On Oracle Inequalities Related to High Dimensional Linear Models*

Yuri Golubev†
golubev.yuri@gmail.com

Abstract

This paper deals with recovering an unknown vector θ from the noisy data $Y = A\theta + \sigma\xi$, where A is a known $m \times n$ - matrix and ξ is a white Gaussian noise. It is assumed that n is large and A may be severely ill-posed. Therefore, in order to estimate θ, a spectral regularization method is used, and our goal is to choose its regularization parameter with the help of the data Y. For spectral regularization methods related to the so-called ordered smoothers [see [Kneip (1994)]] we propose new penalties in the principle of empirical risk minimization. The heuristical idea behind these penalties is related to balancing excess risks. Based on this approach, we derive a sharp oracle inequality controlling the mean square risks of data-driven spectral regularization methods.

1 Introduction and main results

In this paper, we consider a classical problem of recovering an unknown vector $\theta = (\theta(1), \dots, \theta(n))^\top \in \mathbb{R}^n$ in the standard linear model

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

where A is a known $m \times n$ - matrix and $\xi = \big(\xi(1), \dots, \xi(m)\big)^\top$ is a standard white Gaussian noise in $\mathbb{R}^m$ with $\mathbf{E}\xi(k) = 0,\ \mathbf{E}\xi^2(k) = 1,\ k = 1, \dots, m.$ The noise level σ in (1.1) is assumed to be known.

We start out by considering the maximum likelihood estimate of θ

$$\hat\theta_0 = \arg\min_{\theta \in \mathbb{R}^n} \|Y - A\theta\|^2,$$

*The research was supported in part by ANR Grant 07-0234-01.

†CNRS, Université de Provence, France and Institute for Information Transmission Problems, Moscow, Russia

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

$$\begin{aligned} \mathbf{E}\|\hat{\theta}_0 - \theta\|^2 =& \sigma^2 \mathbf{E}\|(A^\top A)^{-1}A^\top \xi\|^2 \\ =& \sigma^2 \mathrm{trace}\big[(A^\top A)^{-1}\big] = \sigma^2 \sum_{k=1}^n \lambda^{-1}(k), \end{aligned} \tag{1.2}$$

where $\lambda(k)$ and $\psi_k \in \mathbb{R}^n$ are eigenvalues and eigenvectors of $A^\top A$

$$A^\top A \psi_k = \lambda(k)\psi_k.$$

In this paper, it is assumed solely that $\lambda(1) \geq \lambda(2) \geq \cdots \geq \lambda(n)$. So, A may be severely ill-posed and (1.2) reveals the principal difficulty in $\hat{\theta}_0$: *Its risk may be very large when n is large or when A has a large condition number.*

The simplest way to improve $\hat{\theta}_0$ is to suppress large $\lambda^{-1}(k)$ in (1.2) with the help of a linear smoother; that is, to estimate θ by $H\hat{\theta}_0$, where H is a properly chosen $n \times n$ - matrix. In what follows, we deal with smoothing matrices admitting the following representation $H = H_\alpha(A^\top A)$, where $H_\alpha(\lambda)$ is a function $\mathbb{R}^+ \to [0,1]$ which depends on a regularization parameter $\alpha \in [0, \bar{\alpha}]$ such that

$$\lim_{\alpha \to 0} H_\alpha(\lambda) = 1, \quad \lim_{\lambda \to 0} H_\alpha(\lambda) = 0.$$

This method is called *spectral regularization* [see [Engl et al. (1996)]] since $A^\top A$ and $H_\alpha(A^\top A)$ have the same eigenvectors. Summarizing, we estimate θ with the help of the following family of linear estimators

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

and our main goal is to choose the best estimator within this family, or equivalently, the best regularization parameter α. Note that α controls the mean square risk of $\hat{\theta}_\alpha$

$$L_\alpha(\theta) \stackrel{\mathrm{def}}{=} \mathbf{E}\|\hat{\theta}_\alpha - \theta\|^2 = \sum_{k=1}^n \big[1 - h_\alpha(k)\big]^2 \langle \theta, \psi_k\rangle^2 + \sigma^2 \sum_{k=1}^n \lambda^{-1}(k) h_\alpha^2(k), \tag{1.3}$$

where here and below we denote for brevity

$$h_\alpha(k) \stackrel{\mathrm{def}}{=} H_\alpha[\lambda(k)], \quad \text{and} \quad \langle \theta, \psi_k\rangle \stackrel{\mathrm{def}}{=} \sum_{l=1}^n \theta(l)\psi_k(l).$$

According to (1.3), the variance of $\hat{\theta}_\alpha$ is always smaller than that of the maximum likelihood estimate, but $\hat{\theta}_\alpha$ has a nonzero bias and adjusting properly α we may improve $\hat{\theta}_0$. Note that this improvement may be significant if $\langle \theta, \psi_k \rangle^2$ are small for large k.

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

The popular Tikhonov-Phillips [see, e.g., [Tikhonov and Arsenin (1977)]] regularization

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

doesn't require SVD. In this case, $\hat{\theta}_\alpha$ is computed as a root of the linear equation

$$(\alpha I + A^\top A)\hat{\theta}_\alpha = A^\top Y$$

and therefore $H_\alpha(\lambda) = \lambda/(\lambda + \alpha)$. It is worth pointing out that this regularization technique is good solely for ill-posed A.

Another widespread regularization technique is due to [Landweber (1951)]. This method is based on a very simple idea: to compute recursively a root of equation

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

Since $A^\top Y = \big[A^\top A - aI\big]\theta + a\theta$ for all $a > 0$, we get $\theta = \big[I - a^{-1}A^\top A\big]\theta + a^{-1}A^\top Y$. This formula motivates Landweber's iterations defined by

$$\hat{\theta}_k = \big[I - a^{-1}A^\top A\big]\hat{\theta}_{k-1} + a^{-1}A^\top Y.$$

Thus we can estimate θ without computing SVD and without solving linear equations. It is easily seen that these iterations converge if $\lambda(1) < a$ and that the corresponding spectral regularization function is given by

$$H_k(\lambda) = 1 - \left(1 - \frac{\lambda}{a}\right)^{k+1}. \tag{1.4}$$

The regularization parameter of the Landweber method is usually defined by $\alpha = 1/k$. Note that in spite of its iterative character, the numerical complexity of the Landweber method may be hight. Indeed, when the noise is very small, $H_k(\lambda)$ should be close to 1, and (1.4) implies that

$$k \gtrsim \mathrm{cond}(A) \stackrel{\mathrm{def}}{=} \frac{\lambda(1)}{\lambda(n)}.$$

This means that 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 by the so-called ν-method [see, e.g., [Engl et al. (1996)] and [Bissantz et al. (2007)]].

All the above mentioned regularization methods are particular cases of the so-called ordered smoothers [see [Kneip (1994)]] defined as follows.

Definition 1 *The family of sequences* $\{h_\alpha(k),\ \alpha \in (0, \bar{\alpha}],\ k \in \mathbb{N}^+\}$ *is called ordered smoother if:*

1. *For any given* $\alpha \in (0, \bar{\alpha}]$, $h_\alpha(k) : \mathbb{N}^+ \to [0, 1]$ *is a monotone function of* k.
2. *If for some* $\alpha_1, \alpha_2 \in (0, \bar{\alpha}]$ *and* $k' \in \mathbb{N}^+$, $h_{\alpha_1}(k') < h_{\alpha_2}(k')$, *then* $h_{\alpha_1}(k) \le h_{\alpha_2}(k)$ *for all* $k \in \mathbb{N}^+$.

It was Kneip who noted that from a probabilistic viewpoint, all ordered smoothers are equivalent to the spectral cut-off with $h_\alpha(k) = \mathbf{1}\{\lambda(k) \ge \alpha\}$. This profound fact plays an essential role in adaptive estimation since it helps to analyze precisely statistical risks of feasible data-driven regularization methods. This is why in this paper we deal solely with the ordered smoothers.

Whatever an inversion method is used, the principal question usually arising in practice is how to choose its regularization parameter. Traditional theoretic approach to this problem is related to the minimax theory, see, for example, [Mair and Ruymgaart (1996)] and [O'Sullivan (1986)]. However, this approach provides the smoothing parameters depending strongly on an a priory information about θ which is hardly available in practice. The only one way to improve this drawback is to use data-driven regularizations. In statistical literature, one can find several general approaches for constructing such methods. We cite here, for instance, the Lepski method which has been adopted to inverse problems in [Mathé (2006)], [Bauer and Hohage (2005)], [Bissantz et al. (2007)], and the model selection technique which was implemented in [Lubes and Ludeña (2008)].

In this paper, we take the classical way related to the famous principle of unbiased risk estimation which goes back to [Akaike (1973)]. The heuristical motivation of this approach is based on the idea that a good data-driven regularization should minimize in some sense the risk $L_\alpha(\theta)$ [see (1.3)]. This idea is put into practice with the help of the empirical risk minimization suggesting to compute data-driven regularization parameters as follows :

$$\hat{\alpha} = \arg\min_{\alpha \in (0,\bar{\alpha}]} R_\alpha[Y, Pen], \tag{1.5}$$

where

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

and $Pen(\alpha) : (0, \bar{\alpha}] \to \mathbb{R}^+$ is a given penalty function. The most important problem in this approach is related to the choice of the penalty. Intuitively, we want that the method mimics the oracle smoothing parameter $\alpha^* = \arg\min_\alpha L_\alpha(\theta)$. This is why we are looking for a minimal penalty that ensures the following inequality

$$L_\alpha(\theta) \lesssim R_\alpha[Y, Pen] + \mathcal{C}, \tag{1.6}$$

where $\mathcal{C}$ is a random variable that doesn't depend on α. It is easily seen that in the considered statistical model,

$$\mathcal{C} = -\|\theta - \hat{\theta}_0\|^2 = -\sigma^2 \sum_{k=1}^{n} \lambda^{-1}(k)\xi^2(k).$$

Traditional approach to solving (1.6) is based on the unbiased risk estimation defining the penalty as a root of the equation

$$L_\alpha(\theta) = \mathbf{E} R_\alpha[Y, Pen] + \mathbf{E}\mathcal{C}.$$

Unfortunately, in spite of its very natural motivation, this penalty is not good for ill-posed problems (see [Cavalier and Golubev (2006)] for more details).

The main idea in this paper is to compute the penalty in a little bit different way, namely as a minimal function assuring the following inequality

$$\mathbf{E} \sup_{\alpha \le \bar{\alpha}} \Big[L_\alpha(\theta) - R_\alpha[Y, Pen] - \mathcal{C} \Big]_+ \le K \mathbf{E} \Big[L_{\bar{\alpha}}(\theta) - R_{\bar{\alpha}}[Y, Pen] - \mathcal{C} \Big]_+, \tag{1.7}$$

where $[x]_+ = \max\{0, x\}$ and $K > 1$ is a constant. The heuristical motivation behind this approach is rather transparent : We are looking for a minimal penalty that balances all excess risks uniformly in $\alpha \in (0, \bar{\alpha}]$. Recall that the excess risk is defined as the difference between the risk of the estimate and its empirical risk. Note that according to (1.6), we may focus on the positive part of the excess risk, and that Equation (1.7) guarantees that for any data driven smoothing parameter $\hat{\alpha}$

$$\mathbf{E} \Big[L_{\hat{\alpha}}(\theta) - R_{\hat{\alpha}}[Y, Pen] - \mathcal{C} \Big]_+ \le K \mathbf{E} \Big[L_{\bar{\alpha}}(\theta) - R_{\bar{\alpha}}[Y, Pen] - \mathcal{C} \Big]_+.$$

In order to explain how one can compute good penalties assuring (1.7), we begin with the spectral representation of the underlying statistical problem. We can check easily that

$$y(k) \stackrel{\text{def}}{=} \langle A^\top Y, \psi_k \rangle \lambda^{-1}(k) = \langle \theta, \psi_k \rangle + \sigma \lambda^{-1/2}(k)\xi(k),$$

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

$$\langle \hat{\theta}_\alpha, \psi_k \rangle = h_\alpha(k) y(k) = h_\alpha(k)\theta(k) + \sigma h_\alpha(k) \lambda^{-1/2}(k)\xi(k),$$

where $\theta(k) = \langle \theta, \psi_k \rangle$, and

$$\begin{aligned} \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 &= \sum_{k=1}^n \bigl[1 - h_\alpha(k)\bigr]^2 y^2(k), \\ \|\theta - \hat{\theta}_\alpha\|^2 &= \sum_{k=1}^n \bigl[\theta(k) - h_\alpha(k) y(k)\bigr]^2. \end{aligned} \tag{1.8}$$

In what follows, it is assumed that the penalty has the following structure

$$Pen(\alpha) = 2 \sum_{k=1}^n \lambda^{-1}(k) h_\alpha(k) + (1+\gamma) Q(\alpha),$$

where γ is a positive number and $Q(\alpha)$, $\alpha > 0$ is a positive function to be defined later on. Then the excess risk is computed as follows :

$$\begin{aligned} &L_\alpha(\theta) - R_\alpha[Y, Pen] - \mathcal{C} \\ &\quad = \sigma^2 \sum_{k=1}^n \lambda^{-1}(k) \bigl[2h_\alpha(k) - h_\alpha^2(k)\bigr](\xi^2(k) - 1) - (1+\gamma)\sigma^2 Q(\alpha) \\ &\qquad - 2\sigma \sum_{k=1}^n \lambda^{-1/2}(k)\bigl[1 - h_\alpha(k)\bigr]^2 \xi(k)\theta(k). \end{aligned} \tag{1.9}$$

Our first idea in solving (1.7) is to use the fact that the absolute value of the cross term

$$2\sigma \sum_{k=1}^n \lambda^{-1/2}(k)\bigl[1 - h_\alpha(k)\bigr]^2 \xi(k)\theta(k)$$

is typically smaller than $L_\alpha(\theta)$ (see for more details Lemma 9 below). Therefore, omitting this term in (1.7), we get the following inequality for $Q(\alpha)$

$$\mathbf{E} \sup_{\alpha \le \bar{\alpha}} [\eta_\alpha - (1+\gamma)Q(\alpha)]_+ \le K \mathbf{E}[\eta_{\bar{\alpha}} - (1+\gamma)Q(\bar{\alpha})]_+, \tag{1.10}$$

where

$$\eta_\alpha = \sum_{k=1}^{n} \lambda^{-1}(k)\big[2h_\alpha(k) - h_\alpha^2(k)\big](\xi^2(k) - 1).$$

Usually, computing the minimal function $Q(\alpha)$ assuring (1.10) is a hard numerical problem. However, when $h_\alpha(k)$ is a family of ordered smoothers it can be solved relatively easily. The main idea is to find a feasible solution $Q^\circ(\alpha)$ of the marginal inequality

$$\mathbf{E}[\eta_\alpha - Q^\circ(\alpha)]_+ \le \mathbf{E}[\eta_{\bar\alpha} - Q^\circ(\bar\alpha)]_+ \tag{1.11}$$

and then to show that $(1+\gamma)Q^\circ(\alpha)$ satisfies (1.10). To solve (1.11), we use the following inequality

$$\mathbf{E}[\eta - x]_+^p \le \Gamma(p+1)\lambda^{-p}\exp(-\lambda x)\mathbf{E}\exp(\lambda\eta) \tag{1.12}$$

which holds for any random variable η and for any $\lambda > 0$. Its proof follows from the Chernoff bound. Without loss of generality we may assume that $Q^\circ(\bar\alpha) = 0$. Therefore, according to the Cauchy-Schwarz inequality

$$\mathbf{E}[\eta_{\bar\alpha} - Q^\circ(\bar\alpha)]_+ \le \sqrt{\mathbf{E}\eta_{\bar\alpha}^2} = D(\bar\alpha),$$

where

$$D(\alpha) = \sqrt{\mathbf{E}\eta_\alpha^2} = \left\{2\sum_{k=1}^{n} \lambda^{-2}(k)\big[2h_\alpha(k) - h_\alpha^2(k)\big]^2\right\}^{1/2}.$$

Hence, $Q^\circ(\alpha)$ is computed as a root of equation

$$\inf_\lambda \exp[-\lambda Q^\circ(\alpha)]\mathbf{E}\exp(\lambda\eta_\alpha) = D(\bar\alpha).$$

It is not difficult to check with a little algebra that

$$Q^\circ(\alpha) = 2D(\alpha)\mu_\alpha \sum_{k=1}^{n} \frac{\rho_\alpha^2(k)}{1 - 2\mu_\alpha\rho_\alpha(k)}, \tag{1.13}$$

where μ_α is a root of equation

$$\sum_{k=1}^{n} F[\mu_\alpha\rho_\alpha(k)] = \log\frac{D(\alpha)}{D(\bar\alpha)}, \tag{1.14}$$

and

$$F(x) = \frac{1}{2}\log(1-2x) + x + \frac{2x^2}{1-2x}$$
$$\rho_\alpha(k) = \sqrt{2}D^{-1}(\alpha)\lambda^{-1}(k)\bigl[2h_\alpha(k) - h_\alpha^2(k)\bigr]. \tag{1.15}$$

The only one numerical difficulty in computing $Q^\circ(\alpha)$ is related to the equation (1.14). However note that in the proof of Lemma 7 it is shown that

$$f(\mu) = \sum_{k=1}^{n} F[\mu\rho_\alpha(k)]$$

is a strictly monotone function and therefore (1.14) may be solved exponentially fast. Note also that Lemma 7 provides lower and upper bounds for $Q^\circ(\alpha)$ and μ_α. In particular, for some constant $C > 0$

$$C^{-1}D(\alpha)\log^{1/2}\frac{D(\alpha)}{D(\bar\alpha)} \le Q^\circ(\alpha) \le CD(\alpha)\log\frac{D(\alpha)}{D(\bar\alpha)}. \tag{1.16}$$

The next theorem shows that $Q^\circ(\alpha)$ computed as a root of the marginal inequality (1.11) satisfies the global inequality (1.10).

Theorem 1 *Let $Q^\circ(\alpha)$ be defined by (1.13–1.15). Then for any $\gamma > r \ge 0$,*

$$\mathbf{E}\sup_{\alpha\le\bar\alpha}\Bigl[\eta_\alpha - (1+\gamma)Q^\circ(\alpha)\Bigr]_+^{1+r} \le \frac{CD^{1+r}(\bar\alpha)}{(\gamma-r)^3},$$

where here and throughout the paper C denotes a generic constant.

The following theorem represents the main result in this paper. It controls the performance of the empirical risk minimization by the penalized oracle risk defined by

$$r(\theta) \stackrel{\text{def}}{=} \inf_{\alpha\le\bar\alpha}\bar R_\alpha(\theta),$$

where

$$\bar R_\alpha(\theta) \stackrel{\text{def}}{=} \mathbf{E}\bigl\{R_\alpha[Y, Pen] + \mathcal{C}\bigr\} = L_\alpha(\theta) + (1+\gamma)\sigma^2 Q^\circ(\alpha).$$

Theorem 2 *Let $Pen(\alpha) = 2\sum_{k=1}^{n}\lambda^{-1}(k)h_\alpha(k) + (1+\gamma)Q^\circ(\alpha)$ with $Q^\circ(\alpha)$ defined by (1.13 – 1.15). Then the mean square risk of $\hat\theta_{\hat\alpha}$ with the data-driven smoothing parameter $\hat\alpha$ defined by (1.5) satisfies the following upper bound*

$$\mathbf{E}\|\theta - \hat\theta_{\hat\alpha}\|^2 \le r(\theta)\left\{1 + \left[\frac{C}{\gamma}\log^{-1/2}\frac{Cr(\theta)}{\sigma^2 D(\bar\alpha)} + \frac{C\sigma^2 D(\bar\alpha)}{\gamma^4 r(\theta)}\right]^{1/2}\right\} \tag{1.17}$$

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

Below, we discuss briefly some statistical aspects of this theorem.

1. Equation (1.17) represents a particular form of the so-called oracle inequality
$$\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \le r(\theta)\left\{1 + \Psi\left[\frac{\sigma^2 D(\bar{\alpha})}{r(\theta)}\right]\right\},$$
where $\Psi(\cdot)$ is a bounded function such that $\lim_{x\to 0}\Psi(x) = 0$. This means that if the ratio $\sigma^2 D(\bar{\alpha})/r(\theta)$ is small, then the risk of the method is close to the risk of the penalized oracle. On the other hand, if this ratio is large, then the risk of the method is of order of the oracle risk.

   Note also that (1.17) is a universal oracle inequality which holds true whatever is the ill-posedness of the underlying inverse problem. It generalizes the corresponding oracle inequalities in [Cavalier and Golubev (2006)] and [Golubev (2004)] obtained for the spectral cut-off method.

2. Theorem 2 reveals some difficulties related to the data-driven choice of the regularization parameter in the Tikhonov-Phillips method. Recalling that for this method $H_\alpha(\lambda) = \lambda/(\alpha + \lambda)$, we obtain
$$\begin{aligned} D^2(\alpha) =& 2\sum_{k=1}^n \lambda^{-2}(k)[2h_\alpha(k) - h_\alpha^2(k)]^2 \ge 2\sum_{k=1}^n \lambda^{-2}(k) h_\alpha^2(k) \\ =& 2\sum_{k=1}^n \frac{1}{[\alpha + \lambda(k)]^2} \ge \frac{2n}{[\alpha + \lambda(n)]^2}. \end{aligned}$$
Since $Q^\circ(\alpha) \ge D(\alpha) \asymp \sqrt{n}$, it is clear that the penalized oracle risk of the Tikhonov-Phillips regularization may be very large compared to the risk of the method computed for given α. This means that in practice, the Tikhonov-Phillips regularization with a data-driven smoothing parameter may fail.

   Note however, that this drawback can be easily improved with the help of high-order Tilhonov-Phillips regularizations computed as follows :
$$\hat{\theta}_\alpha^{(k+1)} = \arg\min_\theta\left\{\|A\hat{\theta}_\alpha^{(k)} - A\theta\|^2 + \alpha\|\theta\|^2\right\},$$
where $\hat{\theta}_\alpha^{(1)}$ stands for the standard Tikhonov-Phillips regularization. One can check with a little algebra that the corresponding smoothers are given by $H_\alpha^{(k)}(\lambda) = \lambda^k(\alpha + \lambda)^{-k}$, $k \ge 2$, and everything goes smoothly in this case.

3. If the inverse problem is not severely ill-posed; that is, $\lambda(k) \geq Ck^{-\beta}$ for some $\beta > 0$, then, according to (1.16), for reasonable spectral regularizations

$$\sum_{k=1}^{n} h_\alpha^2(k)\lambda^{-1}(k) \gg Q^\circ(\alpha) \tag{1.18}$$

when α is small. This means that the risk of the penalized oracle is close to the risk of the ideal oracle $\inf_{\alpha\leq\bar{\alpha}} L_\alpha(\theta)$.

This remark together with the famous [Pinsker (1980)] minimax theorem shows that our method results in adaptive asymptotically (as $\sigma \to 0$) minimax regularizations. To demonstrate this, suppose for simplicity that $n = \infty$ and that θ belongs to the following ellipsoidal body

$$\Theta(W) = \left\{\theta \in l_2(1,\infty) : \sum_{k=1}^{\infty} \langle\theta, \psi_k\rangle^2 b^2[\lambda^{-1}(k)] \leq W\right\}.$$

where $|b(x)|$ is a nondecreasing function such that for some $p, q \in (0,\infty)$ $Cx^p \leq b(x) \leq Cx^q$. Then it follows from [Pinsker (1980)] that as $\sigma \to 0$

$$\begin{aligned}&\inf_{\hat{\theta}} \sup_{\theta\in\Theta(W)} \|\theta - \hat{\theta}\|^2 = (1 + o(1)) \inf_{h} \sup_{\theta\in\Theta(W)} L[\theta, h]\\ &= (1 + o(1)) \sup_{\theta\in\Theta(W)} \inf_{h} L[\theta, h] = (1 + o(1)) \sup_{\theta\in\Theta(W)} \inf_{\alpha\in(0,\bar{\alpha}]} L[\theta, h_\alpha^*],\end{aligned}$$

where inf at the left-hand side is taken over all estimators,

$$L[\theta, h] = \sum_{k=1}^{n} \big[1 - h(k)\big]^2 \langle\theta, \psi_k\rangle^2 + \sigma^2 \sum_{k=1}^{n} \lambda^{-1}(k)h^2(k)$$

and $h_\alpha^*(k) = \big[1 - \alpha\big|b[\lambda^{-1}(k)]\big|\big]_+$. Recall that from a statistical viewpoint, the main drawback in this minimax result is that the optimal smoothing parameter

$$\alpha^*(W) = \arg\min_{\alpha\in(0,\bar{\alpha}]} \sup_{\theta\in\Theta(W)} L[\theta, h_\alpha^*]$$

depends on the size W of $\Theta(W)$ which is hardly known in practice. In order to overcome this difficulty, one may use the data-driven regularization $\hat{\alpha}$ with $h_\alpha(k) = h_\alpha^*(k)$. Noticing that this family of smoothers is ordered, we get according to Theorem 2 and (1.18)

$$\sup_{\theta\in\Theta(W)} \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 = (1 + o(1)) \inf_{\hat{\theta}} \sup_{\theta\in\Theta(W)} \|\theta - \hat{\theta}\|^2, \text{ as } \sigma \to 0.$$

Another interesting situation is related to the case when the inverse problem is severely ill-posed; that is, when the eigenvalues of $AA^\top$ are exponentially decreasing, $\lambda(k) \approx \exp(-\beta k)$ with some $\beta > 0$. Then for small α

$$\sum_{k=1}^{n} h_\alpha^2(k)\lambda^{-1}(k) \ll Q^\circ(\alpha).$$

This means that the risk of the penalized oracle is essentially greater than that of the ideal oracle. In this situation, Theorem 2 provides an upper bound similar to [Golubev (2004)]. It is worth pointing out that neither (1.17) nor the extra penalty can be improved in this case [see for more details [Golubev (2004)]].

2 Proofs

2.1 An exponential chaining inequality.

Let ξ_t be a separable zero mean random process on $\mathbb{R}^+$. Denote for brevity

$$\Delta_\xi(t_1, t_2) = \xi_{t_1} - \xi_{t_2}.$$

We begin with a general fact similar to Dudley's entropy bound [see, e.g., [Van der Vaart and Wellner (1996)]].

Lemma 1 *Let σ_u^2, $u \in \mathbb{R}^+$, be a continuous strictly increasing function with $\sigma_0^2 = 0$. Then for any $\lambda > 0$,*

$$\begin{aligned}\log \mathbf{E} \exp\biggl\{\lambda \max_{0<u\le t} \frac{\Delta_\xi(u,t)}{\sigma_t}\biggr\} \le \frac{\log(2)\sqrt{2}}{\sqrt{2}-1} + \\ + \max_{0<u<v\le t} \max_{|z|\le\sqrt{2}/(\sqrt{2}-1)} \log \mathbf{E} \exp\biggl\{z\lambda \frac{\Delta_\xi(u,v)}{\bar{\Delta}_\sigma(v,u)}\biggr\}\end{aligned} \tag{2.1}$$

where $\bar{\Delta}_\sigma(v,u) = \sqrt{|\sigma_v^2 - \sigma_u^2|}$.

Proof The proof of (2.1) is based on the standard chaining argument (see for more details [Van der Vaart and Wellner (1996)]). Denote for brevity by $t_-(B)$ and $t_+(B)$ left and right elements of a closed subset B in $\mathbb{R}^+$. First, we construct a dyadic partition of $[0, t]$. Let

$$\mathcal{T}_1^1 = \Bigl\{u \ge 0 : \sigma_u^2 \le \frac{\sigma_t^2}{2}\Bigr\},\ \mathcal{T}_2^1 = \Bigl\{u \le t : \sigma_u^2 > \frac{\sigma_t^2}{2}\Bigr\}.$$

Next, we partition $\mathcal{T}_1^1$ and $\mathcal{T}_2^1$ as follows :

$$\begin{aligned}
\mathcal{T}_1^2 &= \Big\{u \in \mathcal{T}_1^1 : \sigma_u^2 \le \frac{\sigma^2_{t_+(\mathcal{T}_1^1)}}{2}\Big\},\\
\mathcal{T}_2^2 &= \Big\{u \in \mathcal{T}_1^1 : \sigma_u^2 > \frac{\sigma^2_{t_+(\mathcal{T}_1^1)}}{2}\Big\},\\
\mathcal{T}_3^2 &= \Big\{u \in \mathcal{T}_2^1 : \sigma_u^2 \le \frac{\bar\Delta^2_\sigma[t_+(\mathcal{T}_2^1), t_-(\mathcal{T}_2^1)]}{2}\Big\},\\
\mathcal{T}_4^2 &= \Big\{u \in \mathcal{T}_2^1 : \sigma_u^2 > \frac{\bar\Delta^2_\sigma[t_+(\mathcal{T}_2^1), t_-(\mathcal{T}_2^1)]}{2}\Big\}.
\end{aligned}$$

Doing so, after p steps, we get partitions $\mathcal{T}_j^p,\ j = 1, \dots, 2^p$, such that for any $x, y \in \mathcal{T}_j^k$

$$\bar\Delta^2_\sigma(y, x) \le 2^{-k}\sigma_t^2. \tag{2.2}$$

With the sets $\mathcal{T}_j^k,\ j = 1, \dots, 2^k$ we associate the set of their right points

$$\tau^k = \bigcup_{j=1}^{2^k} \{t_+(\mathcal{T}_j^k)\},$$

and for any point $x \in \tau^k$ we denote by $\tau_{k-1}(x)$ the nearest point in τ^{k-1}. So, by (2.2), for any $v \in \tau^p$

$$\bar\Delta^2_\sigma(\tau_{p-1}(v), v) \le 2^{-p+1}\sigma_t^2. \tag{2.3}$$

With these notations, for any $u \in \tau^p$, setting $\tau_0(v) = t$, we obtain

$$\begin{aligned}
\xi_u - \xi_t &= \sum_{k=1}^p \big[\xi_{\tau_k(u)} - \xi_{\tau_{k-1}(u)}\big] \le \sum_{k=1}^p \sup_{v\in\tau^k} [\xi_v - \xi_{\tau_{k-1}(v)}]\\
&= \sum_{k=1}^p \sup_{v\in\tau^k} \bar\Delta_\sigma[v, \tau_{k-1}(v)] \times \frac{\Delta_\xi[v, \tau_{k-1}(v)]}{\bar\Delta_\sigma[v, \tau_{k-1}(v)]}.
\end{aligned} \tag{2.4}$$

To bound the right-hand side at the above display, we use the elementary inequality

$$\log \mathbf{E} \exp\Big[\sum_k q(k)\eta(k)\Big] \le \sum_k q(k) \log \mathbf{E} \exp[\eta(k)] \tag{2.5}$$

which holds for any random variables $\eta(k)$ and any given $q(k) \ge 0$ with $\sum_k q(k) = 1$. The proof of (2.5) follows immediately from the convexity of

$\exp(x)$ which implies

$$\mathbf{E}\exp\Big\{\sum_k q(k)\big\{\eta(k) - \log\mathbf{E}\exp[\eta(k)]\big\}\Big\}$$
$$\leq \Big\{\sum_k q(k)\mathbf{E}\exp\Big[\eta(k) - \log\mathbf{E}\exp[\eta(k)]\Big]\Big\} = 1.$$

Applying (2.5) with

$$q(k) = \frac{2^{-k/2}}{\sum_{l=1}^{p} 2^{-l/2}},\ \eta(k) = \frac{\lambda}{\sigma_t q(k)}\sup_{v\in\tau^k}\bar{\Delta}_\sigma[v,\tau_{k-1}(v)]\times\frac{\Delta_\xi[v,\tau_{k-1}(v)]}{\bar{\Delta}_\sigma[v,\tau_{k-1}(v)]},$$

we obtain by (2.3, 2.4)

$$\log\mathbf{E}\exp\Big\{\lambda\sup_{u\in\tau^p}\frac{\Delta_\xi(u,t)}{\sigma_t}\Big\} \leq \sum_{k=1}^{p}\lambda(k)\log\mathbf{E}\exp[\eta(k)]. \tag{2.6}$$

It is easily seen that for any $\lambda > 0$

$$\begin{aligned}
&\mathbf{E}\exp[\eta(k)] \\
&\quad = \mathbf{E}\exp\Big\{\frac{\lambda}{q(k)\sigma_t}\sup_{v\in\tau^k}\bar{\Delta}_\sigma[\tau_{k-1}(v),v]\times\frac{\Delta_\xi[v,\tau_{k-1}(v)]}{\bar{\Delta}_\sigma[\tau_{k-1}(v),v]}\Big\} \\
&\quad \leq \sum_{v\in\tau^k}\mathbf{E}\exp\Big\{\frac{\lambda}{q(k)\sigma_t}\bar{\Delta}_\sigma[\tau_{k-1}(v),v]\times\frac{\Delta_\xi[v,\tau_{k-1}(v)]}{\bar{\Delta}_\sigma[\tau_{k-1}(v),v]}\Big\} \\
&\quad \leq 2^k\sup_{u<v}\sup_{|z|\leq\sqrt{2}/(\sqrt{2}-1)}\mathbf{E}\exp\Big\{\lambda z\frac{\Delta_\xi(u,v)}{\bar{\Delta}_\sigma(v,u)}\Big\}.
\end{aligned} \tag{2.7}$$

In the above equation, it was used that $\sum_{s=1}^{\infty}2^{-s/2} = 1/(\sqrt{2}-1)$.

Finally, substituting (2.7) into (2.6), we arrive at

$$\begin{aligned}
\log\mathbf{E}\exp\Big\{\lambda\sup_{u\in\tau^p}\frac{\Delta_\xi(u,t)}{\sigma_t}\Big\} &\leq \log(2)\sum_{k=1}^{p}kq(k) \\
&\quad + \sup_{u<v}\sup_{|z|\leq\sqrt{2}/(\sqrt{2}-1)}\mathbf{E}\exp\Big\{\lambda z\frac{\Delta_\xi(u,v)}{\bar{\Delta}_\sigma(v,u)}\Big\}.
\end{aligned}$$

thus completing the proof. ■

2.2 Ordered processes

Definition 2 *A zero mean process* $\xi_t,\ t\in\mathbb{R}^+$ *is called ordered if there exists a continuous strictly monotone scaling function* $\sigma_t^2,\ t\in\mathbb{R}^+$ *and some* $\Lambda>0$ *such that*

$$\sup_{u,v\in\mathbb{R}^+:\,u\neq v}\mathbf{E}\exp\left[\Lambda\frac{\Delta_\xi(u,v)}{\bar{\Delta}_\sigma(v,u)}\right]<\infty. \tag{2.8}$$

A banal example of an ordered process is a standard Wiener process w_t. In this case, $\sigma_t^2=t$ and obviously

$$\mathbf{E}\exp\left[\lambda\frac{\Delta_w(u,v)}{\sqrt{|v-u|}}\right]=\exp(\lambda^2/2).$$

Lemma 2 *Let* ξ_t *be an ordered process with* $\xi_0=0$*. Then there exists a constant* C *such that for all* $1\le p,q\le 2$*, uniformly in* $z>0$

$$\mathbf{E}\sup_{t\ge 0}\big[\xi_t-z\sigma_t^q\big]_+^p\le\frac{C}{z^{p/(q-1)}}, \tag{2.9}$$

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

Proof. Without loss of generality, we may assume that that $\lim_{t\to\infty}\sigma_t^2=\infty$. For any integer $k\ge 0$, define $t_k(z)$ as a root of the equation

$$\sigma_{t_k(z)}^{q-1}=\frac{2^{1/(q-1)}pk}{\Lambda z}.$$

Then we have

$$\begin{aligned}\mathbf{E}\sup_{t\ge 0}\big[\xi_t-z\sigma_t^q\big]_+^p&\le\sum_{k=0}^{\infty}\mathbf{E}\sup_{t\in[t_k(\gamma),t_{k+1}(z)]}\big[\xi_t-z\sigma_t^q\big]_+^p\\ &\le\sum_{k=0}^{\infty}\mathbf{E}\sup_{t\in[t_k(z),t_{k+1}(z)]}\big[\xi_t-z\sigma_{t_k}^q\big]_+^p\\ &\le\mathbf{E}\sup_{0\le t\le t_1(z)}|\xi_t|^p+\sum_{k=1}^{\infty}\mathbf{E}\Big[\sup_{0\le t\le t_{k+1}(z)}\xi_t-z\sigma_{t_k(z)}^q\Big]_+^p.\end{aligned} \tag{2.10}$$

According to Lemma 1, the first term at the right-hand side of the above inequality is bounded as follows :

$$\mathbf{E}\sup_{0\le t\le t_1(z)}|\xi_t|^p\le Cp^p\,\sigma_{t_1(z)}^p=\frac{Cp^p}{z^{p/(q-1)}}, \tag{2.11}$$

whereas the second one, by (1.12) and Lemma 1, is controlled by

$$\mathbf{E}\Big[\sup_{t\le t_{k+1}(\gamma)} \xi_t - z\sigma^q_{t_k(z)}\Big]^p_+ = \sum_{k=1}^{\infty} \sigma^p_{t_{k+1}(z)} \mathbf{E}\Big[\sup_{t\le t_{k+1}(z)} \frac{\xi_t}{\sigma_{t_{k+1}(z)}} \ge \frac{z\sigma^q_{t_k(z)}}{\sigma_{t_{k+1}(z)}}\Big]^p_+$$
$$\le C\sum_{k=1}^{\infty} \sigma^p_{t_{k+1}(z)} \exp\Big[-\Lambda \frac{z\sigma^q_{t_k(z)}}{\sigma_{t_{k+1}(z)}}\Big]$$
$$\le \frac{C}{z^{p/(q-1)}\lambda^{1/(q-1)}} \sum_{k=1}^{\infty} (k+1)^{p/(q-1)} \exp\Big[-\frac{2^{1/(q-1)}pk}{(1+1/k)^{1/(q-1)}}\Big] \le \frac{C}{z^{p/(q-1)}}.$$

So, combining the above inequality with (2.10) and (2.11), we arrive at (2.9). ■

The next very simple lemma is useful for understanding the fact that the ordered process is controlled by its variance σ_t^2 .

Lemma 3 *Let $\xi_t,\ t\in\mathbb{R}^+$, be a random process such that*

$$\mathbf{E}\sup_{t\ge 0}\big[\xi_t - z\sigma^q_t\big]^p_+ \le \frac{C}{z^{p/(q-1)}}, \tag{2.12}$$

for any $z>0$ and some $p\ge 1$, $q>1$. Then there exists a constant C' such that for any random variable $\tau\in\mathbb{R}^+$

$$\big[\mathbf{E}|\xi_\tau|^p\big]^{1/p} \le C'\big[\mathbf{E}\sigma_\tau^{qp}\big]^{1/(pq)}.$$

Proof. According to (2.12) and Minkowski's inequality, we obviously have

$$\begin{aligned}\big[\mathbf{E}|\xi_\tau|^p\big]^{1/p} \le& \big\{\mathbf{E}|\xi_\tau - z\sigma^q_\tau + z\sigma^q_\tau|^p\big\}^{1/p}\\ \le& \big\{\mathbf{E}\max_t[\xi_t - z\sigma^q_t]^p\big\}^{1/p} + z\big[\mathbf{E}\sigma_\tau^{pq}\big]^{1/p}\\ \le& z\big[\mathbf{E}\sigma_\tau^{pq}\big]^{1/p} + \frac{C}{z^{1/(q-1)}},\end{aligned}$$

and minimizing the right-hand side in z we finish the proof. ■

2.3 Ordered processes related to the spectral regularization

In this section we focus on typical ordered processes related to the empirical risk minimization.

For given $\alpha^\circ \in (0,\bar\alpha]$, define the following Gaussian processes

$$\xi_\alpha^+ = \sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ+\alpha}(k)\bigr] b(k)\xi(k), \quad 0<\alpha\le \bar\alpha-\alpha^\circ,$$
$$\xi_\alpha^- = \sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ-\alpha}(k)\bigr] b(k)\xi(k), \quad 0\le\alpha\le \alpha^\circ,$$

where $\xi(k)$ are i.i.d. $\mathcal{N}(0,1)$ and $\sum_{k=1}^n b^2(k) < \infty$. It is easily seen that ξ_α^+ and ξ_α^- are ordered processes. Indeed, since they are Gaussian, we can choose

$$\sigma_\alpha^\pm = \sqrt{\mathbf{E}(\xi_\alpha^\pm)^2}$$

and it suffices to check that

$$\bigl|\mathbf{E}(\xi_{\alpha_1}^\pm)^2 - \mathbf{E}(\xi_{\alpha_2}^\pm)^2\bigr| \ge \mathbf{E}(\xi_{\alpha_1}^\pm - \xi_{\alpha_2}^\pm)^2,$$

or equivalently,

$$\mathbf{E}\xi_{\alpha_1}^\pm \xi_{\alpha_2}^\pm \ge \min\bigl\{(\sigma_{\alpha_1}^\pm)^2, (\sigma_{\alpha_2}^\pm)^2\bigr\}.$$

If $\alpha_1 \le \alpha_2$, then we have

$$\begin{aligned}\mathbf{E}[\xi_{\alpha_1}^\pm]^2 &= \sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ\pm\alpha_1}(k)\bigr]\bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ\pm\alpha_1}(k)\bigr] b^2(k)\\ &\le \sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ\pm\alpha_1}(k)\bigr]\bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ\pm\alpha_2}(k)\bigr] b^2(k)\\ &= \mathbf{E}\xi_{\alpha_1}^\pm \xi_{\alpha_2}^\pm .\end{aligned}$$

Therefore according to Lemma 2, we get

$$\mathbf{E}\sup_{0\le\alpha\le\bar\alpha-\alpha^\circ}\left[\xi_\alpha^+ - z\left(\sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ+\alpha}(k)\bigr]^2 b^2(k)\right)^{q/2}\right]_+^p \le \frac{C(p,q)}{z^{p/(q-1)}},$$
$$\mathbf{E}\sup_{0\le\alpha\le\alpha^\circ}\left[\xi_\alpha^- - z\left(\sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_{\alpha^\circ-\alpha}(k)\bigr]^2 b^2(k)\right)^{q/2}\right]_+^p \le \frac{C(p,q)}{z^{p/(q-1)}},$$

and combining these inequalities we arrive at the following lemma

Lemma 4 *For any $z>0$,*

$$\mathbf{E}\sup_{0\le\alpha\le\bar\alpha}\Biggl\{\sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_\alpha(k)\bigr] b(k)\xi(k) - z\left[\sum_{k=1}^n \bigl[h_{\alpha^\circ}(k) - h_\alpha(k)\bigr]^2 b^2(k)\right]^{q/2}\Biggr\}_+^p \le \frac{C(p,q)}{z^{p/(q-1)}}.$$

The next fact is essential for bounding cross terms in the empirical risk.

Lemma 5 *Let α° be a given smoothing parameter. Then for any $p \in [1,2)$, there exists a constant $C(p)$ such that for any data-driven smoothing parameter $\hat{\alpha}$,*

$$\begin{aligned} &\mathbf{E}\bigg|\sum_{k=1}^{\infty}\big[h_{\hat{\alpha}}(k) - h_{\alpha^\circ}(k)\big]\lambda^{-1/2}(k)\theta(k)\xi(k)\bigg|^p \\ &\quad\le C(p)\Big\{\mathbf{E}\max_k \lambda^{-1}(k)h_{\hat{\alpha}}^2(k)\Big\}^{p/2}\bigg\{\sum_{k=1}^{\infty}\big[1 - h_{\alpha^\circ}(k)\big]^2\theta^2(k)\bigg\}^{p/2} \\ &\qquad + C(p)\Big\{\max_k \lambda^{-1}(k)h_{\alpha^\circ}^2(k)\Big\}^{p/2}\bigg\{\mathbf{E}\sum_{k=1}^{\infty}\big[1 - h_{\hat{\alpha}}(k)\big]^2\theta^2(k)\bigg\}^{p/2}. \end{aligned} \tag{2.13}$$

Proof. From Lemmas 4 and 3, it follows that

$$\begin{aligned} &\mathbf{E}\bigg|\sum_{k=1}^{\infty}\big[h_{\hat{\alpha}}(k) - h_{\alpha^\circ}(k)\big]\lambda^{-1/2}(k)\theta(k)\xi(k)\bigg|^p \\ &\quad\le C(p)\bigg\{\mathbf{E}\sum_{k=1}^{\infty}\big[h_{\alpha^\circ}(k) - h_{\hat{\alpha}}(k)\big]^2\lambda^{-1}(k)\theta^2(k)\bigg\}^{p/2}. \end{aligned} \tag{2.14}$$

To bound from above the right-hand side at the above display, we use that $h_\alpha(\cdot)$ is a family of ordered smoothers. With this in mind, let us assume for definiteness that $h_{\alpha_1}(k) \ge h_{\alpha_2}(k)$ for $\alpha_1 \le \alpha_2$. Then, if $\hat{\alpha} \ge \alpha^\circ$, we get

$$\frac{h_{\hat{\alpha}}(k)}{h_{\alpha^\circ}(k)} \le 1, \quad \frac{h_{\hat{\alpha}}(k)}{h_{\alpha^\circ}(k)} \ge h_{\hat{\alpha}}(k)$$

and thus we obtain

$$\begin{aligned} &\sum_{k=1}^{\infty}\big[h_{\alpha^\circ}(k) - h_{\hat{\alpha}}(k)\big]^2\lambda^{-1}(k)\theta^2(k) \\ &\quad= \sum_{k=1}^{\infty}H_{\alpha^\circ}^2(k)\bigg[1 - \frac{h_{\hat{\alpha}}(k)}{h_{\alpha^\circ}(k)}\bigg]^2\lambda^{-2}(k)\theta^2(k) \\ &\quad\le \max_k \lambda^{-1}(k)h_{\alpha^\circ}^2(k)\sum_{k=1}^{\infty}\big[1 - h_{\hat{\alpha}}(k)\big]^2\theta^2(k). \end{aligned} \tag{2.15}$$

Similarly, if $\hat{\alpha} < \alpha^\circ$, then

$$\begin{aligned}\sum_{k=1}^{\infty}\bigl[h_{\alpha^\circ}(k) - h_{\hat{\alpha}}(k)\bigr]^2\lambda^{-1}(k)\theta^2(k)& \\ \le \max_k \lambda^{-1}(k)h_{\hat{\alpha}}^2(k)\sum_{k=1}^{\infty}\bigl[1 - h_{\alpha^\circ}(k)\bigr]^2\theta^2(k)& \\ + \max_k \lambda^{-1}(k)h_{\alpha^\circ}^2(k)\sum_{k=1}^{\infty}\bigl[1 - h_{\hat{\alpha}}(k)\bigr]^2\theta^2(k).&\end{aligned} \tag{2.16}$$

Therefore combining (2.15–2.16), we get (2.13). ■

The following important ordered process is defined by

$$\zeta_\alpha = \sum_{k=1}^{n} h_\alpha(k)b(k)[\xi^2(k) - 1],$$

where $\xi(k)$ are i.i.d. $\mathcal{N}(0,1)$. Let

$$\sigma_\alpha = \left[2\sum_{k=1}^{n} h_\alpha^2(k)b^2(k)\right]^{1/2}.$$

It is easy to check that $\min\bigl\{\mathbf{E}\zeta_{\alpha_1}^2, \mathbf{E}\zeta_{\alpha_2}^2\bigr\} \le \mathbf{E}\zeta_{\alpha_2}\zeta_{\alpha_1}$ and thus

$$|\sigma_{\alpha_1}^2 - \sigma_{\alpha_1}^2| \ge \|h_{\alpha_1} - h_{\alpha_2}\|_b^2,$$

where

$$\|h_{\alpha_1} - h_{\alpha_2}\|_b^2 = \sum_{k=1}^{n} b^2(k)\bigl[h_{\alpha_1}(k) - h_{\alpha_2}(k)\bigr]^2.$$

Hence, in order to apply Lemma 2, it remains to check that for some $\Lambda > 0$

$$\sup_{\alpha_1,\alpha_2} \mathbf{E}\exp\left[\Lambda\frac{\Delta_\zeta(\alpha_1,\alpha_2)}{\|h_{\alpha_1} - h_{\alpha_2}\|_b}\right] < \infty.$$

We have

$$\begin{aligned}&\mathbf{E}\exp\left[\Lambda\frac{\Delta_\eta(\alpha_1,\alpha_2)}{\|h_{\alpha_1} - h_{\alpha_2}\|_b}\right] \\ &= \exp\biggl\{-\frac{\Lambda}{\sqrt{2}\|h_{\alpha_1} - h_{\alpha_2}\|_b}\sum_{k=1}^{n} b(k)\bigl[h_{\alpha_1}(k) - h_{\alpha_2}(k)\bigr] \\ &- \frac{1}{2}\sum_{k=1}^{n}\log\left[1 - \sqrt{2}\Lambda\frac{b(k)\bigl[h_{\alpha_1}(k) - h_{\alpha_2}(k)\bigr]}{\|h_{\alpha_1} - h_{\alpha_2}\|_b}\right]\biggr\}.\end{aligned} \tag{2.17}$$

Since obviously

$$\max_k \{|b(k)||h_{\alpha_1}(k) - h_{\alpha_2}(k)|\} \le \|h_{\alpha_1} - h_{\alpha_2}\|_b,$$

then using the Taylor expansion for $\log(1-\cdot)$ at the right-hand side of (2.17), we get for $\Lambda \le 1/2$

$$\mathbf{E}\exp\left(\Lambda \frac{\Delta_\eta(\alpha_1,\alpha_2)}{\|h_{\alpha_1} - h_{\alpha_2}\|_b}\right) \le \exp(C\lambda^2),$$

thus proving (2.8). Hence, with the help of Lemma 2 we obtain the following fact.

Lemma 6 *For any $z > 0$,*

$$\mathbf{E} \sup_{\alpha\in(0,\bar\alpha]} \left[\sum_{k=1}^n h_\alpha(k)b(k)[\xi^2(k)-1] - z\left[2\sum_{k=1}^n h_\alpha^2(k)b^2(k)\right]^{q/2}\right]_+^p \le \frac{C(p,q)}{z^{p/(q-1)}}.$$

2.4 Proof of Theorem 1

The next lemma describes some basic properties of the universal penalty defined by (1.13–1.15).

Lemma 7 *For any $\alpha \in (0,\bar\alpha]$,*

$$\log \frac{D(\alpha)}{D(\bar\alpha)} \le \mu_\alpha \frac{Q^\circ(\alpha)}{D(\alpha)}, \tag{2.18}$$

$$\mu_\alpha \ge \min\left\{\frac{1}{2}\sqrt{\log\frac{D(\alpha)}{D(\bar\alpha)}}, \frac{1}{4}\right\}, \tag{2.19}$$

$$\frac{Q^\circ(\alpha)}{D(\bar\alpha)} \ge \frac{D(\alpha)}{D(\bar\alpha)}\left(\log\frac{D(\alpha)}{D(\bar\alpha)}\right)^{1/2}, \tag{2.20}$$

$$\frac{D(\alpha)}{D(\bar\alpha)} \ge \frac{\mu_\alpha Q^\circ(\alpha)}{D(\bar\alpha)} \Big/ \log\frac{\mu_\alpha Q^\circ(\alpha)}{D(\bar\alpha)}, \quad \textit{if} \quad D(\alpha) \ge \exp(2)D(\bar\alpha), \tag{2.21}$$

$$\frac{D(\alpha_1)}{D(\alpha_2)} \le \frac{Q^\circ(\alpha_1)}{Q^\circ(\alpha_2)}, \quad \alpha_1 \le \alpha_2. \tag{2.22}$$

Proof. It follows from (1.13–1.15) that

$$\mu_\alpha \frac{Q^\circ(\alpha)}{D(\alpha)} + \sum_{k=1}^n \left\{\frac{1}{2}\log[1-2\mu_\alpha\rho_\alpha(k)] + \mu_\alpha\rho_\alpha(k)\right\} = \log\frac{D(\alpha)}{D(\bar\alpha)}$$

and together with the following inequality $\log(1-2x)/2 + x \le 0$, we get (2.18).

To verify (2.19), note that

$$F(x) \le \frac{2x^2}{1-2x}.$$

and therefore for any $\mu \in [0, 1/4]$,

$$\sum_{k=1}^{n} F[\mu\rho_\alpha(k)] \le 4\mu^2.$$

So, if $\mu_\alpha \le 1/4$, then

$$\mu_\alpha^2 \ge \frac{1}{4}\sum_{k=1}^{n} F[\mu_\alpha\rho_\alpha(k)] = \frac{1}{4}\log\frac{D(\alpha)}{D(\bar{\alpha})},$$

thus proving (2.19).

Next, note that the following inequality holds

$$F(x) \ge \frac{x^2}{1-2x} \tag{2.23}$$

since

$$f(x) = F(x) - \frac{x^2}{1-2x} = \frac{1}{2}\log(1-2x) + x + \frac{x^2}{1-2x}$$

is a nonnegative function for $x \ge 0$ because

$$f'(x) = \frac{2x^2}{(1-2x)^2} \ge 0$$

and $f(0) = 0$.

According to (2.23), $F(x) \ge x^2$ and therefore by (1.14)

$$\mu_\alpha^2 \le \frac{D(\alpha)}{D(\bar{\alpha})}.$$

Substituting this inequality into (2.18), we get (2.20).

We now turn to the proof of (2.21). Again, combining (2.23) with (1.13–1.15), we arrive at

$$\begin{aligned} Q^\circ(\alpha) =& 2D(\alpha)\mu_\alpha \sum_{k=1}^{n} \frac{\rho_\alpha^2(k)}{1-2\mu_\alpha\rho_\alpha(k)} \le \frac{2D(\alpha)}{\mu_\alpha}\sum_{k=1}^{n} F[\mu_\alpha\rho_\alpha(k)] \\ =& \frac{2D(\alpha)}{\mu_\alpha}\log\frac{D(\alpha)}{D(\bar{\alpha})} \end{aligned} \tag{2.24}$$

and to get (2.21) it remains to invert this equation. We proceed to show that if $x \geq \exp(2)$, then inequality $x \log(x) \geq y$ implies

$$x \geq \frac{y}{\log(y)}. \tag{2.25}$$

It is clear that $G(y) = y/\log(y)$ is an increasing function when $y \geq \exp(1)$ and (2.25) holds since

$$x \geq \frac{x\log(x)}{\log(x\log(x))} = \frac{x}{1 + \log\log(x)/\log(x)}.$$

Inverting (2.24) with the help of (2.25), we finish the proof of (2.21).

Finally, (2.22) follows from the fact that

$$g(\alpha) = \frac{Q^\circ(\alpha)}{D(\alpha)} = 2\mu_\alpha \sum_{k=1}^n \frac{\rho_\alpha^2(k)}{1 - 2\mu_\alpha \rho_\alpha(k)}$$

is a decreasing function in $\alpha > 0$. To check this, let us note that $g(\alpha)$ is a root of the equation

$$\exp\Big\{\inf_{\mu \geq 0}[-G_\alpha(\mu) - \mu g(\alpha)]\Big\} = \frac{D(\bar\alpha)}{D(\alpha)},$$

where

$$G_\alpha(\mu) = \sum_{k=1}^n \bigg[\mu\rho_\alpha(k) + \frac{1}{2}\log[1 - 2\mu\rho_\alpha(k)]\bigg].$$

However $\inf_{\mu\geq 0}[-G_\alpha(\mu) - \mu x]$ is obviously a decreasing function in x and therefore if $D(\bar\alpha)/D(\alpha)$ is decreasing in α, then $g(\alpha)$ is decreasing in α too. ■

We are now in a position to prove Theorem 1. Let $\alpha_k,\ k = 0, \ldots,$ be the decreasing sequence defined as follows:

$$\alpha_0 = \bar\alpha, \quad Q^\circ(\alpha_k) = (1+\delta)^{k-1} Q^\circ(\alpha_1),$$

where $\delta < 1/2$ is a small positive number which will be chosen later on, and α_1 is a root of equation

$$D(\alpha_1) = D(\bar\alpha)\exp(2).$$

Denote for brevity $D_k = D(\alpha_k)$ and $Q_k = Q^\circ(\alpha_k)$.

We begin with the simple inequality

$$\mathbf{E}\sup_{\alpha\le\bar\alpha}\Big[\eta_\alpha-(1+\gamma)Q^\circ(\alpha)\Big]_+^{1+r}\le\sum_{k=1}^n\mathbf{E}\sup_{\alpha_k\le\alpha<\alpha_{k-1}}\Big[\zeta_\alpha-(1+\gamma)Q_{k-1}\Big]_+^{1+r}$$
$$=\sum_{k=1}^n\mathbf{E}\Big[\zeta_{\alpha_k}-\epsilon\gamma Q_{k-1}+\sup_{\alpha_k\le\alpha<\alpha_{k-1}}[\zeta_\alpha-\zeta_{\alpha_k}]-(1+\gamma-\epsilon\gamma)Q_{k-1}\Big]_+^{1+r}.$$

Using that $[x+y]_+^{1+r}\le 2^r[x]_+^{1+r}+2^r[y]_+^{1+r}$, we can continue the above equation as follows :

$$\mathbf{E}\sup_{\alpha\le\bar\alpha}\Big[\eta_\alpha-(1+\gamma)Q^\circ(\alpha)\Big]_+^{1+r}\le 2^r\sum_{k=1}^n\mathbf{E}\Big[\zeta_{\alpha_k}-(1+\gamma-\epsilon\gamma)Q_{k-1}\Big]_+^{1+r}$$
$$+2^r\sum_{k=1}^n\mathbf{E}\Big[\sup_{\alpha_k\le\alpha<\alpha_{k-1}}[\zeta_\alpha-\zeta_{\alpha_k}]-\epsilon\gamma Q_{k-1}\Big]_+^{1+r}. \tag{2.26}$$

We control the first term

$$\Delta_1(\gamma,\epsilon)\stackrel{\mathrm{def}}{=}\sum_{k=1}^n\mathbf{E}\Big[\zeta_{\alpha_k}-(1+\gamma-\epsilon\gamma)Q_{k-1}\Big]_+^{1+r}$$

at the right-hand side of (2.26) with the help of (1.12). Thus, we obtain for any $\tilde\lambda_k>0$

$$\Delta_1(\gamma,\epsilon)\le\mathbf{E}|\zeta_{\alpha_1}|^{1+r}+\sum_{k=2}^n D_k^{1+r}\mathbf{E}\left[\frac{\zeta_{\alpha_k}}{D_k}-(1+\gamma-\epsilon\gamma)\frac{Q_{k-1}}{D_k}\right]_+^{1+r}$$
$$\le CD^{1+r}(\bar\alpha)+\Gamma(1+r)\sum_{k=2}^n\tilde\lambda_k^{-1-r}D_k^r\exp\bigg\{-\tilde\lambda_k\frac{Q_k}{D_k}\bigg[\frac{\gamma(1-\epsilon)Q_{k-1}}{Q_k}$$
$$-\left(1-\frac{Q_{k-1}}{Q_k}\right)\bigg]\bigg\}D_k\mathbf{E}\exp\left(\tilde\lambda_k\frac{\zeta_{\alpha_k}}{D_k}-\tilde\lambda_k\frac{Q_k}{D_k}\right). \tag{2.27}$$

According to (1.13–1.14), we have with $\tilde\lambda_k=\mu_{\alpha_k}$,

$$D_k\mathbf{E}\exp\left(\tilde\lambda_k\frac{\zeta_{\alpha_k}}{D_k}-\tilde\lambda_k\frac{Q_k}{D_k}\right)=D_0,$$

and substituting this into (2.27), we get

$$\Delta_1(\gamma,\epsilon)\le CD_0^{1+r}$$
$$\times\left\{1+\sum_{k=2}^n\tilde\lambda_k^{-1-r}\exp\left[-\tilde\lambda_k\frac{(\gamma(1-\epsilon)-\delta)Q_k}{(1+\delta)D_k}+r\log\frac{D_k}{D_0}\right]\right\}. \tag{2.28}$$

Since by (2.18)

$$\tilde{\lambda}_k \frac{Q_k}{D_k} \geq \log \frac{D_k}{D_0} \tag{2.29}$$

and according to (2.19), $\tilde{\lambda}_k$ is bounded from below by a constant, we obtain from (2.28)

$$\Delta_1(\gamma, \epsilon) \leq C D_0^{1+r} \bigg\{ 1 + \sum_{k=2}^{n} \exp\bigg[-\bigg(\frac{\gamma(1-\epsilon)-\delta}{1+\delta} - r \bigg) \log \frac{D_k}{D_0} \bigg] \bigg\}. \tag{2.30}$$

Next, according to (2.21) and (2.19), we get

$$\log \frac{D_k}{D_0} \geq \frac{CQ_k}{Q_0} \log^{-1} \frac{CQ_k}{Q_0} \tag{2.31}$$

and with this inequality we obtain from (2.30)

$$\begin{aligned} \Delta_1(\gamma, \epsilon) &\leq C D_0^{1+r} \sum_{k=2}^{n} \exp\bigg\{ -\bigg(\frac{\gamma(1-\epsilon)-\delta}{1+\delta} - r \bigg) \log \bigg[\frac{CQ_k}{Q_0} \log^{-1} \frac{CQ_k}{Q_0} \bigg] \bigg\} \\ &\leq C D_0^{1+r} \sum_{k=2}^{n} \exp\bigg\{ -\bigg(\frac{\gamma(1-\epsilon)-\delta}{1+\delta} - r \bigg) \big[k \log(1+\delta) - \log(k\delta) \big] \bigg\}. \end{aligned}$$

Finally, one can check by the Laplace method that

$$\sum_{k=1}^{\infty} \exp[-z_1 k + z_2 \log(k)] \leq \frac{C}{z_1} \bigg(\frac{z_2}{z_1} \bigg)^{z_2}, \; z_1, z_2 > 0$$

thus yielding

$$\Delta_1(\gamma, \epsilon) \leq \frac{C D_0^{1+r}}{\delta[\gamma(1-\epsilon) - \delta - (1+\delta) r]_+}. \tag{2.32}$$

Our next step is to bound from above the last term in (2.26), namely,

$$\Delta_2(\gamma, \epsilon) \stackrel{\text{def}}{=} \sum_{k=1}^{n} \mathbf{E} \bigg[\sup_{\alpha_k \leq \alpha < \alpha_{k-1}} [\zeta_\alpha - \zeta_{\alpha_k}] - \epsilon \gamma Q_{k-1} \bigg]_+^{1+r}.$$

Consider the following random processes

$$\tilde{\zeta}_\alpha(k) = \zeta_\alpha - \zeta_{\alpha_k}, \; t \in [0, \alpha_{k-1} - \alpha_k].$$

Denote for brevity $\tilde{\sigma}_\alpha(k) = \sqrt{\mathbf{E}\tilde{\zeta}_\alpha^2(k)}$. Noticing that $\tilde{h}_\alpha(k) = 2h_\alpha(k) - h_\alpha^2(k)$ is a family of ordered smoothers, it is easy to check that

$$\tilde{\sigma}_u^2(k) - \tilde{\sigma}_v^2(k) \geq \mathbf{E}[\tilde{\zeta}_u(k) - \tilde{\zeta}_v(k)]^2, \ u \leq v.$$

According to the Taylor formula, for all $u \geq v$ and all $\lambda \geq 0$,

$$\mathbf{E}\exp\left\{\lambda \frac{\tilde{\zeta}_u(k) - \tilde{\zeta}_v(k)}{\mathbf{E}^{1/2}[\tilde{\zeta}_u(k) - \tilde{\zeta}_v(k)]^2}\right\} \leq \exp\left(\frac{\lambda^2}{2}\right),$$

and applying Lemma 1 and (1.12), we obtain for any $\lambda_k \geq 0$

$$\begin{aligned}\Delta_2(\gamma,\epsilon) &= \sum_{k=1}^n \mathbf{E}\left[\sup_{\alpha\in[0,\alpha_{k-1}-\alpha_k]} \tilde{\zeta}_\alpha(k) - \epsilon\gamma Q_{k-1}\right]_+^{1+r} \leq CD_0^{1+r} \\ &+ C\sum_{k=2}^n \tilde{\sigma}_{\alpha_{k-1}-\alpha_k}^{1+r}(k)\lambda_k^{-1-r}\exp\left[-\frac{\lambda_k\epsilon\gamma Q_{k-1}}{\tilde{\sigma}_{\alpha_{k-1}-\alpha_k}(k)} + \frac{\lambda_k^2}{(\sqrt{2}-1)^2}\right].\end{aligned}$$

Substituting

$$\lambda_k = \frac{(\sqrt{2}-1)^2\epsilon\gamma Q_{k-1}}{2\tilde{\sigma}_{\alpha_{k-1}-\alpha_k}(k)}$$

into the above equation and noticing that $Q_k \geq D_k$, we obtain

$$\begin{aligned}\Delta_2(\gamma,\epsilon) &\leq CD_0^{1+r} \\ &+ \frac{C}{(\epsilon\gamma)^{1+r}}\sum_{k=2}^n \tilde{\sigma}_{\alpha_{k-1}-\alpha_k}^{2(1+r)}(k)Q_{k-1}^{-1-r}\exp\left[-\frac{(\sqrt{2}-1)^2\epsilon^2\gamma^2 Q_{k-1}^2}{4\tilde{\sigma}_{\alpha_{k-1}-\alpha_k}^2(k)}\right] \\ &\leq CD_0^{1+r} \\ &+ \frac{C}{(\epsilon\gamma)^{1+r}}\sum_{k=2}^n [D_k^2 - D_{k-1}^2]^{1+r}Q_{k-1}^{-1-r}\exp\left[-\frac{(\sqrt{2}-1)^2\epsilon^2\gamma^2 Q_{k-1}^2}{4(D_k^2 - D_{k-1}^2)}\right].\end{aligned} \tag{2.33}$$

According to (2.22),

$$\frac{D_k^2}{D_{k-1}^2} \leq \frac{Q_k^2}{Q_{k-1}^2} \leq (1+\delta)^2$$

and with this inequality we continue (2.33) as follows

$$\begin{aligned}\Delta_2(\gamma,\epsilon) &\leq CD_0^{1+r} \\ &+ \frac{C\delta^{1+r}}{(\epsilon\gamma)^{1+r}}\sum_{k=2}^n \left(\frac{D_{k-1}}{Q_{k-1}}\right)^{1+r} D_{k-1}^{1+r}\exp\left[-\frac{(\sqrt{2}-1)^2\epsilon^2\gamma^2 Q_{k-1}^2}{8\delta D_{k-1}^2}\right]\end{aligned}$$

Next, substituting (2.29) and (2.31) into this equation, we get

$$\begin{aligned}
\Delta_2(\gamma,\epsilon) &\le CD_0^{1+r} \\
&+ \frac{C\delta^{1+r}}{(\epsilon\gamma)^{1+r}} \sum_{k=2}^{n} Q_{k-1}^{1+r} \exp\biggl[-\frac{(\sqrt{2}-1)^2\epsilon^2\gamma^2}{8\delta\tilde{\mu}^2_{\alpha_{k-1}}} \log^2 \frac{D_{k-1}}{D_0}\biggr] \\
&\le CD_0^{1+r} + \frac{C\delta^{1+r}}{(\epsilon\gamma)^{1+r}} \sum_{k=2}^{n} Q_{k-1}^{1+r} \exp\biggl[-\frac{C\epsilon^2\gamma^2}{\delta} \log^2 \frac{Q_{k-1}}{D_0}\biggr] \\
&\le \frac{C\delta^{1+r}D_0^{1+r}}{(\epsilon\gamma)^{1+r}} \sum_{k=2}^{n} \exp\biggl[\delta(r+1)k - \frac{C\epsilon^2\gamma^2}{\delta}[k\log(1+\delta) - \log(\delta k)]^2\biggr].
\end{aligned}$$

Bounding the last sum in this display with the help of the Laplace method, we get

$$\begin{aligned}
&\sum_{k=2}^{n} \exp\biggl\{\delta(r+1)k - \frac{C\epsilon^2\gamma^2}{\delta}[k\log(1+\delta) - \log(\delta k)]^2\biggr\} \\
&\qquad\qquad \le \exp\biggl[\frac{C\delta^3}{\epsilon^2\gamma^2\log^2(1+\delta)}\biggr] \frac{C\sqrt{\delta}}{\epsilon\gamma\log(1+\delta)}.
\end{aligned}$$

and therefore with $\delta = \epsilon^2\gamma^2$ we obtain

$$\Delta_2(\gamma,\epsilon) \le \frac{CD_0^{1+r}}{(\epsilon\gamma)^{1-r}}. \tag{2.34}$$

With this δ the equation (2.32) becomes

$$\Delta_1(\gamma,\epsilon) \le \frac{CD^{1+r}(\bar{\alpha})}{(\gamma\epsilon)^2[\gamma - r - \gamma\epsilon - (1+r)(\gamma\epsilon)^2]_+}. \tag{2.35}$$

Let ϵ be a positive root of the equation

$$\gamma - r - \gamma\epsilon - (1+r)(\gamma\epsilon)^2 = \gamma\epsilon,$$

i.e.

$$\epsilon = -\frac{1}{(r+1)\gamma} + \frac{1}{\gamma}\sqrt{\frac{1}{(1+r)^2} + \frac{\gamma - r}{1+r}}.$$

Substituting this ϵ into (2.34–2.35) and combining thus obtained inequalities with (2.26), we finish the proof.

2.5 Proof of Theorem 2

The first step in the proof of this theorem is to show that the data-driven parameter $\hat{\alpha}$ defined by (1.5) cannot be very small, or equivalently, that the ratio $D(\hat{\alpha})/D(\bar{\alpha})$ is not large.

Lemma 8 *For any given $\alpha^\circ \le \bar{\alpha}$ and $\gamma > 0$ the following upper bound holds*

$$\left\{\mathbf{E}\left[\frac{D(\hat{\alpha})}{D(\bar{\alpha})}\right]^{1+\gamma/4}\right\}^{1/(1+\gamma/4)} \le \frac{C\bar{R}_{\alpha^\circ}(\theta)}{\sigma^2\gamma D(\bar{\alpha})}\log^{-1/2}\frac{C\bar{R}_{\alpha^\circ}(\theta)}{\sigma^2 D(\bar{\alpha})} + \frac{C}{\gamma^4}. \tag{2.36}$$

Proof. According to the definition of the empirical risk minimization, for any given α°, $R_{\hat{\alpha}}[Y, Pen] \le R_{\alpha^\circ}[Y, Pen]$. One can check with a little algebra that this inequality is equivalent to (see (1.8))

$$\begin{aligned}
&\sum_{k=1}^{n}\big[1-h_{\hat{\alpha}}(k)\big]^2\langle\theta,\psi_k\rangle^2 + \sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)h_{\hat{\alpha}}^2(k)\\
&\quad - \sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)\tilde{h}_{\hat{\alpha}}(k)[\xi^2(k)-1] + (1+\gamma)\sigma^2 Q^\circ(\hat{\alpha})\\
&\quad + 2\sigma\sum_{k=1}^{n}\lambda^{-1/2}(k)\big[1-h_{\hat{\alpha}}(k)\big]^2\xi(k)\theta(k)\\
\le&\sum_{k=1}^{n}\big[1-h_{\alpha^\circ}(k)\big]^2\langle\theta,\psi_k\rangle^2 + \sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)h_{\alpha^\circ}^2(k)\\
&\quad - \sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)\tilde{h}_{\alpha^\circ}(k)[\xi^2(k)-1] + (1+\gamma)\sigma^2 Q^\circ(\alpha^\circ)\\
&\quad + 2\sigma\sum_{k=1}^{n}\lambda^{-1/2}(k)\big[1-h_{\alpha^\circ}(k)\big]^2\xi(k)\theta(k),
\end{aligned} \tag{2.37}$$

where $\tilde{h}_\alpha(k) = 2h_\alpha(k) - h_\alpha^2(k)$. Next, representing

$$(1+\gamma)Q^\circ(\hat{\alpha}) = \left(1+\frac{\gamma}{2}\right)Q^\circ(\hat{\alpha}) + \frac{\gamma}{2}Q^\circ(\hat{\alpha}),$$

we obtain from (2.37)

$$
\begin{aligned}
\frac{\gamma\sigma^2}{2} Q^\circ(\hat\alpha) &\le \bar R_{\alpha^\circ}(\theta) + \sigma^2 \sum_{k=1}^n \lambda^{-1}(k)\tilde h_{\alpha^\circ}(k)[\xi^2(k)-1] \\
&+ \sigma^2 \sup_{\alpha\le\bar\alpha}\Big[\sum_{k=1}^n \lambda^{-1}(k)\tilde h_\alpha(k)[\xi^2(k)-1] - \Big(1+\frac{\gamma}{2}\Big)Q^\circ(\alpha)\Big]_+ \\
&+ 2\sigma \sum_{k=1}^n \lambda^{-1/2}(k)\big[\tilde h_{\hat\alpha}(k) - \tilde h_{\alpha^\circ}(k)\big]\xi(k)\theta(k) - L_{\hat\alpha}(\theta).
\end{aligned}
\tag{2.38}
$$

Since α° is fixed, we get by Jensen's inequality

$$
\begin{aligned}
\mathbf{E}\Big|\sum_{k=1}^n \lambda^{-1}(k)\tilde h_{\alpha^\circ}(k)[\xi^2(k)-1]\Big|^{1+\gamma/4} &\le C\Big[\sum_{k=1}^n \lambda^{-2}(k)\tilde h^2_{\alpha^\circ}(k)\Big]^{1/2+\gamma/8} \\
&= C\big[D(\alpha^\circ)\big]^{1+\gamma/4} \le C\big[\sigma^{-2}\bar R_{\alpha^\circ}(\theta)\big]^{1+\gamma/4}.
\end{aligned}
\tag{2.39}
$$

Next, by Theorem 1,

$$
\begin{aligned}
\mathbf{E}\sup_{\alpha\le\bar\alpha}\Big[\sum_{k=1}^n \lambda^{-1}(k)\tilde h_\alpha(k)[\xi^2(k)-1] - \Big(1+\frac{\gamma}{2}\Big)Q^\circ(\alpha)\Big]_+^{1+\gamma/4} \\
\le \frac{CD^{1+\gamma/4}(\bar\alpha)}{\gamma^3}.
\end{aligned}
\tag{2.40}
$$

The upper bound for the last line in (2.38) is a little bit more tricky. Noticing that $\tilde h_\alpha(\cdot)$ is a family of ordered smoothers, we get by Lemma 4 that for any $\epsilon > 0$ and given $p \in (1,2)$

$$
\begin{aligned}
\mathbf{E}\Big|2\sigma \sum_{k=1}^n \lambda^{-1/2}(k)\big[\tilde h_{\hat\alpha}(k) - \tilde h_{\alpha^\circ}(k)\big]\xi(k)\theta(k) \\
- \epsilon\Big[4\sigma^2 \sum_{k=1}^n \lambda^{-1}(k)\big[\tilde h_{\hat\alpha}(k) - \tilde h_{\alpha^\circ}(k)\big]^2\theta^2(k)\Big]^{p/2}\Big|^{1+\gamma/4} \\
\le \frac{C(p)}{\epsilon^{(1+\gamma/4)/(p-1)}}.
\end{aligned}
\tag{2.41}
$$

To continue this inequality, note that if $\hat\alpha \ge \alpha^\circ$, then

$$
\frac{\tilde h_{\hat\alpha}(k)}{\tilde h_{\alpha^\circ}(k)} \le 1, \quad \frac{\tilde h_{\hat\alpha}(k)}{\tilde h_{\alpha^\circ}(k)} \ge \tilde h_{\hat\alpha}(k)
$$

and therefore

$$\begin{aligned}
&\sum_{k=1}^{\infty}\big[\tilde h_{\alpha^\circ}(k)-\tilde h_{\hat\alpha}(k)\big]^2\lambda^{-1}(k)\theta^2(k)\\
&\qquad=\sum_{k=1}^{\infty}\tilde h_{\alpha^\circ}^2(k)\bigg[1-\frac{\tilde h_{\hat\alpha}(k)}{\tilde h_{\alpha^\circ}(k)}\bigg]^2\lambda^{-2}(k)\theta^2(k)\\
&\qquad\le\max_k\lambda^{-1}(k)\tilde h_{\alpha^\circ}^2(k)\sum_{k=1}^{\infty}\big[1-\tilde h_{\hat\alpha}(k)\big]^2\theta^2(k)\\
&\qquad\le 4\max_k\lambda^{-1}(k)h_{\alpha^\circ}^2(k)\sum_{k=1}^{\infty}\big[1-h_{\hat\alpha}(k)\big]^2\theta^2(k).
\end{aligned}\tag{2.42}$$

Analogously, if $\hat\alpha<\alpha^\circ$, then

$$\begin{aligned}
&\sum_{k=1}^{n}\big[\tilde h_{\alpha^\circ}(k)-\tilde h_{\hat\alpha}(k)\big]^2\lambda^{-1}(k)\theta^2(k)\\
&\qquad\le\max_k\lambda^{-1}(k)\tilde h_{\hat\alpha}^2(k)\sum_{k=1}^{n}\big[1-\tilde h_{\alpha^\circ}(k)\big]^2\theta^2(k)\\
&\qquad\le 4\max_k\lambda^{-1}(k)h_{\hat\alpha}^2(k)\sum_{k=1}^{n}\big[1-h_{\alpha^\circ}(k)\big]^2\theta^2(k).
\end{aligned}\tag{2.43}$$

Next, combining (2.41–2.43) with Young's inequality

$$yx^q-x\le y^{-1/(q-1)}\Big[q^{-q/(q-1)}-q^{1/(q-1)}\Big],\quad x,y\ge 0,\ q<1,\tag{2.44}$$

gives

$$
\begin{aligned}
&\mathbf{E}\Big|2\sigma\sum_{k=1}^{n}\lambda^{-1/2}(k)\big[\tilde h_{\hat\alpha}(k)-\tilde h_{\alpha^\circ}(k)\big]\xi(k)\theta(k)-L_{\hat\alpha}(\theta)\Big|^{1+\gamma/4}\\
&\le C\mathbf{E}\Big|2\sigma\sum_{k=1}^{n}\lambda^{-1/2}(k)\big[\tilde h_{\hat\alpha}(k)-\tilde h_{\alpha^\circ}(k)\big]\xi(k)\theta(k)\\
&\qquad-\epsilon\Big[4\sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)\big[\tilde h_{\hat\alpha}(k)-\tilde h_{\alpha^\circ}(k)\big]^2\theta^2(k)\Big]^{p/2}\Big|^{1+\gamma/4}\\
&\qquad+C\mathbf{E}\Big|\epsilon\Big[4\sigma^2\sum_{k=1}^{n}\lambda^{-1}(k)\big[\tilde h_{\hat\alpha}(k)-\tilde h_{\alpha^\circ}(k)\big]^2\theta^2(k)\Big]^{p/2}-L_{\hat\alpha}(\theta)\Big|^{1+\gamma/4}\\
&\le\frac{C}{\epsilon^{(1+\gamma/4)/(p-1)}}+\frac{C}{\epsilon^{2(1+\gamma/4)/(p-2)}}\Big[\sigma^2\max_k\lambda^{-1}(k)h^2_{\alpha^\circ}(k)\Big]^{\frac{p(1+\gamma/4)}{2-p}}\\
&\qquad+\frac{C}{\epsilon^{2(1+\gamma/4)/(p-2)}}\Big\{\sum_{k=1}^{\infty}\big[1-h_{\alpha^\circ}(k)\big]^2\theta^2(k)\Big\}^{\frac{p(1+\gamma/4)}{2-p}}.
\end{aligned}
$$

Therefore, minimizing the right-hand side at the above equation in $\epsilon>0$, we get

$$
\begin{aligned}
&\mathbf{E}\Big|2\sigma\sum_{k=1}^{n}\lambda^{-1}(k)\big[\tilde h_{\hat\alpha}(k)-\tilde h_{\alpha^\circ}(k)\big]\xi(k)\theta(k)-L_{\hat\alpha}(\theta)\Big|^{1+\gamma/4}\\
&\le C\Big\{\sum_{k=1}^{\infty}\big[1-h_{\alpha^\circ}(k)\big]^2\theta^2(k)+\sigma^2\max_k\lambda^{-1}(k)h^2_{\alpha^\circ}(k)\Big\}^{1+\gamma/4}.
\end{aligned}
$$

This equation and (2.38–2.40) imply

$$
\gamma^{1+\gamma/4}\mathbf{E}[\sigma^2Q^\circ(\hat\alpha)]^{1+\gamma/4}\le C\bar R^{1+\gamma/4}_{\alpha^\circ}(\theta)+\frac{C[\sigma^2D(\bar\alpha)]^{1+\gamma/4}}{\gamma^3}
$$

and by (2.20) we get

$$
\gamma^{1+\gamma/4}\mathbf{E}\Big[\frac{D(\hat\alpha)}{D(\bar\alpha)}\log^{1/2}\frac{D(\hat\alpha)}{D(\bar\alpha)}\Big]^{1+\gamma/4}\le C\Big[\frac{\bar R_{\alpha^\circ}(\theta)}{\sigma^2D(\bar\alpha)}\Big]^{1+\gamma/4}+\frac{C}{\gamma^3}. \tag{2.45}
$$

It is easily seen that

$$
\begin{aligned}
&\mathbf{E}\Big[\frac{D(\hat\alpha)}{D(\bar\alpha)}\log^{1/2}\frac{D(\hat\alpha)}{D(\bar\alpha)}\Big]^{1+\gamma/4}\\
&=\frac{1}{(1+\gamma/4)^{1/2+\gamma/8}}\mathbf{E}\Big[\frac{D(\hat\alpha)}{D(\bar\alpha)}\Big]^{1+\gamma/4}\Big[\log\Big(\frac{D(\hat\alpha)}{D(\bar\alpha)}\Big)^{1+\gamma/4}\Big]^{1/2+\gamma/8}.
\end{aligned} \tag{2.46}
$$

To finish the proof, let us consider the function $f(x) = x \log^{1/2+\gamma/8}(x)$, $x \geq 1$. Computing its second order derivative, one can easily check that $f(x)$ is convex for all $x \geq \exp(1) = \mathrm{e}$. So, $f(x + \mathrm{e} - 1)$ is convex for $x \geq 1$. Note also that there exists a constant $C > 0$ such that for all $x \geq 1$

$$f(x) \geq \frac{1}{2} f(x + \mathrm{e} - 1) - C.$$

Therefore according to (2.46) and Jensen's inequality,

$$\begin{aligned} \mathbf{E}\biggl(\frac{D(\hat{\alpha})}{D(\bar{\alpha})} \log^{1/2} \frac{D(\hat{\alpha})}{D(\bar{\alpha})}\biggr)^{1+\gamma/4} &\geq C\biggl[\mathbf{E}\biggl(\frac{D(\hat{\alpha})}{D(\bar{\alpha})}\biggr)^{1+\gamma/4} + \mathrm{e} - 1\biggr] \\ \times \biggl\{\log\biggl[\mathbf{E}\biggl(\frac{D(\hat{\alpha})}{D(\bar{\alpha})}\biggr)^{1+\gamma/4} &+ \mathrm{e} - 1\biggr]\biggr\}^{1/2+\gamma/8} - C. \end{aligned}$$

Finally, substituting this inequality into (2.45) and inverting $f(x)$, we arrive at (2.36). ■

The next lemma controls the cross term in the empirical risk.

Lemma 9 *Let* $\tilde{h}^{\epsilon}_{\alpha}(k) = \bigl[(1+2\epsilon)h_\alpha(k) - \epsilon h^2_\alpha(k)\bigr]/(1+\epsilon)$. *Then for any given* $\epsilon \geq 0$ *and* $\alpha^\circ \in (0, \bar{\alpha}]$

$$\begin{aligned} 2\sigma\mathbf{E}\biggl|\sum_{k=1}^n &\bigl[1 - \tilde{h}^{\epsilon}_{\hat{\alpha}}(k)\bigr]\theta(k)\lambda^{-1/2}(k)\xi(k)\biggr| \\ &\leq \biggl[\frac{C\bar{R}_{\alpha^\circ}(\theta)}{\gamma} \log^{-1/2} \frac{C\bar{R}_{\alpha^\circ}(\theta)}{\sigma^2 D(\bar{\alpha})} + \frac{C\sigma^2 D(\bar{\alpha})}{\gamma^4}\biggr]^{1/2} \\ &\times \biggl[\mathbf{E}\sum_{k=1}^n \bigl[1 - h_{\hat{\alpha}}(k)\bigr]^2\theta^2(k) + \sum_{k=1}^n \bigl[1 - h_{\alpha^\circ}(k)\bigr]^2\theta^2(k)\biggr]^{1/2}. \end{aligned} \tag{2.47}$$

Proof. Since $\tilde{h}^{\epsilon}_{\alpha}(k)$ is a family of ordered smoothers, combining Lemma 5 with the obvious inequalities $\max_k \lambda^{-1}(k)h^2_\alpha(k) \leq D(\alpha)$ and $\tilde{h}^{\epsilon}_{\alpha}(k) \geq h_\alpha(k)$, we obtain

$$\begin{aligned} 2\sigma\mathbf{E}\biggl|\sum_{k=1}^n &\bigl[1 - \tilde{h}^{\epsilon}_{\hat{\alpha}}(k)\bigr]\theta(k)\lambda^{-1/2}(k)\xi(k)\biggr| \\ &= 2\sigma\mathbf{E}\biggl|\sum_{k=1}^n \bigl[h^{\epsilon}_{\alpha^\circ}(k) - h^{\epsilon}_{\hat{\alpha}}(k)\bigr]\theta(k)\lambda^{-1/2}(k)\xi(k)\biggr| \\ &\leq C\sigma\biggl[\mathbf{E}D(\hat{\alpha})\sum_{k=1}^n \bigl[1 - h_{\alpha^\circ}(k)\bigr]^2\theta^2(k)\biggr]^{1/2} \\ &\quad + C\sigma\biggl[D(\alpha^\circ)\mathbf{E}\sum_{k=1}^n \bigl[1 - h_{\hat{\alpha}}(k)\bigr]^2\theta^2(k)\biggr]^{1/2}. \end{aligned} \tag{2.48}$$

Next, according to (2.20), $Q^\circ(\alpha) \ge D(\alpha)\sqrt{\log[D(\alpha)/D(\bar\alpha)]}$, and we get

$$D(\alpha^\circ) \le C\sigma^{-2}\bar R_{\alpha^\circ}(\theta)\log^{-1/2}\frac{\bar R_{\alpha^\circ}(\theta)}{\sigma^2 D(\bar\alpha)}.$$

Substituting this inequality and (2.36) in (2.48), we obtain (2.47). ■

We are now in a position to prove Theorem 2. Let $\epsilon \in (0,1]$ be a given number to be defined later on. According to (1.8) and (1.9), we obtain the following equation for the skewed excess risk

$$\begin{aligned}
&\mathcal{E}(\epsilon) \stackrel{\text{def}}{=} \sup_{\theta\in\mathbb{R}^n} \mathbf{E}\Big\{\|\theta-\hat\theta_{\hat\alpha}\|^2 - (1+\epsilon)\big\{R_{\hat\alpha}[Y,Pen]+\mathcal{C}\big\}\Big\} \\
&= \sup_{\theta\in\mathbb{R}^n}\mathbf{E}\Big\{-\epsilon\sum_{k=1}^n\big[1-h_{\hat\alpha}(k)\big]^2\theta^2(k) - \epsilon\sigma^2\sum_{k=1}^n \lambda^{-1}(k)h^2_{\hat\alpha}(k) \\
&\quad - (1+\epsilon)(1+\gamma)\sigma^2 Q^\circ(\hat\alpha) \\
&\quad - 2\sigma\sum_{k=1}^n\big\{1+\epsilon-\big[(1+2\epsilon)h_{\hat\alpha}(k)-\epsilon h^2_{\hat\alpha}(k)\big]\big\}\theta(k)\lambda^{-1/2}(k)\xi(k) \\
&\quad + \sigma^2\sum_{k=1}^n\lambda^{-1}(k)\big[2(1+\epsilon)h_{\hat\alpha}(k)-\epsilon h^2_{\hat\alpha}(k)\big][\xi^2(k)-1]\Big\}.
\end{aligned} \tag{2.49}$$

To control the last line at the right-hand side of this equation, we use that $h^\epsilon_\alpha(k) = \big[2(1+\epsilon)h_\alpha(k)-\epsilon h^2_\alpha(k)\big]/(2+\epsilon)$ is a family of ordered smoothers. Hence, Lemmas 3, 6, and 8 imply

$$\begin{aligned}
&\sigma^2\mathbf{E}\sum_{k=1}^n\lambda^{-1}(k)\big[2(1+\epsilon)h_{\hat\alpha}(k)-\epsilon h^2_{\hat\alpha}(k)\big][\xi^2(k)-1] \\
&\qquad \le C\frac{\bar R_{\alpha^\circ}(\theta)}{\gamma}\log^{-1/2}\frac{C\bar R_{\alpha^\circ}(\theta)}{\sigma^2\gamma D(\bar\alpha)} + \frac{C\sigma^2 D(\bar\alpha)}{\gamma^4}.
\end{aligned} \tag{2.50}$$

Next, substituting (2.50) and (2.47) into (2.49), we obtain the following upper bound for the skewed excess risk

$$\mathcal{E}(\epsilon) \le \epsilon\bar R_{\alpha^\circ}(\theta) + \frac{C}{\epsilon}\bigg[\frac{\bar R_{\alpha^\circ}(\theta)}{\gamma}\log^{-1/2}\frac{C\bar R_{\alpha^\circ}(\theta)}{\sigma^2 D(\bar\alpha)} + \frac{\sigma^2 D(\bar\alpha)}{\gamma^4}\bigg].$$

Finally, substituting this upper bound into

$$\mathbf{E}\|\theta-\hat\theta_{\hat\alpha}\|^2 \le (1+\epsilon)\bar R_{\alpha^\circ}(\theta) + \mathcal{E}(\epsilon)$$

and minimizing thus obtained inequality in ϵ, we get

$$\begin{aligned}\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \le& r(\theta) + Cr(\theta) \inf_{\epsilon \ge 0} \left\{ \epsilon + \frac{1}{\epsilon} \left[\frac{1}{\gamma} \log^{-1/2} \frac{Cr(\theta)}{\sigma^2 D(\bar{\alpha})} + \frac{C\sigma^2 D(\bar{\alpha})}{\gamma^4 r(\theta)} \right] \right\} \\ \le& r(\theta) \left\{ 1 + \left[\frac{C}{\gamma} \log^{-1/2} \frac{Cr(\theta)}{\sigma^2 D(\bar{\alpha})} + \frac{C\sigma^2 D(\bar{\alpha})}{\gamma^4 r(\theta)} \right]^{1/2} \right\},\end{aligned}$$

thus finishing the proof. ■

References

[Akaike (1973)] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Intern. Symp. Inf. Theory, Petrov P.N. and Csaki F. eds. Budapest.* 267-281. MR483125

[Bauer and Hohage (2005)] Bauer, F. and Hohage, T. (2005). A Lepski-type stopping rule for regularized Newton methods. *Inverse Problems.* **21** 1975–1991. MR2183662

[Bissantz et al. (2007)] Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* **45** no. 6 2610–2636. MR2361904

[Cavalier and Golubev (2006)] Cavalier, L. and Golubev, Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34** 1653–1677. MR2283712

[Engl et al. (1996)] Engl, H.W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems. Mathematics and its Applications, 375.* Kluwer Academic Publishers Group. Dordrecht. MR1408680

[Golubev (2004)] Golubev, Yu. (2004). The principle of penalized empirical risk in severely ill-posed problems. *Probab. Theory Related Fields.* **130** 18–38. MR2092871

[Kneip (1994)] Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866. MR1292543

[Landweber (1951)] Landweber, L. (1951). An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.* **73** 615–624. MR0043348

[Lubes and Ludeña (2008)] LOUBES J.-M. AND LUDEÑA, C. (2008). Adaptive complexity regularization for linear inverse problems, *Electron. J. Stat.* **2** 661–677. MR2426106

[Mair and Ruymgaart (1996)] MAIR, B. AND RUYMGAART, F.H. (1996). Statistical inverse estimation in Hilbert scale. *SIAM J. Appl. Math.* **56** no. 5 1424–1444. MR1409127

[Mathé (2006)] MATHÉ, P. (2006). The Lepskii principle revised. *Inverse Problemes.* **22** no. 3 L11-L15. MR2235633

[O'Sullivan (1986)] O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** no. 4 501–527. MR874480

[Pinsker (1980)] PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission.* **16** 120–133. MR0624591

[Tikhonov and Arsenin (1977)] TIKHONOV, A. N. AND ARSENIN, V. A. (1977). *Solution of Ill-posed Problems Translated from the Russian. Preface by translation editor Fritz John. Scripta Series in Mathematics.* V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York. MR0455365

[Van der Vaart and Wellner (1996)] VAN DER VAART, A. AND WELLNER, J. A. (1996). *Weak convergence and empirical processes.* Springer-Verlag, New York MR1385671