Bernoulli **9**(5), 2003, 783–807

On adaptive inverse estimation of linear functionals in Hilbert scales

ALEXANDER GOLDENSHLUGER[1] and SERGEI V. PEREVERZEV[2]

[1]*Department of Statistics, University of Haifa, Haifa 31905, Israel.*
E-mail: goldensh@stat.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*

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 propose an estimation procedure that adapts to unknown smoothness of x, of f, and of the noise covariance operator. It is shown that accuracy of this adaptive estimator is worse only by a logarithmic factor 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.

Keywords: adaptive estimation; Hilbert scales; inverse problems; linear functionals; minimax risk; regularization

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 being 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 (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

1350–7265

and variance $\|\phi\|^2$. Denoting by $\mathbb{E}$ the expectation with respect to $\mathbb{P}$, we have in addition $\mathbb{E}[\xi(\phi)\xi(\psi)] = \langle \phi, \psi \rangle$, for all $\phi, \psi \in \mathbb{Y}$. 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 independent and identically distributed (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 us to construct i.i.d. statistics $\hat{y}_1, \ldots, \hat{y}_n$ 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

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

This observation scheme corresponds to (2) with $\xi = \sqrt{n}(y_\varepsilon - y)$ and $\varepsilon = n^{-1/2}$; here ξ is a zero-mean random element of $\mathbb{Y}$ with compact covariance operator. 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 set-up, see Ruymgaart (1993), Mair and Ruymgaart (1996) and van Rooij *et al*. (2000).

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. The problem of minimax estimation of linear functionals from noisy data is a subject with a considerable literature. Ibragimov and Khas'minskii (1984) study the model with direct white noise observations where A is the identity operator. The case of arbitrary Gaussian noise is considered in Ibragimov and Khas'minskii (1987). For models with indirect observations we refer to Donoho (1994) and references therein.

It is well known that achievable accuracy in estimating $l_f(x)$ is essentially determined by: (i) the ill-posedness of the problem (1), (2); (ii) the smoothness of the representer f; (iii) the smoothness of the unknown solution x. In addition, it is important to realize that noise properties also influence the estimation accuracy. Specifically, if ξ is a zero-mean Gaussian noise, then smoothing properties of the covariance operator are important (see Ibragimov and Khas'minskii 1987; Nussbaum and Pereverzev 1999). Typically (i), (ii) and (iii) are characterized using specific properties of the operator A involved. For example, if a singular value decomposition (SVD) of A is known and ξ is the Gaussian white noise, then (2) can be equivalently represented as a sequence space model with Gaussian errors. In this case both the ill-posedness index of A and the smoothness of f and x are naturally measured by the rate at which corresponding coefficients of the SVD representation decrease (see Cavalier and Tsybakov 2002).

In this paper we adopt a different approach. In order to quantify the effect of factors (i)–(iii) and of noise properties on estimation accuracy, we embed the problem in a Hilbert scale. This approach to statistical inverse estimation has been advocated by Mair and Ruymgaart (1996) and Mathé and Pereverzev (2001). A typical example of a Hilbert scale is the scale of Sobolev spaces containing periodic functions on [0, 1] with square-integrable derivatives. The derivative order, defined via Fourier coefficients, plays the role of the scale index. The operator A should act along the Hilbert scale. For example, if A is the a-fold integration operator then it acts boundedly from a Sobolev space with index μ to the

Sobolev space with index $\mu + a$; both spaces belong to the same scale. Thus the ill-posedness index of A can be naturally measured in the Hilbert scale framework. If the operator A does not fit some standard Hilbert scale such as the Sobolev scale, a Hilbert scale associated with A can be always constructed using the generating operator $L := (A^*A)^{-1}$ (Natterer 1984; Hegland 1995). In what follows we call the Hilbert scale generated by $(A^*A)^{-1}$ *natural*. The smoothness of x and f, the ill-posedness of (1) and (2), and the smoothing properties of the noise covariance operator are to be characterized with respect to a chosen Hilbert scale.

A clear advantage of the Hilbert scale approach is that a unified regularized estimator can be developed irrespective of the type of operator A involved. Typically, in order to choose the regularization parameter one needs to know the smoothness of x and f and the smoothing properties of the noise covariance operator measured with respect to the chosen Hilbert scale (see Mair and Ruymgaart 1996). We note that characterizing the smoothness properties of x and f and of the noise covariance operator relative to a particular Hilbert scale may be a difficult task. In particular, this is the case when the properties of A are poorly understood, and the natural Hilbert scale is used. Therefore developing a general inverse estimator that does not use prior information on the smoothness of x and f and of the noise covariance operator is of keen interest. This is the main objective of the present paper.

It is important to realize that even though f is completely known (in contrast to x and the noise covariance operator), its regularity with respect to a particular Hilbert scale may be unknown. In addition, as we shall see, our estimator also works even if f is not directly available but is given implicitly via an oracle which, for given x, reports on the corresponding value of the linear functional $l_f(x)$. This assumption is quite reasonable and is in the spirit of the information-based complexity theory (see Traub *et al*. 1988).

Our first example illustrates issues discussed in the preceding paragraphs.

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, let $z = \rho(\varphi)\exp\{\mathrm{i}\varphi\}$, 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

$$Y_j = \rho(\varphi_j)\exp\{\mathrm{i}\varphi_j\} + w_j, \qquad j = 1, \ldots, n, \tag{5}$$

where the 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 $\varphi_0 \in [0, 2\pi]$. It will be shown in Section 5 that if $\rho(\theta) \neq \text{const.}$ and $x \in \mathbb{L}_2(0, 2\pi)$, then x satisfies the integral equation

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

Here $J_0(\cdot)$ is the Bessel function of order 0, A: $\mathbb{L}_2(0, 2\pi) \to \mathbb{L}_2(0, r)$, and y is a function that can be estimated from the data at the parametric rate. Thus this formulation is in the framework of the density observation model. We can embed the problem in the natural Hilbert scale and apply a general regularized inverse estimator. We note, however, that the

smoothness indices of x and f and the smoothing properties of the covariance operator with respect to the natural Hilbert scale are unknown. Therefore this information cannot be used in the estimator construction.

Let $\hat{l}^\varepsilon = \hat{l}^\varepsilon(x)$ be an estimate of $l_f(x)$ based on the available data. The accuracy of an estimate $\hat{l}^\varepsilon(x)$ is measured by its uniform risk with respect to W,

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

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

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

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

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

Typically, order-optimal estimators are based on prior knowledge of the *solution set* W which is often not available. Recent progress in nonparametric estimation is related to developing adaptive estimators. An estimator $\hat{l}^\varepsilon(x)$ is said to be *adaptive* with respect to a collection $\mathcal{W}_\varepsilon$ (possibly growing as $\varepsilon \to 0$) of solution sets W if

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

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

An inverse estimator of a linear functional $l_f(x)$ in Hilbert scales, adaptive to the unknown smoothness of x, has been developed recently by Goldenshluger and Pereverzev (2000) for the white noise model. It was assumed there that the smoothness of the representer f with respect to the corresponding Hilbert scale is known. The adaptive estimator in Goldenshluger and Pereverzev (2000) is based on the general adaptation scheme proposed by Lepski (1990). It is our goal in this paper to develop a single estimation method for which (7) holds uniformly over a wide collection $\mathcal{F}$ of representers f. In addition, in the density observation model, under some natural assumptions on the noise, we show that our estimator is adaptive uniformly over a wide collection $\mathcal{K}$ of noise covariance operators. In other words, our procedure can estimate adaptively any linear functional $l_f(x)$ such that $f \in \mathcal{F}$, provided that the covariance operator of the noise belongs to the collection $\mathcal{K}$. This substantially extends the results in Goldenshluger and Pereverzev (2000) where the representer f is fixed, its smoothness is completely known, and only the white noise model is considered. The proposed adaptive estimator may be viewed as a two-stage Lepski scheme with additional adaptation to unknown order of the variance term. Such a procedure may be of interest in other nonparametric estimation problems. For the white noise model we show that the accuracy of the proposed estimator is worse only by a logarithmic factor than one could achieve for a fixed representer f in the case where W is

known exactly. It is well known that this extra logarithmic factor often cannot be avoided in adaptive estimating linear functionals (cf. Lepski 1990; 1992; Brown and Low 1996). In these situations our estimator has the best possible adaptation properties. In order to illustrate the general results we consider the problem of deconvolving a bivariate density with singular support described in Example 1.

The rest of the paper is organized as follows. In Section 2 we introduce notation and describe a general method for constructing inverse estimators in Hilbert scales. Section 3 defines our adaptive estimator and establishes our main results for the white noise model. The results are extended to the density observation model in Section 4. In Section 5 we consider the problem of inverse estimation of a bivariate density with singular support. Some concluding remarks are made in Section 6.

2. Hilbert scale set-up

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, $s \geqslant 0$, 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.

Assumption A. *There exist positive constants a, d and D such that*

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

Examples of operators A satisfying (8) can be found in Vainikko and Hämarik (1985), Neubauer (1988), Mair and Ruymgaart (1996) and Mathé and Pereverzev (2001). As already mentioned, even if the operator A does not fit some standard Hilbert scale (e.g., as in Example 1), one can always construct a scale adapted to A: any compact injective operator A meets condition (8) for $a = \frac{1}{2}$ and the Hilbert scale generated by the operator $L = (A^*A)^{-1}$, where A^* is the adjoint of the operator A in $\mathbb{X}$, that is, $A^* : \mathbb{Y} \to \mathbb{X}$.

Within the Hilbert scale set-up 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

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

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

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

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 $\mathbb{X}$, problem (1) is ill-posed.

It is well known (Tautenhahn 1996) that a wide variety of regularization methods for ill-posed problems 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:

$$\sup_{\lambda\in[0,D^2]} |\lambda^\gamma g_\alpha(\lambda)| \leqslant c_\gamma \alpha^{\gamma-1}, \qquad 0 \leqslant \gamma \leqslant 1,$$

$$\sup_{\lambda\in[0,D^2]} |\lambda^\beta [1 - \lambda g_\alpha(\lambda)]| \leqslant c_\beta \alpha^\beta, \qquad 0 \leqslant \beta \leqslant 1,$$

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 $s \geqslant -\nu$ 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

$$b_{\alpha,s}(f, x) = \langle f, (I - L^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}A^*A)x\rangle,$$

$$v_{\alpha,s}(f, \xi) = -\langle \xi, AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\rangle. \tag{11}$$

The next statement was proved in Goldenshluger and Pereverzev (2000).

Lemma 1. *Let Assumption A hold, $f \in W_\nu(N)$, $\nu < a$, and $\hat{l}^\varepsilon_{\alpha,s}(x)$ be associated with $s \geqslant \max\{0, -\nu\}$. Then, for every $\mu \in (-\nu, 2s + a]$,*

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

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

We observe that Lemma 1 holds both for the white noise and the density observation model.

3. White noise model

3.1. Preliminaries

In this section we consider the white noise model 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

$$\mathbb{E}v^2_{\alpha,s}(f, \xi) = \|AL^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\|^2. \tag{13}$$

The following lemma was established in Goldenshluger and Pereverzev (2000).

Lemma 2. *Suppose that Assumption A is satisfied and that* $s \geqslant \max\{0, -\nu\}$. *Then, for* $f \in W_\nu(N)$, $\nu < a$,

$$\mathbb{E}v^2_{\alpha,s}(f, \xi) \leqslant c_2\alpha^{(\nu-a)/(a+s)}, \tag{14}$$

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

Lemmas 1 and 2 allow us to establish upper bounds on the uniform risk of $\hat{l}^\varepsilon_{\alpha,s}(x)$. In particular, the uniform risk of the estimate $\hat{l}^\varepsilon_{\alpha,s}(x)$ associated with $s \geqslant \max\{0, -\nu\}$ and $\alpha > 0$ admits the upper bound

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

where $c_3 = c_3(\nu, s, a, N, d, D)$. Thus, with the optimal choice $\alpha_* \asymp (M^{-1}\varepsilon)^{2(a+s)/(\mu+a)}$ we have

$$\mathcal{R}[\hat{l}^\varepsilon_{\alpha_*,s}; W_\mu(M)] \leqslant c_3 M^{-2(\nu-a)/(\mu+a)}\varepsilon^{2(\mu+\nu)/(\mu+a)}, \forall\mu \in (-\nu, 2s + a], \tag{15}$$

where '$\asymp$' means equivalent in the sense of the order. It follows from the renormalization argument of Donoho and Low (1992) that the rate of convergence on the right-hand side of (15) cannot be improved for estimating linear functionals with representer $f \in W_\nu(N)$. Thus, the estimate $\hat{l}^\varepsilon_{\alpha_*,s}(x)$ is order-optimal for any pair of balls $W_\nu(N)$ and $W_\mu(M)$ satisfying $\mu \in (-\nu, 2s + a]$, $\nu < a$. We observe, however, that prior knowledge of the smoothness of the unknown solution x is needed in order to choose the regularization parameter optimally.

An inverse estimator of a linear functional $l_f(x) = \langle f, x\rangle$, $f \in W_\nu(N)$, which automatically adapts to the unknown smoothness of x was developed in Goldenshluger and Pereverzev (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. We note that this construction depends crucially on the smoothness index ν of f because the 'typical value' of the stochastic term in the error decomposition depends on ν (cf. (14)).

In view of Lemma 2, it is reasonable to characterize the *effective smoothness* of the representer f via the behaviour of the stochastic error of the regularized inverse estimator $\hat{l}^\varepsilon_{\alpha,s}(x)$. We say that f has the effective smoothness ν if the following assumption is valid:

Assumption F1. *$f \in \mathbb{X}^\nu$, and there exists a constant $c_* = c_*(\nu, a, s)$ such that, for all sufficiently small α,*

$$\|AL^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\| \geqslant c_*\alpha^{(\nu-a)/2(a+s)}. \tag{16}$$

Condition (16) is quite natural. It means that for the functionals considered the order of the variance indicated in (14) cannot be improved in the power scale. Using the same argument as in Neubauer (1997), one can show that (14) implies only $f \in \mathbb{X}^r$ for all $r < \nu$. Thus, (16) allows the smoothness of f to be specified through the properties of a fixed estimator determined by concrete g_α. For this reason we treat ν in Assumption F1 as the effective smoothness of the representer f.

We now we define a two-stage estimator which is adaptive in the sense of (7) over a wide collection of solution sets $W_\mu(M)$ and of representer sets $W_\nu(N)$.

3.2. Adaptive estimator

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

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

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

The basic idea underlying the construction of our adaptive estimator is the following. 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, then the threshold in the adaptation scheme is small, and on a set of 'large' probability the adaptation procedure yields 'too small' a 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. Thus the 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 from among these the minimal regularization parameter which is not 'too small'.

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

$$\underline{\alpha} = \varepsilon^p, \qquad p = \frac{4(a+s)}{\delta}\left(\frac{a-\underline{\nu}}{a-\overline{\nu}}\right). \tag{17}$$

For $q > 1$, define

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

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 $\kappa \geqslant 1$, let $\hat{\alpha}_j$ denote the maximal α from Δ_α such that

$$|\hat{l}_\gamma - \hat{l}_\eta| \leqslant 2\kappa\varepsilon[r_{\nu_j}(\gamma) + r_{\nu_j}(\eta)], \qquad \forall \gamma, \eta \leqslant \alpha, \gamma, \eta \in \Delta_\alpha. \tag{18}$$

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:

$$\hat{\alpha}_0 \geqslant \hat{\alpha}_1 \geqslant \ldots \geqslant \hat{\alpha}_m. \tag{19}$$

Indeed, the threshold on the right-hand side of (18) decreases monotonically as ν_j grows. Therefore the set of estimates satisfying the inequality becomes smaller as ν_j increases.

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

$$j_+ := \max\{j : j \in \mathcal{J}_\tau\}, \qquad \mathcal{J}_\tau := \{j : \hat{\alpha}_j \geqslant \tau, j = 0, \ldots, m\}. \tag{20}$$

Define $\hat{\alpha}_+ = \hat{\alpha}_{j_+}$ if the set $\mathcal{J}_\tau$ is not empty; otherwise 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 κ, s and q; in what follows κ will be chosen as a function of ε, s and q, and other known parameters of the problem.

3.3. Main result

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

Theorem 1. *Suppose that Assumptions* A *and* F1 *are met, and that* (9) *is satisfied for some* $\nu \in \Delta_\nu$ *and* $N \leqslant \overline{N}$, *where* $\overline{N}$ *is known. Let* ε *be small enough that, for some constant* $C_1 = C_1(\nu, a, s, d, D, q)$,

$$\varepsilon\sqrt{\ln \varepsilon^{-1}} \leqslant C_1 \min\{M, M^{-(a-\overline{\nu})/(\mu+\overline{\nu})}\}. \tag{21}$$

Assume also that

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

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

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

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

Note that C_2 depends only on known parameters of the problem so that the choice of κ can be implemented. If we knew in advance the 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 the accuracy of our adaptive estimator coincides to within a factor logarithmic in ε^{-1} with the minimax rate of convergence, that is,

$$\mathcal{R}[\hat{l}_+;\, W_\mu(M)] \asymp \mathcal{R}_*[\varepsilon;\, W_\mu(M)](\ln \varepsilon^{-1})^{(\mu+\nu)/(\mu+a)}, \qquad \varepsilon \to 0.$$

We stress, however, that the upper bound (23) holds simultaneously over 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 $N \leqslant \overline{N}$ satisfying (22), (16) and $s \geqslant \max\{0, -\nu\}$. One can also argue that in many important cases the estimate $\hat{l}_+(x)$ has the best possible adaptive properties; for a discussion of this issue, see Goldenshluger and Pereverzev (2000).

4. Density model

In this section we consider the density observation model. Here the noisy data are represented by y_ε, a random element of $\mathbb{Y}$ defined in (4).

4.1. Preliminaries

Consider i.i.d. $\mathbb{Y}$-valued random elements $\theta_i = \hat{y}_i - y$, $i = 1, \ldots, n$, 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). We will need the following conditions on P_θ.

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

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

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

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

Assumption P2. *There exist positive constants b_1 and H_1 such that*

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

Assumption P3. *There exist positive constants b_2 and H_2 such that*

$$\mathbb{P}\{\|\theta\| \leqslant t\} \leqslant b_2 t, \qquad 0 \leqslant t \leqslant H_2. \tag{25}$$

It is well known that K_θ is a self-adjoint positive operator with finite trace $\mathrm{tr}(K_\theta) = \int \|\theta\|^2 P_\theta(\mathrm{d}\theta) < \infty$ (Vakhania *et al.* 1987, p. 177); hence the operator $K_\theta^{1/2}$ such that $(K_\theta^{1/2})^2 = K_\theta$ is well defined. Assumptions P2 and P3 require existence of the exponential moment and some form of regularity of P_θ near zero, respectively. We note that Assumption P3 is only one of the possible assumptions on the regularity of the distribution of $\|\theta\|$ near zero. In fact, any polynomial decrease in $\mathbb{P}\{\|\theta\| \leqslant t\}$ as $t \to 0$ will be appropriate.

Consider the general regularized inverse estimator $\hat{l}^{\varepsilon}_{\alpha,s}(x)$ defined in Section 2. Lemma 1 establishes an upper bound on its bias. For the density observation model the stochastic error $\upsilon_{\alpha,s}(f, \xi)$ is a zero-mean random variable with variance $\mathbb{E}\upsilon^2_{\alpha,s} = \langle K_\theta \phi_\alpha, \phi_\alpha \rangle$, where $\phi_\alpha = AL^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f$ (see (13)). Because K_θ is a self-adjoint operator with finite trace, the order of the variance term is, in general, smaller than that in the white noise model. This suggests that faster rates of convergence are achievable in the density observation model (see Nussbaum and Pereverzev 1999). The next assumption is an analogue of Assumption F1.

Assumption F2. *There exist a real number* $\tilde{\nu} \in [\nu, a)$ *and positive constants* $\underline{\lambda}$ *and* $\bar{\lambda}$ *such that, for* $s \geq \max\{0, -\nu\}$ *and for all sufficiently small* α*,*

$$\underline{\lambda}\alpha^{(\tilde{\nu}-a)/2(a+s)} \leqslant \|K_\theta^{1/2} AL^{-s} g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f\| \leqslant \bar{\lambda}\alpha^{(\tilde{\nu}-a)/2(a+s)}. \tag{26}$$

Following our terminology for the white noise model, we could call $\tilde{\nu}$ the effective smoothness of the representer f. Similarly to the white noise model, the effective smoothness of the representer f is characterized through the properties of the estimate $\hat{l}^{\varepsilon}_{\alpha,s}(x)$. However, it is important to emphasize that $\tilde{\nu}$ depends on the actual smoothness ν of the representer, on the ill-posedness index a of the operator A, and on smoothing properties of the covariance operator K_θ. In order to illustrate this relationship we consider the following example (cf. Nussbaum and Pereverzev 1999).

Example 2. Let $\mathbb{X} = \mathbb{L}_2(0, 1)$, and let $\mathbb{X}^\mu$ be the Sobolev space of periodic functions on [0, 1] with norm

$$\|x\|_\mu = |\hat{x}(0)|^2 + \sum_{k=-\infty}^{\infty} |k|^{2\mu} |\hat{x}(k)|^2 < \infty,$$

where $\hat{x}(k)$, $k = 0, \pm 1, \pm 2, \ldots$ are Fourier coefficients: $x(t) = \sum_{k=-\infty}^{\infty} \hat{x}(k) \mathrm{e}^{\mathrm{i}k2\pi t}$. Thus $\{\mathbb{X}^\mu\}$ is the Hilbert scale generated by the operator

$$Lx(t) = \hat{x}(0) + \sum_{k=-\infty}^{\infty} |k| \hat{x}(k) \mathrm{e}^{\mathrm{i}k2\pi t}.$$

Let $A : \mathbb{X} \to \mathbb{X}$ be the a-fold integration operator

$$Ax(t) = \frac{1}{\Gamma(a+1)} \int_0^1 (t - \tau)^{a-1} x(\tau) \mathrm{d}\tau.$$

Assume that ξ is a zero-mean $\mathbb{L}_2(0, 1)$-valued Gaussian random noise defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and having the series representation

$$\xi(t) = \sum_{k=-\infty}^{\infty} \langle \xi, \mathrm{e}^{\mathrm{i}k2\pi\cdot}\rangle \mathrm{e}^{\mathrm{i}k2\pi t} = \sum_{k=-\infty}^{\infty} \lambda_k \xi_k \mathrm{e}^{\mathrm{i}k2\pi t},$$

which converges almost surely in $\mathbb{L}_2(0, 1)$. Here $\lambda_k \in \mathbb{R}$, and ξ_k are i.i.d. $\mathcal{N}(0, 1)$ random variables. Assume that $\lambda_k \asymp |k|^{-\beta}$ for some $\beta \in [0, a]$. Under these assumptions it is easily verified that the variance term of $\hat{l}^{\varepsilon}_{\alpha,s}(x)$ is not greater than $O(\alpha^{(\nu-a+\beta)/(a-\beta+s)})$ provided that $\nu < a - \beta$. Thus the covariance operator K_θ acting along the same Hilbert scale reduces ill-posedness index a of the underlying operator A. The upper bound of (26) is satisfied with

$$\tilde{\nu} = a + \left(\frac{a+s}{a-\beta+s}\right)(\nu - a + \beta).$$

We note also that $\tilde{\nu} > \nu$ whenever $\beta > 0$; that is, effectively the representer f is smoother when the noise is correlated.

It is important to realize that our definition of the effective smoothness $\tilde{\nu}$ in (26) is rather arbitrary. For example, we could define the upper and lower bounds in (26) to be of the order $\alpha^{\tilde{\nu}}$ for some $\tilde{\nu} < 0$. The only important requirement for our purposes is that the unknown rate parameter $\tilde{\nu}$ belongs to a known discrete set such as Δ_ν. In this case the two-stage adaptive estimator of Section 3 can be applied; it adapts to the unknown 'true' value of the effective smoothness index. We intentionally consider (26) (cf. (13)) in order to define the adaptive estimator in the same way for the both models, and to give a unified proof of main results.

4.2. Main result

Lemma 1 and Assumption F2 allow us to establish an upper bound on the uniform risk of $\hat{l}^{\varepsilon}_{\alpha,s}(x)$. In particular, for the estimate $\hat{l}^{\varepsilon}_{\alpha,s}(x)$ associated with $s \geqslant \max\{0, -\nu\}$ and small enough α,

$$\mathcal{R}[\hat{l}^{\varepsilon}_{\alpha,s}(x);\, W_\mu(M)] \leqslant c_4\big(M^2\alpha^{(\mu+\nu)/(a+s)} + n^{-1}\alpha^{(\tilde{\nu}-a)/(a+s)}\big),$$

where $c_4 = c_4(\nu, s, a, d, D, N, \bar{\lambda})$. With the optimal choice $\alpha_* \asymp (M\sqrt{n})^{-2(a+s)/(\mu+a+\nu-\tilde{\nu})}$, we have

$$\mathcal{R}[\hat{l}^{\varepsilon}_{\alpha_*,s}(x);\, W_\mu(M)] \leqslant c_4 M^{2(a-\tilde{\nu})/(\mu+a+\nu-\tilde{\nu})} n^{-(\mu+\nu)/(\mu+a+\nu-\tilde{\nu})}, \qquad \forall \mu \in (-\nu, 2s + a].$$

Let $\hat{l}_+(x)$ be the adaptive estimator of Section 3, where $\varepsilon = n^{-1/2}$. We formulate an analogue of Theorem 1 for the density observation model under Assumptions P1–P3 and F2. Unless explicitly stated otherwise, the constants appearing in the following theorem may depend on a, s, d, D, q, $\underline{\lambda}$, $\bar{\lambda}$, and on all parameters involved in Assumptions P2 and P3.

Theorem 2. *Suppose that Assumptions* A, P1–P3 *and* F2 *are valid. Assume that* (9) *is*

fulfilled with $N \leqslant \overline{N}$, *where* $\overline{N}$ *is known, and* (22) *is valid. Let* $\tilde{\nu}$ *defined in Assumption* F2 *be such that*

$$\nu \leqslant \tilde{\nu} < \nu + \frac{\delta}{2}\left(\frac{a-\overline{\nu}}{a-\underline{\nu}}\right). \tag{27}$$

If $\tilde{\nu} \in \Delta_\nu$, *and*

$$\sqrt{\frac{n}{\ln n}} \geqslant C_4 \max\{1,\, M,\, M^{(a-\overline{\nu})/(\mu+\overline{\nu})}\},$$

for some constant C_4, *then there exists a constant* $C_5 = C_5(a, s, \underline{\nu}, \overline{\nu}, \overline{N}, d, D, \overline{\lambda})$ *such that, for the estimate* $\hat{l}_+(x)$ *associated with* $\kappa = C_5\sqrt{\ln n}$ *and* $s \geqslant \max\{0, -\nu\}$, *we have*

$$\mathcal{R}[\hat{l}_+;\, W_\mu(M)] \leqslant C_6\left[M^{2(a-\overline{\nu})/(\mu+a+\nu-\tilde{\nu})}\left(\frac{\ln n}{n}\right)^{(\mu+\nu)/(\mu+a+\nu-\tilde{\nu})} + \frac{m}{n}(\ln n)^{3/2}\right].$$

Observe also that C_5 is defined only in terms of known parameters of the problem. Therefore the choice of κ can be implemented. The performance of the proposed estimator is worse only by a logarithmic factor than that of the estimator knowing the effective smoothness $\tilde{\nu}$ of the representer f. However, in the density model this parameter is hardly known, since the smoothing properties of the covariance operator are usually unknown. The upper bound of Theorem 2 holds uniformly for the collection of representers f and covariance operators K_θ satisfying Assumption F2 and (27).

We note that (27) can be viewed as a restriction on the smoothness of the noise covariance operator. In particular, in the context of Example 2, (27) is written as

$$0 \leqslant \beta < \left(\frac{a+s}{s+\nu+\tilde{p}}\right)\tilde{p},\ \tilde{p} = \frac{\delta}{2}\left(\frac{a-\overline{\nu}}{a-\underline{\nu}}\right),$$

that is, the covariance operator should not be too smooth. We use (27) along with Assumption P2 in order to derive an exponential inequality for the stochastic error of the estimator (see Lemma 5 below). A similar exponential inequality can be established without (27) if a more stringent assumption than Assumption P2 is imposed. For example, if instead of Assumption P2 we suppose that

$$\mathbb{E}\exp\left\{t\|K_\theta^{1/2}\phi_\alpha\|^{-1}\langle\theta,\, \phi_\alpha\rangle\right\} \leqslant b_1, \qquad \text{for } |t| \leqslant H_1, \tag{28}$$

where $\phi_\alpha = AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f$, then the statement of Theorem 2 is valid without (27). However, in our view, (28) is less natural and is hard to verify. Theorem 2 shows that under (27) one cannot expect that the order of the uniform risk for the density model can be much improved in comparison to the white noise model (see Theorem 1 with the usual rescaling $\varepsilon = n^{-1/2}$). At this point one can observe that in the framework of the density model the noise belongs to the observation space, and properties of such a noise are close to those of deterministic noise. On the other hand, it follows from Donoho (1994) that the optimal risk of linear functional estimation has the same order for the model with white noise as well as with deterministic noise.

5. Deconvolving bivariate densities with singular support

To illustrate general results of Section 4.2 we consider the problem of estimating a bivariate density with singular support from indirect observations on the plane.

Let $\Psi_V(u) = \mathbb{E}\exp\{\mathrm{i}V^{\mathrm{T}}u\}$ be the characteristic function of a random variable $V \in \mathbb{R}^2 = \mathbb{C}$. For $u = t\exp\{\mathrm{i}\zeta\}$ we write $\Psi_V(u) = \Psi_V(t, \zeta)$. Then in the context of Example 1, $\Psi_Y(u) = \Psi_z(u)\Psi_w(u)$ where $z = \rho(\varphi)\exp\{\mathrm{i}\varphi\}$. For any $u \in \mathbb{R}^2$, we have

$$\tilde{y}(u) = \tilde{y}(t, \zeta) := \Psi_w^{-1}(t, \zeta)\Psi_Y(t, \zeta)$$

$$= \mathbb{E}\exp\{\mathrm{i}\rho(\varphi)t\cos(\varphi - \zeta)\} = \int_0^{2\pi} \exp\{\mathrm{i}\rho(\varphi)t\cos(\varphi - \zeta)\}x(\varphi)\mathrm{d}\varphi.$$

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

$$y(t) := \frac{1}{2\pi}\int_0^{2\pi} \tilde{y}(t, \zeta)\mathrm{d}\zeta = \int_0^{2\pi} J_0(t\rho(\varphi))x(\varphi)\mathrm{d}\varphi, \tag{29}$$

where $J_0(\cdot)$ is the Bessel function of zero order. If $\rho(\varphi) \neq \text{const.}$, the integral on the right-hand side of (29) can be considered as an integral operator acting from $\mathbb{L}_2(0, 2\pi)$ to $\mathbb{L}_2(0, r)$ for some $r > 0$.

The function $y(\cdot)$ on the left-hand side of (29) can be estimated from the observations Y_j, $j = 1, \ldots, n$, given by (5). By definition,

$$\tilde{y}(u) = \tilde{y}(t, \zeta) = \mathbb{E}\Big[\exp\{\sigma^2|u|^2/2\}\exp\{\mathrm{i}Y_j^{\mathrm{T}}u\}\Big]$$

$$= \exp\{\sigma^2 t^2/2\}\mathbb{E}\big[\exp\{\mathrm{i}t|Y_j|\cos(\arg(Y_j) - \zeta)\}\big],$$

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

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

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

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

because $|J_0(t)| \leqslant 1$, for all t. Thus we are in the framework of the density observation model, and our goal is to apply general results of Section 4.2 to this particular estimation problem.

First, we verify Assumptions P1–P3 for our problem. Assumption P1 is trivially satisfied, and Assumption P2 holds because

$$\|\theta_j\|^2 = \|\hat{y}_j - y\|^2$$

$$= \int_0^r \exp\{\sigma^2 t^2\}\big[J_0(t|Y_j|) - \mathbb{E}J_0(t|Y_j|)\big]^2\mathrm{d}t \tag{30}$$

$$\leqslant 4r\exp\{\sigma^2 r^2\} < \infty.$$

In order to check Assumption P3 we observe that the integrand in (30) is a bounded

continuous and positive function of t. Therefore, it is sufficient to verify condition (25) for the random variable $|J_0(k_*|Y|)|$, where $k_* \in [0, r]$ is a constant. Because $|J_0(k_*|Y|)|$ is a smooth function of $|Y|$ with uniformly bounded first derivative, Assumption P3 will follow if the distribution of $|Y|$ has property (25). But this is an immediate consequence of the fact that Y has a bounded infinitely differentiable density function on the plane.

Clearly, one cannot expect that for a given contour the operator A from (6) will satisfy (8) with some standard Hilbert scale such as the Sobolev scale. Therefore it is reasonable to embed the problem in the natural Hilbert scale generated by the operator $L = (A^*A)^{-1}$. Of course, the smoothness of the representer f of the functional $l_f(x) = x(\varphi_0)$, which we are interested in, and the smoothness of the solution x relative to such a scale are generally unknown. In addition, the noise covariance structure cannot be specified precisely. We note, however, that in this situation Assumption F2 is quite natural, and we can apply our adaptive estimator from Section 3.2 because (8) is satisfied automatically with $a = \frac{1}{2}$ and $d = D = 1$.

From the Taylor expansion of the Bessel function one can see that the singular values of the operator A tend to zero at exponential rate. Moreover, if $\rho(\cdot)$ is a continuous function then all eigenfunctions of A^*A are continuous too. Therefore, for any positive μ, all terms of the SVD of $(A^*A)^\mu$ are continuous functions and the corresponding series converges uniformly on $[0, 2\pi]$. This means that even for small positive μ the spaces $\mathbb{X}^\mu$ from the natural Hilbert scale consist of continuous functions, and the linear functional $l_f(x) = x(\varphi_0)$ is well defined on such a space. For such a scale the representer of this linear functional belongs to a ball $W_\nu(N)$ with negative ν which is close to zero. Moreover, the operator A is automatically injective in its natural Hilbert scale. In this case it is reasonable to consider the regularized estimator (10) associated with $s = 0$. We choose $g_\alpha(\lambda) = (\lambda + \alpha)^{-1}$; this corresponds to the 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

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

where

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

Then the next statement is an immediate consequence of Theorem 2.

Theorem 3. *Let Assumption* F2 *with* (27) *hold for the covariance operator* K_θ *of* $\theta_i = \hat{y}_i - y$, $i = 1, 2, \ldots, n$. *Assume that the effective smoothness* $\tilde{\nu}$ *of the representer of functional* $l_f(x) = x(\varphi_0)$ *relative to the natural Hilbert scale belongs to the set* Δ_ν, $\bar{\nu} = 0$, *and* $\underline{\mu} \in (-\nu, \frac{1}{2}]$. *Then, for sufficiently large n, there exists a constant* C_1 *depending on* $\underline{\nu}$, φ_0 *and* $\bar{\lambda}$ *such that, for* $\hat{\alpha}_+$ *defined by* (18), (20) *with* $\kappa = C_1\sqrt{\ln n}$,

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

Some comments regarding Theorem 3 are in order. Assumptions F2 and P1–P3 determine the class of noise covariance operators K_θ to which the proposed procedure adapts. We emphasize that the adaptive estimator does not use a priori information on this class. While Assumptions P1–P3 are easily verifiable in the context of deconvolving singular densities on the plane, Assumption F2 is much more difficult to check. Nevertheless, we can argue that Assumption F2 is fulfilled in several important cases of interest, and Theorem 3 is valid in these cases.

Let $A = \sum_{k=1}^{\infty} s_k u_k \langle v_k, \cdot\rangle$ be the SVD of the operator from (29). It is natural to assume that for random noise $\theta = y_\varepsilon - y$ the Fourier coefficients $\langle u_k, \theta\rangle$ are independent random variables. This assumption allows to treat the covariance operator K_θ as a diagonal one. Keeping in mind that the representer of the functional $l_f(x) = x(\varphi_0)$ can formally be written as $f(\varphi) = \sum_{k=1}^{\infty} v_k(\varphi_0) v_k(\varphi)$, we obtain

$$\|K_\theta^{1/2} A g_\alpha(A^*A) f\|^2 = \sum_{k=1}^{\infty} \frac{\varkappa_k^2 s_k^2}{(\alpha + s_k^2)^2} |v_k(\varphi_0)|^2,$$

where $\varkappa_k^2 = \mathbb{E}|\langle u_k, \theta\rangle|^2$, $k = 1, 2, \ldots$. As mentioned above, $f \in W_\nu(N)$ with negative ν which is close to zero, that is,

$$\sum_{k=1}^{\infty} s_k^{-4\nu} |v_k(\varphi_0)|^2 < \infty.$$

This means that there is a negative $\tilde{\nu}$ such that $\nu < \tilde{\nu} < 0$, and $s_k^{-4\tilde{\nu}}|v_k(\varphi_0)|^2 \leqslant c_1$, $k = 1, 2, \ldots$. Then

$$\|K_\theta^{1/2} A g_\alpha(A^*A) f\|^2 \leqslant c_1 \operatorname{tr}(K_\theta) \sup_k \frac{s_k^{2+4\tilde{\nu}}}{(\alpha + s_k^2)^2} \leqslant (\bar{\lambda})^2 \alpha^{2\tilde{\nu}-1},$$

which gives us the upper bound (26) for $a = \frac{1}{2}$, $s = 0$. To check the lower bound, we observe that the singular values of the operator A from (29) tend to zero very fast (such that only finitely many of them can be taken into account). Using the Taylor coefficients of the Bessel function, one can estimate the lower bound as

$$s_k \leqslant (c_2 k)^{-\omega_1 k}, \qquad \frac{s_k}{s_{k-1}} \geqslant k^{-\omega_2},$$

where ω_1, $\omega_2 > 0$, $0 < c_2 < 1$. Then it is natural to assume that $\varkappa_k > s_k^{-2\nu_1}$, $k = 1, 2, \ldots$, for some small negative ν_1, because the covariance operator is usually not so smooth. Moreover, for any sufficiently small $\alpha > 0$, one can find $m < \ln(1/\alpha)$ such that $s_{m-1} \geqslant \sqrt{\alpha} \geqslant s_m$. For simplicity we assume that $|v_m(\varphi_0)| > c_3$ (in view of the negative smoothness of f, $|v_m(\varphi_0)|$ can even increase with m). Then

$$\sqrt{\alpha} \geqslant s_m \geqslant s_{m-1} m^{-\omega_2} \geqslant \sqrt{\alpha}\, m^{-\omega_2} \geqslant \sqrt{\alpha} \ln^{-\omega_2} \frac{1}{\alpha}$$

and

$$\|K_\theta^{1/2} A g_\alpha(A^*A) f\|^2 \geqslant \frac{\varkappa_m^2 s_m^2}{(\alpha + s_m^2)^2} |\upsilon_m(\varphi_0)|^2 \geqslant c_3^2 \frac{s_m^{2-4\nu_1}}{(\alpha + s_m^2)^2}$$
$$\geqslant \frac{\alpha^{-1-2\nu_1}}{\ln^{\omega_2(2-4\nu_1)}(1/\alpha)} (\underline{\lambda})^2 \geqslant (\underline{\lambda})^2 \alpha^{2\bar{\nu}-1+\delta},$$

where $\delta > -2(\tilde{\nu} + \nu_1) > 0$ can be sufficiently small due to small size of $\tilde{\nu}$, ν_1. Thus, in the case considered we have a small gap between the upper and lower bounds in (26) in the sense of the order. This gap could be reduced under an additional assumption concerning the rate of increase of $|\upsilon_m(\varphi_0)|$, which depends on the parametrization ρ. Nevertheless, Theorem 3 is still valid even with such a gap if δ is chosen as a minimal step for the set Δ_ν (see also Remark 2 in Section 6).

6. Concluding remarks

Remark 1. In our main results we assume that the discrete set Δ_ν is such that δ is fixed and positive. It can be seen from the proofs 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 effective smoothness index belongs to the set Δ_ν eventually as $\varepsilon \to 0$. The statements of Theorems 1 and 2 remain valid under these circumstances.

Remark 2. Inspection of the proofs in the Appendix reveals that Theorems 1 and 2 also hold under relaxed Assumptions F1 and F2. For example, the right-hand side of (16) may have order $\alpha^{(\nu-a+\delta/2)/(2(a+s))}$, that is, there is a gap between the upper and lower bounds of the variance term. In particular, this situation occurs when $f \in \mathbb{X}^\nu$, but it is effectively smoother, and this cannot be described in the terms of the chosen Hilbert scale. In this case the estimator should be modified: one should put $2p$ instead of p in (17). The resulting accuracy will be the best that one can achieve with fixed Hilbert scale and the fixed set Δ_ν.

Remark 3. The Lepski adaptive scheme is based on the assumption that the order of the error variance term is completely known. Our adaptive procedure can be viewed as a two-stage Lepski adaptive scheme with additional adaptation to unknown order of the variance term. As is shown in Section 4, the advantage of the proposed two-stage estimator is fully revealed in the density model where the noise covariance structure is unknown. Such a procedure may be of interest in other nonparametric estimation problems. It seems that certain inverse problems of the convolution type with imprecisely specified convolution kernel can be treated using these ideas.

Appendix

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

A. Proofs for the white noise model

The goal of this section is to prove Theorem 1. Denote

$$B_\alpha(x) = c_1 \|x\|_\mu \|f\|_\nu \alpha^{(\mu+\nu)/(2(a+s))}, \tag{31}$$

where c_1 is the constant appearing on the right-hand side of (12). Define $\alpha_* = \max\{\alpha \in \Delta_\alpha : B_\alpha(x) \leqslant \kappa\varepsilon r_\nu(\alpha)\}$. In fact, α_* is the ideal regularization parameter that balances the squared bias and variance. Consider the event

$$\Omega_\kappa := \Big\{\omega \in \Omega: \max_{\alpha\in\Delta_\alpha}[r_\nu^{-1}(\alpha)|\upsilon_{\alpha,s}(f, \xi)|] \leqslant \kappa\Big\}. \tag{32}$$

The event Ω_κ corresponds to the 'typical' behaviour of the stochastic term $\upsilon_{\alpha,s}(f, \xi)$. Also, for notational convenience, we denote $t := \min\{j \in \{0, \ldots, m\} : \nu_j > \nu\}$, that is, $\nu = \nu_{t-1}$ for some $t \in \{1, \ldots, m\}$.

First, we establish some auxiliary lemmas. The next statement shows that, conditionally on Ω_κ, 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 3. *Suppose the Assumption* F1 *is satisfied and that* $\alpha_* \geqslant \underline{\alpha}^{(a-\nu_t)/(a-\nu)}$. *Then, for every* $\eta \in \Delta_\alpha$ *satisfying*

$$\eta \geqslant \underline{\alpha}^{(a-\nu_t)/(a-\nu)} \tag{33}$$

and large enough $\kappa \asymp \sqrt{\ln \varepsilon^{-1}}$,

$$\mathbb{P}\{\hat\alpha_t \geqslant \eta \,|\, \Omega_\kappa\} \leqslant k_1 \kappa \underline{\alpha}^{(\nu_t-\nu)/2(a+s)}, \tag{34}$$

where $k_1 = k_1(c_*, a, s, \underline{\nu}, \overline{\nu}, d, D, \overline{N}, q)$ *and* c_* *is defined in* (16).

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

First, assume that $\eta \leqslant \alpha_*$. We have

$$\begin{aligned}\mathbb{P}\{\hat\alpha_t \geqslant \eta|\Omega_\kappa\} &\leqslant \mathbb{P}\{|\hat l_\eta - \hat l_{\underline{\alpha}}| \leqslant 2\kappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\kappa\} \\ &= 1 - \mathbb{P}\{|\hat l_\eta - \hat l_{\underline{\alpha}}| > 2\kappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\kappa\}. \end{aligned} \tag{35}$$

On the set Ω_κ,

$$
\begin{aligned}
|\hat{l}_\eta - \hat{l}_{\underline{\alpha}}| &\geqslant |\hat{l}_{\underline{\alpha}} - l| - |\hat{l}_\eta - l| \\
&\geqslant \varepsilon |\upsilon_{\underline{\alpha},s}(f, \xi)| - B_{\underline{\alpha}}(x) - 2\kappa\varepsilon r_\nu(\eta) \\
&\geqslant \varepsilon |\upsilon_{\underline{\alpha},s}(f, \xi)| - 3\kappa\varepsilon r_\nu(\eta).
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{P}\{|\hat{l}_\eta - \hat{l}_{\underline{\alpha}}| > 2\kappa\varepsilon[r_{\nu_t}(\eta) + r_{\nu_t}(\underline{\alpha})] \,|\, \Omega_\kappa\} \\
&\quad \geqslant \mathbb{P}\{\varepsilon |\upsilon_{\underline{\alpha},s}(f, \xi)| > \kappa\varepsilon[2r_{\nu_t}(\underline{\alpha}) + 2r_{\nu_t}(\eta) + 3r_\nu(\eta)] \,|\, \Omega_\kappa\} \qquad (36) \\
&\quad \geqslant \mathbb{P}\{\varepsilon |\upsilon_{\underline{\alpha},s}(f, \xi)| > \kappa\varepsilon[4r_{\nu_t}(\underline{\alpha}) + 3r_\nu(\eta)] \,|\, \Omega_\kappa\}.
\end{aligned}
$$

By (33), $r_{\nu_t}(\underline{\alpha})$ dominates $r_\nu(\eta)$: $r_\nu(\eta) \leqslant r_{\nu_t}(\underline{\alpha})$. Therefore the last probability in (36) is bounded from below by

$$
\begin{aligned}
\mathbb{P}\{|\upsilon_{\underline{\alpha},s}(f, \xi)| > 7\kappa r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\kappa\} &= 1 - \mathbb{P}\{|\upsilon_{\underline{\alpha},s}(f, \xi)| \leqslant 7\kappa r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\kappa\} \\
&\geqslant 1 - \mathbb{P}\{|\mathcal{N}(0, 1)| \leqslant 7\kappa[\mathbb{E}\upsilon^2_{\underline{\alpha},s}(f, \xi)]^{-1/2} r_{\nu_t}(\underline{\alpha})\}[\mathbb{P}\{\Omega_\kappa\}]^{-1},
\end{aligned}
$$

where $\mathcal{N}(0, 1)$ denotes the standard Gaussian random variable. Arguing as in the proof of Theorem 1 in Goldenshluger and Pereverzev (2000), we obtain $\mathbb{P}\{\overline{\Omega}_\kappa\} \leqslant S_{\underline{\alpha}} \exp\{-\tilde{k}\kappa^2/2\}$, where $S_{\underline{\alpha}} = \mathrm{card}(\Delta_\alpha)$ and $\tilde{k} = \tilde{k}(\nu, \overline{N}, a, s, d, D)$. Hence, for large enough $\kappa \asymp \sqrt{\ln \varepsilon^{-1}}$,

$$
\begin{aligned}
\mathbb{P}\{|\upsilon_{\underline{\alpha},s}(f, \xi)| > 7\kappa r_{\nu_t}(\underline{\alpha}) \,|\, \Omega_\kappa\} &\geqslant 1 - \frac{14\kappa}{\sqrt{2\pi}} \frac{r_{\nu_t}(\underline{\alpha})[\mathbb{E}\upsilon^2_{\underline{\alpha},s}(f, \xi)]^{-1/2}}{1 - S_{\underline{\alpha}} \exp\{-\tilde{k}\kappa^2/2\}} \\
&\geqslant 1 - k_1 \kappa \underline{\alpha}^{(\nu_t - \nu)/2(a+s)}, \qquad (37)
\end{aligned}
$$

Combining (37) and (36) with (35), we obtain (34) under assumption that $\eta \leqslant \alpha_*$.

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

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

The proof now proceeds along the same lines as in the case with η replaced by α_*. □

An immediate consequence of (19) and Lemma 3 is that the same bound holds for all estimates $\hat{\alpha}_j$ associated with $\nu_j > \nu$; namely, under conditions (33) we have

$$
\mathbb{P}\{\hat{\alpha}_j \geqslant \eta | \Omega_\kappa\} \leqslant k_1 \kappa \underline{\alpha}^{(\nu_t - \nu)/2(a+s)}, \qquad \forall j \geqslant t = \min\{j : \nu_j > \nu\}. \qquad (38)
$$

Thus Lemma 3 shows that if we misspecify ν in the threshold of the procedure (18) by choosing a value greater than ν, the scheme will yield a regularization parameter which with 'large' probability is less than $\underline{\alpha}^{(\nu_t - \nu)/(a-\nu)}$.

The above considerations motivated our rule (20). 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 Ω_κ.

Lemma 4. *Suppose that* $\underline{\alpha}^{(a-\nu_t)/(a-\nu)} \leqslant \tau \leqslant \alpha_*$ *and inequality* (16) *hold. If the event* Ω_κ *occurs then* $j_+ \geqslant t-1$. *In addition, for large enough* $\kappa \asymp \sqrt{\ln \varepsilon^{-1}}$,

$$\mathbb{P}(j_+ = t-1 \,|\, \Omega_\kappa) \geqslant 1 - k_1 \kappa \underline{\alpha}^{(\nu_t-\nu)/2(a+s)}, \tag{39}$$

where k_1 *is defined in Lemma 3.*

Proof. The first statement follows immediately from (19) 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 (18), then on the set Ω_κ the resulting regularization parameter $\hat{\alpha}_{t-1}$ will be greater than α_* by construction (see Goldenshluger and Pereverzev 2000). In view of $\tau \leqslant \alpha_*$ and monotonicity of $\{\hat{\alpha}_j\}$, on the set Ω_κ, $\mathcal{J}_\tau \supseteq \{0, 1, \ldots, t-1\}$. Hence $j_+ \geqslant t-1$ as claimed.

To prove (39) we note that the event $\{j_+ \geqslant t\}$ means that there exists an estimate $\hat{\alpha}_j$ associated with $\nu_j > \nu$ such that $\hat{\alpha}_j \geqslant \tau$. But by Lemma 3 (see also (38)), for $j \geqslant t$,

$$\mathbb{P}\{\hat{\alpha}_j \geqslant \tau \,|\, \Omega_\kappa\} \leqslant k_1 \kappa \underline{\alpha}^{(\nu_t-\nu)/2(a+s)}.$$

This completes the proof. □

Proof of Theorem 1. Write

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

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

It follows immediately from (12) and (9) that

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

and

$$\alpha_* \geqslant q^{-1}[(c_1 M\|f\|_\nu)^{-1}\kappa\varepsilon]^{2(a+s)/(\mu+a)} \geqslant [k_2 M^{-1}\kappa\varepsilon]^{2(a+s)/(\mu+a)} \tag{40}$$

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 (17), we have $\tau \geqslant \underline{\alpha}^{(a-\nu_t)/(a-\nu)}$. In order to show this it is sufficient to verify that

$$p\left(\frac{a-\nu_t}{a-\nu}\right) \geqslant \frac{2(a+s)}{a-\overline{\nu}}\,.$$

This follows because

$$\frac{2}{\delta}(a-\underline{\nu})\left(\frac{a-\nu_t}{a-\nu}\right) \geqslant \frac{2}{\delta}(a-\nu_t) \geqslant 2.$$

Thus, Lemma 4 is applicable with our choice of τ and $\underline{\alpha}$.

First we bound I_1. Assume that the event Ω_κ occurs; then, by Lemma 4, $j_+ \geqslant t-1$, that is, $\Omega_\kappa \subseteq \{j_+ \geqslant t-1\}$. Consider the events $B_j = \{j_+ = j\}$, $j = t-1, t, \ldots, m$. Lemma 4 implies that

$$\mathbb{P}(B_{t-1}|\Omega_\kappa) \geqslant 1 - k_1\kappa\underline{\alpha}^{(\nu_t-\nu)/2(a+s)},$$

$$\mathbb{P}(B_j\,|\,\Omega_\kappa) \leqslant k_1\kappa\underline{\alpha}^{(\nu_t-\nu)/2(a+s)}, \qquad \forall j = t, \ldots, m, \tag{41}$$

with k_1 defined in Lemma 3. Write

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

On the set $\Omega_\kappa \cap B_{t-1}$ 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 $\hat{\alpha}_+ = \alpha_{t-1} \geqslant \alpha_*$. Standard calculations (see Goldenshluger and Pereverzev 2000, p. 178) yield $|\hat{l}_{\hat{\alpha}_+} - l| \leqslant 6\kappa\varepsilon r_\nu(\alpha_*)$, showing that

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

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

$$\begin{aligned} |\hat{l}_{\hat{\alpha}_+} - l| = |\hat{l}_{\hat{\alpha}_j} - l| &\leqslant |\hat{l}_{\hat{\alpha}_j} - \hat{l}_\tau| + |\hat{l}_\tau - l| \\ &\leqslant 4\kappa\varepsilon r_{\nu_j}(\tau) + B_\tau(x) + \varepsilon|\upsilon_{\tau,s}(f, \xi)| \\ &\leqslant 4\kappa\varepsilon r_{\nu_j}(\tau) + 2\kappa\varepsilon r_\nu(\tau) \leqslant 6\kappa\varepsilon r_\nu(\tau). \end{aligned} \tag{42}$$

Here we have taken into account that:

(i) $\hat{\alpha}_+ = \hat{\alpha}_j \geqslant \tau$ 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) $\tau \leqslant \alpha_*$ by the premise of the theorem. Therefore on the set $\Omega_\kappa \cap B_j$ the typical value of the stochastic error dominates the bias.

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

It follows from (41) and (42) that, for any $j \geqslant t$,

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

Thus we have the following bound on I_1:

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

Using (17), we obtain

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

so that

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

and substituting the expression for α_* (see (40)) we finally obtain

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

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

Now we consider the case where the event $\overline{\Omega}_\kappa$ occurs. Here our algorithm will choose a value that is not less than τ. The rest of the proof follows the proof given in Goldenshluger and Pereverzev (2000), with obvious modifications ($\underline{\alpha}$ in the paper should be replaced by τ). We emphasize only that in this case κ is chosen as $k_4\sqrt{\ln \varepsilon^{-1}}$, where k_4 depends on a, s, $\underline{\nu}$, $\bar{\nu}$, d, D and $\overline{N}$. □

B. Proofs for the density model

Basically, the proof of Theorem 2 follows similar lines to that of Theorem 1. The main difference is that the noise ξ is a random element of the Hilbert space $\mathbb{Y}$ with compact covariance operator K_θ. We indicate how, using Assumptions P1–P3 and F2, one can modify the arguments in the proof of Theorem 1 in order to prove Theorem 2.

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

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

where $\upsilon_{\alpha,s}(f, \xi) = -\langle \xi, \phi_\alpha\rangle$, $\xi = n^{-1/2}\sum_{i=1}^n \theta_i = n^{-1/2}\sum_{i=1}^n(\hat{y}_i - y)$ and

$$\phi_\alpha := AL^{-s}g_\alpha(L^{-s}A^*AL^{-s})L^{-s}f. \tag{43}$$

By (24) and independence of θ_i, $i = 1, \ldots, n$, we have

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

Assumption F2 provides evident upper and lower bounds on $\mathbb{E}\upsilon^2_{\alpha,s}(f, \xi)$.

Similarly to the proof of Theorem 1, we define

$$\alpha_* = \max\{\alpha \in \Delta_\alpha : B_\alpha(x) \leqslant \kappa\varepsilon r_{\tilde{\nu}}(\alpha)\},$$

where $B_\alpha(x)$ is given by (31). In this case α_* is bounded from below by $[\kappa n^{-1/2}]^{2(a+s)/(\mu+a+\nu-\tilde{\nu})}$ multiplied by a constant (compare with (40)). Also we denote $t := \min\{j \in \{0, \ldots, m\} : \nu_j > \tilde{\nu}\}$, that is, $\tilde{\nu} = \nu_{t-1}$ for some $t \in \{1, \ldots, m\}$. For a fixed $\kappa \geqslant 1$, the event Ω_κ is defined by (32) with $r_\nu^{-1}(\alpha)$ replaced by $r_{\tilde{\nu}}^{-1}(\alpha)$. The proof of Theorem 1 is essentially based on the fact that the constant κ can be chosen in such a way that the event Ω_κ is of 'large' probability. This is easily proved for the white noise model because $\upsilon_{\alpha,s}(f, \xi)$ is a Gaussian normal variable. We now establish a similar exponential inequality for the density observation model.

Lemma 5. *Suppose that Assumptions* P1–P3 *and* F2 *are satisfied, and* $\nu \leqslant \tilde{\nu} < \nu + \frac{1}{2}\delta(a - \bar{\nu})/(a - \underline{\nu})$, where δ, $\underline{\nu}$ *and* $\bar{\nu}$ *are the same as in* (17). *Then there exists a constant* $k_5 = k_5(\bar{\lambda}, \nu, a, s, N, b_1, H_1)$ *such that, for any positive* $\kappa \leqslant k_5 n^\beta$, $0 < \beta < 1 - 2\delta^{-1}(\tilde{\nu} - \nu)(a - \underline{\nu})/(a - \bar{\nu})$,

$$\mathbb{P}(\overline{\Omega}_\kappa) \leqslant 2\, S_{\underline{\alpha}} \exp\left\{-\frac{\kappa^2}{2k_6}\right\},$$

where $S_{\underline{\alpha}} = \mathrm{card}(\Delta_\alpha)$*, and* $k_6 = k_6(\overline{\lambda}, \nu, a, s, N)$.

Proof. Write

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

where ϕ_α is defined in (43). Note that the random variables $\langle \theta_i, r_{\tilde{\nu}}^{-1}(\alpha)\phi_\alpha \rangle$ have zero mean. In view of Assumption P2,

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

for $0 \leqslant t \leqslant H_1 \underline{\lambda}^{-1} \alpha^{(\tilde{\nu}-\nu)/2(a+s)}$, where $\underline{\lambda}$ is defined in Assumption F2. Thus, for fixed ϕ_α, the random variable $r_{\tilde{\nu}}^{-1}(\alpha)\langle \theta, \phi_\alpha \rangle$ has moments of all orders and the following relation holds:

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

Taking into account the upper bound in (26), we obtain that, for any constant $k_6 \geqslant \overline{\lambda}/2$, the inequality

$$\log \left[\mathbb{E} \exp\left\{ t \langle \theta, r_{\tilde{\nu}}^{-1}(\alpha)\phi_\alpha \rangle \right\} \right] \leqslant \tfrac{1}{2} k_7 t^2$$

holds for sufficiently small t. In other words, there exist positive constants $k_6 = k_6(\overline{\lambda}, \nu, a, s, N)$ and $\tilde{H}_1$ such that

$$\mathbb{E} \exp\left\{ t \langle \theta, r_{\tilde{\nu}}^{-1}(\alpha)\phi_\alpha \rangle \right\} \leqslant \exp\{ k_6 t^2/2 \}, \qquad \text{for } 0 \leqslant t \leqslant \tilde{H}_1 \alpha^{(\tilde{\nu}-\nu)/2(a+s)}.$$

Then Theorem 2.6 in Petrov (1995) implies that

$$\mathbb{P}\{ \tilde{\upsilon}_{\alpha,s}(f, \xi) \geqslant \kappa \} = \mathbb{P}\left\{ \sum_{i=1}^{n} \langle \theta_i, r_{\tilde{\nu}}^{-1}(\alpha)\phi_\alpha \rangle \geqslant \sqrt{n}\kappa \right\} \leqslant \exp\left\{ -\frac{\kappa^2}{2k_6} \right\}$$

for $0 \leqslant \kappa \leqslant k_6 \tilde{H}_1 \sqrt{n} \alpha^{(\tilde{\nu}-\nu)/2(a+s)}$. The bound on $\mathbb{P}\{ \tilde{\upsilon}_{\alpha,s}(f, \xi) \leqslant -\kappa \}$ is derived similarly, so that

$$\mathbb{P}\{ |\tilde{\upsilon}_{\alpha,s}(f, \xi)| \geqslant \kappa \} \leqslant 2 \exp\left\{ -\frac{\kappa^2}{2k_6} \right\}, \qquad 0 \leqslant \kappa \leqslant k_6 \tilde{H}_1 \sqrt{n} \alpha^{(\tilde{\nu}-\nu)/2(a+s)}.$$

The statement of the lemma is an immediate consequence of these results and the fact that α is bounded from below by $\underline{\alpha}$ given in (17). □

Proof of Theorem 2. The proof is identical to that of Theorem 1 in every detail, with only the following differences.

First, we note that, under Assumptions P1–P3 and F2, Lemma 3 remains valid with ν replaced by $\tilde{\nu}$, and the bound (34) holds provided n is large enough. Here, in order to prove

(37), we use Lemma 5 and Assumption P3. Then Lemma 4 follows with obvious modifications.

The proof of Theorem 2 on the set Ω_κ coincides with the proof of Theorem 1. To bound the error on the complimentary event $\overline{\Omega}_\kappa$ we use the exponential inequality of Lemma 5. □

Acknowledgement

The authors thank Peter Mathé and two anonymous referees for careful reading and helpful comments that have greatly improved the presentation of the manuscript. The support of the Weierstrass Institute (Berlin) is also gratefully acknowledged.

References

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

Cavalier, L. and Tsybakov, A. (2002) Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, **123**, 323–354.

Donoho, D.L. (1994) Statistical estimation and optimal recovery. *Ann. Statist.*, **22**, 238–270.

Donoho, D. and Low, M. (1992) Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, **20**, 944–970.

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.

Hegland, M. (1995) Variable Hilbert scales and their interpolation inequalities with applications to Tikhonov regularization. *Appl. Anal.*, **59**, 207–223.

Ibragimov, I.A. and Khas'minskii, R.Z. (1984) On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.*, **29**, 18–32.

Ibragimov, I.A. and Khas'minskii, R.Z. (1987) Estimation of linear functionals in Gaussian noise. *Theory Probab. Appl.*, **32**, 30–39.

Krein, S.G., Petunin, Ju.I. and Semenov, E.M. (1982) *Interpolation of Linear Operators*, Transl. Math. Monogr., **54**, American Mathematical Society, Providence, RI.

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

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

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

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

Mathé, P. and 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.

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

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

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

Nussbaum, M. and Pereverzev, S. (1999) The degrees of ill-posedness in stochastic and deterministic noise models. WIAS Preprint no. 509, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

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

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

Skorohod, A.V. (1974) *Integration in Hilbert Space*. Berlin: Springer-Verlag.

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

Traub, J.F., Wasilkowski, G.W. and Wozniakowski, H. (1988) *Information-Based Complexity*. Boston: Academic Press.

Vainikko, G.M. and Hämarik, U.A. (1985) Projection methods and self-regularization of ill-posed problems. *Izv. Vyssh. Uchebn. Zaved. Mat.*, **29**, 1–17.

Vakhania, N.N., Tarieladze, V.I. and Chobanyan, S.A. (1987) *Probability Distributions on Banach Spaces*. Dordrecht: Reidel.

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

Received August 2001 and revised January 2003