Weierstraß–Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

On adaptive inverse estimating linear functionals of unknown smoothness in Hilbert scales

A. Goldenshluger[1], S.V. Pereverzev[2]

submitted: 5th October 2001

[1] Department of Statistics
University of Haifa
Haifa 31905
Israel
E-Mail: goldensh@rstat.haifa.ac.il

[2] Institute of Mathematics
Ukrainian Academy of Sciences
Tereshenkivska str. 3
252601 Kiev-4
Ukraine
E-Mail: serg-p@mail.kar.net

Preprint No. 686
Berlin 2001

2000 *Mathematics Subject Classification.* 62G05, 62G20, 65R30.

Key words and phrases. adaptive estimation, Hilbert scales, inverse problems, linear functionals, minimax risk, regularization.

Edited by
Weierstraß–Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail (X.400): c=de;a=d400-gw;p=WIAS-BERLIN;s=preprint
E-Mail (Internet): preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

We address the problem of estimating the value of a linear functional $\langle f, x\rangle$ from random noisy observations of $y = Ax$ in Hilbert scales. Both the white noise and density observation models are considered. We develop an inverse estimator of $\langle f, x\rangle$ which automatically adapts to unknown smoothness of x and f. It is shown that accuracy of this adaptive estimator is only by a logarithmic factor worse than one could achieve in the case of known smoothness. As an illustrative example, the problem of deconvolving a bivariate density with singular support is considered.

1 Introduction

Consider an operator equation

$$Ax = y, \tag{1}$$

where A is a linear compact injective operator from some real Hilbert space $\mathbb{X}$ into a real Hilbert space $\mathbb{Y}$. We denote the inner products in the Hilbert spaces $\mathbb{X}, \mathbb{Y}$ by $\langle \cdot, \cdot \rangle$ and corresponding norms by $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$. It will be always clear from the context which space is considered. The problem of inverse statistical estimation is to reconstruct x or a functional of it, provided that the right hand side of (1) is observed with a random error. The statistical model can be written in the form

$$y_\varepsilon = Ax + \varepsilon\xi, \tag{2}$$

where ξ is a random noise, and ε is a small positive number measuring the noise level.

Two typical models of observations have been considered in the statistical literature. One can assume that ξ is the Gaussian white noise of the intensity ε [e.g., Skorohod (1974)]. This specifically means that for every element $\phi \in \mathbb{Y}$ we can observe

$$y_\varepsilon(\phi) = \langle Ax, \phi\rangle + \varepsilon\xi(\phi), \tag{3}$$

where $\xi(\phi)$ is a Gaussian random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with zero mean and variance $\|\phi\|^2$. Denoting $\mathbb{E}$ the expectation with respect to $\mathbb{P}$, we have in addition $\mathbb{E}[\xi(\phi)\xi(\psi)] = \langle \phi, \psi\rangle$, $\forall \phi, \psi \in \mathbb{Y}$. We refer to such a model as the *white noise model.*

On the other hand, in some practical situations it is natural to assume that we are given an i.i.d. sample $Y_1, \ldots, Y_n$ of random elements on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

which are in some sense directly related to y. In this case the data allow to construct i.i.d. statistics $\hat{y}_1, \dots, \hat{y}_n$ defined on the same probability space and taking values in $\mathbb{Y}$ with the following properties: $\mathbb{E}\hat{y}_i = y$ and $\mathbb{E}\|\hat{y}_i\|^2 < \infty$. Then a sensible estimator of y is given by

$$y_\varepsilon = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i, \quad \text{and} \quad \mathbb{E}\|y_\varepsilon - y\|^2 = O(n^{-1}), \quad n \to \infty. \tag{4}$$

This observation scheme corresponds to (2) with $\xi = \sqrt{n}(y_\varepsilon - y)$ and $\varepsilon = n^{-1/2}$. We will refer to such a model as the *density observation model* because it is typical in applications related to density estimation from indirect data. For examples of inverse estimation in this setup see Ruymgaart (1993), Mair & Ruymgaart (1996), and van Rooij, Ruymgaart & van Zwet (1999).

In this paper we consider the problem of estimating a linear functional $l_f(x) = \langle f, x\rangle$ for the two aforementioned models of indirect observations. Accuracy in estimating $l_f(x)$ is essentially determined by: (i) ill–posedness of the operator A, (ii) smoothness of the representer f, and (iii) smoothness of the unknown solution x. These factors can be characterized in several ways. For example, if a singular value decomposition (SVD) of A is known, then the inverse problem can be represented in a sequence space. In this case both ill–posedness of A and smoothness of f and x are naturally measured by the rate at which corresponding coefficients of the SVD representation decrease [e.g., Cavalier & Tsybakov (2000)]. In this paper we adopt a different approach. In order to quantify the effect of smoothness of x and f, and ill–posedness of A on the estimation accuracy, we embed the problem into a Hilbert scale. This approach to statistical inverse estimation has been advocated by Mair & Ruymgaart (1996), and Mathe & Pereverzev (2001). The operator A may fit some standard Hilbert scale such as a Sobolev scale. If this is not the case, the Hilbert scale can be always constructed using the generating operator $L := (A^*A)^{-1}$ [e.g., Natterer (1984), Hegland (1995)]. In what follows we call the Hilbert scale generated by $(A^*A)^{-1}$ *natural.* Thus within this framework smoothness of x and f, and ill–posedness of A are measured with respect to a particular Hilbert scale.

An inverse estimator of a linear functional $l_f(x)$ in Hilbert scales, adaptive to unknown smoothness of x, has been developed recently by Goldenshluger & Pereverzev (2000) [henceforth G&P (2000)] for the white noise model. It was assumed there that smoothness of the representer f with respect to the corresponding Hilbert scale is known. However, it is often difficult to characterize smoothness of f relative to a particular Hilbert scale. For example, if the operator A is not well studied then smoothness of the representer f with respect to the natural Hilbert scale is usually unknown. Therefore developing inverse estimators that are simultaneously adaptive to unknown smoothness of x and f is of prime interest.

We consider the following basic example that motivated this paper.

Example 1 Let z be a bivariate random variable that has a singular distribution on the plane with mass concentrated on a contour with given parametric representation.

In particular, assume that $z = \rho(\varphi)\exp\{\imath\varphi\}$, $\imath = \sqrt{-1}$, where $\rho(\cdot)$ is a given positive periodic function on $[0, 2\pi]$, and φ is a random variable with unknown density x on $[0, 2\pi]$. Suppose we observe

$$Y_j = \rho(\varphi_j)\exp\{\imath\varphi_j\} + w_j, \quad j = 1, \dots, n, \tag{5}$$

where w_j are bivariate Gaussian normal variables with zero mean and covariance matrix $\sigma^2 I$. Here we identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$. The objective is to estimate the density x at a single point $\varphi_0 \in [0, 2\pi]$. The case of $\rho(\varphi) = \text{const}$ has been considered recently by Goldenshluger (2001). It will be shown in Section 4 that, if $\rho(\theta) \neq \text{const}$ and $x \in \mathbb{L}_2(0, 2\pi)$, then x satisfies the following integral equation

$$Ax(t) := \int_0^{2\pi} J_0(t\rho(\varphi))x(\varphi)\,d\varphi = y(t), \quad t \in [0, \varrho], \quad \text{for any } \varrho > 0. \tag{6}$$

Here $J_0(\cdot)$ is the Bessel function of the order 0, $A : \mathbb{L}_2(0, 2\pi) \to \mathbb{L}_2(0, \varrho)$, and y is a function that can be estimated from the data at the parametric rate. Thus we are in the framework of the density observation model. Clearly, smoothness of $l_f(x) = x(\varphi_0)$ relative to the natural Hilbert scale is unknown, and this information cannot be used in estimator construction.

Given a functional f, let $\hat{l}^\varepsilon = \hat{l}^\varepsilon(f, x)$ be an estimate of $l_f(x)$ based on the available data. Accuracy of an estimate $\hat{l}^\varepsilon(f, x)$ is measured by its uniform with respect to W risk

$$\mathcal{R}[\hat{l}^\varepsilon; W] := \sup_{x \in W} \mathbb{E}|l_f(x) - \hat{l}^\varepsilon(f, x)|^2$$

where W is a prespecified subset of $\mathbb{X}$ reflecting prior knowledge on smoothness of $x = A^{-1}y$. The minimax risk is defined by

$$\mathcal{R}_*[\varepsilon; f, W] := \inf_{\hat{l}^\varepsilon} \mathcal{R}[\hat{l}^\varepsilon; W],$$

where inf is taken over all possible estimates $\hat{l}^\varepsilon$. The objective is to construct an *optimal in order* estimate $\hat{l}^\varepsilon = \hat{l}^\varepsilon(f, x)$ of a functional $l_f(x) = \langle f, x \rangle$ satisfying

$$\mathcal{R}[\hat{l}^\varepsilon; W] \leq O(1)\mathcal{R}_*[\varepsilon; f, W], \quad \varepsilon \to 0.$$

It turns out that dependence of the minimax risk $\mathcal{R}_*[\varepsilon; f, W]$ on f is expressed only through the smoothness properties of f. Therefore, for a given representer smoothness set F, the minimax risk can be measured uniformly over $f \in F$ by

$$\mathcal{R}_*[\varepsilon; F, W] := \sup_{f \in F} \mathcal{R}_*[\varepsilon; f; W].$$

Typically construction of optimal in order estimators requires prior information on the solution set W and on the representer smoothness set F. As Example 1 indicates, the last requirement is particularly restrictive when dealing with inverse estimation in Hilbert scales. Let $\mathcal{F}$ denote a family of representer sets F, and $\mathcal{W}$ be a family of solution sets W. We say that an estimator $\hat{l}^{\varepsilon} = \hat{l}^{\varepsilon}(f, x)$ is adaptive with respect to $\mathcal{W}$ and $\mathcal{F}$, or *fully adaptive*, if

$$\sup_{F \in \mathcal{F}} \sup_{W \in \mathcal{W}} \left\{ \mathcal{R}[\hat{l}^{\varepsilon}; W] / \mathcal{R}_*[\varepsilon; F, W] \right\} \le C(\varepsilon), \tag{7}$$

where $\sup_{\varepsilon} C(\varepsilon) < \infty$, or $C(\varepsilon)$ grows slowly as ε goes to 0, i.e. $\lim_{\varepsilon \to 0} [C(r\varepsilon)/C(\varepsilon)] = 1$ for any $r > 0$.

The goal of the present paper is develop an inverse estimator of a linear functional in Hilbert scales which automatically adapts to unknown smoothness of x and f. The proposed estimator satisfies (7) with both $\mathcal{F}$ and $\mathcal{W}$ being wide collections of balls in the Hilbert scale. This substantially extends the results of G&P (2000) where F is a fixed known ball in the Hilbert scale. We show that accuracy of our adaptive estimator is only by a logarithmic factor worse than one could achieve in the case where F and W are known exactly. It is well known that often this extra logarithmic factor cannot be avoided in adaptive estimating linear functionals [cf. Lepski (1990, 1992), Brown & Low (1996)]. In these situations our estimator has the best possible adaptation properties. Using the general results we treat the problem of deconvolving a bivariate density with singular support described in Example 1. A fully adaptive estimator is developed and its properties are studied.

The rest of the paper is organized as follows. In Section 2 we introduce our notation and main assumptions. Section 3 defines our adaptive estimator and establishes main results for the white noise and density observation models. In Section 4 we consider the problem of inverse estimating a bivariate density with singular support, and Section 5 contains the proofs.

2 Hilbert scale setup

Recall that a Hilbert scale $\{\mathbb{X}^{\lambda}\}_{\lambda \in \mathbb{R}}$ is a family of Hilbert spaces $\mathbb{X}^{\lambda}$ with inner products $\langle u, v \rangle_{\lambda} := \langle L^{\lambda} u, L^{\lambda} v \rangle$, where L is a given unbounded strictly positive self-adjoint operator in a dense domain of the initial Hilbert space $\mathbb{X}$. More precisely, $\mathbb{X}^{\lambda}$ is the completion of the intersection of domains of the operators L^s, $s \ge 0$, endowed with the norm $\|\cdot\|_{\lambda}$ defined by $\|\cdot\|_{\lambda} := \langle \cdot, \cdot \rangle_{\lambda}^{1/2}$. Here $\mathbb{X}^0 = \mathbb{X}$ and $\|\cdot\|_0 = \|\cdot\|$.

Following Natterer (1984) we assume that A is adapted to the Hilbert scale in the following sense.

(A) There exist positive constants a,d, and D such that

$$d\|u\|_{-a} \le \|Au\| \le D\|u\|_{-a}, \quad \forall u \in \mathbb{X}. \tag{8}$$

Examples of operators A satisfying (8) can be found in Neubauer (1988), Mair and Ruymgaart (1996), and Mathe and Pereverzev (2001). As it was already mentioned, even if the operator A does not fit some standard Hilbert scale (for instance, as in

Example 1), one can always construct a scale adapted to A. Namely, any compact injective operator A meets the condition (8) for $a = 1/2$ and the Hilbert scale generated by the operator $L = (A^*A)^{-1}$, where A^* is the adjoint of the operator A in $\mathbb{X}$, i.e. $A^* : \mathbb{Y} \to \mathbb{X}$.

Within the Hilbert scale setup the natural assumption on the linear functional $l_f(x) = \langle f, x\rangle$ is that both the representer f and the unknown solution x belong to some balls in the Hilbert scale. In particular, suppose that

$$x \in W_\mu(M), \quad W_\mu(M) := \{x \in \mathbb{X}^\mu : \|x\|_\mu \le M\}$$

for some index $\mu > 0$ and constant $M > 0$. Since the dual space of $\mathbb{X}^\mu$ is $\mathbb{X}^{-\mu}$ [e.g., Krein et al. (1982, p. 237)], and $\mathbb{X}^r$ is embedded in $\mathbb{X}^s$ for $r > s$, we need also the condition $f \in \mathbb{X}^\nu$, $\nu \ge -\mu$, to ensure that the linear functional $l_f(x) = \langle f, x\rangle$ is well–defined. To be more specific, we assume that

$$f \in W_\nu(N), \quad \nu \ge -\mu. \tag{9}$$

The condition (8) implies that the inverse operator A^{-1} acts boundedly from $\mathbb{Y}$ into $\mathbb{X}^{-a}$. Since the norm $\|\cdot\|_{-a}$ is weaker than the norm $\|\cdot\|$ of the initial Hilbert space $\mathbb{X}$, the problem (1) is ill-posed. Therefore some kind of regularization is required for estimating the value of $l_f(x) = \langle f, x\rangle$. It is well known [e.g., Tautenhann (1996)] that a wide variety of regularization methods can be constructed in the following way.

Let $g_\alpha(\cdot)$ be a piecewise continuous function on $[0, D^2]$ depending on a *regularization parameter* $\alpha > 0$ and satisfying the following conditions:

$$\begin{aligned} \sup_{\lambda\in[0,D^2]} |\lambda^\gamma g_\alpha(\lambda)| &\le c_\gamma \alpha^{\gamma-1}, \quad 0 \le \gamma \le 1, \\ \sup_{\lambda\in[0,D^2]} |\lambda^\beta[1 - \lambda g_\alpha(\lambda)]| &\le c_\beta \alpha^\beta, \qquad 0 \le \beta \le 1, \end{aligned}$$

where D is given in (8), and c_γ, c_β are positive constants. Fix a non-negative number s and define the regularized estimator $\hat{l}^\varepsilon_{\alpha,s}(x)$ of $l_f(x) = \langle f, x\rangle$ by

$$\hat{l}^\varepsilon_{\alpha,s}(x) = \langle y_\varepsilon, AL^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\rangle\,, \tag{10}$$

where y_ε is given by (3) for the white noise model and by (4) for the density observation model. Observe that if $s \ge -\nu$ then $AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f \in \mathbb{Y}$, and the estimate is well–defined. The well-known Tikhonov-Phillips regularization method is characterized by (10) with $g_\alpha(\lambda) = (\lambda + \alpha)^{-1}$ and $s = 0$.

The mean squared error of the estimate $\hat{l}^\varepsilon_{\alpha,s}(x)$ admits the following standard bias–variance decomposition:

$$\mathbb{E}|l_f(x) - \hat{l}^\varepsilon_{\alpha,s}(x)|^2 = b^2_{\alpha,s}(f, x) + \varepsilon^2 \mathbb{E} v^2_{\alpha,s}(f, \xi),$$

where

$$l_f(x) - \hat{l}^\varepsilon_{\alpha,s} = b_{\alpha,s}(f, x) + \varepsilon v_{\alpha,s}(f, \xi)$$

and

$$\begin{aligned} b_{\alpha,s}(f,x) &= \langle f, (I - L^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}A^*A)x\rangle, \\ v_{\alpha,s}(f,\xi) &= -\langle \xi, AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\rangle. \end{aligned} \tag{11}$$

The following statements have been proved in G&P (2000).

Lemma 2.1 *Let Assumption (A) hold, $f \in W_\nu(N)$, $\nu < a$, and $\hat{l}^\varepsilon_{\alpha,s}(x)$ be associated with $s \geq \max\{0, -\nu\}$. Then for every $\mu \in (-\nu, 2s+a]$ one has*

$$\sup_{x \in W_\mu(M)} |b_{\alpha,s}(f,x)| \leq c_1 M \|f\|_\nu \, \alpha^{\frac{\mu+\nu}{2(a+s)}}, \tag{12}$$

where $c_1 = c_1(\nu, a, s, d, D)$ depends on ν, a, s, d, D only.

Lemma 2.2 *Let Assumption (A) hold, and $s \geq \max\{0, -\nu\}$. Then for the white noise observation model*

$$\mathbb{E}v^2_{\alpha,s}(f,\xi) = \|AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\|^2$$

and for $f \in W_\nu(N)$, $\nu < a$,

$$\mathbb{E}v^2_{\alpha,s}(f,\xi) \leq c_2 \, \alpha^{\frac{\nu-a}{a+s}}, \tag{13}$$

where $c_2 = c_2(\nu, a, s, N)$ depends on ν, a, s and N only.

We note in passing that condition $\nu < a$ in the above statement implies that f is a generalized function relative to the Hilbert space $\mathbb{X}$. In Section 5, under some natural assumptions on the distribution of ξ, we establish the same upper bound (13) on the variance term for the density observation model. In this case the constant c_2 depends also on some properties of the error distribution.

The above results allow to establish upper bounds on the uniform risk. In particular, for the white noise observation model it follows from Lemmas 2.1, 2.2 that the uniform risk of the estimate $\hat{l}^\varepsilon_{\alpha,s}(x)$ associated with $s \geq \max\{0, -\nu\}$ and $\alpha > 0$ admits the following upper bound

$$\mathcal{R}[\hat{l}^\varepsilon_{\alpha,s}; W_\mu(M)] \leq c_3 \left(M^2 \alpha^{\frac{\mu+\nu}{a+s}} + \varepsilon^2 \alpha^{\frac{\nu-a}{a+s}}\right), \quad \forall \mu \in (-\nu, 2s+a],$$

where $c_3 = c_3(\nu, s, a, N, d, D)$. Thus, with the optimal choice $\alpha \asymp (M^{-1}\varepsilon)^{\frac{2(a+s)}{\mu+a}}$ one has

$$\mathcal{R}[\hat{l}^\varepsilon_{\alpha,s}; W_\mu(M)] \leq c_3 M^{-\frac{2(\nu-a)}{\mu+a}} \varepsilon^{\frac{2(\mu+\nu)}{\mu+a}}, \quad \forall \mu \in (-\nu, 2s+a], \tag{14}$$

where "$\asymp$ means equivalent in the sense of the order. It follows from the renormalization argument of Donoho & Low (1992) that, for the white noise model, the rate of convergence on the right hand side of (14) cannot be improved for estimating linear functionals $f \in W_\nu(N)$. More precisely, it is shown that

$$\mathcal{R}_*[\varepsilon; W_\nu(N), W_\mu(M)] \asymp M^{-\frac{\nu-a}{\mu+a}} \varepsilon^{\frac{2(\mu+\nu)}{\mu+a}}, \quad -\mu \leq \nu \leq a. \tag{15}$$

Thus, the estimate $\hat{l}^{\varepsilon}_{\alpha,s}(x)$ is optimal in order for any pair of balls $W_\nu(N)$ and $W_\mu(M)$ satisfying $\mu \in (-\nu, 2s+a]$, $\nu < a$. Observe, however, that prior knowledge on smoothness of the unknown solution x is needed in order to choose the regularization parameter α optimally.

An inverse estimator of a linear functional $l_f(x) = \langle f, x\rangle$, $f \in W_\nu(N)$, which automatically adapts to unknown smoothness of x has been developed by G&P (2000). The adaptation procedure there is a particular implementation of the Lepski (1991) general adaptation scheme for the case where the class $W_\nu(N)$ is completely specified. Note that this construction depends crucially on the actual smoothness ν of f because the 'typical value' of the stochastic term in the error decomposition depends on ν. In the next section we introduce our estimator which is adaptive in the sense of (7) over a wide collection of representer sets $W_\nu(N)$ and solution sets $W_\mu(M)$.

3 Adaptive estimator

In this section we define our adaptive estimator and study its properties both for the white noise and density observation models.

First, introduce assumptions on the collection of possible representer sets $W_\nu(N)$. We assume that $f \in W_\nu(N)$, where ν is unknown and belongs to the discrete set

$$\Delta_\nu := \{\nu_0, \dots, \nu_m\}, \quad \underline{\nu} = \nu_0 < \nu_1 < \cdots < \nu_m = \overline{\nu} < a.$$

Let $\delta := \min\{\nu_i - \nu_{i-1} : i = 1, \dots, m+1\}$, where $\nu_{m+1} := a$ by definition.

The basic idea underlying construction of our adaptive estimator is the following. We consider a discrete ordered set Δ_α of possible regularization parameters, and a family of estimates $\hat{l}^{\varepsilon}_{\alpha,s}(x)$ associated with $\alpha \in \Delta_\alpha$. For every fixed smoothness index ν_j from Δ_ν we can choose adaptively the regularization parameter from Δ_α using the Lepski adaptation procedure. In this way we obtain a family of $m+1$ estimates corresponding to different thresholds in the adaptation scheme. If a parameter $\nu_j \in \Delta_\nu$ is greater than the actual smoothness index ν, the threshold in the adaptation scheme is small, and on a set of 'large' probability the adaptation procedure yields a 'too small' value for the regularization parameter. It turns out that this can be detected very precisely from the data using a special construction of the set Δ_α. Our adaptive estimator is defined in two steps. First, using the Lepski adaptation scheme we obtain a sequence of regularization parameters corresponding to different smoothness indices $\nu_j \in \Delta_\nu$. Second, we select among these the minimal regularization parameter which is not 'too small'. Note that our construction is the same for both white noise and density observation models. In the last case ε is equal to $n^{-1/2}$.

Fix $\overline{\alpha} = 1$ and let

$$\underline{\alpha} = \varepsilon^p, \quad \text{where } p = \frac{4(a+s)}{\delta}\left(\frac{a-\underline{\nu}}{a-\overline{\nu}}\right). \tag{16}$$

For $q > 1$ define

$$\Delta_\alpha = \{\alpha \in [\underline{\alpha}, \overline{\alpha}] : \alpha = \alpha_j = q^j \underline{\alpha},\ j = 0, 1, \ldots\}.$$

Let $r_\nu(\gamma) = \gamma^{-(a-\nu)/(2(a+s))}$, and, for brevity, write $\hat{l}_\gamma$ for $\hat{l}^\varepsilon_{\gamma,s}(x)$. For a given $\varkappa \geq 1$, let $\hat{\alpha}_j$ denote the maximal α from Δ_α such that

$$|\hat{l}_\gamma - \hat{l}_\eta| \leq 2\varkappa\varepsilon[r_{\nu_j}(\gamma) + r_{\nu_j}(\eta)],\ \ \forall \gamma, \eta \leq \alpha,\ \gamma, \eta \in \Delta_\alpha. \tag{17}$$

In other words, $\hat{\alpha}_j$ denotes the regularization parameter chosen by the Lepski procedure with the threshold associated with the smoothness index $\nu_j \in \Delta_\nu$. The basic property of the sequence $\hat{\alpha}_j$, $j = 0, 1, \ldots, m$ is that it is monotone non-increasing:

$$\hat{\alpha}_0 \geq \hat{\alpha}_1 \geq \cdots \geq \hat{\alpha}_m. \tag{18}$$

Indeed, the threshold on the right hand of (17) decreases monotonically as ν_j grows. Therefore the set of estimates satisfying the inequality becomes smaller as ν_j decreases.

Let $\tau = \varepsilon^{2(a+s)/(a-\overline{\nu})}$, and

$$j_+ := \max\{j : j \in \mathcal{J}_\tau\}, \quad \mathcal{J}_\tau := \{j : \hat{\alpha}_j \geq \tau,\ \ j = 0, \ldots, m\}. \tag{19}$$

Define $\hat{\alpha}_+ = \hat{\alpha}_{j_+}$ if the set $\mathcal{J}_\tau$ is not empty; otherwise we set $\hat{\alpha}_+ = \tau$. The estimate we are interested in is defined as

$$\hat{l}_+(x) = \hat{l}^\varepsilon_{\hat{\alpha}_+,s}(x).$$

We stress that the parameters μ and M of the solution set $W_\mu(M)$, and the parameter ν of the representer set $W_\nu(N)$ are not involved in the construction of the estimator $\hat{l}_+(x)$. Note that $\hat{l}_+(x)$ depends on the three design parameters $\varkappa$, s and q; in what follows $\varkappa$ will be chosen as a function of ε, s, and q, and other known parameters of the problem.

Below we establish our main results on accuracy of the estimator $\hat{l}_+(x)$ for the white noise and density observation models.

3.1 White noise model

Here we assume the model is given by (3). Under this assumption, the stochastic error $v_{\alpha,s}(f, \xi)$ defined in (11) is a Gaussian random variable with zero mean and variance

$$\mathbb{E}v^2_{\alpha,s}(f, \xi) = \|AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\|^2.$$

In view of Lemma 2.2 it is reasonable to consider the so-called *effective smoothness* of the representer f. We will say that f has the effective smoothness ν if

(F) $f \in \mathbb{X}^\nu$, and there exists a constant $c_* = c_*(\nu, a, s)$ such that for all sufficiently small α

$$\|AL^{-s}g_\alpha\left(L^{-s}A^*AL^{-s}\right)L^{-s}f\| \geq c_*\alpha^{\frac{\nu-a}{2(a+s)}}. \tag{20}$$

The condition (20) is rather natural. It simply means that for considered functionals the order of the variance indicated in (13) cannot be improved in the power scale. Using the same argument as in Neubauer (1997) one can show that from (13) it follows only that $f \in \mathbb{X}^r$ for all $r < \nu$. Therefore, (20) allows to specify smoothness of f through the properties of a fixed estimator determined by concrete g_α. For this reason we treat ν in (F) as the effective smoothness of the representer f.

The next theorem establishes an upper bound on the risk of the estimate $\hat{l}_+(x) = \hat{l}^{\varepsilon}_{\hat{\alpha}_+,s}(x)$.

Theorem 3.1 *Let Assumptions (A) and (F) hold, and (9) be fulfilled with some $\nu \in \Delta_\nu$ and $N \le \bar{N}$, where $\bar{N}$ is known. Let ε be small enough such that for some constant $C_1 = C_1(\nu, a, s, d, D, q)$*

$$\varepsilon\sqrt{\ln \varepsilon^{-1}} \le C_1 \min\{M, M^{-\frac{a-\overline{\nu}}{\mu+\overline{\nu}}}\}. \tag{21}$$

Assume also that

$$\mu \in (-\nu, 2s + a]. \tag{22}$$

Then there exists a constant $C_2 = C_2(a, s, \underline{\nu}, \overline{\nu}, d, D, \bar{N})$ such that for the estimate $\hat{l}_+(x)$ associated with $\varkappa = C_2\sqrt{\ln \varepsilon^{-1}}$ and $s \ge \max\{0, -\nu\}$ one has

$$\mathcal{R}[\hat{l}_+; W_\mu(M)] \le C_3 \Big[M^{\frac{a-\nu}{\mu+a}} (\varepsilon^2 \ln \varepsilon^{-1})^{\frac{\mu+\nu}{\mu+a}} + m\varepsilon^2 (\ln \varepsilon^{-1})^{3/2} \Big], \tag{23}$$

where $C_3 = C_3(a, s, q, \nu, N, d, D, c_*)$.

Proof is given in Section 5.

Note that C_2 depends only on known parameters of the problem so that the choice of $\varkappa$ can be implemented. If we knew in advance parameters μ, M of the solution set $W_\mu(M)$ and the smoothness index ν of the representer f we could achieve the minimax rate of convergence given in (15). Thus accuracy of our adaptive estimator coincides within a logarithmic in ε^{-1} factor with the minimax rate of convergence. We stress, however, that the upper bound (23) holds simultaneously over a wide collections $\mathcal{W}$ and $\mathcal{F}$ of solution sets $W_\mu(M)$ and representer sets $W_\nu(N)$. Both $\mathcal{W}$ and $\mathcal{F}$ are defined by the conditions of the theorem. In particular, $\mathcal{W}$ is the family of balls $W_\mu(M)$ with parameters μ and M satisfying (21) and (22), while $\mathcal{F}$ is the family of sets $W_\nu(N)$ with $\nu \in \Delta_\nu$ and $N \le \bar{N}$ satisfying (22), (20), and $s \ge \max\{0, -\nu\}$. One can argue also that in many important cases the estimate $\hat{l}_+(x)$ has the best possible adaptive properties; for discussion of this issue we refer to G&P (2000). In Theorem 3.1 we assumed that the discrete set Δ_ν is such that δ is fixed and positive. It can be seen from the proof that this assumption may be relaxed. In particular, one can assume that δ tends slowly to zero as $\varepsilon \to 0$. In this case the only important requirement is that the true smoothness index ν belongs to the set Δ_ν for every ε. The statement of Theorem 3.1 remains valid under these circumstances.

3.2 Density model

In this subsection we establish an upper bound on estimation accuracy of the estimator $\hat{l}_+(x)$ for the density observation model. Here the noisy data are represented by y_ε, a random element of $\mathbb{Y}$ defined in (4). Consider i.i.d. $\mathbb{Y}$-valued random elements $\theta_i = \hat{y}_i - y$, $i = 1, \dots, n$, and let P_θ denote the probability distribution of θ (we write θ for a generic random element in $\mathbb{Y}$ with the same distribution as θ_i). The following conditions on P_θ will be used in the sequel.

(P1) P_θ is a Radon probability measure on $\mathbb{Y}$ which has the strong second order, i.e. $\mathbb{E}||\theta||^2 = \int ||\theta||^2 P_\theta(d\theta) < \infty$. In addition, the mean value of the probability measure P_θ is equal to 0, i.e.

$$\mathbb{E}\langle\theta, \phi\rangle = \int \langle\theta, \phi\rangle P_\theta(d\theta) = 0, \quad \forall\phi \in \mathbb{Y}. \tag{24}$$

The covariance operator $K_\theta : \mathbb{Y} \to \mathbb{Y}$ of P_θ is defined by the relation

$$\langle K_\theta\phi_1, \phi_2\rangle = \int \langle\theta, \phi_1\rangle\langle\theta, \phi_2\rangle P_\theta(d\theta), \quad \phi_1, \phi_2 \in \mathbb{Y}.$$

It is well known that K_θ is a self–adjoint positive operator with the finite trace $\mathrm{tr}(K_\theta) = \int ||\theta||^2 P_\theta(d\theta) < \infty$ [e.g., Vakhania et. al (1986, p. 177)]. We require a slightly stronger condition on K_θ.

(P2) The covariance operator K_θ is strictly positive and

$$0 < \underline{\lambda} = \inf_{\phi\in\mathbb{Y}} \frac{\langle K_\theta\phi, \phi\rangle}{\langle\phi, \phi\rangle} \le \sup_{\phi\in\mathbb{Y}} \frac{\langle K_\theta\phi, \phi\rangle}{\langle\phi, \phi\rangle} = \overline{\lambda} < \infty \tag{25}$$

for some absolute constants $\underline{\lambda}$ and $\overline{\lambda}$.

The next assumptions require existence of the exponential moment and some regularity of P_θ near zero.

(P3) There exist positive constants b_1 and H_1 such that

$$\mathbb{E}e^{t||\theta||} = \int e^{t||\theta||} P_\theta(d\theta) \le b_1 < \infty, \quad \text{for } |t| \le H_1.$$

(P4) There exist positive constants b_2 and H_2 such that

$$\mathbb{P}\{||\theta|| \le t\} \le b_2 t, \qquad 0 \le t \le H_2. \tag{26}$$

We note that (P4) is only one of possible conditions on regularity of the distribution of $||\theta||$ near zero. In fact, the distribution of $||\theta||$ should not have an atom at zero; for example, any polynomial decrease of $\mathbb{P}\{||\theta|| \le t\}$ as $t \to 0$ will be appropriate. We have intentionally considered (26) in order to avoid additional technicalities and to give a unified proof of main results for both the white noise and density observation models.

We formulate an analog of Theorem 3.1 for the density observation model under assumptions (P1)-(P4). The constants appearing in the below theorem, unless it is said explicitly, may depend on a, s, d, D, q, c_* and on all parameters involved in Assumptions (P1)-(P4).

Theorem 3.2 *Let Assumptions (A), (F) and (P1)-(P4) hold. Assume that (9) is fulfilled with some $\nu \in \Delta_\nu$ and $N \le \bar{N}$, where $\bar{N}$ is known. Let n be large enough such that for some constant C_1*

$$\sqrt{\frac{\ln n}{n}} \ge C_1 \max\{1, M, M^{-\frac{a-\overline{\nu}}{\mu+\overline{\nu}}}\},$$

Assume also that (22) is satisfied. Then there exists a constant C_2 depending on a, s, $\underline{\nu}$, $\overline{\nu}$, $\bar{N}$, d, D, and $\overline{\lambda}$ such that for the estimate $\hat{l}_+(x)$ associated with $\varkappa = C_2\sqrt{\ln n}$ and $s \ge \max\{0, -\nu\}$ one has

$$\mathcal{R}[\hat{l}_+; W_\mu(M)] \le C_3 \Big[M^{\frac{a-\nu}{\mu+a}} \Big(\frac{\ln n}{n}\Big)^{\frac{\mu+\nu}{\mu+a}} + \frac{m}{n}(\ln n)^{3/2} \Big].$$

Thus the theorem shows that, similarly to the white noise model, the proposed estimator adapts automatically to unknown smoothness of x and f. Observe also that the constant C_2 is defined in terms of known parameters of the problem only. Therefore the choice of $\varkappa$ can be implemented.

Remark Inspection of the proofs of Section 5 reveals that Theorems 3.1 and 3.2 also hold under relaxed Assumption (F). For example, the right hand side of (20) may have the order $\alpha^{(\nu-a+\delta/2)/(2(a+s))}$. In this case the estimator should be modified: one should put in (16) $2p$ instead of p.

4 Deconvolving bivariate densities with singular support

To illustrate general results of Section 3 we consider the problem of estimating a bivariate density with singular support from indirect observations on the plane. The context is that of Example 1.

Let $\Psi_z(u) = \mathbb{E}\exp\{\imath z'u\}$ and $\Psi_w(u) = \mathbb{E}\exp\{\imath w'u\}$ be the characteristic functions of random variables $z = \rho(\varphi)\exp\{\imath\varphi\}$ and w. Here $z, w \in \mathbb{R}^2 = \mathbb{C}$, and $'$ denotes transpose. For $u = t\exp\{i\phi\}$ write $\Psi_z(u) = \Psi_z(t,\phi)$ and $\Psi_w(u) = \Psi_w(t,\phi)$. For any $u \in \mathbb{R}^2$ we have

$$\begin{aligned}\tilde{y}(u) &= \tilde{y}(t,\phi) := \Psi_w^{-1}(t,\phi)\Psi_z(t,\phi)\\ &= \mathbb{E}\exp\{\imath\rho(\varphi)t\cos(\varphi-\phi)\} = \int_0^{2\pi} \exp\{\imath\rho(\varphi)r\cos(\varphi-\phi)\}x(\varphi)d\varphi\ .\end{aligned}$$

Integrating the both sides of the last equality over $\phi \in [0, 2\pi]$ we obtain

$$y(t) := \frac{1}{2\pi} \int_0^{2\pi} \tilde{y}(t, \phi) d\phi = \int_0^{2\pi} J_0(t\rho(\varphi)) x(\varphi) d\varphi \ , \tag{27}$$

where $J_0(\cdot)$ is the Bessel function of the zero order. If $\rho(\varphi) \neq \text{const}$, the integral on the right hand side of (27) can be considered as an integral operator acting from $\mathbb{L}_2(0, 2\pi)$ to $\mathbb{L}_2(0, \varrho)$ for some $\varrho > 0$.

The function $y(\cdot)$ on the left hand side of (27) can be estimated from the observations Y_j, $j = 1, \ldots, n$, given by (5). By definition

$$\begin{aligned} \tilde{y}(u) &= \tilde{y}(t, \phi) = \mathbb{E}\Big[\exp\{\sigma^2 |u|^2/2\} \exp\{\imath Y_j' u\}\Big] \\ &= \exp\{\sigma^2 t^2/2\} \mathbb{E}\Big[\exp\{\imath t |Y_j| \cos(\arg(Y_j) - \phi)\}\Big], \end{aligned}$$

and therefore, $y(t) = \exp\{\sigma^2 t^2/2\} \mathbb{E}[J_0(t|Y_j|)]$. Now setting for $j = 1, \ldots, n$

$$\hat{y}_j(t) = \exp\{\sigma^2 t^2/2\} J_0(t|Y_j|), \qquad t \in [0, \varrho]$$

we have $\mathbb{E}\hat{y}_j = y$ and $\mathbb{E}\|\hat{y}_j\|^2 \leq \varrho \exp\{\sigma^2 \varrho^2\} < \infty$, where $\|\cdot\|$ denotes the norm in $\mathbb{L}_2(0, \varrho)$. In addition, for $y_\varepsilon = n^{-1} \sum_{j=1}^n \hat{y}_j$

$$\mathbb{E}\|y_\varepsilon - y\|^2 \leq 4\varrho \exp\{\sigma^2 \varrho^2\} n^{-1}$$

because $|J_0(t)| \leq 1$, $\forall t$. Thus we are in the framework of the density observation model, and our goal is to apply general results of Section 3 to this particular estimation problem.

First, we verify Assumptions (P1)-(P4) for our problem. Assumption (P1) is trivially fulfilled, and Assumption (P3) holds because

$$\begin{aligned} \|\theta_j\|^2 &= \|y_j - y\|^2 \\ &= \int_0^{\varrho} \exp\{\sigma^2 t^2\} \Big[J_0(t|Y_j|) - \mathbb{E}J_0(t|Y_j|)\Big]^2 dt \\ &\leq 4\varrho \exp\{\sigma^2 \varrho^2\} < \infty. \end{aligned} \tag{28}$$

In order to check Assumption (P4) we observe that the integrand in (28) is a bounded continuous and positive function of t. Therefore, it is sufficient to verify the condition (26) for the random variable $|J_0(k_*|Y|)|$, where $k_* \in [0, \rho]$ is a constant. Because $|J_0(k_*|Y|)|$ is a smooth function of $|Y|$ with uniformly bounded first derivative, Assumption (P4) will follow if the distribution of $|Y|$ has the property (26). But this is an immediate consequence of the fact that Y has a bounded infinitely differentiable density function on the plane. As for Assumption (P2), the upper bound in (25) follows immediately from boundedness of the random variable $\|\theta\|$. The lower bound, however, should be checked for each concrete contour.

Clearly, one cannot expect that for a given contour the operator A from (6) will satisfy (8) with some standard Hilbert scale as, e.g., the Sobolev one. Therefore it is reasonable to embed the problem into the natural Hilbert scale generated by operator $L = (A^*A)^{-1}$. Of course, smoothness of the representer f of the functional $l_f(x) = x(\varphi_0)$, that we are interested in, and smoothness of the solution x relative to such a scale are generally unknown. Note, however, that (8) is satisfied automatically with $a = 1/2$, $d = D = 1$. Thus we can apply our adaptive estimator from Section 3.

Observe that for infinitely differentiable $\rho(\cdot)$ the natural Hilbert scale corresponding to the operator A in (27) defines sets of infinitely differentiable functions. For such a scale the representer f of the linear functional $x(\varphi_0)$ belongs to a ball $W_\nu(N)$ with negative ν that is close to zero. This follows from the fact that even for small positive μ the ball $W_\mu(M)$ contains continuous functionals. In this case it is reasonable to consider the regularized estimator (10) associated with $s = 0$. The reasonable choice of g_α is $g_\alpha(\lambda) = (\lambda + \alpha)^{-1}$ that corresponds to Tikhonov-Phillips regularization method. With such a choice the regularized estimator $\hat{l}_\alpha$ of $l_f(x) = x(\varphi_0)$ is defined as $\hat{l}_\alpha(x) = x_\alpha(\varphi_0)$, where $x_\alpha(\varphi)$ is the solution of the Fredholm integral equation of the second kind

$$\alpha x_\alpha(\varphi) + \int_0^{2\pi} a(\varphi, \psi) x_\alpha(\psi) d\psi = g_n(\varphi), \quad \varphi \in [0, 2\pi],$$

where

$$a(\varphi, \psi) = \int_0^\rho J_0(t\rho(\varphi)) J_0(t\rho(\psi)) dt, \quad g_n(\varphi) = \frac{1}{n} \sum_{j=1}^n \int_0^\rho J_0(t\rho(\varphi))\, \hat{y}_j(t) dt.$$

Then the next statement is an immediate consequence of Theorem 3.2.

Theorem 4.1 *Let the lower bound of (25) hold true for the covariance operator K_θ of $\theta_i = \hat{y}_i - y$, $i = 1, 2, ..., n$. Assume that the effective smoothness ν of the representer of functional $l_f(x) = x(\varphi_0)$ relative to the natural Hilbert scale belongs to the set Δ_ν, $\overline{\nu} = 0$, and $\mu \in (-\nu, 1/2]$. Then for sufficiently large n there exists a constant C_1 depending on $\underline{\nu}$, φ_0 and $\overline{\lambda}$ such that for $\hat{\alpha}_+$ defined by (17), (19) with $\varkappa = C_1\sqrt{\ln n}$ one has*

$$\mathbb{E}|x(\varphi_0) - x_{\hat{\alpha}_+}(\varphi_0)|^2 \le C_2 \left[\|x\|_\mu^{\frac{1-2\nu}{2\mu+1}} \left(\frac{\ln n}{n} \right)^{\frac{2(\mu+\nu)}{2\mu+1}} + \frac{m}{n} (\ln n)^{3/2} \right].$$

5 Proofs

In this section we prove the main results of the paper, Theorems 3.1 and 3.2. The following notation is used in the proofs both for the white noise and the density observation models.

Denote $B_\alpha(x) = c_1 \|x\|_\mu \|f\|_\nu \alpha^{(\mu+\nu)/(2(a+s))}$, and c_1 is the constant appearing on the right hand side of (12). Define

$$\alpha_* = \max\{\alpha \in \Delta_\alpha : B_\alpha(x) \le \varkappa \varepsilon r_\nu(\alpha)\}.$$

In fact, α_* is the ideal regularization parameter that balances the squared bias and variance. Consider the event

$$\Omega_\varkappa := \{\omega \in \Omega : \max_{\alpha \in \Delta_\alpha} [r_\nu^{-1}(\alpha) |v_{\alpha,s}(f,\xi)|] \le \varkappa\}. \tag{29}$$

The event $\Omega_\varkappa$ corresponds to the 'typical' behavior of the stochastic term $v_{\alpha,s}(f,\xi)$. Also, for notational convenience, we denote $t := \min\{j \in \{0,\dots,m\} : \nu_j > \nu\}$, i.e. $\nu = \nu_{t-1}$ for some $t \in \{1,\dots,m\}$.

In the proofs below $c_1, c_2, \dots$ and $k_1, k_2, \dots$ stand for constants depending on parameters of the problem. They may be different on different occasions.

5.1 White noise model

The goal of this subsection is to prove Theorem 3.1. First, we establish some auxiliary lemmas.

The next statement shows that conditionally on $\Omega_\varkappa$ the regularization parameter $\hat\alpha_t$ given by the adaptive scheme is typically 'small'. Recall that $\hat\alpha_t$ corresponds to the threshold with $\nu_t > \nu = \nu_{t-1}$; here ν is the effective smoothness of f.

Lemma 5.1 *Let Assumption (F) hold and $\alpha_* \ge \underline{\alpha}^{(a-\nu_t)/(a-\nu)}$. Then for every $\eta \in \Delta_\alpha$ satisfying*

$$\eta \ge \underline{\alpha}^{\frac{a-\nu_t}{a-\nu}} \tag{30}$$

one has

$$\mathbb{P}\{\hat\alpha_t \ge \eta \,|\, \Omega_\varkappa\} \le k_1 \varkappa \underline{\alpha}^{\frac{\nu_t-\nu}{2(a+s)}}, \tag{31}$$

where $k_1 = k_1(c_*)$ depends only on c_* from (20).

Proof We prove the lemma considering the cases $\eta \le \alpha_*$ and $\eta > \alpha_*$ separately.

1^0. Assume that $\eta \le \alpha_*$. We have

$$\begin{aligned} \mathbb{P}\{\hat\alpha_t \ge \eta \,|\, \Omega_\varkappa\} &\le \mathbb{P}\{|\hat l_\eta - \hat l_{\underline{\alpha}}| \le 2\varkappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\varkappa\} \\ &= 1 - \mathbb{P}\{|\hat l_\eta - \hat l_{\underline{\alpha}}| > 2\varkappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\varkappa\}. \end{aligned} \tag{32}$$

On the set $\Omega_\varkappa$

$$\begin{aligned} |\hat l_\eta - \hat l_{\underline{\alpha}}| &\ge |\hat l_{\underline{\alpha}} - l| - |\hat l_\eta - l| \\ &\ge \varepsilon |v_{\underline{\alpha},s}(f,\xi)| - B_{\underline{\alpha}}(x) - 2\varkappa\varepsilon r_\nu(\eta) \\ &\ge \varepsilon |v_{\underline{\alpha},s}(f,\xi)| - 3\varkappa\varepsilon r_\nu(\eta). \end{aligned}$$

Here we have used the fact that $|\hat l_\eta - l| \le 2\varkappa\varepsilon r_\nu(\eta)$ on the set $\Omega_\varkappa$, and $B_{\underline{\alpha}}(x) \le B_\eta(x) \le \varkappa\varepsilon r_\nu(\eta)$ because $\eta \le \alpha_*$. Thus we have

$$\begin{gathered} \mathbb{P}\{|\hat l_\eta - \hat l_{\underline{\alpha}}| > 2\varkappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\varkappa\} \ge \\ \mathbb{P}\{\varepsilon |v_{\underline{\alpha},s}(f,\xi)| > \varkappa\varepsilon[2r_{\nu_t}(\underline{\alpha}) + 2r_{\nu_t}(\eta) + 3r_\nu(\eta)] \,|\, \Omega_\varkappa\} \ge \\ \mathbb{P}\{\varepsilon |v_{\underline{\alpha},s}(f,\xi)| > \varkappa\varepsilon[4r_{\nu_t}(\underline{\alpha}) + 3r_\nu(\eta)] \,|\, \Omega_\varkappa\}. \end{gathered} \tag{33}$$

By (30), $r_{\nu_t}(\underline{\alpha})$ dominates $r_\nu(\eta)$: $r_\nu(\eta) \le r_{\nu_t}(\underline{\alpha})$. Therefore the last probability in (33) is bounded from below by

$$\begin{aligned}
\mathbb{P}\{|v_{\underline{\alpha},s}(f,\xi)| > 7\varkappa r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\varkappa\} &= 1 - \mathbb{P}\{|v_{\underline{\alpha},s}(f,\xi)| \le 7\varkappa r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\varkappa\} \\
&= 1 - \mathbb{P}\{|\mathcal{N}(0,1)| \le 7\varkappa [\mathbb{E}v^2_{\underline{\alpha},s}(f,\xi)]^{-1/2} r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\varkappa\} \\
&\ge 1 - \frac{14\varkappa}{\sqrt{2\pi}} [\mathbb{E}v^2_{\underline{\alpha},s}(f,\xi)]^{-1/2} r_{\nu_t}(\underline{\alpha}) \\
&\ge 1 - k_1 \varkappa \underline{\alpha}^{\frac{\nu_t-\nu}{2(a+s)}} \qquad (34)
\end{aligned}$$

where $\mathcal{N}(0,1)$ denotes the standard Gaussian random variable, and by (20) the constant k_1 depends on c_* only. Combining (34) and (33) with (32) we obtain (31) under assumption that $\eta \le \alpha_*$.

2^0. Now consider the case where $\eta > \alpha_*$. Here, by definition of $\hat{\alpha}_t$,

$$\mathbb{P}\{\hat{\alpha}_t \ge \eta \,|\, \Omega_\varkappa\} \le \mathbb{P}\{|\hat{l}_{\alpha_*} - \hat{l}_{\underline{\alpha}}| \le 2\varkappa\varepsilon[r_{\nu_t}(\alpha_*) + r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\varkappa\}.$$

The further proof goes along the same lines as in the previous case with η replaced by α_*. ■

An immediate consequence of (18) and Lemma 5.1 is that the same bound holds for all estimates $\hat{\alpha}_j$ associated with $\nu_j > \nu$; namely, under conditions (30) we have

$$\mathbb{P}\{\hat{\alpha}_j \ge \eta \,|\, \Omega_\varkappa\} \le k_1 \varkappa \underline{\alpha}^{\frac{\nu_t-\nu}{2(a+s)}}, \quad \forall j \ge t = \min\{j : \nu_j > \nu\}. \qquad (35)$$

Thus Lemma 5.1 shows that if we misspecify ν in the threshold of the procedure (17) by choosing the value greater than ν, the scheme will yield a regularization parameter which with 'large' probability less than $\underline{\alpha}^{(\nu_t-\nu)/(a-\nu)}$.

The above considerations motivated our rule (19). We show that under some natural conditions on τ, the quantity j_+ determined there detects the 'right' value of ν with 'large' probability conditionally on $\Omega_\varkappa$.

Lemma 5.2 *Let* $\underline{\alpha}^{(a-\nu_t)/(a-\nu)} \le \tau \le \alpha_*$ *and (20) hold. If the event* $\Omega_\varkappa$ *holds then* $j_+ \ge t-1$. *In addition,*

$$\mathbb{P}(j_+ = t-1 \,|\, \Omega_\varkappa) \ge 1 - k_1 \varkappa \underline{\alpha}^{\frac{\nu_t-\nu}{2(a+s)}}, \qquad (36)$$

where k_1 *is defined in Lemma 5.1.*

Proof The first statement follows immediately from (18) and the standard properties of the Lepski adaptation scheme. Indeed, if we put the 'right' value of $\nu = \nu_{t-1}$ in the threshold on the right hand side of (17), then on the set $\Omega_\varkappa$ the resulting regularization parameter $\hat{\alpha}_{t-1}$ will be greater than α_* by construction [see, e.g., G&P (2000)]. In view of $\tau \le \alpha_*$ and monotonicity of $\{\hat{\alpha}_j\}$, on the set $\Omega_\varkappa$, $\mathcal{J}_\tau \supseteq \{0,1,\ldots,t-1\}$. Hence $j_+ \ge t-1$ as claimed.

To prove (36) we note that the event $\{j_+ \geq t\}$ means that there exists an estimate $\hat{\alpha}_j$ associated with $\nu_j > \nu$ such that $\hat{\alpha}_j \geq \tau$. But by Lemma 5.1 [see also (35)] for $j \geq t$

$$\mathbb{P}\{\hat{\alpha}_j \geq \tau \,|\, \Omega_\varkappa\} \leq k_1 \varkappa \underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}} .$$

This completes the proof. ∎

Proof of Theorem 3.1

Write

$$\mathbb{E}|\hat{l}_+ - l|^2 = I_1 + I_2 := \mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\Omega_\varkappa)] + \mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\bar{\Omega}_\varkappa)],$$

where $l = l_f(x) = \langle f, x \rangle$. We bound I_1 and I_2 separately.

1^0. It follows immediately from (12) and (9) that

$$\varkappa \varepsilon r_\nu(q\alpha_*) < B_{q\alpha_*}(x) \leq c_1 M \|f\|_\nu (q\alpha_*)^{\frac{\mu+\nu}{2(a+s)}}$$

and

$$\alpha_* \geq q^{-1}[(c_1 M \|f\|_\nu)^{-1} \varkappa \varepsilon]^{\frac{2(a+s)}{\mu+a}} \geq [k_2 M^{-1} \varkappa \varepsilon]^{\frac{2(a+s)}{\mu+a}} \tag{37}$$

for some constant $k_2 = k_2(a, s, \nu, d, N, D, q)$. Note that (21) ensures that $\alpha_* \in [\tau, \overline{\alpha}]$ for small enough ε. Further, for our choice of $\tau = \varepsilon^{2(a+s)/(a-\overline{\nu})}$ and $\underline{\alpha}$ given by (16) we have $\tau \geq \underline{\alpha}^{(a-\nu_t)/(a-\nu)}$. In order to show this it is sufficient to verify that

$$p\Big(\frac{a - \nu_t}{a - \nu}\Big) \geq \frac{2(a+s)}{a - \overline{\nu}} .$$

This follows because

$$\frac{2}{\delta}(a - \underline{\nu})\Big(\frac{a - \nu_t}{a - \nu}\Big) \geq \frac{2}{\delta}(a - \nu_t) \geq 2 .$$

Thus, Lemma 5.2 is applicable with our choice of τ and $\underline{\alpha}$.

2^0. First we bound I_1. Let the event $\Omega_\varkappa$ hold; then, by Lemma 5.2, $j_+ \geq t-1$, i.e. $\Omega_\varkappa \subseteq \{j_+ \geq t-1\}$. Consider the events $B_j = \{j_+ = j\}$, $j = t-1, t, \ldots, m$. Lemma 5.2 implies

$$\begin{aligned} \mathbb{P}(B_{t-1} \,|\, \Omega_\varkappa) &\geq 1 - k_1 \varkappa \underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}} \quad \text{and} \\ \mathbb{P}(B_j \,|\, \Omega_\varkappa) &\leq k_1 \varkappa \underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}} \quad \forall j = t, \ldots, m, \end{aligned} \tag{38}$$

with $k_1 = k_1(c_*)$ [see Lemma 5.1]. Write

$$I_1 = \sum_{j=t-1}^{m} \mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\Omega_\varkappa \cap B_j)].$$

On the set $\Omega_\varkappa \cap B_{t-1}$ we have $\hat{\alpha}_+ = \hat{\alpha}_{t-1}$, and the adaptive procedure runs with the 'right' value of $\nu = \nu_{t-1}$. In this case $\hat{\alpha}_+ = \alpha_{t-1} \geq \alpha_*$. The standard calculations [see G&P (2000, p. 178)] yield $|\hat{l}_{\hat{\alpha}_+} - l| \leq 6\varkappa \varepsilon r_\nu(\alpha_*)$ showing that

$$\mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\Omega_\varkappa \cap B_{t-1})] \leq [6\varkappa \varepsilon r_\nu(\alpha_*)]^2 .$$

Now assume that the event $\Omega_\varkappa \cap B_j$ holds for $j \geq t$. This means that the algorithm chooses $\hat{\alpha}_+ = \hat{\alpha}_j \geq \tau$ corresponding to some $\nu_j > \nu$. In this case

$$
\begin{aligned}
|\hat{l}_{\hat{\alpha}_+} - l| &= |\hat{l}_{\hat{\alpha}_j} - l| \leq |\hat{l}_{\hat{\alpha}_j} - \hat{l}_\tau| + |\hat{l}_\tau - l| \\
&\leq 4\varkappa\varepsilon r_{\nu_j}(\tau) + B_\tau(x) + \varepsilon |v_{\tau,s}(f,\xi)| \\
&\leq 4\varkappa\varepsilon r_{\nu_j}(\tau) + 2\varkappa\varepsilon r_\nu(\tau) \leq 6\varkappa\varepsilon r_\nu(\tau). \qquad (39)
\end{aligned}
$$

Here we have taken into account that:

(i) $\hat{\alpha}_+ = \hat{\alpha}_j \geq \tau$ by construction. Hence the distance between $\hat{l}_{\hat{\alpha}_+}$ and $\hat{l}_\tau$ can be bounded in terms of the threshold corresponding to $\nu = \nu_j$;

(ii) $\tau \leq \alpha_*$ by the premise of the theorem. Therefore on the set $\Omega_\varkappa \cap B_j$ the typical value of the stochastic error dominates the bias;

(iii) $r_\nu(\tau)$ decreases when ν grows and τ is fixed, i.e. $r_{\nu_j}(\tau) \leq r_\nu(\tau)$.

It follows from (38) and (39) that for any $j \geq t$

$$
\begin{aligned}
\mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\Omega_\varkappa \cap B_j)] &\leq [6\varkappa\varepsilon r_\nu(\tau)]^2 \mathbb{P}(\Omega_\varkappa \cap B_j) \\
&\leq [6\varkappa\varepsilon r_\nu(\tau)]^2 k_1 \varkappa \underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}}.
\end{aligned}
$$

Thus we have the following bound on I_1:

$$
I_1 = \mathbb{E}[|\hat{l}_+ - l|^2 \mathbf{1}(\Omega_\varkappa)] \leq k_1 m \varkappa \underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}} [6\varkappa\varepsilon r_\nu(\tau)]^2 + [6\varkappa\varepsilon r_\nu(\alpha_*)]^2 .
$$

Using (16) we obtain

$$
\underline{\alpha}^{\frac{\nu_t - \nu}{2(a+s)}} [r_\nu(\tau)]^2 \leq \varepsilon^{\frac{2(\nu - \underline{\nu})}{a - \nu}} \leq 1
$$

so that

$$
I_1 \leq [6\varkappa\varepsilon r_\nu(\alpha_*)]^2 + k_1 m \varkappa^3 \varepsilon^2,
$$

and substituting the expression for α_* [see (37)] we finally obtain

$$
I_1 \leq k_3 \Big[M^{\frac{a-\nu}{\mu+a}} (\varepsilon^2 \ln \varepsilon^{-1})^{\frac{\mu+\nu}{\mu+a}} + m\varepsilon^2 (\ln \varepsilon^{-1})^{3/2} \Big],
$$

where $k_3 = k_3(a, s, q, \nu, N, d, D, c_*)$.

3^0. Now we consider the case where the event $\bar{\Omega}_\varkappa$ holds. Here our algorithm will choose a value that is not less than τ. The further proof coincides with the proof given in G&P (2000) with evident modifications ($\underline{\alpha}$ in the paper should be replaced by τ). We emphasize only that in this case $\varkappa$ is chosen as $k_4\sqrt{\ln \varepsilon^{-1}}$, where k_4 depends on a, s, $\underline{\nu}$, $\overline{\nu}$, d, D and $\bar{N}$. ■

5.2 Density model

Basically, the proof of Theorem 3.2 goes along the same lines as the proof of Theorem 3.1. The main difference is that the stochastic term $v_{\alpha,s}(f,\xi)$ for the density observation model is non–Gaussian. Nevertheless, Assumptions (P1)-(P4) allow to establish similar results. Below we indicate how the arguments in the proof of Theorem 3.1 should be modified in order to prove Theorem 3.2.

Recall that the estimate $\hat{l}_{\alpha,s}(x)$ associated with the regularization parameter α is defined by (10), and

$$l_f(x) - \hat{l}_{\alpha,s}(x) = b_{\alpha,s}(f,x) + n^{-1/2} v_{\alpha,s}(f,\xi)$$

where $\xi = n^{-1/2}\sum_{i=1}^{n}\theta_i = n^{-1/2}\sum_{i=1}^{n}(\hat{y}_i - y)$.

First, we establish an analog of Lemma 2.2 for the density observations model.

Lemma 5.3 *Let Assumptions (A), (P1) and (P2) hold. Assume that $f \in W_\nu(N)$ with $\nu < a$, and $s \geq \max\{0, -\nu\}$. Then*

$$\mathbb{E} v_{\alpha,s}^2(f,\xi) \leq k_4 \alpha^{\frac{\nu-a}{a+s}}, \tag{40}$$

where $k_4 = k_4(\overline{\lambda}, \nu, a, s, N)$. In addition, if Assumption (F) holds then there exists a constant $k_5 = k_5(\underline{\lambda}, \nu, a, s, c_*)$ such that

$$\mathbb{E} v_{\alpha,s}^2(f,\xi) \geq k_5 \alpha^{\frac{\nu-a}{a+s}} \ .$$

Proof Denote for brevity $\phi_\alpha := Q_\alpha f$, where $Q_\alpha = AL^{-s} g_\alpha\left(L^{-s}A^*AL^{-s}\right)L^{-s}$. By (24) and independence of θ_i, $i = 1, \ldots, n$ we have

$$\mathbb{E} v_{\alpha,s}^2(f,\xi) = \mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle\theta_i, \phi_\alpha\rangle\right|^2 = \mathbb{E}|\langle\theta_i,\phi_\alpha\rangle|^2 = \langle K_\theta \phi_\alpha, \phi_\alpha\rangle.$$

Then Assumption (P2) implies that

$$\underline{\lambda}\|\phi_\alpha\|^2 \leq \mathbb{E} v_{\alpha,s}^2(f,\xi) \leq \overline{\lambda}\|\phi_\alpha\|^2 \ .$$

The upper bound (40) follows from the same considerations as in Lemma 2.2. The lower bound is a consequence of (20). ■

Recall that for a fixed $\varkappa \geq 1$ the event $\Omega_\varkappa$ is defined by (29). The proof of Theorem 3.1 is essentially based on the fact that the constant $\varkappa$ can be chosen in such a way that the event $\Omega_\varkappa$ is of 'large' probability. This is easily proved for the white noise model because $v_{\alpha,s}(f,\xi)$ is a Gaussian normal variable. Now we establish a similar exponential inequality for the density observation model.

Lemma 5.4 *Let Assumptions (P1)-(P3) and (F) hold. Assume that $\varkappa \leq k_6\sqrt{n}$ for some constant $k_6 = k_6(\overline{\lambda}, \nu, a, s, N, b_1, H_1)$. Then*

$$\mathbb{P}(\bar{\Omega}_\varkappa) \leq 2\, S_{\underline{\alpha}} \exp\left\{-\frac{\varkappa^2}{2k_7}\right\},$$

where $S_{\underline{\alpha}} = \mathrm{card}(\Delta_\alpha)$, and $k_7 = k_7(\overline{\lambda}, \nu, a, s, N)$.

Proof Write

$$\tilde{v}_{\alpha,s}(f,\xi) = r_\nu^{-1}(\alpha)v_{\alpha,s}(f,\xi) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle\theta_i, r_\nu^{-1}(\alpha)\phi_\alpha\rangle,$$

where ϕ_α is as in the proof of Lemma 5.3. Note that the random variables $\langle\theta_i, r_\nu^{-1}(\alpha)\phi_\alpha\rangle$ have zero mean. In view of Assumption (P3)

$$\mathbb{E}\exp\{t\langle\theta, r_\nu^{-1}(\alpha)\phi_\alpha\rangle\} \le \mathbb{E}\exp\{tr_\nu^{-1}(\alpha)\|\theta\|\,\|\phi_\alpha\|\} \le b_1 < \infty$$

for $0 \le t \le H_1 c_*^{-1}$, where c_* is the constant appearing on the right hand side of (20); c_* provides a lower bound on $\|\phi_\alpha\|$. Thus, for fixed ϕ_α, the random variable $\langle\theta, r_\nu^{-1}(\alpha)\phi_\alpha\rangle$ has the moments of all orders and the following relation holds

$$\log[\mathbb{E}\exp\{t\langle\theta, r_\nu^{-1}(\alpha)\phi_\alpha\rangle\}] = \frac{1}{2}t^2 r_\nu^{-2}(\alpha)\mathbb{E}|\langle\theta,\phi_\alpha\rangle|^2 + o(t^2), \quad \text{as } t \to 0.$$

Taking into account (40) we obtain that for any constant $k_7 \ge k_4/2$ the inequality

$$\log[\mathbb{E}\exp\{t\langle\theta, r_\nu^{-1}(\alpha)\phi_\alpha\rangle\}] \le \frac{1}{2}k_7 t^2$$

holds for sufficiently small t. In other words, there exist positive constants $k_7 = k_7(\overline{\lambda}, \nu, a, s, N)$ and $\tilde{H}_1$ such that

$$\mathbb{E}\exp\{t\langle\theta, r_\nu^{-1}(\alpha)\phi_\alpha\rangle\} \le \exp\{k_7 t^2/2\}, \quad \text{for } 0 \le t \le \tilde{H}_1.$$

Then Theorem 2.6 from Petrov (1995) implies that

$$\mathbb{P}\{\tilde{v}_{\alpha,s}(f,\xi) \ge \varkappa\} = \mathbb{P}\Big\{\sum_{i=1}^{n}\langle\theta_i, r_\nu^{-1}(\alpha)\phi_\alpha\rangle \ge \sqrt{n}\varkappa\Big\} \le \exp\Big\{-\frac{\varkappa^2}{2k_7}\Big\}$$

for $0 \le \varkappa \le k_7\tilde{H}_1\sqrt{n}$. Similarly the bound on $\mathbb{P}\{\tilde{v}_{\alpha,s}(f,\xi) \le -\varkappa\}$ is derived so that

$$\mathbb{P}\{|\tilde{v}_{\alpha,s}(f,\xi)| \ge \varkappa\} \le 2\exp\Big\{-\frac{\varkappa^2}{2k_7}\Big\}, \quad 0 \le \varkappa \le k_7\tilde{H}_1\sqrt{n}.$$

The statement of the lemma is an immediate consequence of these results. ■

Proof of Theorem 3.2

The proof goes along the same lines as the proof of Theorem 3.1. Below we indicate only differences.

First, we note that, under Assumptions (P1)-(P4) and (F), Lemma 5.1 remains valid. The difference only is that now $k_1 = k_1(c_*, b_2)$, and the bound (31) holds provided n is greater than some constant depending on $\underline{\lambda}$, $\overline{\lambda}$, a, s, H_2 only. Here in order to prove (34) we use Assumption (P4). Then Lemma 5.2 follows with evident modifications.

The proof of Theorem 3.2 on the set $\Omega_\varkappa$ coincides with the proof of Theorem 3.1. To bound the error on the complimentary event $\bar{\Omega}_\varkappa$ we use the exponential inequality of Lemma 5.4. Other details of the proof remain unchanged. ■

Acknowledgment

The work was partially written during the visits of the second author to the University of Haifa, Israel (June, 2001), and the Weierstrass-Institut, Berlin (October, 2001). The support of both these institutions is gratefully acknowledged.

References

[1] Brown L., Low, M.G. (1996) A constrained risk inequality with application to nonparametric estimation. *Ann. Statist.* **24**, 2524-2535.

[2] Cavalier, L. & Tsybakov, A. (2000) Sharp adaptation for inverse problems with random noise. *Prépublication n^o 559 du Laboratoire de Probabilités & Modèles Aléatoires, Université Paris VI.*

[3] Goldenshluger, A. (2001) Density deconvolution in the circular structural model. Accepted to *J. Multivariate Anal.*

[4] Goldenshluger, A. & Pereverzev, S.V. (2000) Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory and Related Fields* **118**, 169-186.

[5] Hegland, M. (1995) Variable Hilbert scale and their interpolation inequalities with applications to Tikhonov regularization, *Appl. Anal.* **59**, 207-223.

[6] Krein, S.G., Petunin, Ju.I. & Semenov, E.M. (1982) *Interpolation of Linear Operators.* American Mathematical Society, Providence, Rhode Island.

[7] Lepski, O. (1990) A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **36**, 454-466.

[8] Lepski, O. (1991) Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36**, 682-697.

[9] Lepski, O. (1992) Asymptotically minimax adaptive estimation II: Schemes without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37**, 433-448.

[10] Mair, B.A. & Ruymgaart, F.H. (1996) Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56**, 1424-1444.

[11] Mathe, P., Pereverzev, S.V. (2001) Optimal discretization of inverse problems in Hilbert scales. Regularization and self–regularization of projection methods. *SIAM J. Numer. Anal.* **38**, 1999-2021.

[12] Natterer, F. (1984) Error bounds for Tikhonov regularization in Hilbert scales. *Appl. Anal.* **19**, 29-37.

[13] Neubauer, A. (1988) On a posteriori parameter choice for Tikhonov regularization in Hilbert scale leading to optimal convergence rates. *SIAM J. Numer. Anal.* **25**, 1313-1326.

[14] Neubauer, A. (1997) On converse and saturation results for Tikhonov regularization of linear ill-posed problems. *SIAM J. Numer. Anal.* **34**, 517-527.

[15] Petrov, V.V. (1995) *Limit Theorems of Probability Theory.* Clarendon Press, Oxford.

[16] van Rooij, A.C.M., Ruymgaart, F.H. & van Zwet, W.R. (2000) Asymptotic efficiency of inverse estimators. *Theory of Probab. Appl.* **44**, 722-738.

[17] Ruymgaart, F.H. (1993) A unified approach to inversion problems in statistics. *Math. Methods Statist.* **2**, 130-146.

[18] Skorohod, A.V. (1974) *Integration in Hilbert space.* Berlin: Springer–Verlag.

[19] Tautenhahn, U. (1996). Error estimates for regularization methods in Hilbert scales. *SIAM J. Numer. Anal.* **33**, 2120-2130.

[20] Vakhania, N.N., Tarieladze, V.I. & Chobanyan, S.A. (1987) *Probability Distributions on Banach Spaces.* Kluwer Academic Publishers, Dordrecht.