x > 0$ , we have

$$\begin{aligned}
& \mathbf{E} \sup_{t \geq 0} [\xi(t) - \gamma \sigma^q(t)]_+^p \\
& \leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\gamma), t_{k+1}(\gamma)]} \xi^p(t) \mathbf{1}\{\xi(t) \geq \gamma \sigma^q(t)\} \\
& \leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\gamma), t_{k+1}(\gamma)]} \xi^p(t) \mathbf{1}\{\xi(t) \geq \gamma \sigma^q(t_k(\gamma))\} \\
& \leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\gamma), t_{k+1}(\gamma)]} \xi^p(t) \mathbf{1}\left\{ \sup_{t \in [t_k(\gamma), t_{k+1}(\gamma)]} \xi(t) \geq \gamma \sigma^q(t_k(\gamma)) \right\} \\
& \leq \mathbf{E} \sup_{0 \leq t \leq t_1(\gamma)} |\xi(t)|^p \\
& \quad + \sum_{k=1}^{\infty} \mathbf{E} \sup_{0 \leq t \leq t_{k+1}(\gamma)} \xi^p(t) \mathbf{1}\left\{ \sup_{0 \leq t \leq t_{k+1}(\gamma)} \xi(t) \geq \gamma \sigma^q(t_k(\gamma)) \right\}.
\end{aligned} \tag{2.7}$$

By Lemma 2.1, the first term at the right-hand side of the above inequality is bounded as follows

$$\mathbf{E} \sup_{0 \leq t \leq t_1(\gamma)} |\xi(t)|^p \leq Cp^p, \sigma^p(t_1(\gamma)) = \frac{Cp^p}{\gamma^{p/(q-1)}} \tag{2.8}$$

whereas the second one, in view of (2.6), is controlled by

$$\begin{aligned}
& \sum_{k=1}^{\infty} \mathbf{E} \sup_{0 \leq t \leq t_{k+1}(\gamma)} \xi^p(t) \mathbf{1}\left\{ \sup_{0 \leq t \leq t_{k+1}(\gamma)} \xi(t) \geq \gamma \sigma^q(t_k(\gamma)) \right\} \\
& \leq C(p+d)^{p+d} \sum_{k=1}^{\infty} \frac{\sigma^{p+d}(t_{k+1}(\gamma))}{[\gamma \sigma^q(t_k(\gamma))]^d} = \frac{C(p+d)^{p+d}}{\gamma^{p/(q-1)}} \sum_{k=1}^{\infty} \frac{(k+1)^{p+d}}{k^{qd/(q-1)}} \\
& \leq \frac{C[2(p+d)]^{p+d}}{\gamma^{p/(q-1)}} \sum_{k=1}^{\infty} \frac{1}{k^{d/(q-1)-p}}.
\end{aligned}$$

Setting  $d = (q-1)(p+2)$  in the above inequality and using (2.7) together with (2.8), we prove (2.5).  $\square$

**2.2. Some examples of ordered processes.** The simplest example of an ordered process is  $\xi(t) = \xi t$ , where  $\xi$  is a zero mean random variable with a finite exponential moment  $\mathbf{E} \cosh(\lambda \xi) < \infty$  for some  $\lambda > 0$ . As we have already mentioned, the Wiener process  $W(t)$  is an ordered process. At the first glance,  $\xi t$  and  $W(t)$  are quite different, but from the viewpoint of Lemma 2.2 they are equivalent. Of course, the distribution of  $\max_{t \geq 0} [W(t) - \gamma t]$  is well-known

$$\mathbf{P}\left\{ \max_{t \geq 0} [W(t) - \gamma t] \geq x \right\} = \exp(-2\gamma x).$$

The next two examples play an essential role in adaptive estimation. Let  $H_t(\cdot)$  be a family of ordered smoothers (see Definition 1.1). Consider the following Gaussian processes

$$\begin{aligned}\xi_+(t) &= \sum_{k=1}^n [H_{t_0}(\lambda_k) - H_{t_0+t}(\lambda_k)] b_k \xi(k), \quad t \geq 0 \\ \xi_-(t) &= \sum_{k=1}^n [H_{t_0}(\lambda_k) - H_{t_0-t}(\lambda_k)] b_k \xi(k), \quad 0 \leq t \leq t_0,\end{aligned}$$

where  $\xi(k)$  are i.i.d.  $\mathcal{N}(0, 1)$  and  $\sum_{i=1}^{\infty} b_i^2 < \infty$ . It is easy to see that  $\xi_+(t)$  and  $\xi_-(t)$  are ordered processes. Indeed, in view of (1.12) we have for  $t_2 \geq t_1$

$$\begin{aligned}\mathbf{E}\xi_+^2(t_1) &= \sum_{k=1}^n [H_{t_0}(\lambda_k) - H_{t_0+t_1}(\lambda_k)][H_{t_0}(\lambda_k) - H_{t_0+t_1}(\lambda_k)] b_k^2 \\ &\leq \sum_{k=1}^n [H_{t_0}(\lambda_k) - H_{t_0+t_1}(\lambda_k)][H_{t_0}(\lambda_k) - H_{t_0+t_2}(\lambda_k)] b_k^2 \\ &= \mathbf{E}\xi_+(t_1)\xi_+(t_2),\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbf{E}\xi_+^2(t_1) &= \sum_{k=1}^n [H_{t_0-t_1}(\lambda_k) - H_{t_0}(\lambda_k)][H_{t_0-t_1}(\lambda_k) - H_{t_0}(\lambda_k)] b_k^2 \\ &\leq \sum_{k=1}^n [H_{t_0-t_1}(\lambda_k) - H_{t_0}(\lambda_k)][H_{t_0-t_2}(\lambda_k) - H_{t_0}(\lambda_k)] b_k^2 \\ &= \mathbf{E}\xi_+(t_1)\xi_+(t_2).\end{aligned}$$

Therefore with Lemma 2.2 we get

$$\begin{aligned}\mathbf{E} \sup_{\alpha \geq \alpha_0} \left[ \xi_+(\alpha) - \gamma \sum_{k=1}^n [H_{\alpha_0}(\lambda_k) - H_{\alpha}(\lambda_k)]^2 b_k^2 \right]_+^p &\leq \frac{C(p)}{\gamma^p}, \\ \mathbf{E} \sup_{\alpha \leq \alpha_0} \left[ \xi_-(\alpha) - \gamma \sum_{k=1}^n [H_{\alpha_0}(\lambda_k) - H_{\alpha}(\lambda_k)]^2 b_k^2 \right]_+^p &\leq \frac{C(p)}{\gamma^p},\end{aligned}$$

thus arriving at

LEMMA 2.3. *Let  $\{H_{\alpha}(\cdot), \alpha \geq 0\}$  be a family of ordered smoothers, then for any  $\gamma > 0$*

$$\begin{aligned}\mathbf{E} \sup_{\alpha \geq 0} \left[ \sum_{k=1}^n [H_{\alpha_0}(\lambda_k) - H_{\alpha}(\lambda_k)] b_k \xi(k) \right. \\ \left. - \gamma \sum_{k=1}^n [H_{\alpha_0}(\lambda_k) - H_{\alpha}(\lambda_k)]^2 b_k^2 \right]_+^p &\leq \frac{C(p)}{\gamma^p}.\end{aligned}\tag{2.9}$$

The next important ordered process is defined by

$$\eta(t) = \sum_{k=1}^n H_{1/t}(\lambda_k)(\xi^2(k) - 1),$$

where  $\xi(k)$  are i.i.d.  $\mathcal{N}(0, 1)$  and  $\{H_t(\cdot), t \geq 0\}$  is a family of ordered smoothers. It is easy to check that

$$\mathbf{E}\eta_2^2(t_1) \leq \mathbf{E}\eta_2(t_2)\eta_2(t_1), \quad t_1 \leq t_2.$$

So, in order to apply Lemma 2.2, it remains to check (2.2). Denoting for brevity

$$\|H_{\alpha_2} - H_{\alpha_1}\|^2 = \sum_{k=1}^n [H_{\alpha_2}(\lambda_k) - H_{\alpha_1}(\lambda_k)]^2,$$

we have

$$\begin{aligned} & \mathbf{E} \exp[\lambda \Delta_\xi(\alpha_2, \alpha_1)] \\ &= \exp\left[-\frac{\lambda}{\sqrt{2}\|H_{\alpha_2} - H_{\alpha_1}\|} \sum_{k=1}^n [H_{\alpha_2}(\lambda_k) - H_{\alpha_1}(\lambda_k)] \right. \\ & \quad \left. - \frac{1}{2} \sum_{k=1}^n \log\left(1 - \sqrt{2}\lambda \frac{H_{\alpha_2}(\lambda_k) - H_{\alpha_1}(\lambda_k)}{\|H_{\alpha_2} - H_{\alpha_1}\|}\right)\right]. \end{aligned} \quad (2.10)$$

Since obviously

$$\max_k [H_{\alpha_2}(\lambda_k) - H_{\alpha_1}(\lambda_k)] \leq \|H_{\alpha_2} - H_{\alpha_1}\|,$$

then using the Taylor expansion for  $\log(1 - \cdot)$  at the right-hand side of (2.10), we get for  $\lambda \leq 1/2$

$$\mathbf{E} \exp[\lambda \Delta_\xi(\alpha_2, \alpha_1)] \leq \exp(C\lambda^2),$$

thus proving (2.2). Therefore using Lemma 2.2, we obtain the following fact.

LEMMA 2.4. *Let  $\{H_\alpha(\cdot), \alpha \geq 0\}$  be a family of ordered smoothers, then for all  $\gamma > 0$*

$$\mathbf{E} \sup_{\alpha > 0} \left[ \sum_{k=1}^{\infty} H_\alpha(\lambda_k) [\xi^2(k) - 1] - \gamma \left( 2 \sum_{k=1}^{\infty} H_\alpha^2(\lambda_k) \right)^q \right]_+^p \leq \frac{C(p)}{\gamma^{p/(1-q)}}. \quad (2.11)$$

**2.3. Proof of Theorem 1.1.** Denote for brevity  $\theta_k = \langle \theta, \phi_k \rangle$ . We begin with a simple auxiliary lemma that is cornerstone for the proof.

LEMMA 2.5. *Let  $\{H_\alpha(\cdot), \alpha \geq 0\}$  be a family of ordered smoothers. Then there exists a constant  $C$  such that for any data-driven smoothing parameter  $\hat{\alpha}$*

$$\begin{aligned} & \left[ \mathbf{E}_\theta \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \right]^2 \\ & \leq C \mathbf{E}_\theta \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \mathbf{E}_\theta \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2. \end{aligned} \quad (2.12)$$

and

$$\begin{aligned} & \left[ \mathbf{E}_\theta \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \sqrt{\lambda_k} \theta_k \xi(k) \right]^2 \\ & \leq C \mathbf{E}_\theta \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \mathbf{E}_\theta \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2. \end{aligned} \quad (2.13)$$

*Proof.* Let  $\alpha_0$  be a given smoothing parameter. We obviously have

$$\begin{aligned} & \mathbf{E}_\theta \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \\ & = \mathbf{E}_\theta \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k). \end{aligned} \quad (2.14)$$

It follows immediately from (2.9) that

$$\begin{aligned} & \left| \mathbf{E}_\theta \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \right| \\ & \leq \gamma \mathbf{E}_\theta \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)]^2 \lambda_k \theta_k^2 + \frac{C}{\gamma}. \end{aligned}$$

Therefore minimizing the right-hand side in  $\gamma$ , we obtain

$$\begin{aligned} & \left| \mathbf{E}_\theta \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)] \lambda_k \theta_k \xi(k) \right| \\ & \leq C \left\{ \mathbf{E}_\theta \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)]^2 \lambda_k \theta_k^2 \right\}^{1/2}. \end{aligned} \quad (2.15)$$

To bound from above the right-hand side at the above display, we use once again that  $H_\alpha(\cdot)$  are ordered smoothers. So, when  $\hat{\alpha} \leq \alpha_0$  we obtain

$$\begin{aligned} \sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)]^2 \lambda_k^2 \theta_k^2 &= \sum_{k=1}^{\infty} H_{\alpha_0}^2(\lambda_k) \left[1 - \frac{H_{\hat{\alpha}}(\lambda_k)}{H_{\alpha_0}(\lambda_k)}\right]^2 \lambda_k^2 \theta_k^2 \\ &\leq \max_k \lambda_k H_{\alpha_0}^2(\lambda_k) \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 \end{aligned}$$

and similarly for  $\hat{\alpha} \geq \alpha_0$

$$\sum_{k=1}^{\infty} [H_{\alpha_0}(\lambda_k) - H_{\hat{\alpha}}(\lambda_k)]^2 \lambda_k \theta_k^2 \leq \max_k \lambda_k H_{\hat{\alpha}}(\lambda_k) \sum_{k=1}^{\infty} [1 - H_{\alpha_0}(\lambda_k)]^2 \theta_k^2.$$

Therefore combining these inequalities with (2.14) and (2.15), we get

$$\begin{aligned} &\left| \mathbf{E}_{\theta} \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \right| \\ &\leq C \left[ \mathbf{E}_{\theta} \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \right]^{1/2} \left[ \sum_{k=1}^{\infty} [1 - H_{\alpha_0}(\lambda_k)]^2 \theta_k^2 \right]^{1/2} \\ &\quad + C \max_k \sqrt{\lambda_k} H_{\alpha_0}(\lambda_k) \left[ \mathbf{E}_{\theta} \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 \right]^{1/2}. \end{aligned}$$

Using the elementary inequality  $2ab \leq \mu a^2 + b^2/\mu$ , we can continue the above display as follows

$$\begin{aligned} &\left| \mathbf{E}_{\theta} \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \right| \\ &\leq \mu \sum_{k=1}^{\infty} [1 - H_{\alpha_0}(\lambda_k)]^2 \theta_k^2 + \frac{C \max_k \lambda_k H_{\alpha_0}^2(\lambda_k)}{\mu} \\ &\quad + \mu \mathbf{E}_{\theta} \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 + \frac{C \mathbf{E}_{\theta} \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu}. \end{aligned}$$

Therefore minimizing the right-hand side in  $\alpha_0$ , we get

$$\begin{aligned} &\left| \mathbf{E}_{\theta} \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)] \sqrt{\lambda_k} \theta_k \xi(k) \right| \\ &\leq \inf_{\alpha_0} \left\{ \mu \sum_{k=1}^{\infty} [1 - H_{\alpha_0}(\lambda_k)]^2 \theta_k^2 + \frac{C \max_k \lambda_k H_{\alpha_0}^2(\lambda_k)}{\mu} \right\} \\ &\quad + \mathbf{E}_{\theta} \left\{ \mu \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 + \frac{C \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} \right\} \\ &\leq 2 \mathbf{E}_{\theta} \left\{ \mu \sum_{k=1}^{\infty} [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 + \frac{C \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} \right\}. \end{aligned}$$

To finish the proof of (2.12) it suffices to minimize the right-hand side in  $\mu$ .

Inequality (2.13) follows from (2.12) since  $\tilde{H}_\alpha(\lambda) = 2H_\alpha(\lambda) - H_\alpha^2(\lambda)$  are ordered smoothers and we can apply (2.12) with  $H_\alpha(\cdot) = \tilde{H}_\alpha(\cdot)$ .  $\square$

In view of the definition of the empirical risk and (1.9), we have

$$\begin{aligned} R_{Pen}[Y, \hat{\alpha}] &= \|\hat{\theta}_0 - \hat{\theta}_{\hat{\alpha}}\|^2 + \sigma^2 Pen(\hat{\alpha}) - \sigma^2 \sum_{k=1}^n \lambda_k \\ &= \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 + \sigma^2 Pen(\hat{\alpha}) \\ &\quad + \sum_{k=1}^n [H_{\hat{\alpha}}^2(\lambda_k) - 2H_{\hat{\alpha}}(\lambda_k)] \lambda_k \xi^2(k) \\ &\quad + 2\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \sqrt{\lambda_k} \theta_k \xi(k) \end{aligned}$$

and

$$\begin{aligned} \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 &= \sum_{k=1}^n [\theta(k) - H_{\hat{\alpha}}(\lambda_k) y(k)]^2 \\ &= \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 + \sigma^2 \sum_{k=1}^{\infty} \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \xi^2(k) \\ &\quad - 2\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)] \theta_k \sqrt{\lambda_k} H_{\hat{\alpha}}(\lambda_k) \xi(k). \end{aligned}$$

Therefore for the excess risk we have

$$\begin{aligned} \Delta_{Pen}(\mu) &= \sup_{\theta \in \mathbb{R}^n} \mathbf{E}_\theta \left\{ \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 - (1 + \mu) R_{Pen}[Y, \hat{\alpha}] \right\} \\ &= \sup_{\theta \in \mathbb{R}^n} \mathbf{E}_\theta \left\{ -\mu \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 - \frac{C\sigma^2 \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} \right. \\ &\quad - 2\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)] \theta_k \sqrt{\lambda_k} \xi(k) \\ &\quad - 2\mu\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k \sqrt{\lambda_k} \xi(k) \\ &\quad + \frac{C \max_k \sigma^2 \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} + 2(1 + \mu)\sigma^2 \sum_{k=1}^n \lambda_k H_{\hat{\alpha}}(\lambda_k) \xi^2(k) \\ &\quad \left. - (1 + \mu)\sigma^2 Pen(\hat{\alpha}) - \mu\sigma^2 \sum_{k=1}^n \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \xi^2(k) \right\}. \end{aligned} \tag{2.16}$$

The last two lines can be bounded by  $\sigma^2 \Delta_{Pen}^C(\mu)$ . Indeed,

$$\begin{aligned}
& \mathbf{E} \left\{ \frac{C\sigma^2 \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} + 2(1+\mu)\sigma^2 \sum_{k=1}^n \lambda_k H_{\hat{\alpha}}(\lambda_k) \xi^2(k) \right. \\
& \quad \left. - (1+\mu)\sigma^2 Pen(\hat{\alpha}) - \mu\sigma^2 \sum_{k=1}^n \lambda_k H_{\hat{\alpha}}^2(\lambda_k) \xi^2(k) \right\} \\
& \leq \mathbf{E} \sup_{\alpha} \left\{ \frac{C\sigma^2 \max_k \lambda_k H_{\alpha}^2(\lambda_k)}{\mu} \right. \\
& \quad \left. + 2(1+\mu)\sigma^2 \sum_{k=1}^n \lambda_k H_{\alpha}(\lambda_k) \xi^2(k) \right. \\
& \quad \left. - (1+\mu)\sigma^2 Pen(\alpha) - \mu\sigma^2 \sum_{k=1}^n \lambda_k H_{\alpha}^2(\lambda_k) \xi^2(k) \right\} \\
& = \sigma^2 \Delta_{Pen}^C(\mu).
\end{aligned} \tag{2.17}$$

Finally, with Lemma 2.5 we obtain

$$\begin{aligned}
& \mathbf{E}_{\theta} \left\{ -\mu \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k^2 - \frac{C\sigma^2 \max_k \lambda_k H_{\hat{\alpha}}^2(\lambda_k)}{\mu} \right. \\
& \quad \left. - 2\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)] \theta_k \sqrt{\lambda_k} \xi(k) - 2\mu\sigma \sum_{k=1}^n [1 - H_{\hat{\alpha}}(\lambda_k)]^2 \theta_k \sqrt{\lambda_k} \xi(k) \right\} \\
& \leq 0.
\end{aligned}$$

This inequality together with (2.16) and (2.17) completes the proof of the theorem.

**2.3.1. Proof of Theorem 1.2.** In view of Theorem 1.1, it suffices to check that

$$\Delta_{Pen}^C(\mu) \leq \frac{C^{1/(1-\kappa)} \lambda_1 \sigma^2}{(1-\kappa)^{1/(1-\kappa)} \mu^{(1+\kappa)/(1-\kappa)}}, \tag{2.18}$$

where

$$\begin{aligned}
\Delta_{Pen}^C(\mu) = & \mathbf{E} \sup_{\alpha} \left\{ \frac{C \max_k \lambda_k H_{\alpha}^2(\lambda_k)}{\mu} - \mu \sum_{k=1}^n \lambda_k H_{\alpha}^2(\lambda_k) \right. \\
& \left. + (2+\mu) \sum_{k=1}^n \frac{2(1+\mu)H_{\alpha}(\lambda_k) - \mu H_{\alpha}^2(\lambda_k)}{2+\mu} \lambda_k [\xi^2(k) - 1] \right\}.
\end{aligned}$$

We begin the proof of (2.18) with the deterministic term. By (1.13) we get

$$\begin{aligned} & \sup_{\alpha > 0} \left\{ \frac{C \max_k \lambda_k H_\alpha^2(\lambda_k)}{\mu} - \frac{\mu}{2} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right\} \\ & \leq \sup_{\alpha > 0} \left\{ \frac{C \lambda_1^{1-\kappa}}{\mu} \left[ \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right]^\kappa - \frac{\mu}{2} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right\} \\ & \leq \sup_{x \geq 0} \left\{ \frac{C \lambda_1^{1-\kappa}}{\mu} x^\kappa - \frac{\mu}{2} x \right\} = C^{1/(1-\kappa)} \lambda_1 \mu^{(\kappa+1)/(\kappa-1)}. \end{aligned} \quad (2.19)$$

Denote for brevity

$$\tilde{H}_\alpha(\lambda) = \frac{2(1+\mu)H_\alpha(\lambda_k) - \mu H_\alpha^2(\lambda_k)}{2+\mu}.$$

Our next step is to show that

$$\begin{aligned} & \mathbf{E} \sup_{\alpha > 0} \left\{ (2+\mu) \sum_{k=1}^n \lambda_k \tilde{H}_\alpha(\lambda_k) [\xi^2(k) - 1] - \frac{\mu}{2} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right\} \\ & \leq C^{1/(1-\kappa)} \lambda_1 (1-\kappa)^{-1/(1-\kappa)} \mu^{(\kappa+1)/(\kappa-1)}. \end{aligned} \quad (2.20)$$

It is easy to see that in view of (1.14)

$$\begin{aligned} \sigma^2(\alpha) &= \mathbf{E} \left[ \sum_{k=1}^n \lambda_k \tilde{H}_\alpha(\lambda_k) [\xi^2(k) - 1] \right]^2 = 2 \sum_{k=1}^n \lambda_k^2 \tilde{H}_\alpha^2(\lambda_k) \\ &\leq \frac{2(2+2\mu)^2}{(2+\mu)^2} \sum_{k=1}^n \lambda_k^2 H_\alpha^2(\lambda_k) \\ &\leq \frac{2\lambda_1^2(2+2\mu)^2}{(1-\kappa)(2+\mu)^2} \left[ \frac{1}{\lambda_1} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right]^{1+\kappa}. \end{aligned}$$

Next notice that if  $\{H_\alpha(\cdot), \alpha \geq 0\}$  is a family of ordered smoothers, then  $\{\tilde{H}_\alpha(\cdot), \alpha \geq 0\}$  is also a family of ordered smoothers. Therefore by Lemma 2.4, for any  $\mu > 0$  we obtain

$$\begin{aligned} & \mathbf{E} \sup_{\alpha} \left\{ (2+\mu) \sum_{k=1}^n \lambda_k \tilde{H}_\alpha(\lambda_k) [\xi^2(k) - 1] - \frac{\mu}{2} \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k) \right\} \\ & \leq (2+\mu) \mathbf{E} \sup_{\alpha} \left\{ \sum_{k=1}^n \lambda_k \tilde{H}_\alpha(\lambda_k) [\xi^2(k) - 1] \right. \\ & \quad \left. - \frac{\mu \lambda_1}{2(2+\mu)} \left[ \frac{\sigma^2(\alpha)(1-\kappa)(2+\mu)^2}{2\lambda_1^2(2+2\mu)^2} \right]^{1/(1+\kappa)} \right\} \\ & \leq C(1+\mu)^{2/(1-\kappa)} 2^{(4+\kappa)/(1-\kappa)} \lambda_1 (1-\kappa)^{-1/(1-\kappa)} \mu^{(\kappa+1)/(\kappa-1)}. \end{aligned}$$

thus proving (2.20). Thus (2.18) follows obviously from (2.19) and (2.20).

## REFERENCES

- [1] L. CAVALIER AND YU. GOLUBEV, *Risk hull method and regularization by projections of ill-posed inverse problems*, Ann. of Stat. (2006), **34**: 1653–1677.
- [2] L. CAVALIER, G.K. GOLUBEV, D. PICARD, AND A.B. TSYBAKOV, *Oracle inequalities for inverse problems*, Ann. of Stat. (2002), **30**: 843–874.
- [3] H.W. ENGL, M. HANKE, AND A. NEUBAUER, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 1996.
- [4] I.A. IBRAGIMOV AND R.Z. KHASHMINSKII, *Statistical Estimation. Asymptotic Theory*, Springer-Verlag, NY, 1981.
- [5] A. KNEIP, *Ordered linear smoothers*, Ann. Statist. (1994), **22**: 835–866.
- [6] B. MAIR AND F.H. RUYMGAART, *Statistical estimation in Hilbert scale*. SIAM J. Appl. Math. (1996), **56**: 1424–1444.
- [7] A.N. TIKHONOV AND V.A. ARSEININ, *Solution of Ill-posed Problems*, Winston & Sons, Washington, 1977.
- [8] A. VAN DER VAART AND J. WELLNER *Weak convergence and empirical processes*. Springer-Verlag, NY, 1996.