The Annals of Statistics
2002, Vol. 30, No. 3, 843–874

ORACLE INEQUALITIES FOR INVERSE PROBLEMS

BY L. CAVALIER, G. K. GOLUBEV, D. PICARD AND A. B. TSYBAKOV

Université Aix-Marseille 1, Université Aix-Marseille 1,
Université Paris 7 and Université Paris 6

We consider a sequence space model of statistical linear inverse problems where we need to estimate a function f from indirect noisy observations. Let a finite set Λ of linear estimators be given. Our aim is to mimic the estimator in Λ that has the smallest risk on the true f. Under general conditions, we show that this can be achieved by simple minimization of an unbiased risk estimator, provided the singular values of the operator of the inverse problem decrease as a power law. The main result is a nonasymptotic oracle inequality that is shown to be asymptotically exact. This inequality can also be used to obtain sharp minimax adaptive results. In particular, we apply it to show that minimax adaptation on ellipsoids in the multivariate anisotropic case is realized by minimization of unbiased risk estimator without any loss of efficiency with respect to optimal nonadaptive procedures.

1. Introduction. Let A be a known linear operator on a Hilbert space H with inner product $(\cdot, \cdot)$ and norm $\|\cdot\|$. Let $f \in H$ be an unknown function that we want to estimate from indirect observations

$$Y(g) = (Af, g) + \varepsilon\xi(g), \qquad g \in H, \tag{1}$$

where $0 < \varepsilon < 1$ and $\xi(g)$ is a Gaussian random variable on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, with mean 0 and variance $\|g\|^2$, such that $\mathbf{E}\{\xi(g)\xi(v)\} = (g, v)$, for any $g, v \in H$, where $\mathbf{E}$ is the expectation w.r.t. $\mathbf{P}$.

Relation (1) defines a Gaussian white noise model. Instead of using all the observations $\{Y(g),\ g \in H\}$ it is usually sufficient to consider the set of values $\{Y(\psi_k)\}$, for some orthonormal basis $\{\psi_k\}_{k=1}^{\infty}$. Then $\xi(\psi_k) = \xi_k$ are i.i.d. standard Gaussian random variables.

We assume that the basis $\{\psi_k\}$ is such that $(Af, \psi_k) = b_k\theta_k$, where b_k are real numbers and $\theta_k = (f, \varphi_k)$ are the Fourier coefficients of f w.r.t. some orthonormal basis $\{\varphi_k\}$ (not necessarily $\varphi_k = \psi_k$). A typical example when it occurs is that the operator A admits a singular value decomposition of the form

$$A\varphi_k = b_k\psi_k, \qquad A^*\psi_k = b_k\varphi_k, \tag{2}$$

where A^* is the adjoint of A, b_k are the singular values, $\{\psi_k\}$ is an orthonormal basis in $\mathrm{Range}(A) \subset H$ and $\{\varphi_k\}$ is the corresponding orthonormal basis in H.

Received July 2000; revised October 2001.
AMS 2000 subject classifications. 62G05, 62G20.
Key words and phrases. Statistical inverse problems, oracle inequalities, adaptive curve estimation, model selection, exact minimax constants.

Under these assumptions (but also in some other situations) one has the equivalent discrete sequence observation model derived from (1):

$$y_k = b_k\theta_k + \varepsilon\xi_k, \qquad k = 1, 2, \ldots, \tag{3}$$

where y_k stands for $Y(\psi_k)$. If $b_k \neq 0$, the model (3) can be written in the form

$$X_k = \theta_k + \varepsilon\sigma_k\xi_k, \qquad k = 1, 2, \ldots, \tag{4}$$

where $X_k = y_k/b_k$ and $\sigma_k = b_k^{-1}$. This can also be viewed as a model with direct observations and correlated data [Johnstone (1999)]. The sequence space formulation (3) or (4) for statistical inverse problems has been studied in a number of papers [see Johnstone and Silverman (1990), Korostelev and Tsybakov (1993), Koo (1993), Donoho (1995), Mair and Ruymgaart (1996), Golubev and Khasminskii (1999, 2001), Johnstone (1999), Goldenshluger and Pereverzev (2000) and Cavalier and Tsybakov (2002) among others]. For ill-posed inverse problems we have $|b_k| \to 0$ and $|\sigma_k| \to \infty$, as $k \to \infty$.

Let $\hat\theta = (\hat\theta_1, \hat\theta_2, \ldots)$ be an estimator of $\theta = (\theta_1, \theta_2, \ldots)$ based on the data (4). Then f is estimated by $\hat f = \sum_k \hat\theta_k\varphi_k$. The mean integrated squared risk (MISE) of $\hat f$ is

$$\mathcal{R}(\hat f, f) = \mathbf{E}_f\|\hat f - f\|^2 = \mathbf{E}_\theta\sum_k(\hat\theta_k - \theta_k)^2 = \mathbf{E}_\theta\|\hat\theta - \theta\|^2,$$

where the notation $\|\cdot\|$ means the ℓ_2-norm when applied to θ-vectors in the sequence space. Here and later $\mathbf{E}_f$ and $\mathbf{E}_\theta$ denote the expectations w.r.t. $\{Y(g),\ g \in H\}$ or $X = (X_1, X_2, \ldots)$ for models (1) and (4), respectively. Analyzing the risk $\mathcal{R}(\hat f, f)$ of the estimator $\hat f$ is equivalent to analyzing the corresponding sequence space risk $\mathbf{E}_\theta\|\hat\theta - \theta\|^2$.

Let $\lambda = (\lambda_1, \lambda_2, \ldots)$ be a sequence of nonrandom weights, also called *a filter*. Every filter λ defines the estimator $\hat\theta(\lambda) = (\hat\theta_1, \hat\theta_2, \ldots)$, where $\hat\theta_k = \lambda_k X_k$. Examples of commonly used weights λ_k are the projection weights $\lambda_k = I(k \le w)$, for some integer w [where $I(\cdot)$ denotes the indicator function], the Tikhonov–Phillips weights

$$\lambda_k = \frac{1}{1 + (k/w)^\alpha}, \qquad w > 0,\ \alpha > 0,$$

and the Pinsker (1980) weights

$$\lambda_k = (1 - (k/w)^\alpha)_+, \qquad w > 0,\ \alpha > 0,$$

where $x_+ = \max(x, 0)$. The estimator $\hat f$ with Tikhonov–Phillips weights for even values α is asymptotically equivalent to a smoothing spline estimator [Wahba (1977, 1990)].

Although λ is not finite dimensional, it is usually determined by a finite number of parameters, as in the above examples. In this paper we discuss how to choose

these parameters optimally in a data-driven way. In particular, a data-driven choice of smoothing parameters w and α for the Tikhonov–Phillips method is interesting.

We suppose that there is a finite set of possible candidate filters $\Lambda = \{\lambda^1, \ldots, \lambda^N\}$, with $\lambda^s = (\lambda_1^s, \lambda_2^s, \ldots)$, $s = 1, \ldots, N$, $N \geq 2$, satisfying some general conditions. These filters can be, for example, any of the three types described above, as well as pooled sets of different kinds of filters. Given the data, $X = (X_1, X_2, \ldots)$, our aim is to select a data-dependent sequence of weights $\lambda^* = \lambda^*(X) = (\lambda_1^*, \lambda_2^*, \ldots)$, with values in Λ, that has asymptotically minimal squared risk for the true θ. We show that λ^* can be defined as a minimizer (with respect to $\lambda \in \Lambda$) of an unbiased estimator of the risk. Optimality properties of such λ^* follow from the oracle inequalities that are the main result of the paper. The oracle inequalities are nonasymptotic, they are obtained under very weak conditions on Λ and they lead to asymptotically exact inequalities of the form

$$\mathcal{R}(f^*, f) \leq \bigl(1 + o(1)\bigr) \min_{\lambda \in \Lambda} \mathcal{R}(f_\lambda, f), \tag{5}$$

as $\varepsilon \to 0$, where $f_\lambda = \sum_{k=1}^{\infty} \lambda_k X_k \varphi_k$, $f^* = \sum_{k=1}^{\infty} \lambda_k^* X_k \varphi_k$ and $o(1)$ does not depend on f but depends on the family Λ.

As a consequence of these inequalities, we can justify the optimal choice of smoothing parameters in the Tikhonov–Phillips and projection methods, as well as in Pinsker's method (the last one yields as a by-product sharp minimaxity of λ^* on Sobolev ellipsoids). The optimality results are valid under the assumption that σ_k is growing as a power of k, as $k \to \infty$. Generality of the oracle inequalities allows us to apply them in various problems. As an example, we consider sharp adaptation on multivariate anisotropic smoothness classes. An interesting conclusion is that the adaptive estimator based on λ^* attains the minimax lower bound for the multivariate anisotropic case, without any loss of efficiency.

Other oracle inequalities for inverse problems have been proposed recently by Johnstone (1999) and Cavalier and Tsybakov (2002). Johnstone (1999) deals with a class of nonlinear estimators based on soft thresholding in a wavelet context. He obtains an asymptotically exact oracle inequality for this class. Cavalier and Tsybakov (2002) consider the model (4) and obtain asymptotically exact oracle inequalities of the form (5), where Λ is the class of all monotone weight sequences λ and f^* (respectively, λ^*) is chosen in a different way, by application of a penalized blockwise Stein's rule. Their method is sharp adaptive in a minimax sense on ellipsoids, but it is not suited for parameter selection in restricted classes of filters, such as the Tikhonov–Phillips one.

2. Main results.

We deal with the model (4) and we assume the following.

ASSUMPTION 1. For any $\lambda \in \Lambda$,

$$0 < \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^2 < \infty$$

and

$$\max_{\lambda\in\Lambda}\sup_i|\lambda_i|\le 1. \tag{6}$$

The risk of the linear estimator $\hat\theta(\lambda)$ is given by

$$R_\varepsilon[\lambda,\theta]=\mathbf{E}_\theta\|\hat\theta(\lambda)-\theta\|^2=\sum_{i=1}^{\infty}(1-\lambda_i)^2\theta_i^2+\varepsilon^2\sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2. \tag{7}$$

The assumption (6) is quite natural. In fact, it follows from (7) that the estimator $\hat\theta(\lambda)$ with at least one $\lambda_i\notin[0,1]$ is inadmissible. However, we included the case of negative bounded λ_i since it corresponds to a number of well-known estimators such as kernel estimators with nonnegative kernels. The results below remain valid (up to a change in constants) if we replace 1 by an arbitrary constant C in (6).

Our data-driven choice of the smoothing parameter λ is based on the principle of unbiased risk estimation. To be more specific, recall briefly a heuristic motivation of the Mallows C_p criterion [Akaike (1973), Mallows (1973)]. The best choice of λ is the filter λ^0 (called the *oracle*) that minimizes $R_\varepsilon[\lambda,\theta]$ over $\lambda\in\Lambda$. The oracle λ^0 cannot be found directly since the functional $R_\varepsilon[\lambda,\theta]$ depends on the unknown θ_i^2. However, an unbiased estimator of θ_i^2 is available in the form $X_i^2-\varepsilon^2\sigma_i^2$. Thus the functional

$$\begin{aligned}U[\lambda,X]&=\sum_{i=1}^{\infty}(\lambda_i^2-2\lambda_i)(X_i^2-\varepsilon^2\sigma_i^2)+\varepsilon^2\sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2\\&=\sum_{i=1}^{\infty}(\lambda_i^2-2\lambda_i)X_i^2+2\varepsilon^2\sum_{i=1}^{\infty}\sigma_i^2\lambda_i\end{aligned} \tag{8}$$

is an unbiased estimator of $R_\varepsilon[\lambda,\theta]-\sum_{i=1}^{\infty}\theta_i^2$:

$$\mathbf{E}_\theta U[\lambda,X]=R_\varepsilon[\lambda,\theta]-\sum_{i=1}^{\infty}\theta_i^2. \tag{9}$$

Under our assumptions the random series in (8) converges in the mean squared sense for any $\lambda\in\Lambda$, and the definition (8), as well as other definitions of random series appearing in the paper, are understood in this sense. Of course, in practice one does not compute infinite series. One either requires that $\lambda_i=0$ for all i large enough (as for the projection or Pinsker filters) or translates the definition of $U[\lambda,X]$ into the function space to make it computable (e.g., for the Tikhonov–Phillips filters with even α the computations are possible in terms of splines).

The principle of unbiased risk estimation suggests that we minimize over $\lambda\in\Lambda$ the functional $U[\lambda,X]$ in place of $R_\varepsilon[\lambda,\theta]$. This leads to the following data-driven choice of λ:

$$\lambda^*=\arg\min_{\lambda\in\Lambda}U[\lambda,X]. \tag{10}$$

We show that this simple and intuitive smoothing parameter selection rule is efficient for inverse problems with power growth of σ_k, in the sense that it satisfies asymptotically precise oracle inequalities.

We define

$$\rho(\lambda) = \sup_k \sigma_k^2 |\lambda_k| \left\{ \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^4 \right\}^{-1/2}$$

and

$$\rho = \max_{\lambda \in \Lambda} \rho(\lambda).$$

Although the main results of this paper hold for general ρ, we will typically think of ρ as being small (for small ε). In particular, this will be explicitly assumed in the asymptotic corollaries.

Let the following assumption hold.

ASSUMPTION 2. There exists a constant $C_1 > 0$ such that, uniformly in $\lambda \in \Lambda$,

$$\sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^2 \le C_1 \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^4.$$

Assumptions 1 and 2 are very mild, and they are satisfied in most of the interesting examples. Since $|\lambda_i| \le 1$, we have

$$\sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^4 \le \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^2,$$

and Assumption 2 means that both sums are of the same order. The sums $\varepsilon^4 \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^4$ and $\varepsilon^4 \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^2$ are the main terms of the variance $\operatorname{Var}\{U[\lambda, X]\}$. On the other hand $R_\varepsilon[\lambda, \theta] \ge \varepsilon^2 \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^2$ and

$$\frac{\left(\varepsilon^4 \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^4\right)^{1/2}}{\varepsilon^2 \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^2} \le \rho. \tag{11}$$

Therefore, the smallness of ρ guarantees the smallness of the ratio of standard deviation to the mean error: $\operatorname{Var}^{1/2}\{U[\lambda, X]\}/R_\varepsilon[\lambda, \theta]$, uniformly over λ and θ.

Note also that under Assumption 2 we have

$$\rho(\lambda) \le \sqrt{C_1} \qquad \forall \lambda \in \Lambda. \tag{12}$$

Define

$$S = \max_{\lambda \in \Lambda} \sup_i \sigma_i^2 \lambda_i^2 \Big/ \min_{\lambda \in \Lambda} \sup_i \sigma_i^2 \lambda_i^2 \tag{13}$$

and

$$M = \sum_{\lambda \in \Lambda} \exp\{-1/\rho(\lambda)\},$$

$$L_\Lambda = \log(NS) + \rho^2 \log^2(MS).$$

We recall that N is the number of elements of the family Λ. Since $M \le N$, we always have $L_\Lambda \le \log(NS) + C_1 \log^2(NS)$, but this bound is rough. In asymptotics, as $\varepsilon \to 0$ and $N \to \infty$, the correct order is typically $\rho = o(1)$, $M \sim$ const and $L_\Lambda \sim \log(NS)$.

Main results of the paper are given in the next two theorems.

THEOREM 1. *Let Assumptions* 1 *and* 2 *hold. Then for every* $\theta \in \ell_2$, *for every* $B > B_0$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ *with* $\theta_i^* = \lambda_i^* X_i$, *where* λ^* *is defined by* (10), *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le (1 + \gamma_1 B^{-1}) \min_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta] + \gamma_2 B \varepsilon^2 L_\Lambda \omega(B^2 L_\Lambda), \tag{14}$$

where

$$\omega(x) = \max_{\lambda \in \Lambda} \sup_k \sigma_k^2 \lambda_k^2 I\left\{ \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^2 \le x \sup_k \sigma_k^2 \lambda_k^2 \right\}, \qquad x > 0,$$

and $B_0 > 0$, $\gamma_1 > 0$, $\gamma_2 > 0$ *are constants depending only on* C_1.

THEOREM 2. *Let Assumptions* 1 *and* 2 *hold. Then there exist constants* $\gamma_3 > 0$, $\gamma_4 > 0$ *depending only on* C_1, *such that for every* $\theta \in \ell_2$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ *with* $\theta_i^* = \lambda_i^* X_i$, *where* λ^* *is defined by* (10), *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \left[1 + \gamma_3 \rho \sqrt{L_\Lambda}\right] \min_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta],$$

provided $\rho \sqrt{L_\Lambda} \le \gamma_4$.

To prove (5), in many examples it is sufficient to use Theorem 2, and even its simplified version that we are going to state now. Assume that

$$\lim_{\varepsilon \to 0} \rho^2 \log(NS) = 0. \tag{15}$$

Then $L_\Lambda = O(\log(NS))$, and we get the following corollary of Theorem 2.

COROLLARY 1. *Let Assumptions* 1 *and* 2 *hold. Then there exist constants* $C_2 > 0$, $C_3 > 0$, *depending only on* C_1, *such that for* $\rho^2 \log(NS) < C_2$ *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \left(1 + C_3 \rho \sqrt{\log(NS)}\right) \min_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta],$$

for every $\theta \in \ell_2$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ *with* $\theta_i^* = \lambda_i^* X_i$.

Thus, condition (15) is sufficient to have asymptotically exact oracle inequalities of the form (5).

Recall that for ill-posed inverse problems $\sigma_k \to \infty$, as $k \to \infty$. A crucial restriction, implicit in Theorems 1 and 2 and Corollary 1 is that σ_k should grow not faster than a power of k. In fact, if σ_k grows exponentially, then even in the simplest case of projection weights λ_k we have $\rho \neq o(1)$, as $\varepsilon \to 0$ (thus, Theorem 2 cannot be applied if N grows with $\varepsilon \to 0$). Theorem 1 in this case can be applied but does not give correct rates.

The oracle inequality (14) is suited to obtaining minimax results or rates of convergence for classes of sequences θ. The remainder term in the right-hand side of (14) is typically of the form $\varepsilon^2 \log^a(1/\varepsilon)$ with $a > 1$. This shows limits of applications of (14): for the classes where the least favorable functions θ satisfy $\min_{\lambda\in\Lambda} R_\varepsilon[\lambda, \theta] \ll \varepsilon^2 \log^a(1/\varepsilon)$, the remainder term is not asymptotically negligible, and thus asymptotically sharp adaptation in a minimax sense is not possible by use of our techniques. This remark concerns the classes of analytical functions, for example. On the other hand, the remainder term is negligible compared to the minimax risk on Sobolev ellipsoids.

Theorem 2 is rather a "class of estimators" than a "class of functions" result. It allows us to get oracle inequalities of the form (5) for any fixed $\theta \in \ell_2$, provided the class Λ of estimators is small enough.

REMARK 1. Theorems 1 and 2 remain valid for non-Gaussian ξ_i such that $\mathbf{E}\exp(C\xi_i^2) < \infty$ for some $C > 0$.

REMARK 2. Theorems 1 and 2 can be used not only for ill-posed, but also for well-posed inverse problems where $\sigma_k \not\to \infty$. For example, both theorems apply if $\sigma_k \sim k^{-\beta}$ with $0 \leq \beta < 1/4$ [allowing $\rho = o(1)$, as $\varepsilon \to 0$]. For faster decreasing σ_k only Theorem 1 works. If N and S grow not faster than a power of ε^{-1}, the remainder term in (14) is $O(\varepsilon^2 \log^a(1/\varepsilon))$ for some $a > 0$, which is only logarithmically worse then the optimal rate ε^2 of the well-posed inverse problems.

REMARK 3. Consider the special case that corresponds to direct observations (i.e., $\sigma_k \equiv 1$). Here several oracle inequalities have been known previously. Theorems 1 and 2 extend these results, especially in what concerns multivariate applications.

The first oracle inequalities for the direct observations model appeared, although implicitly, in the proofs of optimality of C_p, cross-validation and related data-driven methods [Li (1986, 1987) and Polyak and Tsybakov (1990, 1992)]. They are also implicit in minimax adaptive constructions [Golubev (1987, 1992) and Golubev and Nussbaum (1992)]. Presumably, the first explicit use of oracle inequalities and its implications for minimax is due to Donoho and Johnstone

(1994) and Kneip (1994). More recent references are Donoho and Johnstone (1995, 1996), Nemirovski (2000), Birgé (2001), Birgé and Massart (2001), Cavalier and Tsybakov (2002) and Goldenshluger and Tsybakov (2001). A link of oracle inequalities to maxisets is discussed by Kerkyacharian and Picard (2002). For linear estimators, the results are often proved in the following model, which can be embedded into ours. Let $Y = (Y_1, \dots, Y_n)^T$ be the vector of observations in the nonparametric regression model

$$Y_k = f(x_k) + \nu_k, \qquad k = 1, \dots, n, \tag{16}$$

where ν_k are i.i.d. $\mathcal{N}(0, 1)$ random errors, $x_k \in [0, 1]$ are nonrandom distinct points and $f(\cdot)$ is the unknown function to be estimated. Consider the linear estimator $\hat{f} = \mathcal{S}Y$ of the vector $f = (f(x_1), f(x_2), \dots, f(x_n))^T$, where $\mathcal{S} = \sum_{i=1}^n \lambda_i u_i u_i^T$ is a symmetric $n \times n$ matrix (smoother matrix) with eigenvalues λ_i, and $\{u_i\}$ is an orthonormal basis in $\mathbb{R}^n$. Denoting $X_i = n^{-1/2} u_i^T Y$, $\theta_i = n^{-1/2} u_i^T f$, $\xi_i = u_i^T \nu$ [where $\nu = (\nu_1, \dots, \nu_n)^T$] and $\varepsilon = n^{-1/2}$, we rewrite the initial regression model in the equivalent form

$$X_i = \theta_i + \varepsilon \xi_i, \qquad i = 1, \dots, n,$$

which is a special case of (4), modulo the fact that the ℓ_2-vectors should contain zeros starting from the $(n+1)$th position: $\theta = (\theta_1, \dots, \theta_n, 0, 0, \dots)$. The linear estimator $\hat{f}$ is translated into $\hat{\theta}(\lambda) = (\hat{\theta}_1, \dots, \hat{\theta}_n, 0, 0, \dots)$, where $\hat{\theta}_i = \lambda_i X_i = n^{-1/2} \lambda_i u_i^T Y$, and the risk is

$$\mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 = \mathbf{E} \frac{1}{n} \sum_{i=1}^n \|\mathcal{S}Y - f\|^2.$$

Kneip (1994) studies this setup assuming that the class of filters Λ contains the sequences λ with monotone nonincreasing coefficients λ_i and such that for any two sequences $\lambda, \lambda' \in \Lambda$ we have either $\lambda_i \le \lambda_i'$, $\forall i$, or $\lambda_i \ge \lambda_i'$, $\forall i$ (ordered linear smoothers). The set Λ in Kneip (1994) need not be finite. With the above translation into our notation, Kneip [(1994), page 844] proves the oracle inequality

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \bigl(1 + O(B^{-1})\bigr) \min_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta] + O(B)\varepsilon^2 \qquad \forall B > 0, \tag{17}$$

where θ^* is the data-driven estimator (10). This is similar to (14) [but recall that Kneip's inequality (17) covers only the case $\sigma_k \equiv 1$]. Another difference is that we assume finiteness of the set Λ (which is not restrictive in view of applications), but we drop the assumption of order. The last point is useful in multivariate models, in particular, anisotropic ones, where the ordering of the filters is not natural (cf. the example in Section 6 below).

We also note that for the regression model (16) there exist several alternatives to Mallows' C_p, for example, cross-validation (CV) and generalized cross-validation (GCV). Asymptotic optimality results [similar to (5)] for CV and GCV

in regression are given by Härdle and Marron (1985), Li (1986, 1987) and Polyak and Tsybakov (1992).

3. Examples. Consider some examples of application of the main results. Typical assumptions on the parameters appearing in Theorems 1 and 2 will be the following:

1. power growth of σ_k, as $k \to \infty$;
2. power behavior of S: $S = O(\varepsilon^{-t})$, for some $t > 0$, as $\varepsilon \to 0$;
3. at most power growth of N: $N = O(\varepsilon^{-\nu})$, for some $\nu > 0$, as $\varepsilon \to 0$ [and $N = O(\log(1/\varepsilon))$ in some examples].

In all the cases we have $\log(NS) = O(\log(1/\varepsilon))$.

EXAMPLE 1 (Projection estimators). Let $1 \le w_1 < \cdots < w_N$ be integers. Consider the projection filters $\lambda^s = (\lambda_1^s, \lambda_2^s, \ldots)$ defined by

$$\begin{aligned} &\lambda_i^1 = I(i \le w_1), \qquad \lambda_i^2 = I(i \le w_2), \ldots, \\ &\lambda_i^N = I(i \le w_N), \qquad i = 1, 2, \ldots . \end{aligned} \tag{18}$$

Throughout this section we assume a power law behavior of σ_k:

$$\sigma_{\min} k^{\beta} \le |\sigma_k| \le \sigma_{\max} k^{\beta}, \tag{19}$$

$k = 1, 2, \ldots$, for some $\sigma_{\max} \ge \sigma_{\min} > 0$, $\beta \ge 0$.

Note that Assumption 2 is satisfied with the equality and $C_1 = 1$. It is easy to see that

$$\rho(\lambda^s) \le C w_s^{-1/2}, \qquad \rho \le \max_{s=1,\ldots,N} C w_s^{-1/2} = C w_1^{-1/2}$$

and

$$S \le C(w_N/w_1)^{2\beta}.$$

Here and later C stands for positive constants (possibly different on different occasions) that do not depend on λ, θ and ε. Using the bound on $\rho(\lambda)$, we conclude that M is bounded by a constant independent of ε for any choice of the sequence w_i. Note also that

$$\omega(x) \le C \sup_k k^{2\beta} I\{k^{2\beta+1} \le C x k^{2\beta}\} \le C x^{2\beta} \tag{20}$$

and

$$L_\Lambda \le C\big(\log(N w_N/w_1) + w_1^{-1} \log^2(w_N/w_1)\big).$$

Since $\rho\sqrt{\log(NS)} \le C w_1^{-1/2} \sqrt{\log(N w_N/w_1)}$, Corollary 1 gives

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \left[1 + C\sqrt{\frac{\log(N w_N/w_1)}{w_1}}\right] \min_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta] \tag{21}$$

provided $w_1^{-1}\log(Nw_N/w_1)$ is small enough. As a consequence, we get an asymptotically exact oracle inequality of the form (5):

COROLLARY 2. *Assume that* $\Lambda = (\lambda^1, \dots, \lambda^N)$ *is the set of projection filters defined by* (18) *and that* (19) *holds. If* $N = N(\varepsilon)$ *and* $w_1 = w_1(\varepsilon)$, $w_N = w_N(\varepsilon)$ *are such that*

$$\lim_{\varepsilon\to 0} \frac{\log(Nw_N/w_1)}{w_1} = 0,$$

then for every $\theta \in \ell_2$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \dots)$ *with* $\theta_i^* = \lambda_i^* X_i$ *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \big(1 + o(1)\big) \inf_{\lambda\in\Lambda} R_\varepsilon[\lambda, \theta],$$

where $o(1) \to 0$ *uniformly in* $\theta \in \ell_2$.

In other words, Corollary 2 states that our adaptively selected filter behaves itself asymptotically at least as well as the best projection estimator in Λ. For the direct case (where $\sigma_k \equiv 1$) such an inequality is obtained by Birgé (2001), who uses the Lepski adaptation method rather than the Mallows C_p.

Next, consider the situation where there is no restriction on w_1 except $w_1 \ge 1$ (i.e., the class Λ can contain the projection filters of order less than or equal to some w_N). Applying (21) we get an inequality with a logarithmic loss of efficiency:

PROPOSITION 1. *Assume that* $\Lambda = (\lambda^1, \dots, \lambda^N)$ *is the set of projection filters defined by* (18) *and that* (19) *holds. For every* $\theta \in \ell_2$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \dots)$ *with* $\theta_i^* = \lambda_i^* X_i$ *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le C\sqrt{\log(Nw_N)} \min_{\lambda\in\Lambda} R_\varepsilon[\lambda, \theta],$$

where $C > 0$ *depends only on* $\sigma_{\min}, \sigma_{\max}, \beta$.

An important special case is wavelet estimators for which we set $w_1 = 2^{j_0}$ (where j_0 is the index of the initial level), $w_j = 2w_{j-1}$. Typically one chooses 2^{-j_0} to be decreasing as a power of ε and $N \sim \log(1/\varepsilon)$. It is easy to see that the result of Corollary 2 remains valid in this case. Thus, a wavelet estimator that uses our data-driven selection of w_j is asymptotically at least as good as the best linear wavelet estimator for any θ. By taking suprema of both sides of the oracle inequality over Besov classes [for definition see Donoho and Johnstone (1994, 1995, 1996)] we

get that our adaptive estimator attains optimal rate of convergence on all the Besov classes where linear wavelet estimators attain optimal rates.

EXAMPLE 2 (Level-wise "keep-or-kill" estimators). Let $m > 1$ and $1 \le w_1 < \cdots < w_m$ be integers and let $e = (e_1, \ldots, e_{m-1})$ be a binary sequence of length $m-1$, $e_k \in \{0, 1\}$. We associate with e a filter $\lambda(e) = (\lambda_1(e), \lambda_2(e), \ldots)$ defined by

$$\lambda_i(e) = I\{i \le w_1\} + \sum_{k=1}^{m-1} e_k I\{w_k < i \le w_{k+1}\}, \qquad i = 1, 2, \ldots .$$

Consider the collection of filters

$$\Lambda = \{\lambda(e) : e \in E\}, \tag{22}$$

where E is the set of all binary sequences e of length $m-1$. The linear estimator with weights $\lambda_i(e)$ "keeps" the blocks of coefficients $\{\theta_i : w_k < i \le w_{k+1}\}$ for which $e_k = 1$ and "kills" the blocks for which $e_k = 0$. Clearly,

$$N = \mathrm{Card}(\Lambda) = 2^{m-1}.$$

As in Example 1, we get that Assumption 2 is satisfied and that $\rho \le C w_1^{-1/2}$, $S \le C(w_m/w_1)^{2\beta}$, provided (19) holds. Therefore, applying Corollary 1 we get the following result.

PROPOSITION 2. *Assume that Λ is the set of levelwise "keep-or-kill" filters defined by* (22) *and that* (19) *holds. If $m = m(\varepsilon)$ and $w_1 = w_1(\varepsilon)$, $w_m = w_m(\varepsilon)$ are such that*

$$\lim_{\varepsilon \to 0} \frac{m + \log(w_m/w_1)}{w_1} = 0, \tag{23}$$

then for every $\theta \in \ell_2$ and for the estimator $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ with $\theta_i^* = \lambda_i^* X_i$ we have

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \bigl(1 + o(1)\bigr) \inf_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta],$$

where $o(1) \to 0$ uniformly in $\theta \in \ell_2$.

Note that for wavelet bases (if we consider global thresholding level-by-level) we have $w_1 = 2^{j_0}$, $w_m = 2^{j_0+m-1}$ with an integer j_0, and condition (23) reduces to

$$\lim_{\varepsilon \to 0} 2^{-j_0} m = 0,$$

which is readily satisfied for the typical situation where 2^{-j_0} decreases as a power of ε and $m \sim \log(1/\varepsilon)$. As a conclusion we get, in particular, that the wavelet keep-or-kill level-by-level estimator that uses our data-driven rule attains the optimal

rate of convergence on all the Besov classes where the linear wavelet level-wise keep-or-kill estimators attain optimal rates.

EXAMPLE 3 (Tikhonov–Phillips estimators). Consider the set of filters

$$\Lambda = \left\{ \lambda = \{\lambda_k\} : \lambda_k = \frac{1}{1 + (k/w)^\alpha},\ w \in \mathcal{W},\ \alpha \in \mathcal{A} \right\}, \tag{24}$$

where $\mathcal{A}$ is a finite set of real numbers, possibly depending on ε:

$$\mathcal{A} = \{\alpha(1), \alpha(2), \ldots, \alpha(N_{\mathcal{A}})\}$$

such that $2\beta + 1/2 < \alpha_{\min} = \alpha(1) < \alpha(2) < \cdots < \alpha(N_{\mathcal{A}}) = \alpha_{\max}$, and $\mathcal{W}$ is a finite subset of the interval $[w_1, w_{\max}]$, with $\mathrm{Card}(\mathcal{W}) = N_{\mathcal{W}}$, where $N_{\mathcal{A}}, N_{\mathcal{W}}$ are integers, $0 < w_1 < w_{\max} < \infty$. Note that $\mathrm{Card}(\Lambda) = N$, where $N = N_{\mathcal{A}} N_{\mathcal{W}}$.

We get by simple algebra

$$\sum_{i=1}^{\infty} \frac{i^{4\beta}}{(1 + (i/w)^\alpha)^4} \geq C_* w^{4\beta+1}, \qquad \sum_{i=1}^{\infty} \frac{i^{4\beta}}{(1 + (i/w)^\alpha)^2} \leq C^* w^{4\beta+1},$$

where C_* and C^* are positive constants depending only on $\alpha_{\min}, \alpha_{\max}, \beta$. This and (19) guarantee that Assumption 2 is satisfied.

Next,

$$\sup_i \frac{i^{2\beta}}{[1 + (i/w)^\alpha]^2} = C(\alpha, \beta) w^{2\beta},$$

where $C(\alpha, \beta)$ is bounded from 0 and ∞ uniformly for $\alpha_{\min} \leq \alpha \leq \alpha_{\max}$. Therefore $S \leq C(w_{\max}/w_1)^{2\beta}$. Furthermore, we get similarly

$$\rho \leq C \max_{\lambda \in \Lambda} \sup_i \frac{i^{2\beta}}{1 + (i/w)^\alpha} \left(\sum_{i=1}^{\infty} \frac{i^{4\beta}}{(1 + (i/w)^\alpha)^4} \right)^{-1/2} \leq C w_1^{-1/2}.$$

Thus,

$$\rho \sqrt{\log(NS)} \leq C w_1^{-1/2} \sqrt{\log(N w_{\max}/w_1)}$$

and, to guarantee an asymptotically exact oracle inequality, it remains to require (in view of Corollary 1) that the parameters $w_{\max}, w_1, N_{\mathcal{W}}, N_{\mathcal{A}}$ are such that (15) holds.

PROPOSITION 3. *Assume that* $\Lambda = (\lambda^1, \ldots, \lambda^N)$ *is the set of Tikhonov–Phillips filters defined by* (24) *and that* (19) *holds. If* $N = N_{\mathcal{A}} N_{\mathcal{W}}$, $w_{\max} = w_{\max}(\varepsilon)$, $w_1 = w_1(\varepsilon)$ *are such that*

$$\lim_{\varepsilon \to 0} \frac{\log(N w_{\max}/w_1)}{w_1} = 0, \tag{25}$$

then for every $\theta \in \ell_2$ *and for the estimator* $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ *with* $\theta_i^* = \lambda_i^* X_i$ *we have*

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le \big(1 + o(1)\big) \inf_{\lambda \in \Lambda} R_\varepsilon[\lambda, \theta],$$

where $o(1) \to 0$ *uniformly in* $\theta \in \ell_2$.

Note that condition (25) is very weak and it is easily checked in typical situations. For example, one can take $N_{\mathcal{W}} = O(\varepsilon^{-a})$, $w_{\max} = O(\varepsilon^{-b})$, for some arbitrary fixed $a > 0$, $b > 0$, and assume that $N_{\mathcal{A}}$, $\alpha_{\max}$ do not depend on ε (it is typical to consider just a small fixed number of integers α, or even one integer α). This should be completed, in order to satisfy (25), by the mild assumption $w_1/\log^2(1/\varepsilon) \to \infty$. Since there is no restriction on the power a, the discrete net $\mathcal{W}$ can be arbitrarily fine, and it is not hard to show that optimality of our discretized rule extends to the set of filters (24) where w varies continuously in the interval.

4. Preliminary lemmas. Let ξ_i be i.i.d. $\mathcal{N}(0, 1)$ random variables and let $v = \{v_i\}_1^\infty \in \ell_2$ be a random sequence measurable w.r.t. $\{\xi_i\}_{i=1}^\infty$. It is assumed that v takes values in a finite set V of ℓ_2-sequences: $v \in V = \{v^1, \ldots, v^N\}$.

LEMMA 1. *For any* $K \ge 1$,

$$\mathbf{E}\left|\sum_{i=1}^\infty v_i \xi_i\right| \le \sqrt{2\log(NK)}\Big(\mathbf{E}\|v\| + \sqrt{2\mathbf{E}\|v\|^2/K}\Big).$$

PROOF. It suffices to consider the case where $\|v^j\| \ne 0$, $\forall\ v^j \in V$. Denote $\zeta_v = \|v\|^{-1} \sum_{i=1}^\infty v_i \xi_i$, where the sums $\|v\|^2$ and $\sum_{i=1}^\infty v_i \xi_i$ are understood in the sense of mean squared convergence. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\mathbf{E}\left|\sum_{i=1}^\infty v_i \xi_i\right| &\le \mathbf{E}\|v\| \max_{v \in V} |\zeta_v| \le \mathbf{E}\|v\| \max_{v \in V} |\zeta_v| I\left\{\max_{v \in V} |\zeta_v| \le \sqrt{2\log(NK)}\right\} \\
&\qquad + \mathbf{E}\|v\| \max_{v \in V} |\zeta_v| I\left\{\max_{v \in V} |\zeta_v| > \sqrt{2\log(NK)}\right\} \\
&\le \sqrt{2\log(NK)}\mathbf{E}\|v\| \\
&\qquad + (\mathbf{E}\|v\|^2)^{1/2}\left(\mathbf{E}\max_{v \in V} |\zeta_v|^2 I\left\{\max_{v \in V} |\zeta_v| > \sqrt{2\log(NK)}\right\}\right)^{1/2}.
\end{aligned}$$

Now, for the function $F(t) = t^2 I\{t > \sqrt{2\log(NK)}\}$ we use that

$$F\left(\max_{v \in V} |\zeta_v|\right) \le \sum_{v \in V} F(|\zeta_v|),$$

and since $\zeta_{v^1}, \ldots, \zeta_{v^N}$ are identically distributed standard Gaussian variables, we get

$$\begin{aligned}\mathbf{E}\left|\sum_{i=1}^{\infty} v_i\xi_i\right| &\le \sqrt{2\log(NK)}\mathbf{E}\|v\| \\ &\quad + (\mathbf{E}\|v\|^2)^{1/2}\Big(N\mathbf{E}|\zeta_{v^1}|^2 I\Big\{|\zeta_{v^1}| > \sqrt{2\log(NK)}\Big\}\Big)^{1/2} \\ &\le \sqrt{2\log(NK)}\Big(\mathbf{E}\|v\| + \sqrt{2\mathbf{E}\|v\|^2/K}\Big),\end{aligned}$$

where for the last inequality we used

$$\mathbf{E}|\zeta_{v^1}|^2 I\{|\zeta_{v^1}| > x\} \le 2(2\pi)^{-1/2}(x + x^{-1})\exp(-x^2/2) \qquad \forall\, x > 0,$$

and $NK \ge 2$. $\square$

LEMMA 2. *Let* $\|v\| \ne 0$, $\forall\, v \in V$, *and denote* $m(v) = \sup_i |v_i|/\|v\|$, $m_V = \max_{v\in V} m(v)$,

$$M(q) = \sum_{v\in V} \exp\{-q/m(v)\}, \tag{26}$$

where $q > 0$. *Then for any* $K \ge 1$ *we have*

$$\mathbf{E}\left|\sum_{i=1}^{\infty} v_i(\xi_i^2 - 1)\right| \le D\Big(\sqrt{\log(NK)} + m_V \log\big(M(q)K\big)\Big)\Big(\mathbf{E}\|v\| + \sqrt{\mathbf{E}\|v\|^2/K}\Big),$$

where D *is a constant depending only on* q.

PROOF. Let $\eta_v = (\sqrt{2}\|v\|)^{-1}\sum_{i=1}^{\infty} v_i(\xi_i^2 - 1)$, where the sums $\|v\|^2$ and $\sum_{i=1}^{\infty} v_i(\xi_i^2 - 1)$ are understood in the sense of mean squared convergence. Using the Markov inequality and the formula

$$-\log(1 - x) = \sum_{k=1}^{\infty} \frac{x^k}{k}$$

one obtains, for any $0 < t < [\sqrt{2}m(v)]^{-1}$,

$$\begin{aligned}\mathbf{P}\{\eta_v > x\} &\le \exp(-tx)\mathbf{E}\exp(t\eta_v) \\ &= \exp(-tx)\prod_{i=1}^{\infty}\exp\Big\{-\frac{tv_i}{\sqrt{2}\|v\|} - \frac{1}{2}\log\Big(1 - \frac{\sqrt{2}tv_i}{\|v\|}\Big)\Big\} \\ &= \exp(-tx)\exp\Big\{\sum_{k=2}^{\infty}\sum_{i=1}^{\infty}\frac{1}{2k}\Big(\frac{\sqrt{2}tv_i}{\|v\|}\Big)^k\Big\}\end{aligned}$$

$$\le \exp(-tx)\exp\left\{\frac{1}{m^2(v)}\sum_{k=2}^{\infty}\frac{1}{2k}\left[\sqrt{2}tm(v)\right]^k\right\}$$

$$\le \exp(-tx)\exp\left\{-\frac{1}{2m^2(v)}\log\left[1-\sqrt{2}tm(v)\right]-\frac{t}{\sqrt{2}m(v)}\right\}.$$

Minimization of the last expression with respect to t yields

$$\mathbf{P}\{\eta_v > x\} \le \exp[\varphi_v(x)], \qquad \varphi_v(x) = \frac{1}{2m^2(v)}\log\left[1+\sqrt{2}xm(v)\right] - \frac{x}{\sqrt{2}m(v)}.$$

Note that for $u \ge 0$ we have

$$\log(1+u) - u = u\int_0^1\left(-\frac{\tau u}{1+\tau u}\right)d\tau \le -\int_0^1\frac{\tau u^2}{1+u}\,d\tau = -\frac{u^2}{2(1+u)}.$$

Thus

$$\varphi_v(x) \le -\frac{x^2}{2(1+\sqrt{2}xm(v))},$$

and we conclude that

$$\mathbf{P}\{|\eta_v| > x\} \le 2\exp\left\{-\frac{x^2}{2(1+\sqrt{2}xm(v))}\right\} \qquad \forall\, x > 0. \tag{27}$$

Using (27), we find, for any $Q > 0$,

$$\mathbf{E}\eta_v^2 I\{|\eta_v| > Q\} = 2\int_Q^{\infty} x\mathbf{P}\{|\eta_v| > x\}\,dx$$

$$\le 4\int_Q^{\infty} x\exp\left\{-\frac{x^2}{2(1+\sqrt{2}xm(v))}\right\}dx.$$

It is easy to see that

$$-\frac{x^2}{2(1+\sqrt{2}xm(v))} \le \begin{cases} -\dfrac{x^2}{4}, & \sqrt{2}m(v)x \le 1, \\ -\dfrac{x}{\sqrt{32}m(v)}, & \sqrt{2}m(v)x > 1, \end{cases}$$

and we get by simple algebra, for $Q > q\sqrt{32}$,

$$\mathbf{E}\eta_v^2 I\{|\eta_v| > Q\} \le C\exp\Big(-\frac{Q^2}{4}\Big) + CQ\exp\Big\{-\frac{Q}{\sqrt{32}m(v)}\Big\}. \tag{28}$$

In view of (26) we have, for any $Q > q\sqrt{32}$,

$$\sum_{v\in V}\exp\Big(-\frac{Q}{\sqrt{32}m(v)}\Big) \le M(q)\exp\Big(-\frac{Q/\sqrt{32}-q}{m_V}\Big). \tag{29}$$

Now, acting as in the proof of Lemma 1 and using (28), (29) we obtain

$$\begin{aligned}
&\mathbf{E}\Big|\sum_{i=1}^{\infty} v_i(\xi_i^2-1)\Big| \\
&\quad\le \sqrt{2}\mathbf{E}\|v\|\max_{v\in V}|\eta_v| \\
&\quad\le \sqrt{2}\mathbf{E}\|v\|\max_{v\in V}|\eta_v| I\Big\{\max_{v\in V}|\eta_v| \le Q\Big\} \\
&\qquad + \sqrt{2}\mathbf{E}\|v\|\max_{v\in V}|\eta_v| I\Big\{\max_{v\in V}|\eta_v| > Q\Big\} \\
&\quad\le \sqrt{2}Q\mathbf{E}\|v\| + \sqrt{2\mathbf{E}\|v\|^2}\Big(\mathbf{E}\max_{v\in V}\eta_v^2 I\Big\{\max_{v\in V}|\eta_v| > Q\Big\}\Big)^{1/2} \\
&\quad\le \sqrt{2}Q\mathbf{E}\|v\| + \sqrt{2\mathbf{E}\|v\|^2}\Big(\sum_{v\in V}\mathbf{E}\eta_v^2 I\{|\eta_v| > Q\}\Big)^{1/2} \\
&\quad\le \sqrt{2}Q\mathbf{E}\|v\| + C\sqrt{\mathbf{E}\|v\|^2}\Big[N\exp\Big(-\frac{Q^2}{4}\Big) + M(q)Q\exp\Big(-\frac{Q}{\sqrt{32}m_V}\Big)\Big]^{1/2},
\end{aligned}$$

for any $Q > q\sqrt{32}$. Choosing $Q = 2\sqrt{\log(NK)} + \sqrt{32}m_V\log(M(q)K) + q\sqrt{32}$ one gets the lemma. □

Consider the estimator

$$\tilde{\theta}_i = \lambda_i(X)X_i,$$

where $\lambda_i = \lambda_i(X) \in [-1,1]$ depends on the data X [not necessarily $\lambda_i(X) = \lambda_i^*(X)$]. We assume that the filter $\lambda(X) = (\lambda_1(X), \lambda_2(X), \ldots)$ takes values in the set of candidate filters Λ. In the next lemma we give a bound for the risk of this estimator. We need the following notation:

$$\Delta^{\varepsilon}[\lambda] = \varepsilon^2 L_{\Lambda}\sup_i \sigma_i^2\lambda_i^2 \qquad \forall\,\lambda\in\Lambda.$$

LEMMA 3. *For any $B > 0$ we have*

$$\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \le (1 + 2B^{-1})\mathbf{E}_\theta R_\varepsilon[\lambda(X), \theta] + CB\mathbf{E}_\theta \Delta^\varepsilon[\lambda(X)],$$

where $C > 0$ is an absolute constant.

PROOF. Write

$$\begin{aligned}\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 &= \mathbf{E}_\theta \left[\sum_{i=1}^\infty (1 - \lambda_i(X))^2 \theta_i^2 + \varepsilon^2 \sum_{i=1}^\infty \sigma_i^2 \lambda_i^2(X)\right] \\ &\quad - 2\varepsilon \mathbf{E}_\theta \sum_{i=1}^\infty (1 - \lambda_i(X))\theta_i \lambda_i(X)\sigma_i \xi_i \\ &\quad + \varepsilon^2 \mathbf{E}_\theta \sum_{i=1}^\infty \sigma_i^2 \lambda_i^2(X)(\xi_i^2 - 1).\end{aligned} \tag{30}$$

Using Lemma 1 with $K = S$ we get

$$\begin{aligned}&\mathbf{E}_\theta \left|\sum_{i=1}^\infty (1 - \lambda_i(X))\theta_i \lambda_i(X)\sigma_i \xi_i\right| \\ &\quad \le \sqrt{2\log(NS)}\mathbf{E}_\theta \left[\sum_{i=1}^\infty (1 - \lambda_i(X))^2 \theta_i^2 \lambda_i^2(X)\sigma_i^2\right]^{1/2} \\ &\qquad + 2\sqrt{\log(NS)}S^{-1/2}\left[\mathbf{E}_\theta \sum_{i=1}^\infty (1 - \lambda_i(X))^2 \theta_i^2 \lambda_i^2(X)\sigma_i^2\right]^{1/2} \\ &\quad \le \sqrt{2\log(NS)}\mathbf{E}_\theta \left[\sup_i |\sigma_i|\,|\lambda_i(X)|\left(\sum_{i=1}^\infty (1 - \lambda_i(X))^2 \theta_i^2\right)^{1/2}\right] \\ &\qquad + 2\sqrt{\log(NS)}S^{-1/2}\left[\mathbf{E}_\theta \sum_{i=1}^\infty (1 - \lambda_i(X))^2 \theta_i^2\right]^{1/2} \max_{\lambda\in\Lambda} \sup_i |\sigma_i|\,|\lambda_i|.\end{aligned}$$

Now (13) and the elementary inequality $2ab \le B^{-1}a^2 + Bb^2$, $\forall\, B > 0$, yield, for any $B > 0$,

$$\begin{aligned}&\varepsilon \mathbf{E}_\theta \left|\sum_{i=1}^\infty [1 - \lambda_i(X)]\theta_i \lambda_i(X)\sigma_i \xi_i\right| \\ &\quad \le B^{-1}\mathbf{E}_\theta \sum_{i=1}^\infty [1 - \lambda_i(X)]^2 \theta_i^2 \\ &\qquad + \varepsilon^2 B\log(NS)\mathbf{E}_\theta \sup_i \sigma_i^2 \lambda_i^2(X) + \varepsilon^2 B\log(NS) \min_{\lambda\in\Lambda} \sup_i \sigma_i^2 \lambda_i^2 \\ &\quad \le B^{-1}\mathbf{E}_\theta \sum_{i=1}^\infty [1 - \lambda_i(X)]^2 \theta_i^2 + 2B\mathbf{E}_\theta \Delta^\varepsilon[\lambda(X)].\end{aligned} \tag{31}$$

To bound the last term in (30) we use Lemma 2 with $v_i = \sigma_i^2\lambda_i^2(X)$, $K = S$, $q = 1$. The inequalities $|\lambda_i(X)| \le 1$ and (12) entail $m(v) \le \rho(\lambda(X)) \le \sqrt{C_1}$. Thus, $M(1) = \sum_{v\in V} \exp\{-1/m(v)\} \le M$. Note also that $m_V \le \rho$ and $\sqrt{\log(NS)} + m_V \log(M(q)S) \le \sqrt{2L_\Lambda}$. Therefore, application of Lemma 2 yields

$$\varepsilon^2\mathbf{E}_\theta\left|\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)(\xi_i^2-1)\right| \le D\sqrt{2L_\Lambda}\varepsilon^2\left\{\mathbf{E}_\theta\left[\sum_{i=1}^\infty \sigma_i^4\lambda_i^4(X)\right]^{1/2} + S^{-1/2}\left[\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^4\lambda_i^4(X)\right]^{1/2}\right\}.$$

Next, by (13),

$$S^{-1/2}\left[\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^4\lambda_i^4(X)\right]^{1/2} \le \min_{\lambda\in\Lambda}\sup_i \sigma_i\lambda_i\left[\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)\right]^{1/2} \le \left[\mathbf{E}_\theta\sup_i \sigma_i^2\lambda_i^2(X)\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)\right]^{1/2}.$$

Hence, for any $B > 0$,

$$\begin{aligned}
&\varepsilon^2\mathbf{E}_\theta\left|\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)(\xi_i^2-1)\right| \\
&\quad\le D\sqrt{2L_\Lambda}\varepsilon^2\mathbf{E}_\theta\left[\sup_i|\sigma_i|\,|\lambda_i(X)|\left(\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)\right)^{1/2}\right] \\
&\qquad + D\sqrt{2L_\Lambda}\varepsilon^2\left[\mathbf{E}_\theta\sup_i \sigma_i^2\lambda_i^2(X)\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X)\right]^{1/2} \qquad (32)\\
&\quad\le 2\varepsilon^2B^{-1}\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X) + D^2L_\Lambda\varepsilon^2B\mathbf{E}_\theta\sup_i \sigma_i^2\lambda_i^2(X) \\
&\quad= 2\varepsilon^2B^{-1}\mathbf{E}_\theta\sum_{i=1}^\infty \sigma_i^2\lambda_i^2(X) + D^2B\mathbf{E}_\theta\,\Delta^\varepsilon[\lambda(X)].
\end{aligned}$$

Combining (30)–(32) we complete the proof. □

5. Proof of the theorems. Denote by $\lambda^0 = (\lambda_1^0, \lambda_2^0, \ldots)$ the oracle $\lambda^0 = \arg\min_{\lambda\in\Lambda} R_\varepsilon[\lambda, \theta]$. We have

$$
\begin{aligned}
\mathbf{E}_\theta U[\lambda^*, X] &= \mathbf{E}_\theta R_\varepsilon[\lambda^*, \theta] - \sum_{i=1}^{\infty}\theta_i^2 \\
&\quad + 2\varepsilon\mathbf{E}_\theta \sum_{i=1}^{\infty}(\lambda_i^{*2} - 2\lambda_i^*)\theta_i\sigma_i\xi_i + \varepsilon^2\mathbf{E}_\theta \sum_{i=1}^{\infty}(\lambda_i^{*2} - 2\lambda_i^*)\sigma_i^2(\xi_i^2 - 1) \\
&= \mathbf{E}_\theta R_\varepsilon[\lambda^*, \theta] - \sum_{i=1}^{\infty}\theta_i^2 \\
&\quad + 2\varepsilon\mathbf{E}_\theta \sum_{i=1}^{\infty}(1 - \lambda_i^*)^2\theta_i\sigma_i\xi_i + \varepsilon^2\mathbf{E}_\theta \sum_{i=1}^{\infty}(\lambda_i^{*2} - 2\lambda_i^*)\sigma_i^2(\xi_i^2 - 1).
\end{aligned} \tag{33}
$$

We now bound the last two terms in (33). First, note that

$$
\begin{aligned}
\big[(1-\lambda_i^*)^2 - (1-\lambda_i^0)^2\big]^2 &= [(1-\lambda_i^*) + (1-\lambda_i^0)]^2[\lambda_i^* - \lambda_i^0]^2 \\
&\le 2\big[(1-\lambda_i^*)^2 + (1-\lambda_i^0)^2\big](\lambda_i^{*2} + \lambda_i^{02}).
\end{aligned} \tag{34}
$$

Then we have, by Lemma 1 with $K = S$ and (34),

$$
\begin{aligned}
&\varepsilon\mathbf{E}_\theta \sum_{i=1}^{\infty}(1-\lambda_i^*)^2\theta_i\sigma_i\xi_i \\
&\quad = \varepsilon\mathbf{E}_\theta \sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 - (1-\lambda_i^0)^2\big]\theta_i\sigma_i\xi_i \\
&\quad \ge -\varepsilon\mathbf{E}_\theta\left|\sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 - (1-\lambda_i^0)^2\big]\theta_i\sigma_i\xi_i\right| \\
&\quad \ge -\varepsilon\sqrt{2\log(NS)}\mathbf{E}_\theta\left\{\sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 - (1-\lambda_i^0)^2\big]^2\theta_i^2\sigma_i^2\right\}^{1/2} \\
&\qquad - 2\varepsilon\sqrt{\log(NS)/S}\left\{\mathbf{E}_\theta \sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 - (1-\lambda_i^0)^2\big]^2\theta_i^2\sigma_i^2\right\}^{1/2} \\
&\quad \ge -2\varepsilon\sqrt{\log(NS)}\mathbf{E}_\theta\left\{\sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 + (1-\lambda_i^0)^2\big](\lambda_i^{*2} + \lambda_i^{02})\theta_i^2\sigma_i^2\right\}^{1/2} \\
&\qquad - 2\varepsilon\sqrt{2\log(NS)/S}\left\{\mathbf{E}_\theta \sum_{i=1}^{\infty}\big[(1-\lambda_i^*)^2 + (1-\lambda_i^0)^2\big](\lambda_i^{*2} + \lambda_i^{02})\theta_i^2\sigma_i^2\right\}^{1/2}.
\end{aligned}
$$

The argument as in the proof of Lemma 3 gives, for any $B > 0$,

$$2\varepsilon\sqrt{\log(NS)}\mathbf{E}_\theta\left\{\sum_{i=1}^{\infty}[(1-\lambda_i^*)^2+(1-\lambda_i^0)^2](\lambda_i^{*2}+\lambda_i^{02})\theta_i^2\sigma_i^2\right\}^{1/2}$$
$$\le \frac{1}{2B}\mathbf{E}_\theta\sum_{i=1}^{\infty}[(1-\lambda_i^*)^2+(1-\lambda_i^0)^2]\theta_i^2$$
$$+2B\varepsilon^2\log(NS)\mathbf{E}_\theta\left\{\sup_i\sigma_i^2\lambda_i^{*2}+\sup_i\sigma_i^2\lambda_i^{02}\right\}$$

and

$$2\varepsilon\sqrt{2\log(NS)/S}\left\{\mathbf{E}_\theta\sum_{i=1}^{\infty}[(1-\lambda_i^*)^2+(1-\lambda_i^0)^2](\lambda_i^{*2}+\lambda_i^{02})\theta_i^2\sigma_i^2\right\}^{1/2}$$
$$\le 2\varepsilon\sqrt{2\log(NS)/S}\left\{\mathbf{E}_\theta\sum_{i=1}^{\infty}[(1-\lambda_i^*)^2+(1-\lambda_i^0)^2]\theta_i^2\right\}^{1/2}$$
$$\times\left(\max_{\lambda\in\Lambda}\sup_i\{\sigma_i^2\lambda_i^2+\sigma_i^2\lambda_i^{02}\}\right)^{1/2}$$
$$\le \frac{1}{2B}\mathbf{E}_\theta\sum_{i=1}^{\infty}[(1-\lambda_i^*)^2+(1-\lambda_i^0)^2]\theta_i^2$$
$$+4B\varepsilon^2\log(NS)\mathbf{E}_\theta\left\{\sup_i\sigma_i^2\lambda_i^{*2}+\sup_i\sigma_i^2\lambda_i^{02}\right\}.$$

Putting together these inequalities, we find

$$\begin{aligned}
&\varepsilon\mathbf{E}_\theta\sum_{i=1}^{\infty}(1-\lambda_i^*)^2\theta_i\sigma_i\xi_i\\
&\quad\ge -B^{-1}\mathbf{E}_\theta\sum_{i=1}^{\infty}(1-\lambda_i^*)^2\theta_i^2\\
&\quad\quad -B^{-1}\sum_{i=1}^{\infty}(1-\lambda_i^0)^2\theta_i^2-6B\varepsilon^2\log(NS)\mathbf{E}_\theta\left\{\sup_i\sigma_i^2\lambda_i^{*2}+\sup_i\sigma_i^2\lambda_i^{02}\right\}\\
&\quad\ge -B^{-1}\mathbf{E}_\theta\sum_{i=1}^{\infty}(1-\lambda_i^*)^2\theta_i^2-B^{-1}R_\varepsilon[\lambda^0,\theta]-6B\mathbf{E}_\theta\Delta^\varepsilon[\lambda^*]-6B\Delta^\varepsilon[\lambda^0].
\end{aligned}\tag{35}$$

Now, we bound the last term in (33) using Lemma 2 with $K = S$, $q = 3$ and $v_i = (\lambda_i^{*2} - 2\lambda_i^*)\sigma_i^2$. Note that

$$\lambda_i^2 \le (\lambda_i^2 - 2\lambda_i)^2 \le 9\lambda_i^2 \qquad \forall\, |\lambda_i| \le 1.$$

This and (12) entail

$$m(v) = \frac{\sup_i |\lambda_i^2 - 2\lambda_i|\sigma_i^2}{\left[\sum_{i=1}^{\infty} \sigma_i^4(\lambda_i^2 - 2\lambda_i)^2\right]^{1/2}} \le 3\rho(\lambda) \le 3\sqrt{C_1}.$$

Hence, $M(3) = \sum_{v \in V} \exp\{-3/m(v)\} \le M$. Furthermore, $m_V \le 3\rho$, and

$$\sqrt{\log(NS)} + m_V \log\bigl(M(3)S\bigr) \le C\sqrt{L_\Lambda}.$$

Therefore, by Lemma 2, we obtain

$$\begin{aligned}
&\varepsilon^2 \mathbf{E}_\theta \sum_{i=1}^{\infty} (1 - \lambda_i^*)^2 \sigma_i^2 (\xi_i^2 - 1) \\
&\quad \ge -C\sqrt{L_\Lambda}\varepsilon^2 \mathbf{E}_\theta \left[\sum_{i=1}^{\infty} (\lambda_i^{*2} - 2\lambda_i^*)^2 \sigma_i^4\right]^{1/2} \\
&\qquad - C\sqrt{L_\Lambda}\varepsilon^2 S^{-1/2} \left[\mathbf{E}_\theta \sum_{i=1}^{\infty} (\lambda_i^{*2} - 2\lambda_i^*)^2 \sigma_i^4\right]^{1/2} \\
&\quad \ge -C\sqrt{L_\Lambda}\varepsilon^2 \left\{\mathbf{E}_\theta \left[\sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^{*4}\right]^{1/2} + S^{-1/2} \left[\mathbf{E}_\theta \sum_{i=1}^{\infty} \sigma_i^4 \lambda_i^{*4}\right]^{1/2}\right\}.
\end{aligned}$$

Here and below in this section C is a generic notation for positive constants that depend only on C_1. Repeating the argument of (32) to bound the last expression we finally get

$$\begin{aligned}
&\varepsilon^2 \mathbf{E}_\theta \sum_{i=1}^{\infty} (1 - \lambda_i^*)^2 \sigma_i^2 (\xi_i^2 - 1) \\
&\quad \ge -2\varepsilon^2 B^{-1} \mathbf{E}_\theta \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^{*2} - C\varepsilon^2 B L_\Lambda \mathbf{E}_\theta \sup_k \sigma_k^2 \lambda_k^{*2} \qquad (36) \\
&\quad = -2\varepsilon^2 B^{-1} \mathbf{E}_\theta \sum_{i=1}^{\infty} \sigma_i^2 \lambda_i^{*2} - CB\mathbf{E}_\theta \Delta_\varepsilon[\lambda^*].
\end{aligned}$$

Now we are ready to complete the proof of Theorems 1 and 2. From (35), (36) we have

$$\begin{aligned}
&2\varepsilon \mathbf{E}_\theta \sum_{i=1}^{\infty} (1 - \lambda_i^*)^2 \theta_i \sigma_i \xi_i + \varepsilon^2 \mathbf{E}_\theta \sum_{i=1}^{\infty} (1 - \lambda_i^*)^2 \sigma_i^2 (\xi_i^2 - 1) \\
&\quad \ge -2B^{-1} R_\varepsilon[\lambda^*, \theta] - 2B^{-1} R_\varepsilon[\lambda^0, \theta] - CB\mathbf{E}_\theta \Delta_\varepsilon[\lambda^*] - CB\Delta_\varepsilon[\lambda^0].
\end{aligned}$$

This and (33) yield

$$\mathbf{E}_\theta R_\varepsilon[\lambda^*,\theta] \le \mathbf{E}_\theta U[\lambda^*,X] + \sum_{i=1}^{\infty}\theta_i^2 + 2B^{-1}\mathbf{E}_\theta R_\varepsilon[\lambda^*,\theta] + CB\mathbf{E}_\theta \Delta_\varepsilon[\lambda^*] + 2B^{-1}R_\varepsilon[\lambda^0,\theta] + CB\Delta^\varepsilon[\lambda^0]. \tag{37}$$

By definition of λ^* and (9),

$$\mathbf{E}_\theta U[\lambda^*,X] \le \mathbf{E}_\theta U[\lambda^0,X] = R_\varepsilon[\lambda^0,\theta] - \sum_{i=1}^{\infty}\theta_i^2.$$

This and (37) imply

$$(1-2B^{-1})\mathbf{E}_\theta R_\varepsilon[\lambda^*,\theta] \le (1+2B^{-1})R_\varepsilon[\lambda^0,\theta] + CB\mathbf{E}_\theta \Delta_\varepsilon[\lambda^*] + CB\Delta^\varepsilon[\lambda^0]. \tag{38}$$

Next, by Lemma 3 for any $B>0$,

$$\mathbf{E}_\theta\|\theta^*-\theta\|^2 \le (1+2B^{-1})\mathbf{E}_\theta R_\varepsilon[\lambda^*,\theta] + CB\mathbf{E}_\theta\Delta^\varepsilon[\lambda^*]. \tag{39}$$

Inequalities (38) and (39) entail Theorems 1 and 2. In fact, to get Theorem 1 we use the following argument. Note that, for any $x>0$,

$$\begin{aligned}
\sup_i \sigma_i^2\lambda_i^2 &= \sup_i \sigma_i^2\lambda_i^2 I\left\{x\sup_i\sigma_i^2\lambda_i^2 < \sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2\right\} \\
&\quad + \sup_i \sigma_i^2\lambda_i^2 I\left\{x\sup_i\sigma_i^2\lambda_i^2 \ge \sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2\right\} \\
&\le \frac{1}{x}\sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2 + \max_{\lambda\in\Lambda}\sup_i \sigma_i^2\lambda_i^2 I\left\{x\sup_i\sigma_i^2\lambda_i^2 \ge \sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2\right\} \\
&= \frac{1}{x}\sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2 + \omega(x) \le \frac{1}{x\varepsilon^2}R_\varepsilon[\lambda,\theta] + \omega(x) \qquad \forall\,\theta\in\ell_2.
\end{aligned}$$

This inequality with $x = B^2L_\Lambda$, the definition of $\Delta^\varepsilon[\lambda]$ and (38), (39) yield Theorem 1.

PROOF OF THEOREM 2. Note first that, in view of (11),

$$\frac{\sup_i \sigma_i^2\lambda_i^2}{\sum_{i=1}^{\infty}\sigma_i^2\lambda_i^2} \le \rho\frac{\sup_i \sigma_i^2\lambda_i^2}{\left(\sum_{i=1}^{\infty}\sigma_i^4\lambda_i^4\right)^{1/2}} \le \rho^2.$$

Thus,

$$\Delta^\varepsilon[\lambda] \le \rho^2 L_\Lambda R_\varepsilon[\lambda,\theta] \qquad \forall\,\lambda\in\Lambda,\ \theta\in\ell_2. \tag{40}$$

Substitution of (40) into (38) and (39) gives, respectively,

$$\mathbf{E}_\theta R_\varepsilon[\lambda^*, \theta] \le \frac{1 + 2B^{-1} + \gamma_5 B\rho^2 L_\Lambda}{1 - 2B^{-1} - \gamma_6 B\rho^2 L_\Lambda} R_\varepsilon[\lambda^0, \theta] \tag{41}$$

(provided the denominator here is positive) and

$$\mathbf{E}_\theta \|\theta^* - \theta\|^2 \le (1 + 2B^{-1} + \gamma_7 B\rho^2 L_\Lambda)\mathbf{E}_\theta R_\varepsilon[\lambda^*, \theta]. \tag{42}$$

The constants γ_i here are positive and depend only on C_1. There exists $\gamma_4 > 0$ small enough such that if $\rho^2 L_\Lambda < \gamma_4$, the choice of $B = (\rho^2 L_\Lambda)^{-1/2}$ satisfies the inequality $2B^{-1} + \gamma_6 B\rho^2 L_\Lambda < 1/2$. With this choice of B, (41) and (42) entail Theorem 2. $\square$

6. Application to sharp adaptive estimation. In this section we apply the oracle inequalities of Section 2 to show that sharp minimax adaptive estimators for inverse problems can be obtained by the principle of unbiased risk estimation. We study the problem where sharp adaptive estimators were not known previously, namely a recovery of *anisotropic* smooth functions from indirect noisy data. For brevity, we restrict the discussion to a specific example (measuring the temperature of the earth). However, the key elements of the proofs are given under general assumptions and the result can be easily extended.

Let $\phi = (\phi_1, \phi_2)$ be the polar coordinates of a point on the surface of the earth (we suppose for simplicity that the earth is a sphere). Consider the problem of measuring the temperature $t(\phi)$ at the point ϕ. The function $t(\phi)$ is sufficiently smooth (since it is a solution of a thermal conductivity equation) and, of course, periodic. More specifically, we assume that $t(\phi)$ belongs the following anisotropic Sobolev ball:

$$W_2^m(p) = \left\{ t : \int_T \left[p_1^2 \left(\frac{\partial^{m_1} t}{\partial \phi_1^{m_1}} \right)^2 + p_2^2 \left(\frac{\partial^{m_2} t}{\partial \phi_2^{m_2}} \right)^2 \right] d\phi \le 1 \right\}, \tag{43}$$

where $T = [0, 2\pi] \times [0, 2\pi]$ and the parameters $m = (m_1, m_2)$, $p = (p_1, p_2)$ characterize the smoothness of the function $t(\phi)$ with respect to ϕ_1 and ϕ_2 ($m_i \in \mathbb{N}$, $p_i > 0$). The anisotropic character of smoothness here is important and reflects the fact that the temperature changes more rapidly along the meridians than along the parallels.

Next, it is assumed that temperature is measured by means of remote infrared detectors located for instance on satellites or on planes. It means that one cannot measure temperature at a given point directly, but one rather measures an average temperature in a vicinity of this point, that is,

$$Et(\phi) = \frac{1}{C(\alpha_0)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\{-2\pi\alpha_0 \|\phi - \phi'\|\} t(\phi_1', \phi_2')\, d\phi_1'\, d\phi_2'.$$

Here $C(\alpha_0)$ is the normalizing constant,

$$C(\alpha_0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\{-2\pi\alpha_0 \|\phi\|\}\, d\phi_1\, d\phi_2$$

and $\|\phi\|^2 = \phi_1^2 + \phi_2^2$. The parameter α_0^{-1} characterizes the size of the vicinity of the point where we measure the temperature. (We consider here and below integration over the whole line instead of integration over T, in order to simplify the technical details and the resulting sequence model.)

The measurements are contaminated by a noise inherent to infrared detection. We may write the resulting model as

$$y(\phi) = Et(\phi) + \varepsilon n(\phi), \qquad \phi \in T, \tag{44}$$

where $n(\phi)$ is a standard white Gaussian noise in the Hilbert space of periodic functions. Writing (44) means that for every $g \in L_2(T)$ we observe the random variable $Y(g) = (g, Et) + \varepsilon\xi(g)$, where $\xi(g) \sim \mathcal{N}(0, \|g\|_2)$ and $\mathbf{E}(\xi(g)\xi(g')) = 0$ if $(g, g') = 0$ [here $(\cdot,\cdot)$ and $\|\cdot\|$ are the scalar product and the norm in $L_2(T)$].

Since the function $t(\phi_1, \phi_2)$ is periodic it is convenient to consider the statistical model (44) in the Fourier domain. Denote by

$$\theta_{kl} = \int_T \exp(ik\phi_1 + il\phi_2)t(\phi)\,d\phi, \qquad k, l = 0, \pm1, \pm2, \ldots,$$

the Fourier coefficients of $t(\phi)$. Then (44) is equivalent to the model

$$z_{kl} = \theta_{kl} + \varepsilon\sigma_{kl}\xi_{kl}, \tag{45}$$

where $z_{kl} = Y(\exp(ik\phi_1 + il\phi_2))$, ξ_{kl} are i.i.d. $\mathcal{N}(0, 1)$ random variables and

$$\sigma_{kl}^{-2} = \frac{1}{C(\alpha_0)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{ik\phi_1 + il\phi_2 - 2\pi\alpha_0\|\phi\|\}\,d\phi = \frac{\alpha_0^3}{(\alpha_0^2 + k^2 + l^2)^{3/2}}.$$

[For the last equality see, for instance, Stein and Weiss (1971), Theorem 1.14.] Note also that the image of Sobolev's ball $W_2^m(p)$ in the Fourier domain has the form

$$\Theta_2^m(p) = \left\{\{\theta_{kl}\}: \sum_{k,l=-\infty}^{\infty} \theta_{kl}^2 a_{kl}^2 \le 1\right\}, \tag{46}$$

where $a_{kl} = [p_1^2(2\pi k)^{2m_1} + p_2^2(2\pi l)^{2m_2}]^{1/2}$. Thus our problem is equivalent to recovering the Fourier coefficients $\theta = \{\theta_{kl}\}$ based on the data (45) under the prior information provided by (46).

Recall that an estimator $\widehat{\theta}^M$ is called asymptotically minimax on $\Theta_2^m(p)$ if, as $\varepsilon \to 0$,

$$r^{\varepsilon}[\Theta_2^m(p)] = \inf_{\widetilde{\theta}} \sup_{\theta\in\Theta_2^m(p)} \mathbf{E}\|\widetilde{\theta} - \theta\|^2 = \big(1 + o(1)\big) \sup_{\theta\in\Theta_2^m(p)} \mathbf{E}\|\widehat{\theta}^M - \theta\|^2,$$

where the infimum is taken over all estimators. The next proposition derived from Pinsker's (1980) theorem gives an asymptotically minimax estimator and its risk.

Denote by μ a root of the equation

$$\frac{\varepsilon^2}{\mu} \sum_{k,l} \sigma_{kl}^2 a_{kl}[1 - \mu a_{kl}]_+ = 1. \tag{47}$$

PROPOSITION 4. *The linear estimator*

$$\widehat{\theta}_{kl}^{M} = [1 - \mu a_{kl}]_+ z_{kl}, \qquad k, l = 0, \pm 1, \pm 2, \dots, \tag{48}$$

is asymptotically minimax on $\Theta_2^m(p)$ *with risk*

$$\sup_{\theta \in \Theta_2^m(p)} \mathbf{E}\|\widehat{\theta}^M - \theta\|^2 = \varepsilon^2 \sum_{k,l} \sigma_{kl}^2 [1 - \mu a_{kl}]_+.$$

Using this proposition one can compute the asymptotic minimax risk with the exact constant. The calculations are cumbersome and we do not reproduce them here, since the knowledge of the exact asymptotic minimax risk is not needed to prove sharp adaptivity of our estimator. We need only to evaluate the order of $r^{\varepsilon}[\Theta_2^m(p)]$:

LEMMA 4. *Let* $m_2 \geq m_1$. *Then, as* $\varepsilon \to 0$,

$$r^{\varepsilon}[\Theta_2^m(p)] \leq C\bigl(\alpha_0^3 p_1^{4/m_1} p_2^{1/m_2}\bigr)^{-\gamma} \varepsilon^{2\gamma}$$

and

$$\mu = O(\varepsilon^{\gamma}),$$

where the constant $C > 0$ *does not depend on* α_0, p_1, p_2 *and*

$$\gamma = \frac{2m_1 m_2}{2m_1 m_2 + m_1 + 4m_2}.$$

REMARK 4. We can write $\gamma = 2s/(2s+1)$, where

$$\frac{1}{s} = \frac{1}{m_2} + \frac{4}{m_1} = \frac{1}{\max(m_1, m_2)} + \frac{4}{\min(m_1, m_2)}.$$

It is remarkable that the respective roles of m_1 and m_2 are not equivalent, unlike the case where $\sigma_{kl} \equiv 1$. The coefficient σ_{kl} is symmetric in k, l, and nevertheless the rate is asymmetric. This effect is due to the anisotropy of the class $\Theta_2^m(p)$.

PROOF OF LEMMA 4. Denote $V_i = (2\pi)^{-1}(p_i \mu)^{-1/m_i}$, $i = 1, 2$. Since σ_{kl}^2 is increasing in k, l we have

$$\varepsilon^2 \sum_{k,l} \sigma_{kl}^2 [1 - \mu a_{kl}]_+ \leq C \varepsilon^2 V_1 V_2 \sigma_{V_1 V_2}^2. \tag{49}$$

Here and later we denote by C positive constants, possibly different on different occasions. It is clear from (47) that $\mu \to 0$ (and thus $V_1 \to \infty$, $V_2 \to \infty$), as

$\varepsilon \to 0$. Consequently, for ε small enough we get

$$
\begin{aligned}
&\sum_{k\ge V_1/2,\, l\ge V_2/2} a_{kl}[1-\mu a_{kl}]_+ \\
&\quad\ge 2^{-m_2}\mu^{-1} \sum_{k\ge V_1/2,\, l\ge V_2/2} \Big[1-\sqrt{(k/V_1)^{2m_1}+(l/V_2)^{2m_2}}\Big]_+ \\
&\quad\ge 2^{-m_2}\mu^{-1} \int_{x\ge V_1/2+1}\int_{y\ge V_2/2+1} \Big[1-\sqrt{(x/V_1)^{2m_1}+(y/V_2)^{2m_2}}\Big]_+ dx\,dy \\
&\quad\ge 2^{-m_2} V_1V_2\mu^{-1} \int_{x\ge 1/2+1/V_1}\int_{y\ge 1/2+1/V_2} \Big[1-\sqrt{x^2+y^2}\Big]_+ dx\,dy \\
&\quad\ge C V_1V_2\mu^{-1},
\end{aligned}
\tag{50}
$$

where we used that $m_1 \ge 1$, $m_2 \ge 1$. Noting that $\sigma^2_{\lfloor k/2\rfloor \lfloor l/2\rfloor} \ge c\sigma^2_{kl}$, where $c>0$ does not depend on k, l, and using (50) one obtains

$$
\begin{aligned}
1 &= \frac{\varepsilon^2}{\mu}\sum_{k,l}\sigma^2_{kl}a_{kl}[1-\mu a_{kl}]_+ \ge \frac{c\varepsilon^2}{\mu}\sum_{|k|\ge V_1/2,|l|\ge V_2/2}\sigma^2_{V_1V_2}a_{kl}[1-\mu a_{kl}]_+ \\
&\ge C\varepsilon^2\mu^{-2}V_1V_2\sigma^2_{V_1V_2} = C\alpha_0^{-3}\varepsilon^2\mu^{-2}V_1V_2(\alpha_0^2+V_1^2+V_2^2)^{3/2}.
\end{aligned}
$$

The proof follows now from this inequality, (49), Proposition 4 and simple algebra. □

The minimax estimator $\widehat{\theta}^M$ defined by (48) depends on the parameters (m, p) of the functional class $\Theta_2^m(p)$, which are usually not known in practice. To overcome this drawback of $\widehat{\theta}^M$ we construct another estimator which is asymptotically minimax but does not depend on the parameters of the Sobolev ball. This estimator is called sharp adaptive. We show that a sharp adaptive estimator can be obtained by using the filter (10) with appropriately chosen set Λ.

Let Λ_0 be the set of all filters with weights λ_{kl} having the form

$$
\lambda_{kl}(W,\beta) = \lambda_{kl}(W_1, W_2, \beta_1, \beta_2) = \Big[1-\sqrt{(k/W_1)^{2\beta_1}+(l/W_2)^{2\beta_2}}\Big]_+,
$$

where $W \in [\log(1/\varepsilon),\infty)\times[\log(1/\varepsilon),\infty)$, $\beta \in [1,\infty)\times[1,\infty)$. The minimax filter belongs to this set for ε small enough [cf. (48) and note that $\mu = O(\varepsilon^\gamma)$ by Lemma 4]. Unfortunately Λ_0 contains infinitely many elements and we cannot apply Theorem 1 for $\Lambda = \Lambda_0$. Note also that the direct numerical minimization of the unbiased risk estimator over Λ_0 is very time consuming. Therefore we look for a finite subfamily which approximates well the filters in Λ_0. To make the estimator numerically feasible we will pick an approximating set containing the "minimal" number of elements.

Let $\delta = \log^{-1}(1/\varepsilon)$. Denote $w_\delta(k) = (1+\delta)^k$, $k = 1, 2, \ldots,$ and $\beta_\delta(k) = (1-\delta k)^{-1}$, $k = 1, 2, \ldots, \lfloor 1/\delta\rfloor$. Define the approximating finite subfamily Λ_δ as

a set of filters of the following form: $\lambda = \{\lambda_{kl}(w_\delta(i), w_\delta(j), \beta_\delta(s), \beta_\delta(p))\}$, where $i, j, s, p \in \mathbb{N}$ are such that $(w_\delta(i), w_\delta(j)) \in [\log(1/\varepsilon), \varepsilon^{-1}] \times [\log(1/\varepsilon), \varepsilon^{-1}]$ and $(\beta_\delta(s), \beta_\delta(p)) \in (1, \sqrt{\log(1/\varepsilon)}] \times (1, \sqrt{\log(1/\varepsilon)}]$.

Although, for the sake of simplicity, we assumed here a specific form of σ_{kl}, our results hold for the general situation where σ_{kl} satisfy the following assumption.

ASSUMPTION 3. The sequence σ_{kl}^2 is positive, monotone nondecreasing in k and l, and there exists $c_* > 0$ such that $\sigma_{\lfloor k/2 \rfloor \lfloor l/2 \rfloor}^2 \geq c_* \sigma_{kl}^2$ for all k, l.

LEMMA 5. *Let Assumption* 3 *hold. For any* $\lambda \in \Lambda_0$ *there exists a filter* $\lambda' \in \Lambda_\delta$ *such that, for all* $\theta \in l_2$,

$$\mathbf{E}_\theta \|\widehat{\theta}(\lambda') - \theta\|^2 \leq (1 + C\delta)\mathbf{E}_\theta \|\widehat{\theta}(\lambda) - \theta\|^2, \tag{51}$$

where $C > 0$ *does not depend on* $\theta, \lambda, \lambda'$.

PROOF. Since

$$\mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 = \sum_{k,l} (1 - \lambda_{kl})^2 \theta_{kl}^2 + \varepsilon^2 \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2$$

it suffices to show that for any $\lambda \in \Lambda_0$ there exists a filter $\lambda' \in \Lambda_\delta$ such that $\lambda_{kl} \leq \lambda'_{kl}$ for all k, l and

$$\sum_{k,l} \sigma_{kl}^2 \lambda_{kl}'^2 \leq (1 + C\delta) \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2.$$

To prove this, first note that the componentwise inequality $(W, \beta) \leq (W', \beta')$ implies

$$\lambda_{kl}(W, \beta) \leq \lambda_{kl}(W', \beta'). \tag{52}$$

Fix $(W, \beta) = (W_1, W_2, \beta_1, \beta_2)$ and define

$$i_0 = \max\{i : w_\delta(i) \leq W_1\}, \qquad j_0 = \max\{j : w_\delta(j) \leq W_2\},$$
$$s_0 = \max\{s : \beta_\delta(s) \leq \beta_1\}, \qquad p_0 = \max\{p : \beta_\delta(p) \leq \beta_2\}.$$

Set

$$(W^0, \beta^0) = \big(w_\delta(i_0), w_\delta(j_0), \beta_\delta(s_0), \beta_\delta(p_0)\big),$$
$$(W^1, \beta^1) = \big(w_\delta(i_0 + 1), w_\delta(j_0 + 1), \beta_\delta(s_0 + 1), \beta_\delta(p_0 + 1)\big).$$

We have $(W^0, \beta^0) \leq (W, \beta) \leq (W^1, \beta^1)$, and thus

$$\lambda_{kl}(W^0, \beta^0) \leq \lambda_{kl}(W, \beta) \leq \lambda_{kl}(W^1, \beta^1). \tag{53}$$

Set $\lambda'_{kl} = \lambda_{kl}(W^1, \beta^1)$. In view of (53) it suffices to show that

$$\sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2(W^1, \beta^1) \leq (1 + C\delta) \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2(W^0, \beta^0). \tag{54}$$

Let $V(W_1, W_2, \beta_1, \beta_2) = \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2(W_1, W_2, \beta_1, \beta_2)$. From Lemma 6 of the Appendix we have

$$V(W^1, \beta^1) - V(W^0, \beta^0) \leq CV(W^0, \beta^0) \max_k \left(\frac{w_\delta(k+1) - w_\delta(k)}{w_\delta(k)} + \frac{\beta_\delta(k+1) - \beta_\delta(k)}{\beta_\delta^2(k+1)} \right).$$

This and the obvious inequalities

$$\beta_\delta(k+1) - \beta_\delta(k) \leq \delta \beta_\delta^2(k+1), \qquad w_\delta(k+1) - w_\delta(k) = \delta w_\delta(k)$$

imply (54). Lemma 5 is proved. $\square$

Now, choose the data-driven filter $\lambda^* = (\lambda_{kl}^*, k, l = 0, \pm 1, \pm 2, \ldots)$ in the finite family of filters $\Lambda = \Lambda_\delta$ using the unbiased risk estimator as in (10):

$$\lambda^* = \arg\min_{\lambda \in \Lambda_\delta} \left\{ \sum_{k,l} (\lambda_{kl}^2 - 2\lambda_{kl}) z_{kl}^2 + 2\varepsilon^2 \sum_{k,l} \sigma_{kl}^2 \lambda_{kl} \right\}.$$

THEOREM 3. *The estimator* $\theta^* = (\theta_{kl}^*, k, l = 0, \pm 1, \pm 2, \ldots)$, *where* $\theta_{kl}^* = \lambda_{kl}^* z_{kl}$, *is minimax sharp adaptive on the scale of Sobolev classes*; *that is*, *for every* (m, p) *such that* $m_i \in \mathbb{N}$, $p_i > 0$ *we have*

$$\sup_{\theta \in \Theta_2^m(p)} \mathbf{E}_\theta \|\theta^* - \theta\|^2 \leq (1 + o(1)) \inf_{\tilde{\theta}} \sup_{\theta \in \Theta_2^m(p)} \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2, \qquad \varepsilon \to 0, \tag{55}$$

where $\inf_{\tilde{\theta}}$ *is the infimum over all estimators*.

PROOF. We apply Theorem 1. First, we check Assumptions 1 and 2. Assumption 1 is obvious; Assumption 2 follows from the inequalities [cf. (59) of the Appendix]

$$\sum_{k,l} \sigma_{kl}^4 \lambda_{kl}^4 \geq c_* \sigma_{W_1 W_2}^4 \sum_{|k| \geq W_1/2, |l| \geq W_2/2} \lambda_{kl}^4 \geq C W_1 W_2 \sigma_{W_1 W_2}^4 \tag{56}$$

and

$$\sum_{k,l} \sigma_{kl}^4 \lambda_{kl}^2 \leq \sum_{|k| \leq W_1, |l| \leq W_2} \sigma_{kl}^4 \leq C W_1 W_2 \sigma_{W_1 W_2}^4,$$

where the constants $C > 0$ do not depend on (W, β). Again using (56), we find

$$\rho(\lambda) \leq \sup_{|k| \leq W_1, |l| \leq W_2} \sigma_{kl}^2 \left(\sum_{k,l} \sigma_{kl}^4 \lambda_{kl}^4 \right)^{-1/2} \leq C(W_1 W_2)^{-1/2} \leq C \log^{-1}(1/\varepsilon),$$

where the constants $C > 0$ do not depend on (W, β). Furthermore, using (59) of the Appendix, we obtain

$$\begin{aligned}\omega(x) &\le \max_{\lambda\in\Lambda_0} \max_{k,l} \sigma_{kl}^2\lambda_{kl}^2 I\left\{\sum_{k,l}\sigma_{kl}^2\lambda_{kl}^2 \le x \max_{k,l}\sigma_{kl}^2\lambda_{kl}^2\right\}\\ &\le \max_{W_1,W_2}\max_{|k|\le W_1,|l|\le W_2}\sigma_{kl}^2 I\left\{CW_1W_2\sigma_{W_1W_2}^2 \le x \max_{|k|\le W_1,|l|\le W_2}\sigma_{kl}^2\right\}\\ &\le C\max_{W_1,W_2}[W_1^2+W_2^2]^{3/2} I\{W_1W_2 \le Cx\} \le Cx^3\log^{-3}(1/\varepsilon),\end{aligned}$$

where we used that $W_i \ge \log(1/\varepsilon)$. Next, $N = \mathrm{Card}(\Lambda_\delta)$ is of the order $\log^6(1/\varepsilon)$ and the parameter S has the order ε^{-3}. Thus by Theorem 1 and Lemma 5 one concludes that for ε small enough the estimator θ^* satisfies the oracle inequality

$$\begin{aligned}\mathbf{E}_\theta\|\theta^*-\theta\|^2 &\le (1+CB^{-1})\min_{\lambda\in\Lambda_\delta}\mathbf{E}_\theta\|\widehat\theta(\lambda)-\theta\|^2 + C\varepsilon^2B^7\log(1/\varepsilon)\\ &\le (1+CB^{-1})(1+C\delta)\min_{\lambda\in\Lambda_0}\mathbf{E}_\theta\|\widehat\theta(\lambda)-\theta\|^2 + C\varepsilon^2B^7\log(1/\varepsilon) \qquad (57)\\ &= (1+CB^{-1})(1+C\delta)\mathbf{E}_\theta\|\widehat\theta^M-\theta\|^2 + C\varepsilon^2B^7\log(1/\varepsilon),\end{aligned}$$

where $\widehat\theta^M$ is the asymptotically minimax estimator defined in (48). For the last equality we used that this estimator belongs to Λ_0 for ε small enough [in view of (48) and the fact that $\mu = O(\varepsilon^\gamma)$ by Lemma 4]. Choose $B = \varepsilon^{-r/4}$, where $r = 1-\gamma > 0$ and γ is defined in Lemma 4. Taking the supremum of (57) w.r.t. $\theta \in \Theta_2^m(p)$, using Proposition 4 and observing that by Lemma 4 the minimax risk over $\Theta_2^m(p)$ has the order ε^{2-2r}, we get (55). □

Note that Theorem 3 remains valid if m in the definition of the ellipsoid $\Theta_2^m(p)$ is real-valued. Furthermore, it is easy to see from the proof that the $o(1)$ in (55) converges to 0 uniformly over a large set of values m, p.

APPENDIX

LEMMA 6. *Let Assumption* 3 *hold. Let* $(W_1, W_2, \beta_1, \beta_2)$ *belong to the set* $\mathcal{W} = [\log(1/\varepsilon),\infty) \times [\log(1/\varepsilon),\infty) \times [1, \sqrt{\log(1/\varepsilon)}] \times [1, \sqrt{\log(1/\varepsilon)}]$. *The partial derivatives of* $\log V$ *satisfy uniformly over* $\mathcal{W}$ *the inequalities*

$$\left|\frac{\partial}{\partial W_i}\log V\right| \le CW_i^{-1}, \quad \left|\frac{\partial}{\partial\beta_i}\log V\right| \le C\beta_i^{-2}, \qquad i = 1,2. \qquad (58)$$

PROOF. Using Assumption 3 and acting as in (50) we get, for $W_i \geq \log(1/\varepsilon)$,

$$
\begin{aligned}
&V(W_1, W_2, \beta_1, \beta_2) \\
&\quad = \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2 (W_1, W_2, \beta_1, \beta_2) \\
&\quad \geq c_* \sigma_{W_1 W_2}^2 \sum_{|k| \geq W_1/2, |l| \geq W_2/2} \lambda_{kl}^2 (W_1, W_2, \beta_1, \beta_2) \\
&\quad \geq c_* \sigma_{W_1 W_2}^2 \int_{|x| \geq W_1/2+1} \int_{|y| \geq W_2/2+1} \Big[1 - \sqrt{(x/W_1)^{2\beta_1} + (y/W_2)^{2\beta_2}} \Big]_+^2 \, dx \, dy \\
&\quad \geq c_* W_1 W_2 \sigma_{W_1 W_2}^2 \int_{|x| \geq 1/2+1/W_1} \int_{|y| \geq 1/2+1/W_2} \Big[1 - \sqrt{x^2 + y^2} \Big]_+^2 \, dx \, dy \\
&\quad \geq C W_1 W_2 \sigma_{W_1 W_2}^2,
\end{aligned}
\tag{59}
$$

where the constant $C > 0$ does not depend on (W, β). For brevity write $A(x, y) = (x^{2\beta_1} + y^{2\beta_2})^{-1/2}$. Since $A(x, y) \leq x^{-\beta_1}$, $\forall\, x > 0$, $y > 0$, one obtains

$$
\begin{aligned}
\left| \frac{\partial}{\partial W_1} V \right| &= \left| \frac{\partial}{\partial W_1} \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2 (W_1, W_2, \beta_1, \beta_2) \right| \\
&= 4\beta_1 W_1^{-1} \sum_{k,l} \sigma_{kl}^2 \lambda_{kl} (W_1, W_2, \beta_1, \beta_2) A\Big(\frac{k}{W_1}, \frac{l}{W_2} \Big) \Big(\frac{k}{W_1} \Big)^{2\beta_1 - 1} \\
&\leq 4\beta_1 W_1^{-1} \sigma_{W_1 W_2}^2 \sum_{|k| \leq W_1, |l| \leq W_2} \Big(\frac{k}{W_1} \Big)^{\beta_1 - 1} \\
&\leq C\beta_1 W_1^{-1} \sigma_{W_1 W_2}^2 W_2 \Big[W_1 \int_0^1 x^{\beta_1 - 1} dx + 1 \Big] \leq C W_2 \sigma_{W_1 W_2}^2,
\end{aligned}
\tag{60}
$$

where the constants $C > 0$ do not depend on (W, β), and we used that $W_1 \geq \log(1/\varepsilon)$, $\beta_1 \leq \sqrt{\log(1/\varepsilon)}$. Combining (59) and (60) we get the first inequality in (58). The second inequality in (58) is checked similarly. In fact, it suffices to use (59) and the relation

$$
\begin{aligned}
\left| \frac{\partial}{\partial \beta_1} V \right| &= \left| \frac{\partial}{\partial \beta_1} \sum_{k,l} \sigma_{kl}^2 \lambda_{kl}^2 (W_1, W_2, \beta_1, \beta_2) \right| \\
&= 4 \sum_{k,l} \sigma_{kl}^2 \lambda_{kl} (W_1, W_2, \beta_1, \beta_2) A\Big(\frac{k}{W_1}, \frac{l}{W_2} \Big) \Big(\frac{k}{W_1} \Big)^{2\beta_1} \log \frac{k}{W_1} \\
&\leq 4 \sigma_{W_1 W_2}^2 \sum_{k,l} \left[1 - \sqrt{\Big(\frac{k}{W_1} \Big)^{2\beta_1} + \Big(\frac{l}{W_2} \Big)^{2\beta_2}} \right]_+ \Big(\frac{k}{W_1} \Big)^{\beta_1} \left| \log \frac{k}{W_1} \right|
\end{aligned}
$$

$$\le CW_2\sigma^2_{W_1W_2}\sum_{|k|\le W_1}\left(\frac{k}{W_1}\right)^{\beta_1}\left|\log\frac{k}{W_1}\right|$$

$$\le CW_2\sigma^2_{W_1W_2}\left[W_1\int_0^1 x^{\beta_1}|\log(x)|\,dx+1\right]\le CW_1W_2\sigma^2_{W_1W_2}\beta_1^{-2},$$

where the constants $C>0$ do not depend on (W,β), and we used that $W_1\ge \log(1/\varepsilon)$, $\beta_1\le\sqrt{\log(1/\varepsilon)}$. □

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Internat. Symp. Inform. Theory*, *Budapest* (B. N. Petrov and F. Csaki, eds.) 267–281. Akademiai Kiado, Budapest.

BIRGÉ, L. (2001). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics. Festschrift for W. R. van Zwet* (M. de Gunst, C. Klaassen and A. van der Vaart, eds.) 113–133. IMS, Hayward, CA.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268.

CAVALIER, L. and TSYBAKOV, A. B. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*. To appear. Available at www.proba.jussieu.fr.

DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.

DONOHO, D. L. and JOHNSTONE, I. M. (1996). Neoclassical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2** 39–62.

GOLDENSHLUGER, A. and PEREVERZEV, S. V. (2000). Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory Related Fields* **118** 169–186.

GOLDENSHLUGER, A. and TSYBAKOV, A. B. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.* **29** 1601–1619.

GOLUBEV, G. K. (1987). Adaptive asymptotically minimax estimates of smooth signals. *Problems Inform. Transmission* **23** 57–67.

GOLUBEV, G. K. (1992). Nonparametric estimation of smooth probability densities in L_2. *Problems Inform. Transmission* **28** 44–54.

GOLUBEV, G. K. and KHASMINSKII, R. Z. (1999). A statistical approach to some inverse boundary problems for partial differential equations. *Problems Inform. Transmission* **35** 136–149.

GOLUBEV, G. K. and KHASMINSKII, R. Z. (2001). A statistical approach to the Cauchy problem for the Laplace equation. In *State of the Art in Probability and Statistics. Festschrift for W. R. van Zwet* (M. de Gunst, C. Klaassen and A. van der Vaart, eds.) 419–433. IMS, Hayward, CA.

GOLUBEV, G. K. and NUSSBAUM, M. (1992). Adaptive spline estimates in a nonparametric regression model. *Theory Probab. Appl.* **37** 521–529.

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

JOHNSTONE, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9** 51–83.

JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280.

KERKYACHARIAN, G. and PICARD, D. (2002). Minimax or maxisets? *Bernoulli* **8** 219–253.

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.

KOO, J.-Y. (1993). Optimal rates of convergence for nonparametric statistical inverse problems. *Ann. Statist.* **21** 590–599.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statist.* **82**. Springer, New York.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

LI, K.-C. (1987). Asymptotic optimality of C_P, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15** 958–976.

MAIR, B. and RUYMGAART, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56** 1424–1444.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

NEMIROVSKI, A. (2000). Topics in nonparametric statistics. *Ecole d'Été de Probabilités de St. Flour XXVIII. Lecture Notes in Math.* **1738** 85–277. Springer, New York.

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16** 120–133.

POLYAK, B. T. and TSYBAKOV, A. B. (1990). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35** 293–306.

POLYAK, B. T. and TSYBAKOV, A. B. (1992). A family of asymptotic optimal methods for choosing the estimate order in orthogonal series regression. *Theory Probab. Appl.* **37** 471–481.

STEIN, E. and WEISS, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces.* Princeton Univ. Press.

WAHBA, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14** 651–667.

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

L. CAVALIER
G. K. GOLUBEV
CMI
UNIVERSITÉ AIX-MARSEILLE 1
39 RUE F. JOLIOT-CURIE
F-13453 MARSEILLE CEDEX 13
FRANCE

D. PICARD
A. B. TSYBAKOV
LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UNIVERSITÉS PARIS 7–PARIS 6
4 PL. JUSSIEU, BP 188
F-75252 PARIS CEDEX 05
FRANCE
E-MAIL: tsybakov@ccr.jussieu.fr